Class Notes in Statistics and Econometrics

Hans G. Ehrbar

Economics Department, University of Utah, 1645 Campus Center Drive, Salt Lake City UT 84112-9300, U.S.A.

URL: `www.econ.utah.edu/ehrbar/ecmet.pdf`

E-mail address: `ehrbar@econ.utah.edu`

Abstract. This is an attempt to make a carefully argued set of class notes freely available. The source code for these notes can be downloaded from `www.econ.utah.edu/ehrbar/ecmet-sources.zip` Copyright Hans G. Ehrbar under the GNU Public License

Contents

Chapter 1. Preface xiii

Chapter 2. Probability Fields 1
2.1. The Concept of Probability 1
2.2. Events as Sets 6
2.3. The Axioms of Probability 10
2.4. Objective and Subjective Interpretation of Probability 13
2.5. Counting Rules 14
2.6. Relationships Involving Binomial Coefficients 16
2.7. Conditional Probability 17
2.8. Ratio of Probabilities as Strength of Evidence 23
2.9. Bayes Theorem 24
2.10. Independence of Events 25
2.11. How to Plot Frequency Vectors and Probability Vectors 28

Chapter 3. Random Variables 33
3.1. Notation 33
3.2. Digression about Infinitesimals 33
3.3. Definition of a Random Variable 35
3.4. Characterization of Random Variables 36
3.5. Discrete and Absolutely Continuous Probability Measures 40
3.6. Transformation of a Scalar Density Function 41
3.7. Example: Binomial Variable 43
3.8. Pitfalls of Data Reduction: The Ecological Fallacy 44
3.9. Independence of Random Variables 45
3.10. Location Parameters and Dispersion Parameters of a Random Variable 46
3.11. Entropy 52

Chapter 4. Random Number Generation and Encryption 63
4.1. Alternatives to the Linear Congruential Random Generator 64
4.2. How to test random generators 65

4.3. The Wichmann Hill generator 66
4.4. Public Key Cryptology 68

Chapter 5. Specific Random Variables 71
5.1. Binomial 71
5.2. The Hypergeometric Probability Distribution 74
5.3. The Poisson Distribution 75
5.4. The Exponential Distribution 78
5.5. The Gamma Distribution 80
5.6. The Uniform Distribution 83
5.7. The Beta Distribution 84
5.8. The Normal Distribution 84
5.9. The Chi-Square Distribution 87
5.10. The Lognormal Distribution 88
5.11. The Cauchy Distribution 88

Chapter 6. Sufficient Statistics and their Distributions 91
6.1. Factorization Theorem for Sufficient Statistics 91
6.2. The Exponential Family of Probability Distributions 92

Chapter 7. Chebyshev Inequality, Weak Law of Large Numbers, and Central Limit Theorem 97
7.1. Chebyshev Inequality 97
7.2. The Probability Limit and the Law of Large Numbers 98
7.3. Central Limit Theorem 100

Chapter 8. Vector Random Variables 103
8.1. Expected Value, Variances, Covariances 105
8.2. Marginal Probability Laws 108
8.3. Conditional Probability Distribution and Conditional Mean 109
8.4. The Multinomial Distribution 111
8.5. Independent Random Vectors 112
8.6. Conditional Expectation and Variance 114
8.7. Expected Values as Predictors 116
8.8. Transformation of Vector Random Variables 121

Chapter 9. Random Matrices 125
9.1. Linearity of Expected Values 125
9.2. Means and Variances of Quadratic Forms in Random Matrices 127

Chapter 10. The Multivariate Normal Probability Distribution 133
10.1. More About the Univariate Case 133

10.2. Definition of Multivariate Normal 134
10.3. Special Case: Bivariate Normal 135
10.4. Multivariate Standard Normal in Higher Dimensions 145
10.5. Higher Moments of the Multivariate Standard Normal 148
10.6. The General Multivariate Normal 152

Chapter 11. The Regression Fallacy 157

Chapter 12. A Simple Example of Estimation 167
12.1. Sample Mean as Estimator of the Location Parameter 167
12.2. Intuition of the Maximum Likelihood Estimator 168
12.3. Variance Estimation and Degrees of Freedom 171

Chapter 13. Estimation Principles and Classification of Estimators 181
13.1. Asymptotic or Large-Sample Properties of Estimators 181
13.2. Small Sample Properties 183
13.3. Comparison Unbiasedness Consistency 184
13.4. The Cramer-Rao Lower Bound 188
13.5. Best Linear Unbiased Without Distribution Assumptions 196
13.6. Maximum Likelihood Estimation 198
13.7. Method of Moments Estimators 201
13.8. M-Estimators 201
13.9. Sufficient Statistics and Estimation 201
13.10. The Likelihood Principle 205
13.11. Bayesian Inference 206

Chapter 14. Interval Estimation 209

Chapter 15. Hypothesis Testing 217
15.1. Duality between Significance Tests and Confidence Regions 221
15.2. The Neyman Pearson Lemma and Likelihood Ratio Tests 221
15.3. The Runs Test 225
15.4. Pearson's Goodness of Fit Test. 228
15.5. Permutation Tests 231
15.6. The Wald, Likelihood Ratio, and Lagrange Multiplier Tests 237

Chapter 16. General Principles of Econometric Modelling 239

Chapter 17. Causality and Inference 241

Chapter 18. Mean-Variance Analysis in the Linear Model 245
18.1. Three Versions of the Linear Model 245
18.2. Ordinary Least Squares 246

18.3. The Coefficient of Determination 254
18.4. The Adjusted R-Square 259

Chapter 19. Digression about Correlation Coefficients 261
19.1. A Unified Definition of Correlation Coefficients 261
19.2. Correlation Coefficients and the Associated Least Squares Problem 264
19.3. Canonical Correlations 264
19.4. Some Remarks about the Sample Partial Correlation Coefficients 266

Chapter 20. Numerical Methods for computing OLS Estimates 269
20.1. QR Decomposition 269
20.2. The LINPACK Implementation of the QR Decomposition 270

Chapter 21. About Computers 273
21.1. General Strategy 273
21.2. The Emacs Editor 276
21.3. How to Enter and Exit SAS 277
21.4. How to Transfer SAS Data Sets Between Computers 278
21.5. Instructions for Statistics 5969, Hans Ehrbar's Section 279
21.6. The Data Step in SAS 284

Chapter 22. Specific Datasets 287
22.1. Cobb Douglas Aggregate Production Function 287
22.2. Houthakker's Data 297
22.3. Long Term Data about US Economy 302
22.4. Dougherty Data 303
22.5. Wage Data 304

Chapter 23. The Mean Squared Error as an Initial Criterion of Precision 321
23.1. Comparison of Two Vector Estimators 321

Chapter 24. Sampling Properties of the Least Squares Estimator 325
24.1. The Gauss Markov Theorem 326
24.2. Digression about Minimax Estimators 328
24.3. Miscellaneous Properties of the BLUE 329
24.4. Estimation of the Variance 339
24.5. Mallow's Cp-Statistic as Estimator of the Mean Squared Error 340
24.6. Optimality of Variance Estimators 341

Chapter 25. Variance Estimation: Should One Require Unbiasedness? 345
25.1. Setting the Framework Straight 346

25.2. Derivation of the Best Bounded MSE Quadratic Estimator of the Variance 348
25.3. Unbiasedness Revisited 351
25.4. Summary 353

Chapter 26. Nonspherical Positive Definite Covariance Matrix 355

Chapter 27. Best Linear Prediction 359
27.1. Minimum Mean Squared Error, Unbiasedness Not Required 359
27.2. The Associated Least Squares Problem 365
27.3. Prediction of Future Observations in the Regression Model 367

Chapter 28. Updating of Estimates When More Observations become Available 373

Chapter 29. Constrained Least Squares 377
29.1. Building the Constraint into the Model 377
29.2. Conversion of an Arbitrary Constraint into a Zero Constraint 378
29.3. Lagrange Approach to Constrained Least Squares 379
29.4. Constrained Least Squares as the Nesting of Two Simpler Models 383
29.5. Solution by Quadratic Decomposition 383
29.6. Sampling Properties of Constrained Least Squares 384
29.7. Estimation of the Variance in Constrained OLS 386
29.8. Inequality Restrictions 390
29.9. Application: Biased Estimators and Pre-Test Estimators 391

Chapter 30. Additional Regressors 393
30.1. Selection of Regressors 405

Chapter 31. Residuals: Standardized, Predictive, "Studentized" 409
31.1. Three Decisions about Plotting Residuals 409
31.2. Relationship between Ordinary and Predictive Residuals 411
31.3. Standardization 414

Chapter 32. Regression Diagnostics 417
32.1. Missing Observations 417
32.2. Grouped Data 418
32.3. Influential Observations and Outliers 418
32.4. Sensitivity of Estimates to Omission of One Observation 420

Chapter 33. Regression Graphics 427
33.1. Scatterplot Matrices 427
33.2. Conditional Plots 429
33.3. Spinning 430

33.4. Sufficient Plots 431

Chapter 34. Asymptotic Properties of the OLS Estimator 435
34.1. Consistency of the OLS estimator 436
34.2. Asymptotic Normality of the Least Squares Estimator 437

Chapter 35. Least Squares as the Normal Maximum Likelihood Estimate 439

Chapter 36. Bayesian Estimation in the Linear Model 445

Chapter 37. OLS With Random Constraint 451

Chapter 38. Stein Rule Estimators 455

Chapter 39. Random Regressors 459
39.1. Strongest Assumption: Error Term Well Behaved Conditionally on Explanatory Variables 459
39.2. Contemporaneously Uncorrelated Disturbances 461
39.3. Disturbances Correlated with Regressors in Same Observation 461

Chapter 40. The Mahalanobis Distance 463
40.1. Definition of the Mahalanobis Distance 463
40.2. The Conditional Mahalanobis Distance 466
40.3. First Scenario: Minimizing relative increase in Mahalanobis distance if distribution is known 466
40.4. Second Scenario: One Additional IID Observation 467
40.5. Third Scenario: one additonal observation in a Regression Model 469

Chapter 41. Interval Estimation 475
41.1. A Basic Construction Principle for Confidence Regions 475
41.2. Coverage Probability of the Confidence Regions 479
41.3. Conventional Formulas for the Test Statistics 480
41.4. Interpretation in terms of Studentized Mahalanobis Distance 480

Chapter 42. Three Principles for Testing a Linear Constraint 485
42.1. Mathematical Detail of the Three Approaches 486
42.2. Examples of Tests of Linear Hypotheses 489
42.3. The F-Test Statistic is a Function of the Likelihood Ratio 497
42.4. Tests of Nonlinear Hypotheses 498
42.5. Choosing Between Nonnested Models 498

Chapter 43. Multiple Comparisons in the Linear Model 499
43.1. Rectangular Confidence Regions 499

43.2. Relation between F-test and t-tests. 502
43.3. Large-Sample Simultaneous Confidence Regions 505

Chapter 44. Sample SAS Regression Output 507

Chapter 45. Flexible Functional Form 511
45.1. Categorical Variables: Regression with Dummies and Factors 511
45.2. Flexible Functional Form for Numerical Variables 513
45.3. More than One Explanatory Variable: Backfitting 520

Chapter 46. Transformation of the Response Variable 523
46.1. Alternating Least Squares and Alternating Conditional Expectations 523
46.2. Additivity and Variance Stabilizing Transformations (avas) 527
46.3. Comparing ace and avas 528

Chapter 47. Density Estimation 529
47.1. How to Measure the Precision of a Density Estimator 529
47.2. The Histogram 529
47.3. The Frequency Polygon 530
47.4. Kernel Densities 530
47.5. Transformational Kernel Density Estimators 531
47.6. Confidence Bands 531
47.7. Other Approaches to Density Estimation 532
47.8. Two-and Three-Dimensional Densities 532
47.9. Other Characterizations of Distributions 532
47.10. Quantile-Quantile Plots 532
47.11. Testing for Normality 534

Chapter 48. Measuring Economic Inequality 535
48.1. Web Resources about Income Inequality 535
48.2. Graphical Representations of Inequality 535
48.3. Quantitative Measures of Income Inequality 536
48.4. Properties of Inequality Measures 538

Chapter 49. Distributed Lags 539
49.1. Geometric lag 544
49.2. Autoregressive Distributed Lag Models 545

Chapter 50. Investment Models 551
50.1. Accelerator Models 551
50.2. Jorgenson's Model 552
50.3. Investment Function Project 554

Chapter 51. Distinguishing Random Variables from Variables Created by a Deterministic Chaotic Process 557
51.1. Empirical Methods: Grassberger-Procaccia Plots. 558

Chapter 52. Instrumental Variables 561

Chapter 53. Errors in Variables 567
53.1. The Simplest Errors-in-Variables Model 567
53.2. General Definition of the EV Model 571
53.3. Particular Forms of EV Models 573
53.4. The Identification Problem 575
53.5. Properties of Ordinary Least Squares in the EV model 580
53.6. Kalman's Critique of Malinvaud 583
53.7. Estimation if the EV Model is Identified 594
53.8. P-Estimation 597
53.9. Estimation When the Error Covariance Matrix is Exactly Known 603

Chapter 54. Dynamic Linear Models 605
54.1. Specification and Recursive Solution 605
54.2. Locally Constant Model 607
54.3. The Reference Model 610
54.4. Exchange Rate Forecasts 613
54.5. Company Market Share 616
54.6. Productivity in Milk Production 619

Chapter 55. Numerical Minimization 623

Chapter 56. Nonlinear Least Squares 627
56.1. The J Test 633
56.2. Nonlinear instrumental variables estimation 634

Chapter 57. Applications of GLS with Nonspherical Covariance Matrix 637
57.1. Cases when OLS and GLS are identical 637
57.2. Heteroskedastic Disturbances 638
57.3. Equicorrelated Covariance Matrix 639

Chapter 58. Unknown Parameters in the Covariance Matrix 643
58.1. Heteroskedasticity 643
58.2. Autocorrelation 648
58.3. Autoregressive Conditional Heteroskedasticity (ARCH) 660

Chapter 59. Generalized Method of Moments Estimators 663

Chapter 60. Bootstrap Estimators 669

Chapter 61. Random Coefficients 671

Chapter 62. Multivariate Regression 677
62.1. Multivariate Econometric Models: A Classification 677
62.2. Multivariate Regression with Equal Regressors 678
62.3. Growth Curve Models 684

Chapter 63. Independent Observations from the Same Multivariate Population 687
63.1. Notation and Basic Statistics 687
63.2. Two Geometries 689
63.3. Assumption of Normality 690
63.4. EM-Algorithm for Missing Observations 691
63.5. Wishart Distribution 694
63.6. Sample Correlation Coefficients 694

Chapter 64. Pooling of Cross Section and Time Series Data 697
64.1. OLS Model 697
64.2. The Between-Estimator 698
64.3. Dummy Variable Model (Fixed Effects) 699
64.4. Relation between the three Models so far: 702
64.5. Variance Components Model (Random Effects) 703

Chapter 65. Disturbance Related (Seemingly Unrelated) Regressions 709
65.1. The Supermatrix Representation 709
65.2. The Likelihood Function 711
65.3. Concentrating out the Covariance Matrix (Incomplete) 714
65.4. Situations in which OLS is Best 715
65.5. Unknown Covariance Matrix 718

Chapter 66. Simultaneous Equations Systems 721
66.1. Examples 721
66.2. General Mathematical Form 725
66.3. Indirect Least Squares 729
66.4. Instrumental Variables (2SLS) 730
66.5. Identification 732
66.6. Other Estimation Methods 734

Chapter 67. Timeseries Analysis 741
67.1. Covariance Stationary Timeseries 741
67.2. Vector Autoregressive Processes 748

67.3. Nonstationary Processes 753
67.4. Cointegration 755

Chapter 68. Seasonal Adjustment 757
68.1. Methods of Seasonal Adjustment 759
68.2. Seasonal Dummies in a Regression 760

Chapter 69. Binary Choice Models 767
69.1. Fisher's Scoring and Iteratively Reweighted Least Squares 767
69.2. Binary Dependent Variable 768
69.3. The Generalized Linear Model 771

Chapter 70. Multiple Choice Models 773

Appendix A. Matrix Formulas 775
A.1. A Fundamental Matrix Decomposition 775
A.2. The Spectral Norm of a Matrix 776
A.3. Inverses and g-Inverses of Matrices 776
A.4. Deficiency Matrices 778
A.5. Nonnegative Definite Symmetric Matrices 781
A.6. Projection Matrices 786
A.7. Determinants 788
A.8. More About Inverses 789
A.9. Eigenvalues and Singular Value Decomposition 793

Appendix B. Arrays of Higher Rank 797
B.1. Informal Survey of the Notation 797
B.2. Axiomatic Development of Array Operations 800
B.3. An Additional Notational Detail 805
B.4. Equality of Arrays and Extended Substitution 805
B.5. Vectorization and Kronecker Product 806

Appendix C. Matrix Differentiation 817
C.1. First Derivatives 817

Appendix. Bibliography 823

CHAPTER 1

Preface

These are class notes from several different graduate econometrics and statistics classes. In the Spring 2000 they were used for Statistics 6869, syllabus on p. **??**, and in the Fall 2000 for Economics 7800, syllabus on p. **??**. The notes give a careful and complete mathematical treatment intended to be accessible also to a reader inexperienced in math. There are 618 exercise questions, almost all with answers. The R-package `ecmet` has many of the datasets and R-functions needed in the examples. P. 279 gives instructions how to download it.

Here are some features by which these notes may differ from other teaching material available:

- A typographical distinction is made between random variables and the values taken by them (page 33).
- Best linear prediction of jointly distributed random variables is given as a second basic building block next to the least squares model (chapter 27).
- Appendix A gives a collection of general matrix formulas in which the g-inverse is used extensively.
- The "deficiency matrix," which gives an algebraic representation of the null space of a matrix, is defined and discussed in Appendix A.4.
- A molecule-like notation for concatenation of higher-dimensional arrays is introduced in Appendix B and used occasionally, see (10.5.7), (64.3.2), (65.0.18).

Other unusual treatments can be found in chapters/sections 3.11, 18.3, 25, 29, 40, 36, 41–42, and 64. There are a number of plots of density functions, confidence ellipses, and other graphs which use the full precision of TEX, and more will be added in the future. Some chapters are carefully elaborated, while others are still in the process of construction. In some topics covered in those notes I am an expert, in others I am still a beginner.

This edition also includes a number of comments from a critical realist perspective, inspired by [**Bha78**] and [**Bha93**]; see also [**Law89**]. There are many situations in the teaching of probability theory and statistics where the concept of totality, transfactual efficacy, etc., can and should be used. These comments are still

at an experimental state, and are the students are not required to know them for the exams. In the on-line version of the notes they are printed in a different color.

After some more cleaning out of the code, I am planning to make the $\mathcal{AMS}$-LATEX source files for these notes publicly available under the GNU public license, and upload them to the TEX-archive network CTAN. Since I am using Debian GNU/Linux, the materials will also be available as a `deb` archive.

The most up-to-date version will always be posted at the web site of the Economics Department of the University of Utah `www.econ.utah.edu/ehrbar/ecmet.pdf`. You can contact me by email at `ehrbar@econ.utah.edu`

Hans Ehrbar

CHAPTER 2

Probability Fields

2.1. The Concept of Probability

Probability theory and statistics are useful in dealing with the following types of situations:

- Games of chance: throwing dice, shuffling cards, drawing balls out of urns.
- Quality control in production: you take a sample from a shipment, count how many defectives.
- Actuarial Problems: the length of life anticipated for a person who has just applied for life insurance.
- Scientific Eperiments: you count the number of mice which contract cancer when a group of mice is exposed to cigarette smoke.
- Markets: the total personal income in New York State in a given month.
- Meteorology: the rainfall in a given month.
- Uncertainty: the exact date of Noah's birth.
- Indeterminacy: The closing of the Dow Jones industrial average or the temperature in New York City at 4 pm. on February 28, 2014.
- Chaotic determinacy: the relative frequency of the digit 3 in the decimal representation of π.
- Quantum mechanics: the proportion of photons absorbed by a polarization filter
- Statistical mechanics: the velocity distribution of molecules in a gas at a given pressure and temperature.

In the probability theoretical literature the situations in which probability theory applies are called "experiments," see for instance [**Rén70**, p. 1]. We will not use this terminology here, since probabilistic reasoning applies to several different types of situations, and not all these can be considered "experiments."

Problem 1. *(This question will not be asked on any exams) Rényi says: "Observing how long one has to wait for the departure of an airplane is an experiment." Comment.*

Answer. Rény commits the epistemic fallacy in order to justify his use of the word "experiment." Not the observation of the departure but the departure itself is the event which can be theorized probabilistically, and the word "experiment" is not appropriate here. □

What does the fact that probability theory is appropriate in the above situations tell us about the world? Let us go through our list one by one:

- Games of chance: Games of chance are based on the sensitivity on initial conditions: you tell someone to roll a pair of dice or shuffle a deck of cards, and despite the fact that this person is doing exactly what he or she is asked to do and produces an outcome which lies within a well-defined universe known beforehand (a number between 1 and 6, or a permutation of the deck of cards), the question *which* number or *which* permutation is beyond their control. The precise location and speed of the die or the precise order of the cards varies, and these small variations in initial conditions give rise, by the "butterfly effect" of chaos theory, to unpredictable final outcomes.

  A critical realist recognizes here the openness and stratification of the world: If many different influences come together, each of which is governed by laws, then their sum total is not determinate, as a naive hyper-determinist would think, but indeterminate. This is not only a condition for the possibility of science (in a hyper-deterministic world, one could not know anything before one knew everything, and science would also not be necessary because one could not do anything), but also for practical human activity: the macro outcomes of human practice are largely independent of micro detail (the postcard arrives whether the address is written in cursive or in printed letters, etc.). Games of chance are situations which deliberately project this micro indeterminacy into the macro world: the micro influences cancel each other out without one enduring influence taking over (as would be the case if the die were not perfectly symmetric and balanced) or deliberate human corrective activity stepping into the void (as a card trickster might do if the cards being shuffled somehow were distinguishable from the backside).

  The experiment in which one draws balls from urns shows clearly another aspect of this paradigm: the set of different possible outcomes is fixed beforehand, and the probability enters in the choice of one of these predetermined outcomes. This is not the only way probability can arise; it is an extensionalist example, in which the connection between success and failure is external. The world is not a collection of externally related outcomes collected in an urn. Success and failure are not determined by a choice between different spacially separated and individually inert balls (or playing cards or faces on a die), but it is the outcome of development and struggle that is internal to the individual unit.

- Quality control in production: you take a sample from a shipment, count how many defectives. Why is statistics and probability useful in production? Because production is work, it is not spontaneous. Nature does not voluntarily give us things in the form in which we need them. Production is similar to a scientific experiment because it is the attempt to create local closure. Such closure can never be complete, there are always leaks in it, through which irregularity enters.
- Actuarial Problems: the length of life anticipated for a person who has just applied for life insurance. Not only production, but also life itself is a struggle with physical nature, it is emergence. And sometimes it fails: sometimes the living organism is overwhelmed by the forces which it tries to keep at bay and to subject to its own purposes.
- Scientific Eperiments: you count the number of mice which contract cancer when a group of mice is exposed to cigarette smoke: There is local closure regarding the conditions under which the mice live, but even if this closure were complete, individual mice would still react differently, because of genetic differences. No two mice are exactly the same, and despite these differences they are still mice. This is again the stratification of reality. Two mice are two different individuals but they are both mice. Their reaction to the smoke is not identical, since they are different individuals, but it is not completely capricious either, since both are mice. It can be predicted probabilistically. Those mechanisms which make them mice react to the smoke. The probabilistic regularity comes from the transfactual efficacy of the mouse organisms.
- Meteorology: the rainfall in a given month. It is very fortunate for the development of life on our planet that we have the chaotic alternation between cloud cover and clear sky, instead of a continuous cloud cover as in Venus or a continuous clear sky. Butterfly effect all over again, but it is possible to make probabilistic predictions since the fundamentals remain stable: the transfactual efficacy of the energy received from the sun and radiated back out into space.
- Markets: the total personal income in New York State in a given month. Market economies are a very much like the weather; planned economies would be more like production or life.
- Uncertainty: the exact date of Noah's birth. This is epistemic uncertainty: assuming that Noah was a real person, the date exists and we know a time range in which it must have been, but we do not know the details. Probabilistic methods can be used to represent this kind of uncertain knowledge, but other methods to represent this knowledge may be more appropriate.

- Indeterminacy: The closing of the Dow Jones Industrial Average (DJIA) or the temperature in New York City at 4 pm. on February 28, 2014: This is ontological uncertainty, not only epistemological uncertainty. Not only do we not know it, but it is objectively not yet decided what these data will be. Probability theory has limited applicability for the DJIA since it cannot be expected that the mechanisms determining the DJIA will be the same at that time, therefore we cannot base ourselves on the transfactual efficacy of some stable mechanisms. It is not known which stocks will be included in the DJIA at that time, or whether the US dollar will still be the world reserve currency and the New York stock exchange the pinnacle of international capital markets. Perhaps a different stock market index located somewhere else will at that time play the role the DJIA is playing today. We would not even be able to ask questions about that alternative index today.

  Regarding the temperature, it is more defensible to assign a probability, since the weather mechanisms have probably stayed the same, except for changes in global warming (unless mankind has learned by that time to manipulate the weather locally by cloud seeding etc.).
- Chaotic determinacy: the relative frequency of the digit 3 in the decimal representation of π: The laws by which the number π is defined have very little to do with the procedure by which numbers are expanded as decimals, therefore the former has no systematic influence on the latter. (It has an influence, but not a systematic one; it is the error of actualism to think that every influence must be systematic.) But it is also known that laws can have remote effects: one of the most amazing theorems in mathematics is the formula $\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{4} + \cdots$ which estalishes a connection between the geometry of the circle and some simple arithmetics.
- Quantum mechanics: the proportion of photons absorbed by a polarization filter: If these photons are already polarized (but in a different direction than the filter) then this is not epistemic uncertainty but ontological indeterminacy, since the polarized photons form a pure state, which is atomic in the algebra of events. In this case, the distinction between epistemic uncertainty and ontological indeterminacy is operational: the two alternatives follow different mathematics.
- Statistical mechanics: the velocity distribution of molecules in a gas at a given pressure and temperature. Thermodynamics cannot be reduced to the mechanics of molecules, since mechanics is reversible in time, while thermodynamics is not. An additional element is needed, which can be modeled using probability.

PROBLEM 2. *Not every kind of uncertainty can be formulated stochastically. Which other methods are available if stochastic means are inappropriate?*

ANSWER. Dialectics. □

PROBLEM 3. *How are the probabilities of rain in weather forecasts to be interpreted?*

ANSWER. Renyi in [**Rén70**, pp. 33/4]: "By saying that the probability of rain tomorrow is 80% (or, what amounts to the same, 0.8) the meteorologist means that in a situation similar to that observed on the given day, there is usually rain on the next day in about 8 out of 10 cases; thus, while it is not certain that it will rain tomorrow, the *degree of certainty* of this event is 0.8." □

Pure uncertainty is as hard to generate as pure certainty; it is needed for encryption and numerical methods.

Here is an encryption scheme which leads to a random looking sequence of numbers (see [**Rao97**, p. 13]): First a string of binary random digits is generated which is known only to the sender and receiver. The sender converts his message into a string of binary digits. He then places the message string below the key string and obtains a coded string by changing every message bit to its alternative at all places where the key bit is 1 and leaving the others unchanged. The coded string which appears to be a random binary sequence is transmitted. The received message is decoded by making the changes in the same way as in encrypting using the key string which is known to the receiver.

PROBLEM 4. *Why is it important in the above encryption scheme that the key string is purely random and does not have any regularities?*

PROBLEM 5. [**Knu81**, pp. 7, 452] *Suppose you wish to obtain a decimal digit at random, not using a computer. Which of the following methods would be suitable?*

• a. *Open a telephone directory to a random place (i.e., stick your finger in it somewhere) and use the unit digit of the first number found on the selected page.*

ANSWER. This will often fail, since users select "round" numbers if possible. In some areas, telephone numbers are perhaps assigned randomly. But it is a mistake in any case to try to get several successive random numbers from the same page, since many telephone numbers are listed several times in a sequence. □

• b. *Same as a, but use the units digit of the* page *number.*

ANSWER. But do you use the left-hand page or the right-hand page? Say, use the left-hand page, divide by 2, and use the units digit. □

• c. *Roll a die which is in the shape of a regular icosahedron, whose twenty faces have been labeled with the digits 0, 0, 1, 1,. . . , 9, 9. Use the digit which appears on top, when the die comes to rest. (A felt table with a hard surface is recommended for rolling dice.)*

ANSWER. The markings on the face will slightly bias the die, but for practical purposes this method is quite satisfactory. See Math. Comp. 15 (1961), 94–95, for further discussion of these dice. □

• d. *Expose a geiger counter to a source of radioactivity for one minute (shielding yourself) and use the unit digit of the resulting count. (Assume that the geiger counter displays the number of counts in decimal notation, and that the count is initially zero.)*

ANSWER. This is a difficult question thrown in purposely as a surprise. The number is *not* uniformly distributed! One sees this best if one imagines the source of radioactivity is very low level, so that only a few emissions can be expected during this minute. If the average number of emissions per minute is λ, the probability that the counter registers k is $e^{-\lambda}\lambda^k/k!$ (the Poisson distribution). So the digit 0 is selected with probability $e^{-\lambda}\sum_{k=0}^{\infty}\lambda^{10k}/(10k)!$, etc. □

• e. *Glance at your wristwatch, and if the position of the second-hand is between* $6n$ *and* $6(n+1)$*, choose the digit* n*.*

ANSWER. Okay, provided that the time since the last digit selected in this way is random. A bias may arise if borderline cases are not treated carefully. A better device seems to be to use a stopwatch which has been started long ago, and which one stops arbitrarily, and then one has all the time necessary to read the display. □

• f. *Ask a friend to think of a random digit, and use the digit he names.*

ANSWER. No, people usually think of certain digits (like 7) with higher probability. □

• g. *Assume 10 horses are entered in a race and you know nothing whatever about their qualifications. Assign to these horses the digits 0 to 9, in arbitrary fashion, and after the race use the winner's digit.*

ANSWER. Okay; your assignment of numbers to the horses had probability 1/10 of assigning a given digit to a winning horse. □

2.2. Events as Sets

With every situation with uncertain outcome we associate its *sample space* U, which represents the set of all possible outcomes (described by the characteristics which we are interested in).

Events are associated with subsets of the sample space, i.e., with bundles of outcomes that are observable in the given experimental setup. The set of all events we denote with $\mathcal{F}$. ($\mathcal{F}$ is a set of subsets of U.)

Look at the example of rolling a die. $U = \{1, 2, 3, 4, 5, 6\}$. The events of getting an even number is associated with the subset $\{2, 4, 6\}$; getting a six with $\{6\}$; not getting a six with $\{1, 2, 3, 4, 5\}$, etc. Now look at the example of rolling two indistinguishable dice. Observable events may be: getting two ones, getting a one and a two, etc. But we cannot distinguish between the first die getting a one and the second a

two, and vice versa. I.e., if we define the sample set to be $U = \{1, \ldots, 6\} \times \{1, \ldots, 6\}$, i.e., the set of all pairs of numbers between 1 and 6, then certain subsets are not observable. $\{(1, 5)\}$ is not observable (unless the dice are marked or have different colors etc.), only $\{(1, 5), (5, 1)\}$ is observable.

If the experiment is measuring the height of a person in meters, and we make the idealized assumption that the measuring instrument is infinitely accurate, then all possible outcomes are numbers between 0 and 3, say. Sets of outcomes one is usually interested in are whether the height falls within a given interval; therefore all intervals within the given range represent observable events.

If the sample space is finite or countably infinite, very often *all* subsets are observable events. If the sample set contains an uncountable continuum, it is not desirable to consider all subsets as observable events. Mathematically one can define quite crazy subsets which have no practical significance and which cannot be meaningfully given probabilities. For the purposes of Econ 7800, it is enough to say that all the subsets which we may reasonably define are candidates for observable events.

The "set of all possible outcomes" is well defined in the case of rolling a die and other games; but in social sciences, situations arise in which the outcome is open and the range of possible outcomes cannot be known beforehand. If one uses a probability theory based on the concept of a "set of possible outcomes" in such a situation, one reduces a process which is open and evolutionary to an imaginary predetermined and static "set." Furthermore, in social theory, the mechanism by which these uncertain outcomes are generated are often internal to the members of the statistical population. The mathematical framework models these mechanisms as an extraneous "picking an element out of a pre-existing set."

From given observable events we can derive new observable events by *set theoretical operations*. (All the operations below involve subsets of the same U.)

Mathematical Note: Notation of sets: there are two ways to denote a set: either by giving a rule, or by listing the elements. (The order in which the elements are listed, or the fact whether some elements are listed twice or not, is irrelevant.)

Here are the formal definitions of set theoretic operations. The letters A, B, etc. denote subsets of a given set U (events), and I is an arbitrary index set. ω stands

for an element, and $\omega \in A$ means that ω is an element of A.

$$A \subset B \iff (\omega \in A \Rightarrow \omega \in B) \qquad (A \text{ is contained in } B) \tag{2.2.1}$$

$$A \cap B = \{\omega \colon \omega \in A \text{ and } \omega \in B\} \qquad (\text{intersection of } A \text{ and } B) \tag{2.2.2}$$

$$\bigcap_{i \in I} A_i = \{\omega \colon \omega \in A_i \text{ for all } i \in I\} \tag{2.2.3}$$

$$A \cup B = \{\omega \colon \omega \in A \text{ or } \omega \in B\} \qquad (\text{union of } A \text{ and } B) \tag{2.2.4}$$

$$\bigcup_{i \in I} A_i = \{\omega \colon \text{ there exists an } i \in I \text{ such that } \omega \in A_i\} \tag{2.2.5}$$

$$U \quad \text{Universal set: all } \omega \text{ we talk about are } \in U. \tag{2.2.6}$$

$$A' = \{\omega \colon \omega \notin A \text{ but } \omega \in U\} \tag{2.2.7}$$

$$\emptyset = \text{ the empty set: } \omega \notin \emptyset \text{ for all } \omega. \tag{2.2.8}$$

These definitions can also be visualized by Venn diagrams; and for the purposes of this class, demonstrations with the help of Venn diagrams will be admissible in lieu of mathematical proofs.

Problem 6. *For the following set-theoretical exercises it is sufficient that you draw the corresponding Venn diagrams and convince yourself by just looking at them that the statement is true. For those who are interested in a precise mathematical proof derived from the definitions of $A \cup B$ etc. given above, should remember that a proof of the set-theoretical identity $A = B$ usually has the form: first you show that $\omega \in A$ implies $\omega \in B$, and then you show the converse.*

- a. *Prove that $A \cup B = B \iff A \cap B = A$.*

Answer. If one draws the Venn diagrams, one can see that either side is true if and only if $A \subset B$. If one wants a more precise proof, the following proof by contradiction seems most illuminating: Assume the lefthand side does not hold, i.e., there exists a $\omega \in A$ but $\omega \notin B$. Then $\omega \notin A \cap B$, i.e., $A \cap B \neq A$. Now assume the righthand side does not hold, i.e., there is a $\omega \in A$ with $\omega \notin B$. This ω lies in $A \cup B$ but not in B, i.e., the lefthand side does not hold either.

□

- b. *Prove that $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$*

Answer. If $\omega \in A$ then it is clearly always in the righthand side and in the lefthand side. If there is therefore any difference between the righthand and the lefthand side, it must be for the $\omega \notin A$: If $\omega \notin A$ and it is still in the lefthand side then it must be in $B \cap C$, therefore it is also in the righthand side. If $\omega \notin A$ and it is in the righthand side, then it must be both in B and in C, therefore it is in the lefthand side.

□

- c. *Prove that $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.*

Answer. If $\omega \notin A$ then it is clearly neither in the righthand side nor in the lefthand side. If there is therefore any difference between the righthand and the lefthand side, it must be for the $\omega \in A$: If $\omega \in A$ and it is in the lefthand side then it must be in $B \cup C$, i.e., in B or in C or in both, therefore it is also in the righthand side. If $\omega \in A$ and it is in the righthand side, then it must be in either B or C or both, therefore it is in the lefthand side. □

- d. *Prove that* $A \cap \left(\bigcup_{i=1}^{\infty} B_i\right) = \bigcup_{i=1}^{\infty}(A \cap B_i)$.

Answer. Proof: If ω in lefthand side, then it is in A and in at least one of the B_i, say it is in B_k. Therefore it is in $A \cap B_k$, and therefore it is in the righthand side. Now assume, conversely, that ω is in the righthand side; then it is at least in one of the $A \cap B_i$, say it is in $A \cap B_k$. Hence it is in A and in B_k, i.e., in A and in $\bigcup B_i$, i.e., it is in the lefthand side. □

Problem 7. *3 points Draw a Venn Diagram which shows the validity of de Morgan's laws:* $(A \cup B)' = A' \cap B'$ *and* $(A \cap B)' = A' \cup B'$. *If done right, the same Venn diagram can be used for both proofs.*

Answer. There is a proof in [**HT83**, p. 12]. Draw A and B inside a box which represents U, and shade A' from the left (blue) and B' from the right (yellow), so that $A' \cap B'$ is cross shaded (green); then one can see these laws. □

Problem 8. *3 points* [**HT83**, Exercise 1.2-13 on p. 14] *Evaluate the following unions and intersections of intervals. Use the notation* (a, b) *for open and* $[a, b]$ *for closed intervals,* $(a, b]$ *or* $[a, b)$ *for half open intervals,* $\{a\}$ *for sets containing one element only, and* $\emptyset$ *for the empty set.*

$$\bigcup_{n=1}^{\infty}\left(\frac{1}{n}, 2\right) = \qquad \bigcap_{n=1}^{\infty}\left(0, \frac{1}{n}\right) = \tag{2.2.9}$$

$$\bigcup_{n=1}^{\infty}\left[\frac{1}{n}, 2\right] = \qquad \bigcap_{n=1}^{\infty}\left[0, 1 + \frac{1}{n}\right] = \tag{2.2.10}$$

Answer.

$$\bigcup_{n=1}^{\infty}\left(\frac{1}{n}, 2\right) = (0, 2) \qquad \bigcap_{n=1}^{\infty}\left(0, \frac{1}{n}\right) = \emptyset \tag{2.2.11}$$

$$\bigcup_{n=1}^{\infty}\left[\frac{1}{n}, 2\right] = (0, 2] \qquad \bigcap_{n=1}^{\infty}\left[0, 1 + \frac{1}{n}\right] = [0, 1] \tag{2.2.12}$$

Explanation of $\bigcup_{n=1}^{\infty}\left[\frac{1}{n}, 2\right]$: for every α with $0 < \alpha \leq 2$ there is a n with $\frac{1}{n} \leq \alpha$, but 0 itself is in none of the intervals. □

The set operations become logical operations if applied to events. Every experiment returns an element $\omega \in U$ as outcome. Here ω is rendered green in the electronic version of these notes (and in an upright font in the version for black-and-white printouts), because ω does not denote a specific element of U, but it depends

on chance which element is picked. I.e., the green color (or the unusual font) indicate that ω is "alive." We will also render the events themselves (as opposed to their set-theoretical counterparts) in green (or in an upright font).

- We say that the event A has *occurred* when $\omega \in A$.
- If $A \subset B$ then event A implies event B, and we will write this directly in terms of events as $\mathsf{A} \subset \mathsf{B}$.
- The set $A \cap B$ is associated with the event that both A and B occur (e.g. an even number smaller than six), and considered as an event, not a set, the event that both A and B occur will be written $\mathsf{A} \cap \mathsf{B}$.
- Likewise, $\mathsf{A} \cup \mathsf{B}$ is the event that either A or B, or both, occur.
- A' is the event that A does not occur.
- U the event that always occurs (as long as one performs the experiment).
- The empty set $\emptyset$ is associated with the impossible event $\emptyset$, because whatever the value ω of the chance outcome ω of the experiment, it is always $\omega \notin \emptyset$.

If $A \cap B = \emptyset$, the set theoretician calls A and B "disjoint," and the probability theoretician calls the events A and B "mutually exclusive." If $\mathsf{A} \cup \mathsf{B} = \mathsf{U}$, then A and B are called "collectively exhaustive."

The set $\mathcal{F}$ of all observable events must be a σ-algebra, i.e., it must satisfy:

$$\emptyset \in \mathcal{F}$$

$$\mathsf{A} \in \mathcal{F} \Rightarrow \mathsf{A}' \in \mathcal{F}$$

$$\mathsf{A}_1, \mathsf{A}_2, \ldots \in \mathcal{F} \Rightarrow \mathsf{A}_1 \cup \mathsf{A}_2 \cup \cdots \in \mathcal{F} \quad \text{which can also be written as} \quad \bigcup_{i=1,2,\ldots} \mathsf{A}_i \in \mathcal{F}$$

$$\mathsf{A}_1, \mathsf{A}_2, \ldots \in \mathcal{F} \Rightarrow \mathsf{A}_1 \cap \mathsf{A}_2 \cap \cdots \in \mathcal{F} \quad \text{which can also be written as} \quad \bigcap_{i=1,2,\ldots} \mathsf{A}_i \in \mathcal{F}.$$

2.3. The Axioms of Probability

A probability measure $\Pr : \mathcal{F} \to \mathbb{R}$ is a mapping which assigns to every event a number, the probability of this event. This assignment must be compatible with the set-theoretic operations between events in the following way:

$$\Pr[\mathsf{U}] = 1 \tag{2.3.1}$$

$$\Pr[\mathsf{A}] \geq 0 \quad \text{for all events } \mathsf{A} \tag{2.3.2}$$

$$\text{If } \mathsf{A}_i \cap \mathsf{A}_j = \emptyset \text{ for all } i, j \text{ with } i \neq j \text{ then} \quad \Pr[\bigcup_{i=1}^{\infty} \mathsf{A}_i] = \sum_{i=1}^{\infty} \Pr[\mathsf{A}_i] \tag{2.3.3}$$

Here an infinite sum is mathematically defined as the limit of partial sums. These axioms make probability what mathematicians call a *measure*, like area or weight.

In a Venn diagram, one might therefore interpret the probability of the events as the *area* of the bubble representing the event.

PROBLEM 9. *Prove that* $\Pr[\mathsf{A}'] = 1 - \Pr[\mathsf{A}]$.

ANSWER. Follows from the fact that A and A' are disjoint and their union U has probability 1. □

PROBLEM 10. *2 points Prove that* $\Pr[\mathsf{A} \cup \mathsf{B}] = \Pr[\mathsf{A}] + \Pr[\mathsf{B}] - \Pr[\mathsf{A} \cap \mathsf{B}]$.

ANSWER. For Econ 7800 it is sufficient to argue it out intuitively: if one adds $\Pr[\mathsf{A}] + \Pr[\mathsf{B}]$ then one counts $\Pr[\mathsf{A} \cap \mathsf{B}]$ twice and therefore has to subtract it again.

The brute force mathematical proof guided by this intuition is somewhat verbose: Define $\mathsf{D} = \mathsf{A} \cap \mathsf{B}'$, $\mathsf{E} = \mathsf{A} \cap \mathsf{B}$, and $\mathsf{F} = \mathsf{A}' \cap \mathsf{B}$. D, E, and F satisfy

$$\mathsf{D} \cup \mathsf{E} = (\mathsf{A} \cap \mathsf{B}') \cup (\mathsf{A} \cap \mathsf{B}) = \mathsf{A} \cap (\mathsf{B}' \cup \mathsf{B}) = \mathsf{A} \cap \mathsf{U} = \mathsf{A}, \tag{2.3.4}$$

$$\mathsf{E} \cup \mathsf{F} = \mathsf{B}, \tag{2.3.5}$$

$$\mathsf{D} \cup \mathsf{E} \cup \mathsf{F} = \mathsf{A} \cup \mathsf{B}. \tag{2.3.6}$$

You may need some of the properties of unions and intersections in Problem 6. Next step is to prove that D, E, and F are mutually exclusive. Therefore it is easy to take probabilities

$$\Pr[\mathsf{A}] = \Pr[\mathsf{D}] + \Pr[\mathsf{E}]; \tag{2.3.7}$$

$$\Pr[\mathsf{B}] = \Pr[\mathsf{E}] + \Pr[\mathsf{F}]; \tag{2.3.8}$$

$$\Pr[\mathsf{A} \cup \mathsf{B}] = \Pr[\mathsf{D}] + \Pr[\mathsf{E}] + \Pr[\mathsf{F}]. \tag{2.3.9}$$

Take the sum of (2.3.7) and (2.3.8), and subtract (2.3.9):

$$\Pr[\mathsf{A}] + \Pr[\mathsf{B}] - \Pr[\mathsf{A} \cup \mathsf{B}] = \Pr[\mathsf{E}] = \Pr[\mathsf{A} \cap \mathsf{B}]; \tag{2.3.10}$$

A shorter but trickier alternative proof is the following. First note that $\mathsf{A} \cup \mathsf{B} = \mathsf{A} \cup (\mathsf{A}' \cap \mathsf{B})$ and that this is a disjoint union, i.e., $\Pr[\mathsf{A} \cup \mathsf{B}] = \Pr[\mathsf{A}] + \Pr[\mathsf{A}' \cap \mathsf{B}]$. Then note that $\mathsf{B} = (\mathsf{A} \cap \mathsf{B}) \cup (\mathsf{A}' \cap \mathsf{B})$, and this is a disjoint union, therefore $\Pr[\mathsf{B}] = \Pr[\mathsf{A} \cap \mathsf{B}] + \Pr[\mathsf{A}' \cap \mathsf{B}]$, or $\Pr[\mathsf{A}' \cap \mathsf{B}] = \Pr[\mathsf{B}] - \Pr[\mathsf{A} \cap \mathsf{B}]$. Putting this together gives the result.

□

PROBLEM 11. *1 point Show that for arbitrary events* A *and* B, $\Pr[\mathsf{A} \cup \mathsf{B}] \leq \Pr[\mathsf{A}] + \Pr[\mathsf{B}]$.

ANSWER. From Problem 10 we know that $\Pr[\mathsf{A} \cup \mathsf{B}] = \Pr[\mathsf{A}] + \Pr[\mathsf{B}] - \Pr[\mathsf{A} \cap \mathsf{B}]$, and from axiom (2.3.2) follows $\Pr[\mathsf{A} \cap \mathsf{B}] \geq 0$. □

PROBLEM 12. *2 points (Bonferroni inequality) Let* A *and* B *be two events. Writing* $\Pr[\mathsf{A}] = 1 - \alpha$ *and* $\Pr[\mathsf{B}] = 1 - \beta$, *show that* $\Pr[\mathsf{A} \cap \mathsf{B}] \geq 1 - (\alpha + \beta)$. *You are allowed to use that* $\Pr[\mathsf{A} \cup \mathsf{B}] = \Pr[\mathsf{A}] + \Pr[\mathsf{B}] - \Pr[\mathsf{A} \cap \mathsf{B}]$ *(Problem 10), and that all probabilities are* ≤ 1.

ANSWER.

$$\Pr[\mathsf{A} \cup \mathsf{B}] = \Pr[\mathsf{A}] + \Pr[\mathsf{B}] - \Pr[\mathsf{A} \cap \mathsf{B}] \leq 1 \tag{2.3.11}$$

$$\Pr[\mathsf{A}] + \Pr[\mathsf{B}] \leq 1 + \Pr[\mathsf{A} \cap \mathsf{B}] \tag{2.3.12}$$

$$\Pr[\mathsf{A}] + \Pr[\mathsf{B}] - 1 \leq \Pr[\mathsf{A} \cap \mathsf{B}] \tag{2.3.13}$$

$$1 - \alpha + 1 - \beta - 1 = 1 - \alpha - \beta \leq \Pr[\mathsf{A} \cap \mathsf{B}] \tag{2.3.14}$$

□

PROBLEM 13. *(Not eligible for in-class exams) Given a rising sequence of events* $\mathsf{B}_1 \subset \mathsf{B}_2 \subset \mathsf{B}_3 \cdots$, *define* $\mathsf{B} = \bigcup_{i=1}^{\infty} \mathsf{B}_i$. *Show that* $\Pr[\mathsf{B}] = \lim_{i \to \infty} \Pr[\mathsf{B}_i]$.

ANSWER. Define $\mathsf{C}_1 = \mathsf{B}_1$, $\mathsf{C}_2 = \mathsf{B}_2 \cap \mathsf{B}'_1$, $\mathsf{C}_3 = \mathsf{B}_3 \cap \mathsf{B}'_2$, etc. Then $\mathsf{C}_i \cap \mathsf{C}_j = \emptyset$ for $i \neq j$, and $\mathsf{B}_n = \bigcup_{i=1}^{n} \mathsf{C}_i$ and $\mathsf{B} = \bigcup_{i=1}^{\infty} \mathsf{C}_i$. In other words, now we have represented every B_n and B as a union of disjoint sets, and can therefore apply the third probability axiom (2.3.3): $\Pr[\mathsf{B}] = \sum_{i=1}^{\infty} \Pr[\mathsf{C}_i]$. The infinite sum is merely a short way of writing $\Pr[\mathsf{B}] = \lim_{n \to \infty} \sum_{i=1}^{n} \Pr[\mathsf{C}_i]$, i.e., the infinite sum is the limit of the finite sums. But since these finite sums are exactly $\sum_{i=1}^{n} \Pr[\mathsf{C}_i] = \Pr[\bigcup_{i=1}^{n} \mathsf{C}_i] = \Pr[\mathsf{B}_n]$, the assertion follows. This proof, as it stands, is for our purposes entirely acceptable. One can make some steps in this proof still more stringent. For instance, one might use induction to prove $\mathsf{B}_n = \bigcup_{i=1}^{n} \mathsf{C}_i$. And how does one show that $\mathsf{B} = \bigcup_{i=1}^{\infty} \mathsf{C}_i$? Well, one knows that $\mathsf{C}_i \subset \mathsf{B}_i$, therefore $\bigcup_{i=1}^{\infty} \mathsf{C}_i \subset \bigcup_{i=1}^{\infty} \mathsf{B}_i = \mathsf{B}$. Now take an $\omega \in \mathsf{B}$. Then it lies in at least one of the B_i, but it can be in many of them. Let k be the smallest k for which $\omega \in \mathsf{B}_k$. If $k = 1$, then $\omega \in \mathsf{C}_1 = \mathsf{B}_1$ as well. Otherwise, $\omega \notin \mathsf{B}_{k-1}$, and therefore $\omega \in \mathsf{C}_k$. I.e., any element in B lies in at least one of the C_k, therefore $\mathsf{B} \subset \bigcup_{i=1}^{\infty} \mathsf{C}_i$. □

PROBLEM 14. *(Not eligible for in-class exams) From problem 13 derive also the following: if* $\mathsf{A}_1 \supset \mathsf{A}_2 \supset \mathsf{A}_3 \cdots$ *is a declining sequence, and* $\mathsf{A} = \bigcap_i \mathsf{A}_i$, *then* $\Pr[\mathsf{A}] = \lim \Pr[\mathsf{A}_i]$.

ANSWER. If the A_i are declining, then their complements $\mathsf{B}_i = \mathsf{A}'_i$ are rising: $\mathsf{B}_1 \subset \mathsf{B}_2 \subset \mathsf{B}_3 \cdots$ are rising; therefore I know the probability of $\mathsf{B} = \bigcup \mathsf{B}_i$. Since by de Morgan's laws, $\mathsf{B} = \mathsf{A}'$, this gives me also the probability of A. □

The results regarding the probabilities of rising or declining sequences are equivalent to the third probability axiom. This third axiom can therefore be considered a continuity condition for probabilities.

If U is finite or countably infinite, then the probability measure is uniquely determined if one knows the probability of every one-element set. We will call $\Pr[\{\omega\}] = p(\omega)$ the probability mass function. Other terms used for it in the literature are probability function, or even probability density function (although it is *not* a density, more about this below). If U has more than countably infinite elements, the probabilities of one-element sets may not give enough information to define the whole probability measure.

Mathematical Note: Not all infinite sets are countable. Here is a proof, by contradiction, that the real numbers between 0 and 1 are not countable: assume there is an enumeration, i.e., a sequence $a_1, a_2, \ldots$ which contains them all. Write

them underneath each other in their (possibly infinite) decimal representation, where $0.d_{i1}d_{i2}d_{i3}\ldots$ is the decimal representation of a_i. Then any real number whose decimal representation is such that the first digit is *not* equal to d_{11}, the second digit is *not* equal d_{22}, the third *not* equal d_{33}, etc., is a real number which is *not* contained in this enumeration. That means, an enumeration which contains all real numbers cannot exist.

On the real numbers between 0 and 1, the length measure (which assigns to each interval its length, and to sets composed of several invervals the sums of the lengths, etc.) is a probability measure. In this probability field, every one-element subset of the sample set has zero probability.

This shows that events other than $\emptyset$ may have zero probability. In other words, if an event has probability 0, this does not mean it is logically impossible. It may well happen, but it happens so infrequently that in repeated experiments the *average* number of occurrences converges toward zero.

2.4. Objective and Subjective Interpretation of Probability

The mathematical probability axioms apply to both objective and subjective interpretation of probability.

The *objective* interpretation considers probability a quasi physical property of the experiment. One cannot simply say: $\Pr[\mathsf{A}]$ is the relative frequency of the occurrence of A, because we know intuitively that this frequency does not necessarily converge. E.g., even with a fair coin it is physically possible that one always gets head, or that one gets some other sequence which does not converge towards $\frac{1}{2}$. The above axioms resolve this dilemma, because they allow to derive the theorem that the relative frequencies converges towards the probability *with probability one*.

Subjectivist interpretation (de Finetti: "probability does not exist") defines probability in terms of people's ignorance and willingness to take bets. Interesting for economists because it uses money and utility, as in expected utility. Call "a lottery on A" a lottery which pays \$1 if A occurs, and which pays nothing if A does not occur. If a person is willing to pay p dollars for a lottery on A and $1-p$ dollars for a lottery on A', then, according to a subjectivist definition of probability, he assigns subjective probability p to A.

There is the presumption that his willingness to bet does not depend on the size of the payoff (i.e., the payoffs are considered to be small amounts).

Problem 15. *Assume A, B, and C are a complete disjunction of events, i.e., they are mutually exclusive and $\mathsf{A}\cup\mathsf{B}\cup\mathsf{C}=\mathsf{U}$, the universal set.*

• a. *1 point Arnold assigns subjective probability p to A, q to B, and r to C. Explain exactly what this means.*

Answer. We know six different bets which Arnold is always willing to make, not only on A, B, and C, but also on their complements. □

• b. *1 point Assume that $p+q+r>1$. Name three lotteries which Arnold would be willing to buy, the net effect of which would be that he loses with certainty.*

Answer. Among those six we have to pick subsets that make him a sure loser. If $p+q+r>1$, then we sell him a bet on A, one on B, and one on C. The payoff is always 1, and the cost is $p+q+r>1$. □

• c. *1 point Now assume that $p+q+r<1$. Name three lotteries which Arnold would be willing to buy, the net effect of which would be that he loses with certainty.*

Answer. If $p+q+r<1$, then we sell him a bet on A', one on B', and one on C'. The payoff is 2, and the cost is $1-p+1-q+1-r>2$. □

• d. *1 point Arnold is therefore only coherent if $\Pr[\mathsf{A}]+\Pr[\mathsf{B}]+\Pr[\mathsf{C}]=1$. Show that the additivity of probability can be derived from coherence, i.e., show that any subjective probability that satisfies the rule: whenever A, B, and C is a complete disjunction of events, then the sum of their probabilities is 1, is additive, i.e., $\Pr[\mathsf{A}\cup\mathsf{B}]=\Pr[\mathsf{A}]+\Pr[\mathsf{B}]$.*

Answer. Since r is his subjective probability of C, $1-r$ must be his subjective probability of $\mathsf{C}'=\mathsf{A}\cup\mathsf{B}$. Since $p+q+r=1$, it follows $1-r=p+q$. □

This last problem indicates that the finite additivity axiom follows from the requirement that the bets be consistent or, as subjectivists say, "coherent" with each other. However, it is not possible to derive the additivity for countably *infinite* sequences of events from such an argument.

2.5. Counting Rules

In this section we will be working in a *finite* probability space, in which all atomic events have equal probabilities. The acts of rolling dice or drawing balls from urns can be modeled by such spaces. In order to compute the probability of a given event, one must *count* the elements of the set which this event represents. In other words, we *count* how many different ways there are to achieve a certain outcome. This can be tricky, and we will develop some general principles how to do it.

Problem 16. *You throw two dice.*

• a. *1 point What is the probability that the sum of the numbers shown is five or less?*

Answer. $\begin{array}{llll}11&12&13&14\\21&22&23&\\31&32&&\\41&&&\end{array}$, i.e., 10 out of 36 possibilities, gives the probability $\frac{5}{18}$. □

• b. *1 point What is the probability that both of the numbers shown are five or less?*

Answer. $\begin{array}{ccccc} 11 & 12 & 13 & 14 & 15 \\ 21 & 22 & 23 & 24 & 25 \\ 31 & 32 & 33 & 34 & 35 \\ 41 & 42 & 43 & 44 & 45 \\ 51 & 52 & 53 & 54 & 55 \end{array}$, i.e., $\frac{25}{36}$. □

• c. *2 points What is the probability that the maximum of the two numbers shown is five? (As a clarification: if the first die shows 4 and the second shows 3 then the maximum of the numbers shown is 4.)*

Answer. $\begin{array}{ccccc} & & & & 15 \\ & & & & 25 \\ & & & & 35 \\ & & & & 45 \\ 51 & 52 & 53 & 54 & 55 \end{array}$, i.e., $\frac{1}{4}$. □

In this and in similar questions to follow, the answer should be given as a fully shortened fraction.

The *multiplication principle* is a basic aid in counting: If the first operation can be done n_1 ways, and the second operation n_2 ways, then the total can be done $n_1 n_2$ ways.

Definition: A permutation of a set is its arrangement in a certain order. It was mentioned earlier that for a set it does not matter in which order the elements are written down; the number of permutations is therefore the number of ways a given set can be written down without repeating its elements. From the multiplication principle follows: the number of permutations of a set of n elements is $n(n-1)(n-2)\cdots(2)(1) = n!$ (n factorial). By definition, $0! = 1$.

If one does not arrange the whole set, but is interested in the number of k-tuples made up of *distinct* elements of the set, then the number of possibilities is $n(n-1)(n-2)\cdots(n-k+2)(n-k+1) = \frac{n!}{(n-k)!}$. (Start with n and the number of factors is k.) (k-tuples are sometimes called *ordered* k-tuples because the order in which the elements are written down matters.) [**Ame94**, p. 8] uses the notation P^n_k for this.

This leads us to the next question: how many k-element subsets does a n-element set have? We already know how many permutations into k elements it has; but always $k!$ of these permutations represent the same subset; therefore we have to divide by $k!$. The number of k-element subsets of an n-element set is therefore

$$\frac{n!}{k!(n-k)!} = \frac{n(n-1)(n-2)\cdots(n-k+1)}{(1)(2)(3)\cdots k} = \binom{n}{k}, \tag{2.5.1}$$

It is pronounced as n choose k, and is also called a "binomial coefficient." It is defined for all $0 \le k \le n$. [**Ame94**, p. 8] calls this number C^n_k.

Problem 17. *5 points Compute the probability of getting two of a kind and three of a kind (a "full house") when five dice are rolled. (It is not necessary to express it as a decimal number; a fraction of integers is just fine. But please explain what you are doing.)*

Answer. See [**Ame94**, example 2.3.3 on p. 9]. Sample space is all ordered 5-tuples out of 6, which has 6^5 elements. Number of full houses can be identified with number of all ordered pairs of distinct elements out of 6, the first element in the pair denoting the number which appears twice and the second element that which appears three times, i.e., $P^6_2 = 6 \cdot 5$. Number of arrangements of a given full house over the five dice is $C^5_2 = \frac{5\cdot 4}{1\cdot 2}$ (we have to specify the two places taken by the two-of-a-kind outcomes.) Solution is therefore $P^6_2 \cdot C^5_2/6^5 = 50/6^4 = 0.03858$. This approach uses counting.

Alternative approach, using conditional probability: probability of getting 3 of one kind and then two of a different kind is $1 \cdot \frac{1}{6} \cdot \frac{1}{6} \cdot \frac{5}{6} \cdot \frac{1}{6} = \frac{5}{6^4}$. Then multiply by $\binom{5}{2} = 10$, since this is the number of arrangements of the 3 and 2 over the five cards. □

Problem 18. *What is the probability of drawing the King of Hearts and the Queen of Hearts if one draws two cards out of a 52 card game? Is it $\frac{1}{52^2}$? Is it $\frac{1}{(52)(51)}$? Or is it $1/\binom{52}{2} = \frac{2}{(52)(51)}$?*

Answer. Of course the last; it is the probability of drawing one special subset. There are two ways of drawing this subset: first the King and then the Queen, or first the Queen and then the King. □

2.6. Relationships Involving Binomial Coefficients

Problem 19. *Show that $\binom{n}{k} = \binom{n}{n-k}$. Give an intuitive argument why this must be so.*

Answer. Because $\binom{n}{n-k}$ counts the complements of k-element sets. □

Assume U has n elements, one of which is $\nu \in U$. How many k-element subsets of U have ν in them? There is a simple trick: Take all $(k-1)$-element subsets of the set you get by removing ν from U, and add ν to each of these sets. I.e., the number is $\binom{n-1}{k-1}$. Now how many k-element subsets of U do *not* have ν in them? Simple; just take the k-element subsets of the set which one gets by removing ν from U; i.e., it is $\binom{n-1}{k}$. Adding those two kinds of subsets together one gets all k-element subsets of U:

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}. \tag{2.6.1}$$

This important formula is the basis of the Pascal triangle:

(2.6.2)

$$\begin{array}{ccccccccccc} & & & & 1 & & & & & & \\ & & & 1 & & 1 & & & & & \\ & & 1 & & 2 & & 1 & & & & \\ & 1 & & 3 & & 3 & & 1 & & & \\ 1 & & 4 & & 6 & & 4 & & 1 & & \\ 1 & & 5 & & 10 & & 10 & & 5 & & 1 \end{array} = \begin{array}{ccccccccccc} & & & & \binom{0}{0} & & & & & & \\ & & & \binom{1}{0} & & \binom{1}{1} & & & & & \\ & & \binom{2}{0} & & \binom{2}{1} & & \binom{2}{2} & & & & \\ & \binom{3}{0} & & \binom{3}{1} & & \binom{3}{2} & & \binom{3}{3} & & & \\ \binom{4}{0} & & \binom{4}{1} & & \binom{4}{2} & & \binom{4}{3} & & \binom{4}{4} & & \\ \binom{5}{0} & & \binom{5}{1} & & \binom{5}{2} & & \binom{5}{3} & & \binom{5}{4} & & \binom{5}{5} \end{array}$$

The binomial coefficients also occur in the Binomial Theorem

$$(a+b)^n = a^n + \tbinom{n}{1}a^{n-1}b + \cdots + \tbinom{n}{n-1}ab^{n-1} + b^n = \sum_{k=0}^{n}\tbinom{n}{k}a^{n-k}b^k \tag{2.6.3}$$

Why? When the n factors $a+b$ are multiplied out, each of the resulting terms selects from each of the n original factors either a or b. The term $a^{n-k}b^k$ occurs therefore $\binom{n}{n-k} = \binom{n}{k}$ times.

As an application: If you set $a = 1$, $b = 1$, you simply get a sum of binomial coefficients, i.e., you get the number of subsets in a set with n elements: it is 2^n (always count the empty set as one of the subsets). The number of all subsets is easily counted directly. You go through the set element by element and about every element you ask: is it in the subset or not? I.e., for every element you have two possibilities, therefore by the multiplication principle the total number of possibilities is 2^n.

2.7. Conditional Probability

The concept of conditional probability is arguably more fundamental than probability itself. Every probability is conditional, since we must know that the "experiment" has happened before we can speak of probabilities. [**Ame94**, p. 10] and [**Rén70**] give axioms for conditional probability which take the place of the above axioms (2.3.1), (2.3.2) and (2.3.3). However we will follow here the common procedure of defining conditional probabilities in terms of the unconditional probabilities:

$$\Pr[\mathsf{B}|\mathsf{A}] = \frac{\Pr[\mathsf{B}\cap\mathsf{A}]}{\Pr[\mathsf{A}]} \tag{2.7.1}$$

How can we motivate (2.7.1)? If we know that A has occurred, then of course the only way that B occurs is when $\mathsf{B}\cap\mathsf{A}$ occurs. But we want to multiply all probabilities of subsets of A with an appropriate proportionality factor so that the probability of the event A itself becomes $= 1$.

Problem 20. *3 points Let A be an event with nonzero probability. Show that the probability conditionally on A, i.e., the mapping $\mathsf{B} \mapsto \Pr[\mathsf{B}|\mathsf{A}]$, satisfies all the axioms of a probability measure:*

$$\Pr[\mathsf{U}|\mathsf{A}] = 1 \tag{2.7.2}$$

$$\Pr[\mathsf{B}|\mathsf{A}] \geq 0 \quad \textit{for all events } \mathsf{B} \tag{2.7.3}$$

$$\Pr[\bigcup_{i=1}^{\infty}\mathsf{B}_i|\mathsf{A}] = \sum_{i=1}^{\infty}\Pr[\mathsf{B}_i|\mathsf{A}] \quad \textit{if } \mathsf{B}_i\cap\mathsf{B}_j = \emptyset \textit{ for all } i,j \textit{ with } i\neq j. \tag{2.7.4}$$

Answer. $\Pr[\mathsf{U}|\mathsf{A}] = \Pr[\mathsf{U}\cap\mathsf{A}]/\Pr[\mathsf{A}] = 1$. $\Pr[\mathsf{B}|\mathsf{A}] = \Pr[\mathsf{B}\cap\mathsf{A}]/\Pr[\mathsf{A}] \geq 0$ because $\Pr[\mathsf{B}\cap\mathsf{A}] \geq 0$ and $\Pr[\mathsf{A}] > 0$. Finally,

$$\Pr[\bigcup_{i=1}^{\infty}\mathsf{B}_i|\mathsf{A}] = \frac{\Pr[(\bigcup_{i=1}^{\infty}\mathsf{B}_i)\cap\mathsf{A}]}{\Pr[\mathsf{A}]} = \frac{\Pr[\bigcup_{i=1}^{\infty}(\mathsf{B}_i\cap\mathsf{A})]}{\Pr[\mathsf{A}]} = \frac{1}{\Pr[\mathsf{A}]}\sum_{i=1}^{\infty}\Pr[\mathsf{B}_i\cap\mathsf{A}] = \sum_{i=1}^{\infty}\Pr[\mathsf{B}_i|\mathsf{A}] \tag{2.7.5}$$

First equal sign is definition of conditional probability, second is distributivity of unions and intersections (Problem 6 d), third because the B_i are disjoint and therefore the $\mathsf{B}_i\cap\mathsf{A}$ are even more disjoint: $\mathsf{B}_i\cap\mathsf{A}\cap\mathsf{B}_j\cap\mathsf{A} = \mathsf{B}_i\cap\mathsf{B}_j\cap\mathsf{A} = \emptyset\cap\mathsf{A} = \emptyset$ for all i,j with $i\neq j$, and the last equal sign again by the definition of conditional probability. □

Problem 21. *You draw two balls without replacement from an urn which has 7 white and 14 black balls.*

If both balls are white, you roll a die, and your payoff is the number which the die shows in dollars.

If one ball is black and one is white, you flip a coin until you get your first head, and your payoff will be the number of flips it takes you to get a head, in dollars again.

If both balls are black, you draw from a deck of 52 cards, and you get the number shown on the card in dollars. (Ace counts as one, J, Q, and K as 11, 12, 13, i.e., basically the deck contains every number between 1 and 13 four times.)

Show that the probability that you receive exactly two dollars in this game is 1/6.

Answer. You know a complete disjunction of events: $\mathsf{U} = \{ww\}\cup\{bb\}\cup\{wb\}$, with $\Pr[\{ww\}] = \frac{7}{21}\frac{6}{20} = \frac{1}{10}$; $\Pr[\{bb\}] = \frac{14}{21}\frac{13}{20} = \frac{13}{30}$; $\Pr[\{bw\}] = \frac{7}{21}\frac{14}{20} + \frac{14}{21}\frac{7}{20} = \frac{7}{15}$. Furthermore you know the conditional probabilities of getting 2 dollars conditonally on each of these events: $\Pr[\{2\}|\{ww\}] = \frac{1}{6}$; $\Pr[\{2\}|\{bb\}] = \frac{1}{13}$; $\Pr[\{2\}|\{wb\}] = \frac{1}{4}$. Now $\Pr[\{2\}\cap\{ww\}] = \Pr[\{2\}|\{ww\}]\Pr[\{ww\}]$ etc., therefore

$$\Pr[\{2\}] = \Pr[\{2\}\cap\{ww\}] + \Pr[\{2\}\cap\{bw\}] + \Pr[\{2\}\cap\{bb\}] \tag{2.7.6}$$

$$= \frac{1}{6}\frac{7}{21}\frac{6}{20} + \frac{1}{4}\left(\frac{7}{21}\frac{14}{20} + \frac{14}{21}\frac{7}{20}\right) + \frac{1}{13}\frac{14}{21}\frac{13}{20} \tag{2.7.7}$$

$$= \frac{1}{6}\frac{1}{10} + \frac{1}{4}\frac{7}{15} + \frac{1}{13}\frac{13}{30} = \frac{1}{6} \tag{2.7.8}$$

□

Problem 22. *2 points A and B are arbitrary events. Prove that the probability of B can be written as:*

$$\Pr[\mathsf{B}] = \Pr[\mathsf{B}|\mathsf{A}]\Pr[\mathsf{A}] + \Pr[\mathsf{B}|\mathsf{A}']\Pr[\mathsf{A}'] \tag{2.7.9}$$

This is the law of iterated expectations (8.6.2) *in the case of discrete random variables: it might be written as* $\Pr[\mathsf{B}] = \mathrm{E}\big[\Pr[\mathsf{B}|\mathsf{A}]\big]$.

Answer. $\mathsf{B} = \mathsf{B}\cap\mathsf{U} = \mathsf{B}\cap(\mathsf{A}\cup\mathsf{A}') = (\mathsf{B}\cap\mathsf{A})\cup(\mathsf{B}\cap\mathsf{A}')$ and this union is disjoint, i.e., $(\mathsf{B}\cap\mathsf{A})\cap(\mathsf{B}\cap\mathsf{A}') = \mathsf{B}\cap(\mathsf{A}\cap\mathsf{A}') = \mathsf{B}\cap\emptyset = \emptyset$. Therefore $\Pr[\mathsf{B}] = \Pr[\mathsf{B}\cap\mathsf{A}] + \Pr[\mathsf{B}\cap\mathsf{A}']$. Now apply definition of conditional probability to get $\Pr[\mathsf{B}\cap\mathsf{A}] = \Pr[\mathsf{B}|\mathsf{A}]\Pr[\mathsf{A}]$ and $\Pr[\mathsf{B}\cap\mathsf{A}'] = \Pr[\mathsf{B}|\mathsf{A}']\Pr[\mathsf{A}']$. □

PROBLEM 23. *2 points Prove the following lemma: If* $\Pr[\mathsf{B}|\mathsf{A}_1] = \Pr[\mathsf{B}|\mathsf{A}_2]$ *(call it c) and* $\mathsf{A}_1 \cap \mathsf{A}_2 = \emptyset$ *(i.e.,* A_1 *and* A_2 *are disjoint), then also* $\Pr[\mathsf{B}|\mathsf{A}_1 \cup A_2] = c$.

ANSWER.

$$\begin{aligned}\Pr[\mathsf{B}|\mathsf{A}_1 \cup A_2] &= \frac{\Pr[\mathsf{B} \cap (\mathsf{A}_1 \cup \mathsf{A}_2)]}{\Pr[\mathsf{A}_1 \cup \mathsf{A}_2]} = \frac{\Pr[(\mathsf{B} \cap \mathsf{A}_1) \cup (\mathsf{B} \cap \mathsf{A}_2)]}{\Pr[\mathsf{A}_1 \cup \mathsf{A}_2]} \\ &= \frac{\Pr[\mathsf{B} \cap \mathsf{A}_1] + \Pr[\mathsf{B} \cap \mathsf{A}_2]}{\Pr[\mathsf{A}_1] + \Pr[\mathsf{A}_2]} = \frac{c\Pr[\mathsf{A}_1] + c\Pr[\mathsf{A}_2]}{\Pr[\mathsf{A}_1] + \Pr[\mathsf{A}_2]} = c. \qquad (2.7.10)\end{aligned}$$

□

PROBLEM 24. *Show by counterexample that the requirement* $\mathsf{A}_1 \cap \mathsf{A}_2 = \emptyset$ *is necessary for this result to hold. Hint: use the example in Problem 38 with* $\mathsf{A}_1 = \{HH, HT\}$, $\mathsf{A}_2 = \{HH, TH\}$, $\mathsf{B} = \{HH, TT\}$.

ANSWER. $\Pr[\mathsf{B}|\mathsf{A}_1] = 1/2$ and $\Pr[\mathsf{B}|\mathsf{A}_2] = 1/2$, but $\Pr[\mathsf{B}|\mathsf{A}_1 \cup A_2] = 1/3$. □

The conditional probability can be used for computing probabilities of intersections of events.

PROBLEM 25. [**Lar82**, exercises 2.5.1 and 2.5.2 on p. 57, solutions on p. 597, but no discussion]. *Five white and three red balls are laid out in a row at random.*

• a. *3 points What is the probability that both end balls are white? What is the probability that one end ball is red and the other white?*

ANSWER. You can lay the first ball first and the last ball second: for white balls, the probability is $\frac{5}{8}\frac{4}{7} = \frac{5}{14}$; for one white, one red it is $\frac{5}{8}\frac{3}{7} + \frac{3}{8}\frac{5}{7} = \frac{15}{28}$. □

• b. *4 points What is the probability that all red balls are together? What is the probability that all white balls are together?*

ANSWER. All red balls together is the same as 3 reds first, multiplied by 6, because you may have between 0 and 5 white balls before the first red. $\frac{3}{8}\frac{2}{7}\frac{1}{6} \cdot 6 = \frac{3}{28}$. For the white balls you get $\frac{5}{8}\frac{4}{7}\frac{3}{6}\frac{2}{5}\frac{1}{4} \cdot 4 = \frac{1}{14}$.

BTW, 3 reds first is same probability as 3 reds last, ie., the 5 whites first: $\frac{5}{8}\frac{4}{7}\frac{3}{6}\frac{2}{5}\frac{1}{4} = \frac{3}{8}\frac{2}{7}\frac{1}{6}$. □

PROBLEM 26. *The first three questions here are discussed in* [**Lar82**, example 2.6.3 on p. 62]*: There is an urn with 4 white and 8 black balls. You take two balls out without replacement.*

• a. *1 point What is the probability that the first ball is white?*

ANSWER. 1/3 □

• b. *1 point What is the probability that both balls are white?*

ANSWER. It is Pr[second ball white|first ball white] Pr[first ball white] $= \frac{3}{3+8}\frac{4}{4+8} = \frac{1}{11}$. □

• c. *1 point What is the probability that the second ball is white?*

ANSWER. It is Pr[first ball white and second ball white]+Pr[first ball black and second ball white] =

$$= \frac{3}{3+8}\frac{4}{4+8} + \frac{4}{7+4}\frac{8}{8+4} = \frac{1}{3}. \qquad (2.7.11)$$

This is the same as the probability that the first ball is white. The probabilities are not dependent on the order in which one takes the balls out. This property is called "exchangeability." One can see it also in this way: Assume you number the balls at random, from 1 to 12. Then the probability for a white ball to have the number 2 assigned to it is obviously $\frac{1}{3}$. □

• d. *1 point What is the probability that both of them are black?*

ANSWER. $\frac{8}{12}\frac{7}{11} = \frac{2}{3}\frac{7}{11} = \frac{14}{33}$ (or $\frac{56}{132}$). □

• e. *1 point What is the probability that both of them have the same color?*

ANSWER. The sum of the two above, $\frac{14}{33} + \frac{1}{11} = \frac{17}{33}$ (or $\frac{68}{132}$). □

Now you take three balls out without replacement.

• f. *2 points Compute the probability that at least two of the three balls are white.*

ANSWER. It is $\frac{13}{55}$. The possibilities are *wwb*, *wbw*, *bww*, and *www*. Of the first three, each has probability $\frac{4}{12}\frac{3}{11}\frac{8}{10}$. Therefore the probability for exactly two being white is $\frac{288}{1320} = \frac{12}{55}$. The probability for *www* is $\frac{4 \cdot 3 \cdot 2}{12 \cdot 11 \cdot 10} = \frac{24}{1320} = \frac{1}{55}$. Add this to get $\frac{312}{1320} = \frac{13}{55}$. More systematically, the answer is $\left(\binom{4}{2}\binom{8}{1} + \binom{4}{3}\right) \Big/ \binom{12}{3}$. □

• g. *1 point Compute the probability that at least two of the three are black.*

ANSWER. It is $\frac{42}{55}$. For exactly two: $\frac{672}{1320} = \frac{28}{55}$. For three it is $\frac{(8)(7)(6)}{(12)(11)(10)} = \frac{336}{1320} = \frac{14}{55}$. Together $\frac{1008}{1320} = \frac{42}{55}$. One can also get is as: it is the complement of the last, or as $\left(\binom{8}{3} + \binom{8}{2}\binom{4}{1}\right) \Big/ \binom{12}{3}$. □

• h. *1 point Compute the probability that two of the three are of the same and the third of a different color.*

ANSWER. It is $\frac{960}{1320} = \frac{40}{55} = \frac{8}{11}$, or $\left(\binom{4}{1}\binom{8}{2} + \binom{4}{2}\binom{8}{1}\right) \Big/ \binom{12}{3}$. □

• i. *1 point Compute the probability that at least two of the three are of the same color.*

ANSWER. This probability is 1. You have 5 black socks and 5 white socks in your drawer. There is a fire at night and you must get out of your apartment in two minutes. There is no light. You fumble in the dark for the drawer. How many socks do you have to take out so that you will have at least 2 of the same color? The answer is 3 socks. □

PROBLEM 27. *If a poker hand of five cards is drawn from a deck, what is the probability that it will contain three aces? (How can the concept of conditional probability help in answering this question?)*

Answer. [**Ame94**, example 2.3.3 on p. 9] and [**Ame94**, example 2.5.1 on p. 13] give two alternative ways to do it. The second answer uses conditional probability: Probability to draw three aces in a row first and then 2 nonaces is $\frac{4}{52}\frac{3}{51}\frac{2}{50}\frac{48}{49}\frac{47}{48}$ Then multiply this by $\binom{5}{3} = \frac{5\cdot 4\cdot 3}{1\cdot 2\cdot 3} = 10$ This gives 0.0017, i.e., 0.17%. □

Problem 28. *2 points A friend tosses two coins. You ask: "did one of them land heads?" Your friend answers, "yes." What's the probability that the other also landed heads?*

Answer. $\mathsf{U} = \{HH, HT, TH, TT\}$; Probability is $\frac{1}{4}/\frac{3}{4} = \frac{1}{3}$. □

Problem 29. *(Not eligible for in-class exams)* [**Ame94**, p. 5] *What is the probability that a person will win a game in tennis if the probability of his or her winning a point is p?*

Answer.

$$p^4\left(1 + 4(1-p) + 10(1-p)^2 + \frac{20p(1-p)^3}{1-2p(1-p)}\right) \tag{2.7.12}$$

How to derive this: $\{ssss\}$ has probability p^4; $\{sssfs\}$, $\{ssfss\}$, $\{sfsss\}$, and $\{fssss\}$ have probability $4p^4(1-p)$; $\{sssffs\}$ etc. (2 f and 3 s in the first 5, and then an s, together $\binom{5}{2} = 10$ possibilities) have probability $10p^4(1-p)^2$. Now $\{sssfff\}$ and $\binom{6}{3} = 20$ other possibilities give deuce at least once in the game, i.e., the probability of deuce is $20p^3(1-p)^3$. Now $\Pr[\text{win}|\text{deuce}] = p^2 + 2p(1-p)\Pr[\text{win}|\text{deuce}]$, because you win either if you score twice in a row (p^2) or if you get deuce again (probablity $2p(1-p)$) and then win. Solve this to get $\Pr[\text{win}|\text{deuce}] = p^2/\big(1-2p(1-p)\big)$ and then multiply this conditional probability with the probability of getting deuce at least once: $\Pr[\text{win after at least one deuce}] = 20p^3(1-p)^3p^2/\big(1-2p(1-p)\big)$. This gives the last term in (2.7.12). □

Problem 30. *(Not eligible for in-class exams) Andy, Bob, and Chris play the following game: each of them draws a card without replacement from a deck of 52 cards. The one who has the highest card wins. If there is a tie (like: two kings and no aces), then that person wins among those who drew this highest card whose name comes first in the alphabet. What is the probability for Andy to be the winner? For Bob? For Chris? Does this probability depend on the order in which they draw their cards out of the stack?*

Answer. Let A be the event that Andy wins, B that Bob, and C that Chris wins.

One way to approach this problem is to ask: what are the chances for Andy to win when he draws a king?, etc., i.e., compute it for all 13 different cards. Then: what are the chances for Bob to win when he draws a king, and also his chances for the other cards, and then for Chris.

It is computationally easier to make the following partitioning of all outcomes: Either all three cards drawn are different (call this event D), or all three cards are equal (event E), or two of the three cards are equal (T). This third case will have to be split into $\mathsf{T} = \mathsf{H} \cup \mathsf{L}$, according to whether the card that is different is higher or lower.

If all three cards are different, then Andy, Bob, and Chris have equal chances of winning; if all three cards are equal, then Andy wins. What about the case that two cards are the same and the

third is different? There are two possibilities. If the card that is different is higher than the two that are the same, then the chances of winning are evenly distributed; but if the two equal cards are higher, then Andy has a $\frac{2}{3}$ chance of winning (when the distribution of the cards Y (lower) and Z (higher) among ABC is is ZZY and ZYZ), and Bob has a $\frac{1}{3}$ chance of winning (when the distribution is YZZ). What we just did was computing the conditional probabilities $\Pr[\mathsf{A}|\mathsf{D}]$, $\Pr[\mathsf{A}|\mathsf{E}]$, etc.

Now we need the probabilities of D, E, and T. What is the probability that all three cards drawn are the same? The probability that the second card is the same as the first is $\frac{3}{51}$; and the probability that the third is the same too is $\frac{2}{50}$; therefore the total probability is $\frac{(3)(2)}{(51)(50)} = \frac{6}{2550}$. The probability that all three are unequal is $\frac{48}{51}\frac{44}{50} = \frac{2112}{2550}$. The probability that two are equal and the third is different is $3\frac{3}{51}\frac{48}{50} = \frac{432}{2550}$. Now in half of these cases, the card that is different is higher, and in half of the cases it is lower.

Putting this together one gets:

		Uncond. Prob.	Cond. Prob.			Prob. of intersection		
			A	B	C	A	B	C
E	all 3 equal	6/2550	1	0	0	6/2550	0	0
H	2 of 3 equal, 3rd higher	216/2550	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	72/2550	72/2550	72/2550
L	2 of 3 equal, 3rd lower	216/2550	$\frac{2}{3}$	$\frac{1}{3}$	0	144/2550	72/2550	0
D	all 3 unequal	2112/2550	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	704/2550	704/2550	704/2550
Sum		2550/2550				926/2550	848/2550	776/2550

I.e., the probability that A wins is $926/2550 = 463/1275 = .363$, the probability that B wins is $848/2550 = 424/1275 = .3325$, and the probability that C wins is $776/2550 = 338/1275 = .304$.

Here we are using $\Pr[\mathsf{A}] = \Pr[\mathsf{A}|\mathsf{E}]\Pr[\mathsf{E}] + \Pr[\mathsf{A}|\mathsf{H}]\Pr[\mathsf{H}] + \Pr[\mathsf{A}|\mathsf{L}]\Pr[\mathsf{L}] + \Pr[\mathsf{A}|\mathsf{D}]\Pr[\mathsf{D}]$. □

Problem 31. *4 points You are the contestant in a game show. There are three closed doors at the back of the stage. Behind one of the doors is a sports car, behind the other two doors are goats. The game master knows which door has the sports car behind it, but you don't. You have to choose one of the doors; if it is the door with the sports car, the car is yours.*

After you make your choice, say door A, the game master says: "I want to show you something." He opens one of the two other doors, let us assume it is door B, and it has a goat behind it. Then the game master asks: "Do you still insist on door A, or do you want to reconsider your choice?"

Can you improve your odds of winning by abandoning your previous choice and instead selecting the door which the game master did not open? If so, by how much?

Answer. If you switch, you will lose the car if you had initially picked the right door, but you will get the car if you were wrong before! Therefore you improve your chances of winning from 1/3 to 2/3. This is simulated on the web, see `www.stat.sc.edu/~west/javahtml/LetsMakeaDeal.html`.

It is counterintuitive. You may think that one of the two other doors always has a goat behind it, whatever your choice, therefore there is no reason to switch. But the game master not only shows you *that* there is another door with a goat, he also shows you one of the other doors with a goat behind it, i.e., he restricts your choice if you switch. This is valuable information. It is as if you could bet on both other doors simultaneously, i.e., you get the car if it is behind one of the doors B or C. I.e., if the quiz master had said: I give you the opportunity to switch to the following: you

get the car if it is behind B or C. Do you want to switch? The only doubt the contestant may have about this is: had I not picked a door with the car behind it then I would not have been offered this opportunity to switch.

□

2.8. Ratio of Probabilities as Strength of Evidence

$\Pr_1$ and $\Pr_2$ are two probability measures defined on the same set $\mathcal{F}$ of events. Hypothesis H_1 says $\Pr_1$ is the true probability, and H_2 says $\Pr_2$ is the true probability. Then the observation of an event A for which $\Pr_1[\mathsf{A}] > \Pr_2[\mathsf{A}]$ is evidence in favor of H_1 as opposed to H_2. [**Roy97**] argues that the *ratio* of the probabilities (also called "likelihood ratio") is the right way to measure the *strength* of this evidence. Among others, the following justification is given [**Roy97**, p. 7]: If H_2 is true, it is usually not *impossible* to find evidence favoring H_1, but it is *unlikely*; and its probability is bounded by the (reverse of) the ratio of probabilities.

This can be formulated mathematically as follows: Let S be the union of all events A for which $\Pr_1[\mathsf{A}] \geq k \Pr_2[\mathsf{A}]$. Then it can be shown that $\Pr_2[\mathsf{S}] \leq 1/k$, i.e., if H_2 is true, the probability to find evidence favoring H_1 with strength k is never greater than $1/k$. Here is a proof in the case that there is only a finite number of possible outcomes $U = \{\omega_1, \ldots, \omega_n\}$: Renumber the outcomes such that for $i = 1, \ldots, m$, $\Pr_1[\{\omega_i\}] < k \Pr_2[\{\omega_i\}]$, and for $j = m+1, \ldots, n$, $\Pr_1[\{\omega_j\}] \geq k \Pr_2[\{\omega_j\}]$. Then $\mathsf{S} = \{\omega_{m+1}, \ldots, \omega_n\}$, therefore $\Pr_2[\mathsf{S}] = \sum_{j=m+1}^n \Pr_2[\{\omega_j\}] \leq \sum_{j=m+1}^n \frac{\Pr_1[\{\omega_j\}]}{k} = \frac{1}{k}\Pr_1[\mathsf{S}] \leq \frac{1}{k}$ as claimed. The last inequality holds because $\Pr_1[\mathsf{S}] \leq 1$, and the equal-sign before this is simply the definition of S.

With more mathematical effort, see [**Rob70**], one can strengthen this simple inequality in a very satisfactory manner: Assume an unscrupulous researcher attempts to find evidence supporting his favorite but erroneous hypothesis H_1 over his rival's H_2 by a factor of at least k. He proceeds as follows: he observes an outcome of the above experiment once, say the outcome is $\omega_{i(1)}$. If $\Pr_1[\{\omega_{i(1)}\}] \geq k \Pr_2[\{\omega_{i(1)}\}]$ he publishes his result; if not, he makes a second independent observation of the experiment $\omega_{i(2)}$. If $\Pr_1[\{\omega_{i(1)}\}]\Pr_1[\{\omega_{i(2)}\}] > k \Pr_2[\{\omega_{i(1)}\}]\Pr_2[\{\omega_{i(2)}\}]$ he publishes his result; if not he makes a third observation and incorporates that in his publication as well, etc. It can be shown that this strategy will not help: if his rival's hypothesis is true, then the probability that he will ever be able to publish results which seem to show that his own hypothesis is true is still $\leq 1/k$. I.e., the sequence of independent observations $\omega_{i(2)}, \omega_{i(2)}, \ldots$ is such that

$$\Pr_2\Big[\prod_{j=1}^n \Pr_1[\{\omega_{i(j)}\}] \geq k \prod_{j=1}^n \Pr_2[\{\omega_{i(1)}\}] \quad \text{for some } n = 1, 2, \ldots\Big] \leq \frac{1}{k} \tag{2.8.1}$$

It is not possible to take advantage of the indeterminacy of a random outcome by carrying on until chance places one ahead, and then to quit. If one fully discloses

all the evidence one is accumulating, then the probability that this accumulated evidence supports one's hypothesis cannot rise above $1/k$.

Problem 32. *It is usually not possible to assign probabilities to the hypotheses H_1 and H_2, but sometimes it is. Show that in this case, the likelihood ratio of event A is the factor by which the ratio of the probabilities of H_1 and H_2 is changed by the observation of A, i.e.,*

$$\frac{\Pr[\mathsf{H}_1|\mathsf{A}]}{\Pr[\mathsf{H}_2|\mathsf{A}]} = \frac{\Pr[\mathsf{H}_1]}{\Pr[\mathsf{H}_2]}\frac{\Pr[\mathsf{A}|\mathsf{H}_1]}{\Pr[\mathsf{A}|\mathsf{H}_2]} \tag{2.8.2}$$

Answer. Apply Bayes's theorem (2.9.1) twice, once for the numerator, once for the denominator. □

A world in which probability theory applies is therefore a world in which the transitive dimension must be distinguished from the intransitive dimension. Research results are not determined by the goals of the researcher.

2.9. Bayes Theorem

In its simplest form Bayes's theorem reads

$$\Pr[\mathsf{A}|\mathsf{B}] = \frac{\Pr[\mathsf{B}|\mathsf{A}]\Pr[\mathsf{A}]}{\Pr[\mathsf{B}|\mathsf{A}]\Pr[\mathsf{A}] + \Pr[\mathsf{B}|\mathsf{A}']\Pr[\mathsf{A}']}. \tag{2.9.1}$$

Problem 33. *Prove Bayes theorem!*

Answer. Obvious since numerator is $\Pr[\mathsf{B} \cap \mathsf{A}]$ and denominator $\Pr[\mathsf{B} \cap \mathsf{A}] + \Pr[\mathsf{B} \cap \mathsf{A}'] = \Pr[\mathsf{B}]$. □

This theorem has its significance in cases in which A can be interpreted as a cause of B, and B an effect of A. For instance, A is the event that a student who was picked randomly from a class has learned for a certain exam, and B is the event that he passed the exam. Then the righthand side expression contains that information which you would know from the cause-effect relations: the unconditional probability of the event which is the cause, and the conditional probabilities of the effect conditioned on whether or not the cause happened. From this, the formula computes the conditional probability of the cause given that the effect happened. Bayes's theorem tells us therefore: if we know that the effect happened, how sure can we be that the cause happened? Clearly, Bayes's theorem has relevance for statistical inference.

Let's stay with the example with learning for the exam; assume $\Pr[\mathsf{A}] = 60\%$, $\Pr[\mathsf{B}|\mathsf{A}] = .8$, and $\Pr[\mathsf{B}|\mathsf{A}'] = .5$. Then the probability that a student who passed the exam has learned for it is $\frac{(.8)(.6)}{(.8)(.6)+(.5)(.4)} = \frac{.48}{.68} = .706$. Look at these numbers: The numerator is the average percentage of students who learned and passed, and the denominator average percentage of students who passed.

PROBLEM 34. *AIDS diagnostic tests are usually over 99.9% accurate on those who do not have AIDS (i.e., only 0.1% false positives) and 100% accurate on those who have AIDS (i.e., no false negatives at all). (A test is called positive if it indicates that the subject has AIDS.)*

• a. *3 points Assuming that 0.5% of the population actually have AIDS, compute the probability that a particular individual has AIDS, given that he or she has tested positive.*

ANSWER. A is the event that he or she has AIDS, and T the event that the test is positive.

$$\Pr[\mathsf{A}|\mathsf{T}] = \frac{\Pr[\mathsf{T}|\mathsf{A}]\Pr[\mathsf{A}]}{\Pr[\mathsf{T}|\mathsf{A}]\Pr[\mathsf{A}] + \Pr[\mathsf{T}|\mathsf{A}']\Pr[\mathsf{A}']} = \frac{1\cdot 0.005}{1\cdot 0.005 + 0.001\cdot 0.995} =$$
$$= \frac{100\cdot 0.5}{100\cdot 0.5 + 0.1\cdot 99.5} = \frac{1000\cdot 5}{1000\cdot 5 + 1\cdot 995} = \frac{5000}{5995} = \frac{1000}{1199} = 0.834028$$

Even after testing positive there is still a 16.6% chance that this person does not have AIDS. □

• b. *1 point If one is young, healthy and not in one of the risk groups, then the chances of having AIDS are not 0.5% but 0.1% (this is the proportion of the applicants to the military who have AIDS). Re-compute the probability with this alternative number.*

ANSWER.

$$\frac{1\cdot 0.001}{1\cdot 0.001 + 0.001\cdot 0.999} = \frac{100\cdot 0.1}{100\cdot 0.1 + 0.1\cdot 99.9} = \frac{1000\cdot 1}{1000\cdot 1 + 1\cdot 999} = \frac{1000}{1000+999} = \frac{1000}{1999} = 0.50025.$$

□

2.10. Independence of Events

2.10.1. Definition of Independence. Heuristically, we want to say: event B is independent of event A if $\Pr[\mathsf{B}|\mathsf{A}] = \Pr[\mathsf{B}|\mathsf{A}']$. From this follows by Problem 23 that the conditional probability is equal to the unconditional probability $\Pr[\mathsf{B}]$, i.e., $\Pr[\mathsf{B}] = \Pr[\mathsf{B}\cap\mathsf{A}]/\Pr[\mathsf{A}]$. Therefore we will adopt as definition of independence the so-called *multiplication rule:*

Definition: B and A are independent, notation $\mathsf{B}\perp\mathsf{A}$, if $\Pr[\mathsf{B}\cap\mathsf{A}] = \Pr[\mathsf{B}]\Pr[\mathsf{A}]$.

This is a symmetric condition, i.e., if B is independent of A, then A is also independent of B. This symmetry is not immediately obvious given the above definition of independence, and it also has the following nontrivial practical implication (this example from [**Daw79a**, pp. 2/3]): A is the event that one is exposed to some possibly carcinogenic agent, and B the event that one develops a certain kind of cancer. In order to test whether $\mathsf{B}\perp\mathsf{A}$, i.e., whether the exposure to the agent does not increase the incidence of cancer, one often collects two groups of subjects, one group which has cancer and one control group which does not, and checks whether the exposure in these two groups to the carcinogenic agent is the same. I.e., the experiment checks

whether $\mathsf{A}\perp\mathsf{B}$, although the purpose of the experiment was to determine whether $\mathsf{B}\perp\mathsf{A}$.

PROBLEM 35. *3 points Given that* $\Pr[\mathsf{B}\cap\mathsf{A}] = \Pr[\mathsf{B}]\cdot\Pr[\mathsf{A}]$ *(i.e.,* B *is independent of* A*), show that* $\Pr[\mathsf{B}\cap\mathsf{A}'] = \Pr[\mathsf{B}]\cdot\Pr[\mathsf{A}']$ *(i.e.,* B *is also independent of* A'*).*

ANSWER. If one uses our heuristic definition of independence, i.e., B is independent of event A if $\Pr[\mathsf{B}|\mathsf{A}] = \Pr[\mathsf{B}|\mathsf{A}']$, then it is immediately obvious since definition is symmetric in A and A'. However if we use the multiplication rule as the definition of independence, as the text of this Problem suggests, we have to do a little more work: Since B is the disjoint union of $(\mathsf{B}\cap\mathsf{A})$ and $(\mathsf{B}\cap\mathsf{A}')$, it follows $\Pr[\mathsf{B}] = \Pr[\mathsf{B}\cap\mathsf{A}] + \Pr[\mathsf{B}\cap\mathsf{A}']$ or $\Pr[\mathsf{B}\cap\mathsf{A}'] = \Pr[\mathsf{B}] - \Pr[\mathsf{B}\cap\mathsf{A}] = \Pr[\mathsf{B}] - \Pr[\mathsf{B}]\Pr[\mathsf{A}] = \Pr[\mathsf{B}](1-\Pr[\mathsf{A}]) = \Pr[\mathsf{B}]\Pr[\mathsf{A}']$. □

PROBLEM 36. *2 points* A *and* B *are two independent events with* $\Pr[\mathsf{A}] = \frac{1}{3}$ *and* $\Pr[\mathsf{B}] = \frac{1}{4}$*. Compute* $\Pr[\mathsf{A}\cup\mathsf{B}]$*.*

ANSWER. $\Pr[\mathsf{A}\cup\mathsf{B}] = \Pr[\mathsf{A}] + \Pr[\mathsf{B}] - \Pr[\mathsf{A}\cap\mathsf{B}] = \Pr[\mathsf{A}] + \Pr[\mathsf{B}] - \Pr[\mathsf{A}]\Pr[\mathsf{B}] = \frac{1}{3} + \frac{1}{4} - \frac{1}{12} = \frac{1}{2}$. □

PROBLEM 37. *3 points You have an urn with five white and five red balls. You take two balls out without replacement.* A *is the event that the first ball is white, and* B *that the second ball is white. a. What is the probability that the first ball is white? b. What is the probability that the second ball is white? c. What is the probability that both have the same color? d. Are these two events independent, i.e., is* $\Pr[\mathsf{B}|\mathsf{A}] = \Pr[\mathsf{A}]$*? e. Are these two events disjoint, i.e., is* $\mathsf{A}\cap\mathsf{B} = \emptyset$*?*

ANSWER. Clearly, $\Pr[\mathsf{A}] = 1/2$. $\Pr[\mathsf{B}] = \Pr[\mathsf{B}|\mathsf{A}]\Pr[\mathsf{A}] + \Pr[\mathsf{B}|\mathsf{A}']\Pr[\mathsf{A}'] = (4/9)(1/2) + (5/9)(1/2) = 1/2$. The events are not independent: $\Pr[\mathsf{B}|\mathsf{A}] = 4/9 \neq \Pr[\mathsf{B}]$, or $\Pr[\mathsf{A}\cap\mathsf{B}] = \frac{5}{10}\frac{4}{9} = 2/9 \neq 1/4$. They would be independent if the first ball had been replaced. The events are also not disjoint: it is possible that both balls are white. □

2.10.2. Independence of More than Two Events. If there are more than two events, we must require that all possible intersections of these events, not only the pairwise intersections, follow the above multiplication rule. For instance,

$$\text{(2.10.1)}\quad \mathsf{A},\mathsf{B},\mathsf{C}\text{ mutually independent} \iff \begin{aligned}\Pr[\mathsf{A}\cap\mathsf{B}] &= \Pr[\mathsf{A}]\Pr[\mathsf{B}];\\ \Pr[\mathsf{A}\cap\mathsf{C}] &= \Pr[\mathsf{A}]\Pr[\mathsf{C}];\\ \Pr[\mathsf{B}\cap\mathsf{C}] &= \Pr[\mathsf{B}]\Pr[\mathsf{C}];\\ \Pr[\mathsf{A}\cap\mathsf{B}\cap\mathsf{C}] &= \Pr[\mathsf{A}]\Pr[\mathsf{B}]\Pr[\mathsf{C}].\end{aligned}$$

This last condition is not implied by the other three. Here is an example. Draw a ball at random from an urn containing four balls numbered 1, 2, 3, 4. Define $\mathsf{A} = \{1,4\}$, $\mathsf{B} = \{2,4\}$, and $\mathsf{C} = \{3,4\}$. These events are pairwise independent but not mutually independent.

PROBLEM 38. *2 points Flip a coin two times independently and define the following three events:*

$$\begin{aligned} \mathsf{A} &= \text{Head in first flip} \\ \mathsf{B} &= \text{Head in second flip} \\ \mathsf{C} &= \text{Same face in both flips.} \end{aligned} \tag{2.10.2}$$

Are these three events pairwise independent? Are they mutually independent?

ANSWER. $\mathsf{U} = \left\{ \begin{smallmatrix} HH & HT \\ TH & TT \end{smallmatrix} \right\}$. $\mathsf{A} = \{HH, HT\}$, $\mathsf{B} = \{HH, TH\}$, $\mathsf{C} = \{HH, TT\}$. $\Pr[\mathsf{A}] = \frac{1}{2}$, $\Pr[\mathsf{B}] = \frac{1}{2}$, $\Pr[\mathsf{C}] = \frac{1}{2}$. They are pairwise independent, but $\Pr[\mathsf{A} \cap \mathsf{B} \cap \mathsf{C}] = \Pr[\{HH\}] = \frac{1}{4} \neq \Pr[\mathsf{A}]\Pr[\mathsf{B}]\Pr[\mathsf{C}]$, therefore the events cannot be mutually independent. □

PROBLEM 39. *3 points* A, B, *and* C *are pairwise independent events whose probabilities are greater than zero and smaller than one, and* $\mathsf{A} \cap \mathsf{B} \subset \mathsf{C}$. *Can those events be mutually independent?*

ANSWER. No; from $\mathsf{A} \cap \mathsf{B} \subset \mathsf{C}$ follows $\mathsf{A} \cap \mathsf{B} \cap \mathsf{C} = \mathsf{A} \cap \mathsf{B}$ and therefore $\Pr[\mathsf{A} \cap \mathsf{B} \cap \mathsf{C}] \neq \Pr[\mathsf{A} \cap \mathsf{B}]\Pr[\mathsf{C}]$ since $\Pr[\mathsf{C}] < 1$ and $\Pr[\mathsf{A} \cap \mathsf{B}] > 0$. □

If one takes unions, intersections, complements of different mutually independent events, one will still end up with mutually independent events. E.g., if A, B, C mutually independent, then A', B, C are mutually independent as well, and $\mathsf{A} \cap \mathsf{B}$ independent of C, and $\mathsf{A} \cup \mathsf{B}$ independent of C, etc. This is not the case if the events are only pairwise independent. In Problem 39, $\mathsf{A} \cap \mathsf{B}$ is not independent of C.

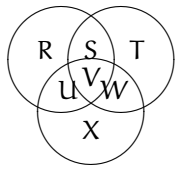


FIGURE 1. Generic Venn Diagram for 3 Events

2.10.3. Conditional Independence. If A and B are independent in the probability measure conditionally on C, i.e., if $\Pr[\mathsf{A} \cap \mathsf{B}|\mathsf{C}] = \Pr[\mathsf{A}|\mathsf{C}]\Pr[\mathsf{B}|\mathsf{C}]$, then they are called conditionally independent given that C occurred, notation $\mathsf{A} \perp \mathsf{B}|\mathsf{C}$. In formulas,

$$\frac{\Pr[\mathsf{A} \cap \mathsf{C}]}{\Pr[\mathsf{C}]} \frac{\Pr[\mathsf{B} \cap \mathsf{C}]}{\Pr[\mathsf{C}]} = \frac{\Pr[\mathsf{A} \cap \mathsf{B} \cap \mathsf{C}]}{\Pr[\mathsf{C}]}. \tag{2.10.3}$$

PROBLEM 40. *5 points Show that* $\mathsf{A} \perp \mathsf{B}|\mathsf{C}$ *is equivalent to* $\Pr[\mathsf{A}|\mathsf{B} \cap C] = \Pr[\mathsf{A}|\mathsf{C}]$. *In other words: independence of* A *and* B *conditionally on* C *means: once we know that* C *occurred, the additional knowledge whether* B *occurred or not will not help us to sharpen our knowledge about* A.

Literature about conditional independence (of random variables, not of events) includes [**Daw79a**], [**Daw79b**], [**Daw80**].

2.10.4. Independent Repetition of an Experiment. If a given experiment has sample space U, and we perform the experiment n times in a row, then this repetition can be considered a single experiment with the sample space consisting of n-tuples of elements of U. This set is called the product set $\mathsf{U}^n = \mathsf{U} \times \mathsf{U} \times \cdots \times \mathsf{U}$ (n terms).

If a probability measure Pr is given on $\mathcal{F}$, then one can define in a unique way a probability measure on the subsets of the product set so that events in different repetitions are always independent of each other.

The *Bernoulli experiment* is the simplest example of such an independent repetition. $\mathsf{U} = \{s, f\}$ (stands for success and failure). Assume $\Pr[\{s\}] = p$, and that the experimenter has several independent trials. For instance, U^5 has, among others, the following possible outcomes:

$$\begin{aligned} \text{If } \omega = &(f,f,f,f,f) \quad \text{then} \quad \Pr[\{\omega\}] = (1-p)^n \\ &(f,f,f,f,s) \qquad\qquad (1-p)^{n-1}p \\ &(f,f,f,s,f) \qquad\qquad (1-p)^{n-1}p \\ &(f,f,f,s,s) \qquad\qquad (1-p)^{n-2}p^2 \\ &(f,f,s,f,f) \qquad\qquad (1-p)^{n-1}p, \text{ etc.} \end{aligned} \tag{2.10.4}$$

One sees, this is very cumbersome, and usually unnecessarily so. If we toss a coin 5 times, the only thing we usually want to know is how many successes there were. As long as the experiments are independent, the question how the successes were distributed over the n different trials is far less important. This brings us to the definition of a random variable, and to the concept of a sufficient statistic.

2.11. How to Plot Frequency Vectors and Probability Vectors

If there are only 3 possible outcomes, i.e., $\mathsf{U} = \{\omega_1, \omega_2, \omega_3\}$, then the set of all probability measures is the set of nonnegative 3-vectors whose components sum up to 1. Graphically, such vectors can be represented as points inside a trilateral triangle with height 1: the three components of the vector are the distances of the point to each of the sides of the triangle. The R/Splus-function `triplot` in the `ecmet` package, written by Jim Ramsay ramsay@ramsay2.psych.mcgill.ca, does this, with optional rescaling if the rows of the data matrix do not have unit sums.

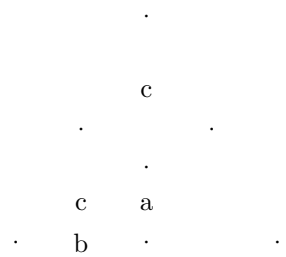


FIGURE 2. Geometry of an equilateral triangle

PROBLEM 41. *In an equilateral triangle, call $a =$ the distance of the sides from the center point, $b =$ half the side length, and $c =$ the distance of the corners from the center point (as in Figure 2). Show that $b = a\sqrt{3}$ and $c = 2a$.*

ANSWER. From $(a + c)^2 + b^2 = 4b^2$, i.e., $(a + c)^2 = 3b^2$, follows $a + c = b\sqrt{3}$. But we also have $a^2 + b^2 = c^2$. Therefore $a^2 + 2ac + c^2 = 3b^2 = 3c^2 - 3a^2$, or $4a^2 + 2ac - 2c^2 = 0$ or $2a^2 + ac - c^2 = (2a - c)(a + c) = 0$. The positive solution is therefore $c = 2a$. This gives $a + c = 3a = b\sqrt{3}$, or $b = a\sqrt{3}$. □

And the function `quadplot`, also written by Jim Ramsey, does *quadrilinear* plots, meaning that proportions for four categories are plotted within a regular tetrahedron. `Quadplot` displays the probability tetrahedron and its points using `XGobi`. Each vertex of the triangle or tetrahedron corresponds to the degenerate probability distribution in which one of the events has probability 1 and the others have probability 0. The labels of these vertices indicate which event has probability 1.

The script `kai` is an example visualizing data from [**Mor65**]; it can be run using the command `ecmet.script(kai)`.

Example: Statistical linguistics.

In the study of ancient literature, the authorship of texts is a perplexing problem. When books were written and reproduced by hand, the rights of authorship were limited and what would now be considered forgery was common. The names of reputable authors were borrowed in order to sell books, get attention for books, or the writings of disciples and collaborators were published under the name of the master, or anonymous old manuscripts were optimistically attributed to famous authors. In the absence of conclusive evidence of authorship, the attribution of ancient texts must be based on the texts themselves, for instance, by statistical analysis of literary style. Here it is necessary to find stylistic criteria which vary from author to author, but are independent of the subject matter of the text. An early suggestion was to use the probability distribution of word length, but this was never acted upon, because it is too dependent on the subject matter. Sentence-length distributions, on the other hand, have proved highly reliable. [**Mor65**, p. 184] says that sentence-length

is "periodic rather than random," therefore the sample should have at least about 100 sentences. "Sentence-length distributions are not suited to dialogue, they cannot be used on commentaries written on one author by another, nor are they reliable on such texts as the fragmentary books of the historian Diodorus Siculus."

PROBLEM 42. *According to* [**Mor65**, p. 184], *sentence-length is "periodic rather than random." What does this mean?*

ANSWER. In a text, passages with long sentences alternate with passages with shorter sentences. This is why one needs at least 100 sentences to get a representative distribution of sentences, and this is why fragments and drafts and commentaries on others' writings do not exhibit an average sentence length distribution: they do not have the melody of the finished text. □

Besides the *length* of sentences, also the number of common words which express a general relation ("and", "in", "but", "I", "to be") is random with the same distribution at least among the same genre. By contrast, the occurrence of the definite article "the" cannot be modeled by simple probabilistic laws because the number of nouns with definite article depends on the subject matter.

Table 1 has data about the epistles of St. Paul. Abbreviations: `Rom` Romans; `Co1` 1st Corinthians; `Co2` 2nd Corinthians; `Gal` Galatians; `Phi` Philippians; `Col` Colossians; `Th1` 1st Thessalonians; `Ti1` 1st Timothy; `Ti2` 2nd Timothy; `Heb` Hebrews. 2nd Thessalonians, Titus, and Philemon were excluded because they were too short to give reliable samples. From an analysis of these and other data [**Mor65**, p. 224] the first 4 epistles (Romans, 1st Corinthians, 2nd Corinthians, and Galatians) form a consistent group, and all the other epistles lie more than 2 standard deviations from the mean of this group (using χ^2 statistics). If Paul is defined as being the author of Galatians, then he also wrote Romans and 1st and 2nd Corinthians. The remaining epistles come from at least six hands.

TABLE 1. Number of Sentences in Paul's Epistles with 0, 1, 2, and ≥ 3 occurrences of *kai*

	Rom	Co1	Co2	Gal	Phi	Col	Th1	Ti1	Ti2	Heb
no *kai*	386	424	192	128	42	23	34	49	45	155
one	141	152	86	48	29	32	23	38	28	94
two	34	35	28	5	19	17	8	9	11	37
3 or more	17	16	13	6	12	9	16	10	4	24

PROBLEM 43. *Enter the data from Table 1 into xgobi and brush the four epistles which are, according to Morton, written by Paul himself. 3 of those points are almost on top of each other, and one is a little apart. Which one is this?*

ANSWER. In R, issue the commands `library(xgobi)` then `data(PaulKAI)` then `quadplot(PaulKAI, normalize = TRUE)`. If you have `xgobi` but not R, this dataset is one of the default datasets coming with `xgobi`.

□

CHAPTER 3

Random Variables

3.1. Notation

Throughout these class notes, lower case bold letters will be used for vectors and upper case bold letters for matrices, and letters that are not bold for scalars. The (i, j) element of the matrix $\boldsymbol{A}$ is a_{ij}, and the ith element of a vector $\boldsymbol{b}$ is b_i; the arithmetic mean of all elements is $\bar{b}$. All vectors are column vectors; if a row vector is needed, it will be written in the form $\boldsymbol{b}^\top$. Furthermore, the on-line version of these notes uses green symbols for random variables, and the corresponding black symbols for the values taken by these variables. If a black-and-white printout of the on-line version is made, then the symbols used for random variables and those used for specific values taken by these random variables can only be distinguished by their grey scale or cannot be distinguished at all; therefore a special monochrome version is available which should be used for the black-and-white printouts. It uses an *upright* math font, called "Euler," for the random variables, and the same letter in the usual slanted italic font for the values of these random variables.

Example: If y is a random vector, then $\boldsymbol{y}$ denotes a particular value, for instance an observation, of the whole vector; y_i denotes the ith element of y (a random scalar), and y_i is a particular value taken by that element (a nonrandom scalar).

With real-valued random variables, the powerful tools of calculus become available to us. Therefore we will begin the chapter about random variables with a digression about infinitesimals

3.2. Digression about Infinitesimals

In the following pages we will recapitulate some basic facts from calculus. But it will differ in two respects from the usual calculus classes. (1) everything will be given its probability-theoretic interpretation, and (2) we will make explicit use of infinitesimals. This last point bears some explanation.

You may say infinitesimals do not exist. Do you know the story with Achilles and the turtle? They are racing, the turtle starts 1 km ahead of Achilles, and Achilles runs ten times as fast as the turtle. So when Achilles arrives at the place the turtle started, the turtle has run 100 meters; and when Achilles has run those 100 meters,

the turtle has run 10 meters, and when Achilles has run the 10 meters, then the turtle has run 1 meter, etc. The Greeks were actually arguing whether Achilles would ever reach the turtle.

This may sound like a joke, but in some respects, modern mathematics never went beyond the level of the Greek philosophers. If a modern mathematicien sees something like

$$\lim_{i\to\infty} \frac{1}{i} = 0, \qquad \text{or} \qquad \lim_{n\to\infty} \sum_{i=0}^{n} \frac{1}{10^i} = \frac{10}{9}, \tag{3.2.1}$$

then he will probably say that the lefthand term in each equation never really reaches the number written on the right, all he will say is that the term on the left comes *arbitrarily close* to it.

This is like saying: I know that Achilles will get as close as 1 cm or 1 mm to the turtle, he will get closer than any distance, however small, to the turtle, instead of simply saying that Achilles reaches the turtle. Modern mathematical proofs are full of races between Achilles and the turtle of the kind: give me an ε, and I will prove to you that the thing will come at least as close as ε to its goal (so-called epsilontism), but never speaking about the moment when the thing will reach its goal.

Of course, it "works," but it makes things terribly cumbersome, and it may have prevented people from seeing connections.

Abraham Robinson in [**Rob74**] is one of the mathematicians who tried to remedy it. He did it by adding more numbers, infinite numbers and infinitesimal numbers. Robinson showed that one can use infinitesimals without getting into contradictions, and he demonstrated that mathematics becomes much more intuitive this way, not only its elementary proofs, but especially the deeper results. One of the elemrntary books based on his calculus is [**HK79**].

The well-know logician Kurt Gödel said about Robinson's work: "I think, in coming years it will be considered a great oddity in the history of mathematics that the first exact theory of infinitesimals was developed 300 years after the invention of the differential calculus."

Gödel called Robinson's theory the *first* theory. I would like to add here the following speculation: perhaps Robinson shares the following error with the "standard" mathematicians whom he criticizes: they consider numbers only in a static way, without allowing them to move. It would be beneficial to expand on the intuition of the inventors of differential calculus, who talked about "fluxions," i.e., quantities in flux, in motion. Modern mathematicians even use arrows in their symbol for limits, but they are not calculating with moving quantities, only with static quantities.

This perspective makes the category-theoretical approach to infinitesimals taken in [**MR91**] especially promising. Category theory considers objects on the same footing with their transformations (and uses lots of arrows).

Maybe a few years from now mathematics will be done right. We should not let this temporary backwardness of mathematics allow to hold us back in our *intuition*. The equation $\frac{\Delta y}{\Delta x} = 2x$ does not hold exactly on a parabola for any pair of given (static) Δx and Δy; but if you take a pair $(\Delta x, \Delta y)$ which is *moving* towards zero then this equation holds *in the moment when they reach zero, i.e., when they vanish.* Writing dy and dx means therefore: we are looking at magnitudes which are in the process of vanishing. If one applies a function to a moving quantity one again gets a moving quantity, and the derivative of this function compares the speed with which the transformed quantity moves with the speed of the original quantity. Likewise, the equation $\sum_{i=1}^{n} \frac{1}{2^n} = 1$ holds *in the moment when n reaches infinity.* From this point of view, the axiom of σ-additivity in probability theory (in its equivalent form of rising or declining sequences of events) indicates that the probability of a vanishing event vanishes.

Whenever we talk about infinitesimals, therefore, we really mean magnitudes which are moving, and which are in the process of vanishing. $dV_{x,y}$ is therefore not, as one might think from what will be said below, a static but small volume element located close to the point (x, y), but it is a volume element which is vanishing into the point (x, y). The probability density function therefore signifies the speed with which the probability of a vanishing element vanishes.

3.3. Definition of a Random Variable

The best intuition of a random variable would be to view it as a numerical variable whose values are not determinate but follow a statistical pattern, and call it x, while possible values of x are called x.

In order to make this a mathematically sound definition, one says: A mapping $x : U \to \mathbb{R}$ of the set U of all possible outcomes into the real numbers $\mathbb{R}$ is called a random variable. (Again, mathematicians are able to construct pathological mappings that cannot be used as random variables, but we let that be their problem, not ours.) The green x is then defined as $\mathsf{x} = x(\omega)$. I.e., all the randomness is shunted off into the process of selecting an element of U. Instead of being an indeterminate function, it is defined as a determinate function of the random ω. It is written here as $x(\omega)$ and not as $\mathsf{x}(\omega)$ because the function itself is determinate, only its argument is random.

Whenever one has a mapping $x : U \to \mathbb{R}$ between sets, one can construct from it in a natural way an "inverse image" mapping between subsets of these sets. Let $\mathcal{F}$, as usual, denote the set of subsets of U, and let $\mathcal{B}$ denote the set of subsets of $\mathbb{R}$. We will define a mapping $x^{-1} : \mathcal{B} \to \mathcal{F}$ in the following way: For any $B \subset \mathbb{R}$, we define $x^{-1}(B) = \{\omega \in U \colon x(\omega) \in B\}$. (This is not the usual inverse of a mapping, which does not always exist. The inverse-image mapping always exists, but the inverse image of a one-element set is no longer necessarily a one-element set; it may have more than one element or may be the empty set.)

This "inverse image" mapping is well behaved with respect to unions and intersections, etc. In other words, we have identities $x^{-1}(A \cap B) = x^{-1}(A) \cap x^{-1}(B)$ and $x^{-1}(A \cup B) = x^{-1}(A) \cup x^{-1}(B)$, etc.

Problem 44. *Prove the above two identities.*

Answer. These are a very subtle proofs. $x^{-1}(A \cap B) = \{\omega \in U \colon x(\omega) \in A \cap B\} = \{\omega \in U \colon x(\omega) \in A \text{ and } x(\omega) \in B = \{\omega \in U \colon x(\omega) \in A\} \cap \{\omega \in U \colon x(\omega) \in B\} = x^{-1}(A) \cap x^{-1}(B)$. The other identity has a similar proof. □

Problem 45. *Show, on the other hand, by a counterexample, that the "direct image" mapping defined by $x(E) = \{r \in \mathbb{R} \colon$ there exists $\omega \in E$ with $x(\omega) = r\}$ no longer satisfies $x(E \cap F) = x(E) \cap x(F)$.*

By taking inverse images under a random variable x, the probability measure on $\mathcal{F}$ is transplanted into a probability measure on the subsets of $\mathbb{R}$ by the simple prescription $\Pr[\mathsf{B}] = \Pr\big[x^{-1}(\mathsf{B})\big]$. Here, B is a subset of $\mathbb{R}$ and $x^{-1}(\mathsf{B})$ one of U, the Pr on the right side is the given probability measure on U, while the Pr on the left is the new probability measure on $\mathbb{R}$ induced by x. This induced probability measure is called the probability law or probability distribution *of the random variable.*

Every random variable induces therefore a probability measure on $\mathbb{R}$, and this probability measure, not the mapping itself, is the most important ingredient of a random variable. That is why Amemiya's first definition of a random variable (definition 3.1.1 on p. 18) is: "A random variable is a variable that takes values acording to a certain distribution." In other words, it is the outcome of an experiment whose set of possible outcomes is $\mathbb{R}$.

3.4. Characterization of Random Variables

We will begin our systematic investigation of random variables with an overview over *all possible probability measures on* $\mathbb{R}$.

The simplest way to get such an overview is to look at the *cumulative distribution functions.* Every probability measure on $\mathbb{R}$ has a cumulative distribution function, but we will follow the common usage of assigning the cumulative distribution not to a probability measure but to the random variable which induces this probability measure on $\mathbb{R}$.

Given a random variable $\mathsf{x} : U \ni \omega \mapsto x(\omega) \in \mathbb{R}$. Then the cumulative distribution function of x is the function $F_{\mathsf{x}} : \mathbb{R} \to \mathbb{R}$ defined by:

$$F_{\mathsf{x}}(a) = \Pr[\{\omega \in U : x(\omega) \leq a\}] = \Pr[\mathsf{x} \leq a]. \tag{3.4.1}$$

This function uniquely defines the probability measure which x induces on $\mathbb{R}$.

Properties of cumulative distribution functions: a function $F\colon \mathbb{R} \to \mathbb{R}$ is a cumulative distribution function if and only if

$$a \le b \Rightarrow F(a) \le F(b) \tag{3.4.2}$$

$$\lim_{a\to-\infty} F(a) = 0 \tag{3.4.3}$$

$$\lim_{a\to\infty} F(a) = 1 \tag{3.4.4}$$

$$\lim_{\varepsilon\to 0,\varepsilon>0} F(a+\varepsilon) = F(a) \tag{3.4.5}$$

Equation (3.4.5) is the definition of *continuity from the right* (because the limit holds only for $\varepsilon \ge 0$). Why is a cumulative distribution function continuous from the right? For every nonnegative sequence $\varepsilon_1, \varepsilon_2, \ldots \ge 0$ converging to zero which also satisfies $\varepsilon_1 \ge \varepsilon_2 \ge \ldots$ follows $\{x \le a\} = \bigcap_i \{x \le a + \varepsilon_i\}$; for these sequences, therefore, the statement follows from what Problem 14 above said about the probability of the intersection of a declining set sequence. And a converging sequence of nonnegative ε_i which is not declining has a declining subsequence.

A cumulative distribution function need not be continuous from the left. If $\lim_{\varepsilon\to 0,\varepsilon>0} F(x-\varepsilon) \ne F(x)$, then x is a jump point, and the height of the jump is the probability that $\mathsf{x} = x$.

It is a matter of convention whether we are working with right continuous or left continuous functions here. If the distribution function were defined as $\Pr[\mathsf{x} < a]$ (some authors do this, compare [**Ame94**, p. 43]), then it would be continuous from the left but not from the right.

Problem 46. *6 points Assume $F_{\mathsf{x}}(x)$ is the cumulative distribution function of the random variable x (whose distribution is not necessarily continuous). Which of the following formulas are correct? Give proofs or verbal justifications.*

$$\Pr[\mathsf{x} = x] = \lim_{\varepsilon>0;\ \varepsilon\to 0} F_{\mathsf{x}}(x+\varepsilon) - F_{\mathsf{x}}(x) \tag{3.4.6}$$

$$\Pr[\mathsf{x} = x] = F_{\mathsf{x}}(x) - \lim_{\delta>0;\ \delta\to 0} F_{\mathsf{x}}(x-\delta) \tag{3.4.7}$$

$$\Pr[\mathsf{x} = x] = \lim_{\varepsilon>0;\ \varepsilon\to 0} F_{\mathsf{x}}(x+\varepsilon) - \lim_{\delta>0;\ \delta\to 0} F_{\mathsf{x}}(x-\delta) \tag{3.4.8}$$

Answer. (3.4.6) does not hold generally, since its rhs is always $= 0$; the other two equations always hold. □

Problem 47. *4 points Assume the distribution of z is symmetric about zero, i.e., $\Pr[\mathsf{z} < -z] = \Pr[\mathsf{z} > z]$ for all z. Call its cumulative distribution function $F_{\mathsf{z}}(z)$. Show that the cumulative distribution function of the random variable $\mathsf{q} = \mathsf{z}^2$ is $F_{\mathsf{q}}(q) = 2F_{\mathsf{z}}(\sqrt{q}) - 1$ for $q \ge 0$, and 0 for $q < 0$.*

Answer. If $q \ge 0$ then

$$F_{\mathsf{q}}(q) = \Pr[\mathsf{z}^2 \le q] = \Pr[-\sqrt{q} \le \mathsf{z} \le \sqrt{q}] \tag{3.4.9}$$

$$= \Pr[\mathsf{z} \le \sqrt{q}] - \Pr[\mathsf{z} < -\sqrt{q}] \tag{3.4.10}$$

$$= \Pr[\mathsf{z} \le \sqrt{q}] - \Pr[\mathsf{z} > \sqrt{q}] \tag{3.4.11}$$

$$= F_{\mathsf{z}}(\sqrt{q}) - (1 - F_{\mathsf{z}}(\sqrt{q})) \tag{3.4.12}$$

$$= 2F_{\mathsf{z}}(\sqrt{q}) - 1. \tag{3.4.13}$$

□

Instead of the cumulative distribution function F_{y} one can also use the *quantile function* F_{y}^{-1} to characterize a probability measure. As the notation suggests, the quantile function can be considered some kind of "inverse" of the cumulative distribution function. The quantile function is the function $(0,1) \to \mathbb{R}$ defined by

$$F_{\mathsf{y}}^{-1}(p) = \inf\{u:\ F_{\mathsf{y}}(u) \ge p\} \tag{3.4.14}$$

or, plugging the definition of F_{y} into (3.4.14),

$$F_{\mathsf{y}}^{-1}(p) = \inf\{u:\ \Pr[\mathsf{y} \le u] \ge p\}. \tag{3.4.15}$$

The quantile function is only defined on the open unit interval, not on the endpoints 0 and 1, because it would often assume the values $-\infty$ and $+\infty$ on these endpoints, and the information given by these values is redundant. The quantile function is continuous from the left, i.e., from the other side than the cumulative distribution function. If F is continuous and strictly increasing, then the quantile function is the inverse of the distribution function in the usual sense, i.e., $F^{-1}(F(t)) = t$ for all $t \in \mathbb{R}$, and $F(F^{-1}((p)) = p$ for all $p \in (0,1)$. But even if F is flat on certain intervals, and/or F has jump points, i.e., F does not have an inverse function, the following important identity holds for every $y \in \mathbb{R}$ and $p \in (0,1)$:

$$p \le F_{\mathsf{y}}(y) \quad \text{iff} \quad F_{\mathsf{y}}^{-1}(p) \le y \tag{3.4.16}$$

Problem 48. *3 points Prove equation* (3.4.16).

Answer. $\Rightarrow$ is trivial: if $F(y) \ge p$ then of course $y \ge \inf\{u:\ F(u) \ge p\}$. $\Leftarrow$: $y \ge \inf\{u:\ F(u) \ge p\}$ means that every $z > y$ satisfies $F(z) \ge p$; therefore, since F is continuous from the right, also $F(y) \ge p$. This proof is from [**Rei89**, p. 318]. □

Problem 49. *You throw a pair of dice and your random variable x is the sum of the points shown.*

- a. *Draw the cumulative distribution function of x.*

Answer. This is Figure 1: the cdf is 0 in $(-\infty, 2)$, 1/36 in [2,3), 3/36 in [3,4), 6/36 in [4,5), 10/36 in [5,6), 15/36 in [6,7), 21/36 in [7,8), 26/36 on [8,9), 30/36 in [9,10), 33/36 in [10,11), 35/36 on [11,12), and 1 in $[12, +\infty)$. □

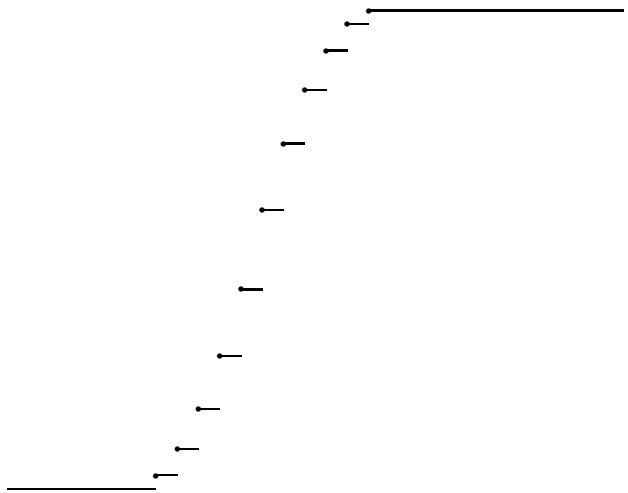


FIGURE 1. Cumulative Distribution Function of Discrete Variable

• b. *Draw the quantile function of* x.

ANSWER. This is Figure 2: the quantile function is 2 in $(0, 1/36]$, 3 in (1/36,3/36], 4 in (3/36,6/36], 5 in (6/36,10/36], 6 in (10/36,15/36], 7 in (15/36,21/36], 8 in (21/36,26/36], 9 in (26/36,30/36], 10 in (30/36,33/36], 11 in (33/36,35/36], and 12 in (35/36,1]. □

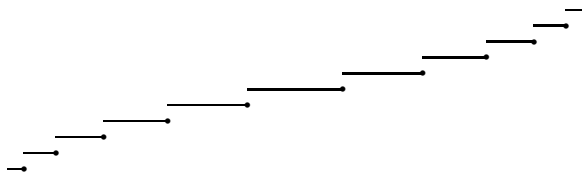


FIGURE 2. Quantile Function of Discrete Variable

PROBLEM 50. *1 point Give the formula of the cumulative distribution function of a random variable which is uniformly distributed between* 0 *and* b.

ANSWER. 0 for $x \leq 0$, x/b for $0 \leq x \leq b$, and 1 for $x \geq b$. □

Empirical Cumulative Distribution Function:

Besides the cumulative distribution function of a random variable or of a probability measure, one can also define the empirical cumulative distribution function of a sample. Empirical cumulative distribution functions are zero for all values below the lowest observation, then $1/n$ for everything below the second lowest, etc. They are step functions. If two observations assume the same value, then the step at that value is twice as high, etc. The empirical cumulative distribution function can be considered an estimate of the cumulative distribution function of the probability distribution underlying the sample. [**Rei89**, p. 12] writes it as a sum of indicator functions:

$$F = \frac{1}{n} \sum_i 1_{[x_i, +\infty)} \tag{3.4.17}$$

3.5. Discrete and Absolutely Continuous Probability Measures

One can define two main classes of probability measures on $\mathbb{R}$:

One kind is concentrated in countably many points. Its probability distribution can be defined in terms of the probability mass function.

PROBLEM 51. *Show that a distribution function can only have countably many jump points.*

ANSWER. Proof: There are at most two with jump height $\geq \frac{1}{2}$, at most four with jump height $\geq \frac{1}{4}$, etc. □

Among the other probability measures we are only interested in those which can be represented by a density function (absolutely continuous). A density function is a nonnegative integrable function which, integrated over the whole line, gives 1. Given such a density function, called $f_{\mathsf{x}}(x)$, the probability $\Pr[\mathsf{x} \in (a, b)] = \int_a^b f_{\mathsf{x}}(x)dx$. The density function is therefore an alternate way to characterize a probability measure. But not all probability measures have density functions.

Those who are not familiar with integrals should read up on them at this point. Start with derivatives, then: the indefinite integral of a function is a function whose derivative is the given function. Then it is an important theorem that the area under the curve is the difference of the values of the indefinite integral at the end points. This is called the definite integral. (The area is considered negative when the curve is below the x-axis.)

The intuition of a density function comes out more clearly in terms of infinitesimals. If $f_{\mathsf{x}}(x)$ is the value of the density function at the point x, then the probability that the outcome of x lies in an interval of infinitesimal length located near the point x is the length of this interval, multiplied by $f_{\mathsf{x}}(x)$. In formulas, for an infinitesimal

dx follows

$$\Pr\big[\mathsf{x}\in[x, x+dx]\big] = f_{\mathsf{x}}(x)\ |dx|\,. \tag{3.5.1}$$

The name "density function" is therefore appropriate: it indicates how densely the probability is spread out over the line. It is, so to say, the quotient between the probability measure induced by the variable, and the length measure on the real numbers.

If the cumulative distribution function has everywhere a derivative, this derivative is the density function.

3.6. Transformation of a Scalar Density Function

Assume x is a random variable with values in the region $A \subset \mathbb{R}$, i.e., $\Pr[\mathsf{x}\notin A] = 0$, and t is a one-to-one mapping $A \to \mathbb{R}$. One-to-one (as opposed to many-to-one) means: if $a, b \in A$ and $t(a) = t(b)$, then already $a = b$. We also assume that t has a continuous nonnegative first derivative $t' \geq 0$ everywhere in A. Define the random variable y by $\mathsf{y} = t(\mathsf{x})$. We know the density function of y, and we want to get that of x. (I.e., t expresses the old variable, that whose density function we know, in terms of the new variable, whose density function we want to know.)

Since t is one-to-one, it follows for all $a, b \in A$ that $a = b \iff t(a) = t(b)$. And recall the definition of a derivative in terms of infinitesimals dx: $t'(x) = \frac{t(x+dx)-t(x)}{dx}$.

In order to compute $f_{\mathsf{x}}(x)$ we will use the following identities valid for all $x \in A$:

$$f_{\mathsf{x}}(x)\ |dx| = \Pr\big[\mathsf{x}\in[x, x+dx]\big] = \Pr\big[t(\mathsf{x})\in[t(x), t(x+dx)]\big] \tag{3.6.1}$$

$$= \Pr\big[t(\mathsf{x})\in[t(x), t(x)+t'(x)\,dx]\big] = f_{\mathsf{y}}(t(x))\ |t'(x)dx| \tag{3.6.2}$$

Absolute values are multiplicative, i.e., $|t'(x)dx| = |t'(x)|\,|dx|$; divide by $|dx|$ to get

$$f_{\mathsf{x}}(x) = f_{\mathsf{y}}\big(t(x)\big)\,|t'(x)|\,. \tag{3.6.3}$$

This is the transformation formula how to get the density of x from that of y. This formula is valid for all $x \in A$; the density of x is 0 for all $x \notin A$.

Heuristically one can get this transformation as follows: write $|t'(x)| = \frac{|dy|}{|dx|}$, then one gets it from $f_{\mathsf{x}}(x)\ |dx| = f_{\mathsf{y}}(t(x))\ |dy|$ by just dividing both sides by $|dx|$.

In other words, this transformation rule consists of 4 steps: (1) Determine A, the range of the new variable; (2) obtain the transformation t which expresses the old variable in terms of the new variable, and check that it is one-to-one on A; (3) plug expression (2) into the old density; (4) multiply this plugged-in density by the absolute value of the derivative of expression (2). This gives the density inside A; it is 0 outside A.

An alternative proof is conceptually simpler but cannot be generalized to the multivariate case: First assume t is monotonically *increasing*. Then $F_{\mathsf{x}}(x) = \Pr[\mathsf{x} \leq$

$x] = \Pr[t(\mathsf{x}) \leq t(i)] = F_{\mathsf{y}}(t(x))$. Now differentiate and use the chain rule. Then also do the monotonically *decresing* case. This is how [**Ame94**, theorem 3.6.1 on pp. 48] does it. [**Ame94**, pp. 52/3] has an extension of this formula to many-to-one functions.

Problem 52. *4 points* [**Lar82**, example 3.5.4 on p. 148] *Suppose* y *has density function*

$$f_{\mathsf{y}}(y) = \begin{cases} 1 & \text{for } 0 < y < 1 \\ 0 & \text{otherwise.} \end{cases} \tag{3.6.4}$$

Obtain the density $f_{\mathsf{x}}(x)$ *of the random variable* $\mathsf{x} = -\log \mathsf{y}$.

Answer. (1) Since y takes values only between 0 and 1, its logarithm takes values between $-\infty$ and 0, the negative logarithm therefore takes values between 0 and $+\infty$, i.e., $A = \{x : 0 < x\}$. (2) Express y in terms of x: $\mathsf{y} = e^{-\mathsf{x}}$. This is one-to-one on the whole line, therefore also on A. (3) Plugging $y = e^{-x}$ into the density function gives the number 1, since the density function does not depend on the precise value of y, as long is we know that $0 < y < 1$ (which we do). (4) The derivative of $y = e^{-x}$ is $-e^{-x}$. As a last step one has to multiply the number 1 by the absolute value of the derivative to get the density inside A. Therefore $f_{\mathsf{x}}(x) = e^{-x}$ for $x > 0$ and 0 otherwise. □

Problem 53. *6 points* [**Dhr86**, p. 1574] *Assume the random variable* z *has the exponential distribution with parameter* λ*, i.e., its density function is* $f_{\mathsf{z}}(z) = \lambda \exp(-\lambda z)$ *for* $z > 0$ *and* 0 *for* $z \leq 0$*. Define* $\mathsf{u} = -\log \mathsf{z}$*. Show that the density function of* u *is* $f_{\mathsf{u}}(u) = \exp\big(\mu - u - \exp(\mu - u)\big)$ *where* $\mu = \log \lambda$*. This density will be used in Problem 151.*

Answer. (1) Since z only has values in $(0, \infty)$, its log is well defined, and $A = \mathbb{R}$. (2) Express old variable in terms of new: $-\mathsf{u} = \log \mathsf{z}$ therefore $\mathsf{z} = e^{-\mathsf{u}}$; this is one-to-one everywhere. (3) plugging in (since $e^{-u} > 0$ for all u, we must plug it into $\lambda \exp(-\lambda z)$) gives .... (4) the derivative of $z = e^{-u}$ is $-e^{-u}$, taking absolute values gives the Jacobian factor e^{-u}. Plugging in and multiplying gives the density of u: $f_{\mathsf{u}}(u) = \lambda \exp(-\lambda e^{-u})e^{-u} = \lambda e^{-u-\lambda e^{-u}}$, and using $\lambda \exp(-u) = \exp(\mu - u)$ this simplifies to the formula above.

Alternative without transformation rule for densities: $F_{\mathsf{u}}(u) = \Pr[\mathsf{u} \leq u] = \Pr[-\log \mathsf{z} \leq u] = \Pr[\log \mathsf{z} \geq -u] = \Pr[\mathsf{z} \geq e^{-u}] = \int_{e^{-u}}^{+\infty} \lambda e^{-\lambda z}\, dz = -e^{-\lambda z}\big|_{e^{-u}}^{+\infty} = e^{-\lambda e^{-u}}$, now differentiate. □

Problem 54. *4 points Assume the random variable* z *has the exponential distribution with* $\lambda = 1$*, i.e., its density function is* $f_{\mathsf{z}}(z) = \exp(-z)$ *for* $z \geq 0$ *and* 0 *for* $z < 0$*. Define* $\mathsf{u} = \sqrt{\mathsf{z}}$*. Compute the density function of* u.

Answer. (1) $A = \{u \colon u \geq 0\}$ since $\sqrt{\ }$ always denotes the nonnegative square root; (2) Express old variable in terms of new: $z = u^2$, this is one-to-one on A (but not one-to-one on all of $\mathbb{R}$); (3) then the derivative is $2u$, which is nonnegative as well, no absolute values are necessary; (4) multiplying gives the density of u: $f_{\mathsf{u}}(u) = 2u \exp(-u^2)$ if $u \geq 0$ and 0 elsewhere. □

3.7. Example: Binomial Variable

Go back to our Bernoulli trial with parameters p and n, and define a random variable x which represents the number of successes. Then the probability mass function of x is

$$p_{\mathsf{x}}(k) = \Pr[\mathsf{x}{=}k] = \binom{n}{k} p^k (1-p)^{(n-k)} \quad k = 0, 1, 2, \ldots, n \tag{3.7.1}$$

Proof is simple, every subset of k elements represents one possibility of spreading out the k successes.

We will call any observed random variable a *statistic*. And we call a statistic t *sufficient for a parameter* θ if and only if for any event A and for any possible value t of t, the conditional probability $\Pr[\mathsf{A}|\mathsf{t}{\leq}t]$ does not involve θ. This means: after observing t no additional information can be obtained about θ from the outcome of the experiment.

Problem 55. *Show that x, the number of successes in the Bernoulli trial with parameters p and n, is a sufficient statistic for the parameter p (the probability of success), with n, the number of trials, a known fixed number.*

Answer. Since the distribution of x is discrete, it is sufficient to show that for any given k, $\Pr[\mathsf{A}|\mathsf{x}{=}k]$ does not involve p whatever the event A in the Bernoulli trial. Furthermore, since the Bernoulli trial with n tries is finite, we only have to show it if A is an elementary event in $\mathcal{F}$, i.e., an event consisting of one element. Such an elementary event would be that the outcome of the trial has a certain given sequence of successes and failures. A general A is the finite disjoint union of all elementary events contained in it, and if the probability of each of these elementary events does not depend on p, then their sum does not either.

Now start with the definition of conditional probability

$$\Pr[\mathsf{A}|\mathsf{x}{=}k] = \frac{\Pr[\mathsf{A} \cap \{\mathsf{x}{=}k\}]}{\Pr[\mathsf{x}{=}k]}. \tag{3.7.2}$$

If A is an elementary event whose number of sucesses is not k, then $\mathsf{A} \cap \{\mathsf{x}{=}k\} = \emptyset$, therefore its probability is 0, which does not involve p. If A is an elementary event which has k successes, then $\mathsf{A} \cap \{\mathsf{x}{=}k\} = \mathsf{A}$, which has probability $p^k(1-p)^{n-k}$. Since $\Pr[\{\mathsf{x}{=}k\}] = \binom{n}{k} p^k (1-p)^{n-k}$, the terms in formula (3.7.2) that depend on p cancel out, one gets $\Pr[\mathsf{A}|\mathsf{x}{=}k] = 1/\binom{n}{k}$. Again there is no p in that formula. □

Problem 56. *You perform a Bernoulli experiment, i.e., an experiment which can only have two outcomes, success s and failure f. The probability of success is p.*

• a. *3 points You make 4 independent trials. Show that the probability that the first trial is successful, given that the total number of successes in the 4 trials is 3, is 3/4.*

Answer. Let $\mathsf{B} = \{sfff, sffs, sfsf, sfss, ssff, ssfs, sssf, ssss\}$ be the event that the first trial is successful, and let $\{\mathsf{x}{=}3\} = \{fsss, sfss, ssfs, sssf\}$ be the event that there are 3 successes,

it has $\binom{4}{3} = 4$ elements. Then

$$\Pr[\mathsf{B}|\mathsf{x}{=}3] = \frac{\Pr[\mathsf{B} \cap \{\mathsf{x}{=}3\}]}{\Pr[\mathsf{x}{=}3]} \tag{3.7.3}$$

Now $\mathsf{B} \cap \{\mathsf{x}{=}3\} = \{sfss, ssfs, sssf\}$, which has 3 elements. Therefore we get

$$\Pr[\mathsf{B}|\mathsf{x}{=}3] = \frac{3 \cdot p^3(1-p)}{4 \cdot p^3(1-p)} = \frac{3}{4}. \tag{3.7.4}$$

□

• b. *2 points Discuss this result.*

Answer. It is significant that this probability is independent of p. I.e., once we know how many successes there were in the 4 trials, knowing the true p does not help us computing the probability of the event. From this also follows that the outcome of the event has no information about p. The value 3/4 is the same as the unconditional probability if $p = 3/4$. I.e., whether we know that the true frequency, the one that holds in the long run, is 3/4, or whether we know that the actual frequency in this sample is 3/4, both will lead us to the same predictions regarding the first throw. But not all conditional probabilities are equal to their unconditional counterparts: the conditional probability to get 3 successes in the first 4 trials is 1, but the unconditional probability is of course not 1. □

3.8. Pitfalls of Data Reduction: The Ecological Fallacy

The nineteenth-century sociologist Emile Durkheim collected data on the frequency of suicides and the religious makeup of many contiguous provinces in Western Europe. He found that, on the average, provinces with greater proportions of Protestants had higher suicide rates and those with greater proportions of Catholics lower suicide rates. Durkheim concluded from this that Protestants are more likely to commit suicide than Catholics. But this is not a compelling conclusion. It may have been that Catholics in predominantly Protestant provinces were taking their own lives. The oversight of this logical possibility is called the "Ecological Fallacy" [**Sel58**].

This seems like a far-fetched example, but arguments like this have been used to discredit data establishing connections between alcoholism and unemployment etc. as long as the unit of investigation is not the individual but some aggregate.

One study [**RZ78**] found a positive correlation between driver education and the incidence of fatal automobile accidents involving teenagers. Closer analysis showed that the net effect of driver education was to put more teenagers on the road and therefore to increase rather than decrease the number of fatal crashes involving teenagers.

Problem 57. *4 points Assume your data show that counties with high rates of unemployment also have high rates of heart attacks. Can one conclude from this that the unemployed have a higher risk of heart attack? Discuss, besides the "ecological fallacy," also other objections which one might make against such a conclusion.*

Answer. Ecological fallacy says that such a conclusion is only legitimate if one has individual data. Perhaps a rise in unemployment is associated with increased pressure and increased workloads among the employed, therefore it is the employed, not the unemployed, who get the heart attacks. Even if one has individual data one can still raise the following objection: perhaps unemployment and heart attacks are both consequences of a third variable (both unemployment and heart attacks depend on age or education, or freezing weather in a farming community causes unemployment for workers and heart attacks for the elderly).

□

But it is also possible to commit the opposite error and rely too much on individual data and not enough on "neighborhood effects." In a relationship between health and income, it is much more detrimental for your health if you are poor in a poor neighborhood, than if you are poor in a rich neighborhood; and even wealthy people in a poor neighborhood do not escape some of the health and safety risks associated with this neighborhood.

Another pitfall of data reduction is Simpson's paradox. According to table 1, the new drug was better than the standard drug both in urban and rural areas. But if you aggregate over urban and rural areas, then it looks like the standard drug was better than the new drug. This is an artificial example from [**Spr98**, p. 360].

Responses in Urban and Rural Areas to Each of Two Drugs				
	Standard Drug		New Drug	
	Urban	Rural	Urban	Rural
No Effect	500	350	1050	120
Cure	100	350	359	180

Table 1. Disaggregated Results of a New Drug

Response to Two Drugs		
	Standard Drug	New Drug
No Effect	850	1170
Cure	450	530

Table 2. Aggregated Version of Table 1

3.9. Independence of Random Variables

The concept of independence can be extended to random variables: x and y are independent if all events that can be defined in terms of x are independent of all events that can be defined in terms of y, i.e., all events of the form $\{\omega \in U\colon\ x(\omega) \in C\}$ are independent of all events of the form $\{\omega \in U\colon\ y(\omega) \in D\}$ with arbitrary (measurable)

subsets $C, D \subset \mathbb{R}$. Equivalent to this is that all events of the sort $\mathsf{x} \leq a$ are independent of all events of the sort $\mathsf{y} \leq b$.

Problem 58. *3 points The simplest random variables are indicator functions, i.e., functions which can only take the values* 0 *and* 1. *Assume* x *is indicator function of the event* A *and* y *indicator function of the event* B, *i.e.,* x *takes the value 1 if* A *occurs, and the value 0 otherwise, and similarly with* y *and* B. *Show that according to the above definition of independence,* x *and* y *are independent if and only if the events* A *and* B *are independent. (Hint: which are the only two events, other than the certain event* U *and the null event* $\emptyset$, *that can be defined in terms of* x*)?*

Answer. Only A and A'. Therefore we merely need the fact, shown in Problem 35, that if A and B are independent, then also A and B' are independent. By the same argument, also A' and B are independent, and A' and B' are independent. This is all one needs, except the observation that every event is independent of the certain event and the null event. □

3.10. Location Parameters and Dispersion Parameters of a Random Variable

3.10.1. Measures of Location. A location parameter of random variables is a parameter which increases by c if one adds the constant c to the random variable.

The *expected value* is the most important location parameter. To motivate it, assume x is a discrete random variable, i.e., it takes the values $x_1, \ldots, x_r$ with probabilities $p_1, \ldots, p_r$ which sum up to one: $\sum_{i=1}^{r} p_i = 1$. x is observed n times independently. What can we expect the average value of x to be? For this we first need a formula for this average: if k_i is the number of times that x assumed the value x_i $(i = 1, \ldots, r)$ then $\sum k_i = n$, and the average is $\frac{k_1}{n}x_1 + \cdots + \frac{k_n}{n}x_n$. With an appropriate definition of convergence, the relative frequencies $\frac{k_i}{n}$ converge towards p_i. Therefore the average converges towards $p_1x_1 + \cdots + p_nx_n$. This limit is the expected value of x, written as

$$\mathrm{E}[\mathsf{x}] = p_1x_1 + \cdots + p_nx_n. \tag{3.10.1}$$

Problem 59. *Why can one not use the usual concept of convergence here?*

Answer. Because there is no guarantee that the sample frequencies converge. It is not physically impossible (although it is highly unlikely) that certain outcome will never be realized. □

Note the difference between the sample mean, i.e., the average measured in a given sample, and the "population mean" or expected value. The former is a random variable, the latter is a parameter. I.e., the former takes on a different value every time the experiment is performed, the latter does not.

Note that the expected value of the number of dots on a die is 3.5, which is not one of the possible outcomes when one rolls a die.

Expected value can be visualized as the center of gravity of the probability mass. If one of the tails has its weight so far out that there is no finite balancing point then the expected value is infinite of minus infinite. If both tails have their weights so far out that neither one has a finite balancing point, then the expected value does not exist.

It is trivial to show that for a function $g(\mathsf{x})$ (which only needs to be defined for those values which x can assume with nonzero probability), $\mathrm{E}[g(\mathsf{x})] = p_1 g(x_1) + \cdots + p_n g(x_n)$.

Example of a countable probability mass distribution which has an infinite expected value: $\Pr[\mathsf{x} = x] = \frac{a}{x^2}$ for $x = 1, 2, \ldots$. (a is the constant $1 \Big/ \sum_{i=1}^{\infty} \frac{1}{i^2}$.) The expected value of x would be $\sum_{i=1}^{\infty} \frac{a}{i}$, which is infinite. But if the random variable is bounded, then its expected value exists.

The expected value of a *continuous* random variable is defined in terms of its density function:

$$\mathrm{E}[\mathsf{x}] = \int_{-\infty}^{+\infty} x f_{\mathsf{x}}(x)\, dx \tag{3.10.2}$$

It can be shown that for any function $g(x)$ defined for all those x for which $f_{\mathsf{x}}(x) \neq 0$ follows:

$$\mathrm{E}[g(\mathsf{x})] = \int_{f_{\mathsf{x}}(x) \neq 0} g(x) f_{\mathsf{x}}(x)\, dx \tag{3.10.3}$$

Here the integral is taken over all the points which have nonzero density, instead of the whole line, because we did not require that the function g is defined at the points where the density is zero.

Problem 60. *Let the random variable x have the Cauchy distribution, i.e., its density function is*

$$f_{\mathsf{x}}(x) = \frac{1}{\pi(1 + x^2)} \tag{3.10.4}$$

Show that x does not have an expected value.

Answer.

$$\int \frac{x\, dx}{\pi(1 + x^2)} = \frac{1}{2\pi} \int \frac{2x\, dx}{1 + x^2} = \frac{1}{2\pi} \int \frac{d(x^2)}{1 + x^2} = \frac{1}{2\pi} \ln(1 + x^2) \tag{3.10.5}$$

□

Rules about how to calculate with expected values (as long as they exist):

$$\mathrm{E}[c] = c \text{ if } c \text{ is a constant} \tag{3.10.6}$$

$$\mathrm{E}[c\mathsf{h}] = c\,\mathrm{E}[\mathsf{h}] \tag{3.10.7}$$

$$\mathrm{E}[\mathsf{h} + \mathsf{j}] = \mathrm{E}[\mathsf{h}] + \mathrm{E}[\mathsf{j}] \tag{3.10.8}$$

and if the random variables h and j are independent, then also

$$\mathrm{E}[\mathsf{h}\mathsf{j}] = \mathrm{E}[\mathsf{h}]\,\mathrm{E}[\mathsf{j}]. \tag{3.10.9}$$

Problem 61. *2 points You make two independent trials of a Bernoulli experiment with success probability θ, and you observe t, the number of successes. Compute the expected value of t^3. (Compare also Problem 197.)*

Answer. $\Pr[t = 0] = (1 - \theta)^2$; $\Pr[t = 1] = 2\theta(1 - \theta)$; $\Pr[t = 2] = \theta^2$. Therefore an application of (3.10.1) gives $\mathrm{E}[t^3] = 0^3 \cdot (1 - \theta)^2 + 1^3 \cdot 2\theta(1 - \theta) + 2^3 \cdot \theta^2 = 2\theta + 6\theta^2$. □

Theorem 3.10.1. *Jensen's Inequality: Let $g : \mathbb{R} \to \mathbb{R}$ be a function which is convex on an interval $B \subset \mathbb{R}$, which means*

$$g(\lambda a + (1 - \lambda)b) \leq \lambda g(a) + (1 - \lambda) g(b) \tag{3.10.10}$$

for all $a, b \in B$. Furthermore let $\mathsf{x} : \mathbb{R} \to \mathbb{R}$ be a random variable so that $\Pr[\mathsf{x} \in B] = 1$. Then $g(\mathrm{E}[\mathsf{x}]) \leq \mathrm{E}[g(\mathsf{x})]$.

Proof. The Jensen inequality holds with equality if $h(x)$ is a linear function (with a constant term), i.e., in this case, $\mathrm{E}[h(\mathsf{x})] = h(\mathrm{E}[\mathsf{x}])$. (2) Therefore Jensen's inequality is proved if we can find a linear function h with the two properties $h(\mathrm{E}[\mathsf{x}]) = g(\mathrm{E}[\mathsf{x}])$, and $h(x) \leq g(x)$ for all other x—because with such a h, $\mathrm{E}[g(\mathsf{x})] \geq \mathrm{E}[h(\mathsf{x})] = h(\mathrm{E}[\mathsf{x}])$. (3) The existence of such a h follows from convexity. Since g is convex, for every point $a \in B$ there is a number β so that $g(x) \geq g(a) + \beta(x - a)$. This β is the slope of g if g is differentiable, and otherwise it is some number between the left and the right derivative (which both always exist for a convex function). We need this for $a = \mathrm{E}[\mathsf{x}]$.

This existence is the deepest part of this proof. We will not prove it here, for a proof see [**Rao73**, pp. 57, 58]. One can view it as a special case of the separating hyperplane theorem. □

Problem 62. *Use Jensen's inequality to show that $(\mathrm{E}[\mathsf{x}])^2 \leq \mathrm{E}[\mathsf{x}^2]$. You are allowed to use, without proof, the fact that a function is convex on B if the second derivative exists on B and is nonnegative.*

Problem 63. *Show that the expected value of the empirical distribution of a sample is the sample mean.*

Other measures of locaction: The *median* is that number m for which there is as much probability mass to the left of m as to the right, i.e.,

$$\Pr[\mathsf{x}\leq m] = \frac{1}{2} \qquad \text{or, equivalently,} \qquad F_{\mathsf{x}}(m) = \frac{1}{2}. \tag{3.10.11}$$

It is much more robust with respect to outliers than the mean. If there is more than one m satisfying (3.10.11), then some authors choose the smallest (in which case the median is a special case of the quantile function $m = F^{-1}(1/2)$), and others the average between the biggest and smallest. If there is no m with property (3.10.11), i.e., if the cumulative distribution function jumps from a value that is less than $\frac{1}{2}$ to a value that is greater than $\frac{1}{2}$, then the median is this jump point.

The *mode* is the point where the probability mass function or the probability density function is highest.

3.10.2. Measures of Dispersion. Here we will discuss *variance*, *standard deviation*, and *quantiles* and *percentiles:* The variance is defined as

$$\operatorname{var}[\mathsf{x}] = \operatorname{E}[(\mathsf{x} - \operatorname{E}[\mathsf{x}])^2], \tag{3.10.12}$$

but the formula

$$\operatorname{var}[\mathsf{x}] = \operatorname{E}[\mathsf{x}^2] - (\operatorname{E}[\mathsf{x}])^2 \tag{3.10.13}$$

is usually more convenient.

How to calculate with variance?

$$\operatorname{var}[a\mathsf{x}] = a^2 \operatorname{var}[\mathsf{x}] \tag{3.10.14}$$

$$\operatorname{var}[\mathsf{x} + c] = \operatorname{var}[\mathsf{x}] \text{ if } c \text{ is a constant} \tag{3.10.15}$$

$$\operatorname{var}[\mathsf{x} + \mathsf{y}] = \operatorname{var}[\mathsf{x}] + \operatorname{var}[\mathsf{y}] \text{ if } \mathsf{x} \text{ and } \mathsf{y} \text{ are independent.} \tag{3.10.16}$$

Note that the variance is additive only when x and y are independent; the expected value is always additive.

Problem 64. *Here we make the simple step from the definition of the variance to the usually more convenient formula* (3.10.13).

• a. *2 points Derive the formula* $\operatorname{var}[\mathsf{x}] = \operatorname{E}[\mathsf{x}^2] - (\operatorname{E}[\mathsf{x}])^2$ *from the definition of a variance, which is* $\operatorname{var}[\mathsf{x}] = \operatorname{E}[(\mathsf{x} - \operatorname{E}[\mathsf{x}])^2]$. *Hint: it is convenient to define* $\mu = \operatorname{E}[\mathsf{x}]$. *Write it down carefully, you will lose points for missing or unbalanced parentheses or brackets.*

Answer. Here it is side by side with and without the notation $\operatorname{E}[\mathsf{x}] = \mu$:

$$\begin{aligned} \operatorname{var}[\mathsf{x}] &= \operatorname{E}[(\mathsf{x} - \operatorname{E}[\mathsf{x}])^2] & \operatorname{var}[\mathsf{x}] &= \operatorname{E}[(\mathsf{x} - \mu)^2] \\ &= \operatorname{E}[\mathsf{x}^2 - 2\mathsf{x}(\operatorname{E}[\mathsf{x}]) + (\operatorname{E}[\mathsf{x}])^2] & &= \operatorname{E}[\mathsf{x}^2 - 2\mathsf{x}\mu + \mu^2] \\ &= \operatorname{E}[\mathsf{x}^2] - 2(\operatorname{E}[\mathsf{x}])^2 + (\operatorname{E}[\mathsf{x}])^2 & &= \operatorname{E}[\mathsf{x}^2] - 2\mu^2 + \mu^2 \\ &= \operatorname{E}[\mathsf{x}^2] - (\operatorname{E}[\mathsf{x}])^2. & &= \operatorname{E}[\mathsf{x}^2] - \mu^2. \end{aligned} \tag{3.10.17}$$

□

• b. *1 point Assume* $\operatorname{var}[\mathsf{x}] = 3$, $\operatorname{var}[\mathsf{y}] = 2$, x *and* y *are independent. Compute* $\operatorname{var}[-\mathsf{x}]$, $\operatorname{var}[3\mathsf{y} + 5]$, *and* $\operatorname{var}[\mathsf{x} - \mathsf{y}]$.

Answer. 3, 18, and 5. □

Problem 65. *If all* y_i *are independent with same variance* σ^2, *then show that* $\bar{\mathsf{y}}$ *has variance* σ^2/n.

The *standard deviation* is the square root of the variance. Often preferred because has same scale as x. The variance, on the other hand, has the advantage of a simple addition rule.

Standardization: if the random variable x has expected value μ and standard deviation σ, then $\mathsf{z} = \frac{\mathsf{x}-\mu}{\sigma}$ has expected value zero and variance one.

An αth quantile or a 100αth percentile of a random variable x was already defined previously to be the smallest number x so that $\Pr[\mathsf{x}\leq x] \geq \alpha$.

3.10.3. Mean-Variance Calculations. If one knows mean and variance of a random variable, one does not by any means know the whole distribution, but one has already some information. For instance, one can compute $\operatorname{E}[\mathsf{y}^2]$ from it, too.

Problem 66. *4 points Consumer M has an expected utility function for money income* $u(x) = 12x - x^2$. *The meaning of an expected utility function is very simple: if he owns an asset that generates some random income* y, *then the utility he derives from this asset is the expected value* $\operatorname{E}[u(\mathsf{y})]$. *He is contemplating acquiring two assets. One asset yields an income of 4 dollars with certainty. The other yields an expected income of 5 dollars with standard deviation 2 dollars. Does he prefer the certain or the uncertain asset?*

Answer. $\operatorname{E}[u(\mathsf{y})] = 12\operatorname{E}[\mathsf{y}] - \operatorname{E}[\mathsf{y}^2] = 12\operatorname{E}[\mathsf{y}] - \operatorname{var}[\mathsf{y}] - (\operatorname{E}[\mathsf{y}])^2$. Therefore the certain asset gives him utility $48 - 0 - 16 = 32$, and the uncertain one $60 - 4 - 25 = 31$. He prefers the certain asset. □

3.10.4. Moment Generating Function and Characteristic Function. Here we will use the exponential function e^x, also often written $\exp(x)$, which has the two properties: $e^x = \lim_{n\to\infty}(1 + \frac{x}{n})^n$ (Euler's limit), and $e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$.

Many (but not all) random variables x have a *moment generating function* $m_{\mathsf{x}}(t)$ for certain values of t. If they do for t in an open interval around zero, then their distribution is uniquely determined by it. The definition is

$$m_{\mathsf{x}}(t) = \mathrm{E}[e^{t\mathsf{x}}] \tag{3.10.18}$$

It is a powerful computational device.

The moment generating function is in many cases a more convenient characterization of the random variable than the density function. It has the following uses:

1. One obtains the moments of x by the simple formula

$$\mathrm{E}[\mathsf{x}^k] = \frac{d^k}{dt^k} m_{\mathsf{x}}(t)\Big|_{t=0}. \tag{3.10.19}$$

Proof:

$$e^{t\mathsf{x}} = 1 + t\mathsf{x} + \frac{t^2\mathsf{x}^2}{2!} + \frac{t^3\mathsf{x}^3}{3!} + \cdots \tag{3.10.20}$$

$$m_{\mathsf{x}}(t) = \mathrm{E}[e^{t\mathsf{x}}] = 1 + t\,\mathrm{E}[\mathsf{x}] + \frac{t^2}{2!}\,\mathrm{E}[\mathsf{x}^2] + \frac{t^3}{3!}\,\mathrm{E}[\mathsf{x}^3] + \cdots \tag{3.10.21}$$

$$\frac{d}{dt} m_{\mathsf{x}}(t) = \mathrm{E}[\mathsf{x}] + t\,\mathrm{E}[\mathsf{x}^2] + \frac{t^2}{2!}\,\mathrm{E}[\mathsf{x}^3] + \cdots \tag{3.10.22}$$

$$\frac{d^2}{dt^2} m_{\mathsf{x}}(t) = \mathrm{E}[\mathsf{x}^2] + t\,\mathrm{E}[\mathsf{x}^3] + \cdots \quad \text{etc.} \tag{3.10.23}$$

2. The moment generating function is also good for determining the probability distribution of linear combinations of independent random variables.

a. it is easy to get the m.g.f. of $\lambda\mathsf{x}$ from the one of x:

$$m_{\lambda\mathsf{x}}(t) = m_{\mathsf{x}}(\lambda t) \tag{3.10.24}$$

because both sides are $\mathrm{E}[e^{\lambda t\mathsf{x}}]$.

b. If x, y independent, then

$$m_{\mathsf{x}+\mathsf{y}}(t) = m_{\mathsf{x}}(t) m_{\mathsf{y}}(t). \tag{3.10.25}$$

The proof is simple:

$$\mathrm{E}[e^{t(\mathsf{x}+\mathsf{y})}] = \mathrm{E}[e^{t\mathsf{x}} e^{t\mathsf{y}}] = \mathrm{E}[e^{t\mathsf{x}}]\,\mathrm{E}[e^{t\mathsf{y}}] \quad \text{due to independence.} \tag{3.10.26}$$

The *characteristic function* is defined as $\psi_{\mathsf{x}}(t) = \mathrm{E}[e^{it\mathsf{x}}]$, where $i = \sqrt{-1}$. It has the disadvantage that it involves complex numbers, but it has the advantage that it always exists, since $\exp(ix) = \cos x + i \sin x$. Since cos and sin are both bounded, they always have an expected value.

And, as its name says, the characteristic function characterizes the probability distribution. Analytically, many of its properties are similar to those of the moment generating function.

3.11. Entropy

3.11.1. Definition of Information. Entropy is the *average* information gained by the performance of the experiment. The *actual* information yielded by an event A with probabbility $\Pr[\mathsf{A}] = p \neq 0$ is defined as follows:

$$I[\mathsf{A}] = \log_2 \frac{1}{\Pr[\mathsf{A}]} \tag{3.11.1}$$

This is simply a transformation of the probability, and it has the dual interpretation of either how unexpected the event was, or the informaton yielded by the occurrense of event A. It is characterized by the following properties [**AD75**, pp. 3–5]:

- $I[\mathsf{A}]$ only depends on the probability of A, in other words, the information content of a message is independent of how the information is coded.
- $I[\mathsf{A}] \geq 0$ (nonnegativity), i.e., after knowing whether A occurred we are no more ignorant than before.
- If A and B are independent then $I[\mathsf{A} \cap \mathsf{B}] = I[\mathsf{A}] + I[\mathsf{B}]$ (additivity for independent events). This is the most important property.
- Finally the (inessential) normalization that if $\Pr[\mathsf{A}] = 1/2$ then $I[\mathsf{A}] = 1$, i.e., a yes-or-no decision with equal probability (coin flip) is one unit of information.

Note that the information yielded by occurrence of the certain event is 0, and that yielded by occurrence of the impossible event is ∞.

But the important information-theoretic results refer to average, not actual, information, therefore let us define now *entropy:*

3.11.2. Definition of Entropy. The entropy of a probability field (experiment) is a measure of the uncertainty prevailing before the experiment is performed, or of the *average* information yielded by the performance of this experiment. If the set U of possible outcomes of the experiment has only a finite number of different elements, say their number is n, and the probabilities of these outcomes are $p_1, \ldots, p_n$, then the Shannon entropy $\mathrm{H}[\mathcal{F}]$ of this experiment is defined as

$$\frac{\mathrm{H}[\mathcal{F}]}{\text{bits}} = \sum_{k=1}^{n} p_k \log_2 \frac{1}{p_k} \tag{3.11.2}$$

This formula uses $\log_2$, logarithm with base 2, which can easily be computed from the natural logarithms, $\log_2 x = \log x / \log 2$. The choice of base 2 is convenient because in this way the most informative Bernoulli experiment, that with success probability $p = 1/2$ (coin flip), has entropy 1. This is why one says: "the entropy is measured in bits." If one goes over to logarithms of a different base, this simply means that one measures entropy in different units. In order to indicate this dependence on the measuring unit, equation (3.11.2) was written as the definition $\frac{\mathrm{H}[\mathcal{F}]}{\text{bits}}$ instead of $\mathrm{H}[\mathcal{F}]$

itself, i.e., this is the number one gets if one measures the entropy in bits. If one uses natural logarithms, then the entropy is measured in "nats."

Entropy can be characterized axiomatically by the following axioms [**Khi57**]:

- The uncertainty associated with a finite complete scheme takes its largest value if all events are equally likely, i.e., $\mathrm{H}(p_1, \dots, p_n) \leq \mathrm{H}(1/n, \dots, 1/n)$.
- The addition of an impossible event to a scheme does not change the amount of uncertainty.
- *Composition Law:* If the possible outcomes are arbitrarily combined into m groups $\mathsf{W}_1 = \mathsf{X}_{11} \cup \cdots \cup \mathsf{X}_{1k_1}$, $\mathsf{W}_2 = \mathsf{X}_{21} \cup \cdots \cup \mathsf{X}_{2k_2}$, $\dots$, $\mathsf{W}_m = \mathsf{X}_{m1} \cup \cdots \cup \mathsf{X}_{mk_m}$, with corresponding probabilities $w_1 = p_{11} + \cdots + p_{1k_1}$, $w_2 = p_{21} + \cdots + p_{2k_2}$, $\dots$, $w_m = p_{m1} + \cdots + p_{mk_m}$, then

$$\begin{aligned}\mathrm{H}(p_1, \dots, p_n) = \mathrm{H}(w_1, \dots, w_n) &+ \\ + w_1 H(p_{11}/w_1 + \cdots + p_{1k_1}/w_1) &+ \\ + w_2 H(p_{21}/w_2 + \cdots + p_{2k_2}/w_2) + \cdots &+ \\ + w_m H(p_{m1}/w_m + \cdots + p_{mk_m}/w_m)&.\end{aligned}$$

Since $p_{ij}/w_j = \Pr[\mathsf{X}_{ij}|W_j]$, the composition law means: if you first learn half the outcome of the experiment, and then the other half, you will in the average get as much information as if you had been told the total outcome all at once.

The entropy of a *random variable* x is simply the entropy of the probability field induced by x on $\mathbb{R}$. It does not depend on the values x takes but only on the probabilities. For discretely distributed random variables it can be obtained by the following "eerily self-referential" prescription: plug the random variable into its own probability mass function and compute the expected value of the negative logarithm of this, i.e.,

$$\frac{\mathrm{H}[\mathsf{x}]}{\mathrm{bits}} = \mathrm{E}[-\log_2 p_{\mathsf{x}}(\mathsf{x})] \tag{3.11.3}$$

One interpretation of the entropy is: it is the average number of yes-or-no questions necessary to describe the outcome of the experiment. For instance, consider an experiment which has 32 different outcomes occurring with equal probabilities. The entropy is

$$\frac{\mathrm{H}}{\mathrm{bits}} = \sum_{i=1}^{32} \frac{1}{32} \log_2 32 = \log_2 32 = 5 \quad \text{i.e.,} \qquad \mathrm{H} = 5\,\mathrm{bits} \tag{3.11.4}$$

which agrees with the number of bits necessary to describe the outcome.

Problem 67. *Design a questioning scheme to find out the value of an integer between 1 and 32, and compute the expected number of questions in your scheme if all numbers are equally likely.*

Answer. In binary digits one needs a number of length 5 to describe a number between 0 and 31, therefore the 5 questions might be: write down the binary expansion of your number minus 1. Is the first binary digit in this expansion a zero, then: is the second binary digit in this expansion a zero, etc. Formulated without the use of binary digits these same questions would be: is the number between 1 and 16?, then: is it between 1 and 8 or 17 and 24?, then, is it between 1 and 4 or 9 and 12 or 17 and 20 or 25 and 28?, etc., the last question being whether it is odd. Of course, you can formulate those questions conditionally: First: between 1 and 16? if no, then second: between 17 and 24? if yes, then second: between 1 and 8? Etc. Each of these questions gives you exactly the entropy of 1 bit. □

Problem 68. [**CT91**, example 1.1.2 on p. 5] *Assume there is a horse race with eight horses taking part. The probabilities for winning for the eight horses are* $\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}$.

- a. *1 point Show that the entropy of the horse race is* 2 bits.

Answer.

$$\begin{aligned}\frac{\mathrm{H}}{\mathrm{bits}} &= \frac{1}{2}\log_2 2 + \frac{1}{4}\log_2 4 + \frac{1}{8}\log_2 8 + \frac{1}{16}\log_2 16 + \frac{4}{64}\log_2 64 = \\ &= \frac{1}{2} + \frac{1}{2} + \frac{3}{8} + \frac{1}{4} + \frac{3}{8} = \frac{4+4+3+2+3}{8} = 2\end{aligned}$$

□

- b. *1 point Suppose you want to send a binary message to another person indicating which horse won the race. One alternative is to assign the bit strings* 000, 001, 010, 011, 100, 101, 110, 111 *to the eight horses. This description requires* 3 bits *for any of the horses. But since the win probabilities are not uniform, it makes sense to use shorter descriptions for the horses more likely to win, so that we achieve a lower expected value of the description length. For instance, we could use the following set of bit strings for the eight horses:* 0, 10, 110, 1110, 111100, 111101, 111110, 111111. *Show that the the expected length of the message you send to your friend is* 2 bits, *as opposed to* 3 bits *for the uniform code. Note that in this case the expected value of the description length is equal to the entropy.*

Answer. The math is the same as in the first part of the question:

$$\frac{1}{2}\cdot 1 + \frac{1}{4}\cdot 2 + \frac{1}{8}\cdot 3 + \frac{1}{16}\cdot 4 + 4\cdot\frac{1}{64}\cdot 6 = \frac{1}{2} + \frac{1}{2} + \frac{3}{8} + \frac{1}{4} + \frac{3}{8} = \frac{4+4+3+2+3}{8} = 2$$

□

Problem 69. [**CT91**, example 2.1.2 on pp. 14/15]*: The experiment has four possible outcomes; outcome* x*=a occurs with probability* 1/2, x*=b with probability* 1/4, x*=c with probability* 1/8, *and* x*=d with probability* 1/8.

- a. *2 points The entropy of this experiment (in* bits*) is one of the following three numbers: 11/8, 7/4, 2. Which is it?*

• b. *2 points Suppose we wish to determine the outcome of this experiment with the minimum number of questions. An efficient first question is "Is* $\mathsf{x}=a$*?" This splits the probability in half. If the answer to the first question is no, then the second question can be "Is* $\mathsf{x}=b$*?" The third question, if it is necessary, can then be: "Is* $\mathsf{x}=c$*?" Compute the expected number of binary questions required.*

• c. *2 points Show that the entropy gained by each question is 1 bit.*

• d. *3 points Assume we know about the first outcome that* $\mathsf{x}\neq a$*. What is the entropy of the remaining experiment (i.e., under the conditional probability)?*

• e. *5 points Show in this example that the composition law for entropy holds.*

Problem 70. *2 points In terms of natural logarithms equation* (3.11.4) *defining entropy reads*

$$\frac{\mathrm{H}}{bits} = \frac{1}{\ln 2}\sum_{k=1}^{n} p_k \ln\frac{1}{p_k}. \tag{3.11.5}$$

Compute the entropy of (i.e., the average informaton gained by) a roll of an unbiased die.

Answer. Same as the actual information gained, since each outcome is equally likely:

$$\frac{\mathrm{H}}{\mathrm{bits}} = \frac{1}{\ln 2}\left(\frac{1}{6}\ln 6 + \cdots + \frac{1}{6}\ln 6\right) = \frac{\ln 6}{\ln 2} = 2.585 \tag{3.11.6}$$

□

• a. *3 points How many questions does one need in the average to determine the outcome of the roll of an unbiased die? In other words, pick a certain questioning scheme (try to make it efficient) and compute the average number of questions if this scheme is followed. Note that this average cannot be smaller than the entropy* $\mathrm{H}/bits$*, and if one chooses the questions optimally, it is smaller than* $\mathrm{H}/bits + 1$*.*

Answer. First question: is it bigger than 3? Second question: is it even? Third question (if necessary): is it a multiple of 3? In this scheme, the number of questions for the six faces of the die are $3, 2, 3, 3, 2, 3$, therefore the average is $\frac{4}{6}\cdot 3 + \frac{2}{6}\cdot 2 = 2\frac{2}{3}$. Also optimal: (1) is it bigger than 2? (2) is it odd? (3) is it bigger than 4? Gives $2, 2, 3, 3, 3, 3$. Also optimal: 1st question: is it 1 or 2? If anser is no, then second question is: is it 3 or 4?; otherwise go directly to the third question: is it odd or even? The steamroller approach: Is it 1? Is it 2? etc. gives $1, 2, 3, 4, 5, 5$ with expected number $3\frac{1}{3}$. Even this is here $< 1 + \mathrm{H}/\mathrm{bits}$. □

Problem 71.

• a. *1 point Compute the entropy of a roll of two unbiased dice if they are distinguishable.*

Answer. Just twice the entropy from Problem 70.

$$\frac{\mathrm{H}}{\mathrm{bits}} = \frac{1}{\ln 2}\left(\frac{1}{36}\ln 36 + \cdots + \frac{1}{36}\ln 36\right) = \frac{\ln 36}{\ln 2} = 5.170 \tag{3.11.7}$$

□

• b. *Would you expect the entropy to be greater or less in the more usual case that the dice are indistinguishable? Check your answer by computing it.*

Answer. If the dice are indistinguishable, then one gets less information, therefore the experiment has less entropy. One has six like pairs with probability $1/36$ and $6\cdot 5/2 = 15$ unlike pairs with probability $2/36 = 1/18$ each. Therefore the average information gained is

$$\frac{\mathrm{H}}{\mathrm{bits}} = \frac{1}{\ln 2}\left(6\cdot\frac{1}{36}\ln 36 + 15\cdot\frac{1}{18}\ln 18\right) = \frac{1}{\ln 2}\left(\frac{1}{6}\ln 36 + \frac{5}{6}\ln 18\right) = 4.337 \tag{3.11.8}$$

□

• c. *3 points Note that the difference between these two entropies is* $5/6 = 0.833$*. How can this be explained?*

Answer. This is the composition law (**??**) in action. Assume you roll two dice which you first consider indistinguishable and afterwards someone tells you which is which. How much information do you gain? Well, if the numbers are the same, then telling you which die is which does not give you any information, since the outcomes of the experiment are defined as: which number has the first die, which number has the second die, regardless of where on the table the dice land. But if the numbers are different, then telling you which is which allows you to discriminate between two outcomes both of which have conditional probability $1/2$ given the outcome you already know; in this case the information you gain is therefore 1 bit. Since the probability of getting two different numbers is $5/6$, the expected value of the information gained explains the difference in entropy. □

All these definitions use the convention $0\log\frac{1}{0} = 0$, which can be justified by the following continuity argument: Define the function, graphed in Figure 3:

$$\eta(w) = \begin{cases} w\log\frac{1}{w} & \text{if } w > 0 \\ 0 & \text{if } w = 0. \end{cases} \tag{3.11.9}$$

η is continuous for all $w \geq 0$, even at the boundary point $w = 0$. Differentiation gives $\eta'(w) = -(1+\log w)$, and $\eta''(w) = -w^{-1}$. The function starts out at the origin with a vertical tangent, and since the second derivative is negative, it is strictly concave for all $w > 0$. The definition of strict concavity is $\eta(w) < \eta(v) + (w-v)\eta'(v)$ for $w \neq v$, i.e., the function lies below all its tangents. Substituting $\eta'(v) = -(1+\log v)$ and simplifying gives $w - w\log w \leq v - w\log v$ for $v, w > 0$. One verifies that this inequality also holds for $v, w \geq 0$.

Problem 72. *Make a complete proof, discussing all possible cases, that for* $v, w \geq 0$ *follows*

$$w - w\log w \leq v - w\log v \tag{3.11.10}$$

Answer. We already know it for $v, w > 0$. Now if $v = 0$ and $w = 0$ then the equation reads $0 \le 0$; if $v > 0$ and $w = 0$ the equation reads $0 \le v$, and if $w > 0$ and $v = 0$ then the equation reads $w - w \log w \le +\infty$. □

3.11.3. How to Keep Forecasters Honest. This mathematical result allows an interesting alternative mathematical characterization of entropy. Assume Anita performs a Bernoulli experiment whose success probability she does not know but wants to know. Clarence knows this probability but is not on very good terms with Anita; therefore Anita is unsure that he will tell the truth if she asks him.

Anita knows "how to keep forecasters honest." She proposes the following deal to Clarence: "you tell me the probability q, and after performing my experiment I pay you the amount $\log_2(q)$ if the experiment is a success, and $\log_2(1-q)$ if it is a failure. If Clarence agrees to this deal, then telling Anita that value q which is the true success probability of the Bernoulli experiment maximizes the expected value of his payoff. And the maximum expected value of this payoff is exactly the negative of the entropy of the experiment.

Proof: Assume the correct value of the probability is p, and the number Clarence tells Tina is q. For every p, q between 0 and 1 we have to show:

$$p \log p + (1-p)\log(1-p) \ge p \log q + (1-p)\log(1-q). \tag{3.11.11}$$

For this, plug $w = p$ and $v = q$ as well as $w = 1-p$ and $v = 1-q$ into equation (3.11.10) and add.

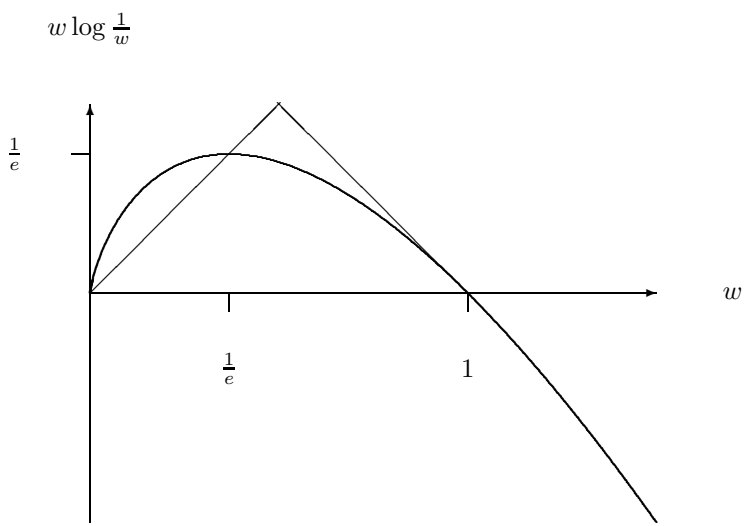


Figure 3. $\eta : w \mapsto w \log \frac{1}{w}$ is continuous at 0, and concave everywhere

3.11.4. The Inverse Problem. Now let us go over to the inverse problem: computing those probability fields which have maximum entropy subject to the information you have.

If you know that the experiment has n different outcomes, and you do not know the probabilities of these outcomes, then the maximum entropy approach amounts to assigning equal probability $1/n$ to each outcome.

Problem 73. *(Not eligible for in-class exams) You are playing a slot machine. Feeding one dollar to this machine leads to one of four different outcomes:* E_1*: machine returns nothing, i.e., you lose \$1.* E_2*: machine returns \$1, i.e., you lose nothing and win nothing.* E_3*: machine returns \$2, i.e., you win \$1.* E_4*: machine returns \$10, i.e., you win \$9. Events* E_i *occurs with probability* p_i*, but these probabilities are unknown. But due to a new "Truth-in-Gambling Act" you find a sticker on the side of the machine which says that in the long run the machine pays out only \$0.90 for every dollar put in. Show that those values of* p_1*,* p_2*,* p_3*, and* p_4 *which maximize the entropy (and therefore make the machine most interesting) subject to the constraint that the expected payoff per dollar put in is \$0.90, are* $p_1 = 0.4473$*,* $p_2 = 0.3158$*,* $p_3 = 0.2231$*,* $p_4 = 0.0138$*.*

Answer. Solution is derived in [**Rie85**, pp. 68/9 and 74/5], and he refers to [**Rie77**]. You have to maximize $-\sum p_n \log p_n$ subject to $\sum p_n = 1$ and $\sum c_n p_n = d$. In our case $c_1 = 0$, $c_2 = 1$, $c_3 = 2$, and $c_4 = 10$, and $d = 0.9$, but the treatment below goes through for arbitrary c_i as long as not all of them are equal. This case is discussed in detail in the answer to Problem 74. □

• a. *Difficult: Does the maximum entropy approach also give us some guidelines how to select these probabilities if all we know is that the expected value of the payout rate is smaller than 1?*

Answer. As shown in [**Rie85**, pp. 68/9 and 74/5], one can give the minimum value of the entropy for all distributions with payoff smaller than 1: $\mathrm{H} < 1.6590$, and one can also give some bounds for the probabilities: $p_1 > 0.4272$, $p_2 < 0.3167$, $p_3 < 0.2347$, $p_4 < 0.0214$. □

• b. *What if you also know that the entropy of this experiment is* 1.5*?*

Answer. This was the purpose of the paper [**Rie85**]. □

Problem 74. *(Not eligible for in-class exams) Let* $p_1, p_2, \ldots, p_n$ *(*$\sum p_i = 1$*) be the proportions of the population of a city living in* n *residential colonies. The cost of living in colony* i*, which includes cost of travel from the colony to the central business district, the cost of the time this travel consumes, the rent or mortgage payments, and other costs associated with living in colony* i*, is represented by the monetary amount* c_i*. Without loss of generality we will assume that the* c_i *are numbered in such a way that* $c_1 \le c_2 \le \cdots \le c_n$*. We will also assume that the* c_i *are not all equal. We assume that the* c_i *are known and that also the average expenditures on travel etc. in the population is known; its value is* d*. One approach to modelling the*

population distribution is to maximize the entropy subject to the average expenditures, i.e., to choose $p_1, p_2, \dots p_n$ such that $\mathrm{H} = \sum p_i \log \frac{1}{p_i}$ is maximized subject to the two constraints $\sum p_i = 1$ and $\sum p_i c_i = d$. This would give the greatest uncertainty about where someone lives.

• a. *3 points Set up the Lagrange function and show that*

$$p_i = \frac{\exp(-\lambda c_i)}{\sum \exp(-\lambda c_i)} \tag{3.11.12}$$

where the Lagrange multiplier λ must be chosen such that $\sum p_i c_i = d$.

Answer. The Lagrange function is

$$L = -\sum p_n \log p_n - \kappa(\sum p_n - 1) - \lambda(\sum c_n p_n - d) \tag{3.11.13}$$

Partial differentiation with respect to p_i gives the first order conditions

$$-\log p_i - 1 - \kappa - \lambda c_i = 0. \tag{3.11.14}$$

Therefore $p_i = \exp(-\kappa - 1)\exp(-\lambda c_i)$. Plugging this into the first constraint gives $1 = \sum p_i = \exp(-\kappa - 1)\sum \exp(-\lambda c_i)$ or $\exp(-\kappa - 1) = \frac{1}{\sum \exp(-\lambda c_i)}$. This constraint therefore defines κ uniquely, and we can eliminate κ from the formula for p_i:

$$p_i = \frac{\exp(-\lambda c_i)}{\sum \exp(-\lambda c_i)} \tag{3.11.15}$$

Now all the p_i depend on the same unknown λ, and this λ must be chosen such that the second constraint holds. This is the Maxwell-Boltzmann distribution if $\mu = kT$ where k is the Boltzmann constant and T the temperature. □

• b. *2 points Here is a mathematical lemma needed for the next part: Prove that for $a_i \geq 0$ and c_i arbitrary follows $\sum a_i \sum a_i c_i^2 \geq (\sum a_i c_i)^2$, and if all $a_i > 0$ and not all c_i equal, then this inequality is strict.*

Answer. By choosing the same subscripts in the second sum as in the first we pair elements of the first sum with elements of the second sum:

$$\sum_i a_i \sum_j c_j^2 a_j - \sum_i c_i a_i \sum_j c_j a_j = \sum_{i,j} (c_j^2 - c_i c_j) a_i a_j \tag{3.11.16}$$

but if we interchange i and j on the rhs we get

$$= \sum_{j,i} (c_i^2 - c_j c_i) a_j a_i = \sum_{i,j} (c_i^2 - c_i c_j) a_i a_j \tag{3.11.17}$$

Now add the righthand sides to get

$$2\Big(\sum_i a_i \sum_j c_j^2 a_j - \sum_i c_i a_i \sum_j c_j a_j\Big) = \sum_{i,j} (c_i^2 + c_j^2 - 2c_i c_j) a_i a_j = \sum_{i,j} (c_i - c_j)^2 a_i a_j \geq 0 \tag{3.11.18}$$

□

• c. *3 points It is not possible to solve equations (3.11.12) analytically for λ, but the following can be shown* [**Kap89**, p. 310/11]*: the function f defined by*

$$f(\lambda) = \frac{\sum c_i \exp(-\lambda c_i)}{\sum \exp(-\lambda c_i)} \tag{3.11.19}$$

is a strictly decreasing function which decreases from c_n to c_1 as λ goes from $-\infty$ to ∞, and $f(0) = \bar{c}$ where $\bar{c} = (1/n)\sum c_i$. We need that λ for which $f(\lambda) = d$, and this equation has no real root if $d < c_1$ or $d > c_n$, it has a unique positive root if $c_1 < d < \bar{c}$ it has the unique root 0 for $d = \bar{c}$, and it has a unique negative root for $\bar{c} < d < c_n$. From this follows: as long as d lies between the lowest and highest cost, and as long as the cost numbers are not all equal, the p_i are uniquely determined by the above entropy maximization problem.

Answer. Here is the derivative; it is negative because of the mathematical lemma just shown:

$$f'(\lambda) = \frac{u'v - uv'}{v^2} = -\frac{\sum \exp(-\lambda c_i) \sum c_i^2 \exp(-\lambda c_i) - \big(\sum c_i \exp(-\lambda c_i)\big)^2}{\big(\sum \exp(-\lambda c_i)\big)^2} < 0 \tag{3.11.20}$$

Since $c_1 \leq c_2 \leq \cdots \leq c_n$, it follows

$$c_1 = \frac{\sum c_1 \exp(-\lambda c_i)}{\sum \exp(-\lambda c_i)} \leq \frac{\sum c_i \exp(-\lambda c_i)}{\sum \exp(-\lambda c_i)} \leq \frac{\sum c_n \exp(-\lambda c_i)}{\sum \exp(-\lambda c_i)} = c_n \tag{3.11.21}$$

Now the statement about the limit can be shown if not all c_j are equal, say $c_1 < c_{k+1}$ but $c_1 = c_k$. The fraction can be written as

$$\frac{kc_1 \exp(-\lambda c_1) + \sum_{i=1}^{n-k} c_{k+i} \exp(-\lambda c_{k+i})}{k \exp(-\lambda c_1) + \sum i = 1^{n-k} \exp(-\lambda c_{k+i})} = \frac{kc_1 + \sum_{i=1}^{n-k} c_{k+i} \exp(-\lambda(c_{k+i} - c_1))}{k + \sum_{i=1}^{n-k} \exp(-\lambda(c_{k+i} - c_1))} \tag{3.11.22}$$

Since $c_{k+i} - c_1 > 0$, this converges towards c_1 for $\lambda \to \infty$. □

• d. *3 points Show that the maximum attained entropy is $\mathrm{H} = \lambda d + k(\lambda)$ where*

$$k(\lambda) = \log\big(\sum \exp(-\lambda c_j)\big). \tag{3.11.23}$$

Although λ depends on d, show that $\frac{\partial \mathrm{H}}{\partial d} = \lambda$, i.e., it is the same as if λ did not depend on d. This is an example of the "envelope theorem," and it also gives an interpretation of λ.

Answer. We have to plug the optimal $p_i = \frac{\exp(-\lambda c_i)}{\sum \exp(-\lambda c_i)}$ into the formula for $\mathrm{H} = -\sum p_i \log p_i$. For this note that $-\log p_i = \lambda c_i + k(\lambda)$ where $k(\lambda) = \log(\sum \exp(-\lambda c_j))$ does not depend on i. Therefore $\mathrm{H} = \sum p_i(\lambda c_i + k(\lambda)) = \lambda \sum p_i c_i + k(\lambda) \sum p_i = \lambda d + k(\lambda)$, and $\frac{\partial \mathrm{H}}{\partial d} = \lambda + d\frac{\partial \lambda}{\partial d} + k'(\lambda)\frac{\partial \lambda}{\partial d}$. Now we need the derivative of $k(\lambda)$, and we discover that $k'(\lambda) = -f(\lambda)$ where $f(\lambda)$ was defined in (3.11.19). Therefore $\frac{\partial \mathrm{H}}{\partial d} = \lambda + (d - f(\lambda))\frac{\partial \lambda}{\partial d} = \lambda$. □

• e. *5 points Now assume d is not known (but the c_i are still known), i.e., we know that (3.11.12) holds for some λ but we don't know which. We want to estimate*

this λ (and therefore all p_i) by taking a random sample of m people from that metropolitan area and asking them what their regional living expenditures are and where they live. Assume x_i people in this sample live in colony i. One way to estimate this λ would be to use the average consumption expenditure of the sample, $\sum \frac{\mathsf{x}_i}{m} c_i$, as an estimate of the missing d in the above procedure, i.e., choose that λ which satisfies $f(\lambda) = \sum \frac{\mathsf{x}_i}{m} c_i$. Another procedure, which seems to make a better use of the information given by the sample, would be to compute the maximum likelihood estimator of λ based on all x_i. Show that these two estimation procedures are identical.

Answer. The x_i have the multinomial distribution. Therefore, given that the proportion p_i of the population lives in colony i, and you are talking a random sample of size m from the whole population, then the probability to get the outcome $x_1, \dots, x_n$ is

$$L = \frac{m!}{x_1! \cdots x_n!} p_1^{x_1} p_2^{x_2} \cdots p_n^{x_n} \tag{3.11.24}$$

This is what we have to maximize, subject to the condition that the p_i are an entropy maximizing population distribution. Let's take logs for computational simplicity:

$$\log L = \log m! - \sum_j \log x_j! + \sum x_i \log p_i \tag{3.11.25}$$

All we know about the p_i is that they must be some entropy maximizing probabilities, but we don't know yet which ones, i.e., they depend on the unknown λ. Therefore we need the formula again $-\log p_i = \lambda c_i + k(\lambda)$ where $k(\lambda) = \log(\sum \exp(-\lambda c_j))$ does not depend on i. This gives

$$\log L = \log m! - \sum_j \log x_j! - \sum x_i(\lambda c_i + k(\lambda)) = \log m! - \sum_j \log x_j! - \lambda \sum x_i c_i + k(\lambda) m \tag{3.11.26}$$

(for this last term remember that $\sum x_i = m$. Therefore the derivative is

$$\frac{1}{m} \frac{\partial}{\partial \lambda} \log L = \sum \frac{x_i}{m} c_i - f(\lambda) \tag{3.11.27}$$

I.e., using the obvious estimate for d is the same as maximum likelihood under the assumption of maximum entropy. □

This is a powerful estimation strategy. An article with sensational image reconstitutions using maximum entropy algorithms is [**SG85**, pp. 111, 112, 115, 116]. And [**GJM96**] applies maximum entropy methods to ill-posed or underdetermined problems in econometrics!

CHAPTER 4

Random Number Generation and Encryption

How can a computer, which is a fully determinate system, be programmed to generate random numbers?

The most frequently used method to generate pseudo-random numbers uniformly distributed between 0 and 1 is the "linear congruential" method. The algorithm is parametrized with four integers, as follows:

μ	the modulus	$\mu > 0$
α	the multiplier	$0 \le \alpha < \mu$
γ	the increment	$0 \le \gamma < \mu$
x_0	the starting value, or seed	$0 \le x_0 < \mu$

If x_n is the current value of the "random seed" then a call to the random number generator first computes

$$x_{n+1} = (\alpha x_n + \gamma) \bmod \mu \tag{4.0.28}$$

as the seed for the next call, and then returns x_{n+1}/μ as independent observation of a pseudo random number which is uniformly distributed in $(0, 1)$.

$a \bmod b$ is the remainder in the integer division of a by b. For instance $13 \bmod 10 = 3$, $16 \bmod 8 = 0$, etc.

The selection of α, γ, and μ is critical here. We need the following criteria:

- The random generator should have a full period, i.e., it should produce all numbers $0 < x < \mu$ before repeating. (Once one number is repeated, the whole cycle is repeated).
- The function should "appear random."
- The function should implement efficiently with 32-bit arithmetic.

If μ is prime and $\gamma = 0$, then for certain values of α the period is $\mu - 1$, with only the value 0 missing. For 32-bit arithmetic, a convenient value of μ is $2^{31} - 1$, which is a prime number. Of the more than 2 billion possible choices for α, only and handful pass all 3 tests.

Problem 75. *Convince yourself by some examples that for all a, b, and μ follows:*

$$a \cdot b \bmod \mu = \big(a \cdot (b \bmod \mu)\big) \bmod \mu \tag{4.0.29}$$

In view of of Question 75, the multiplicative congruential random generator is based on the following procedure: generate two sequences of integers a_i and b_i as follows: $a_i = x_1 \cdot \alpha^i$ and $b_i = i \cdot \mu$. a_i is multiplicative and b_i additive, and μ is a prime number which is not a factor of α. In other words, a_i and b_i have very little to do with each other. Then for each i find $a_i - b_{s(i)}$ where $b_{s(i)}$ is the largest b which is smaller than or equal to a_i, and then form form $\frac{a_i - b_{s(i)}}{\mu}$ to get a number between 0 and 1. This is a measure of relationship between two processes which have very little to do with each other, and therefore we should not be surprised if this interaction turns out to look "random." Knuth writes [**Knu81**, p. 10]: "taking the remainder $\bmod \mu$ is somewhat like determining where a ball will land in a spinning roulette wheel." Of course, this is a heuristic argument. There is a lot of mathematical theory behind the fact that linear congruential random number generators are good generators.

If $\gamma = 0$ then the period is shorter: then the maximum period is $\mu - 1$ because any sequence which contains 0 has 0 everywhere. But not having to add γ at every step makes computation easier.

Not all pairs α and μ give good random number generators, and one should only use random number generators which have been thoroughly tested. There are some examples of bad random number generators used in certain hardware or software programs.

Problem 76. *The dataset located at www.econ.utah.edu/ehrbar/data/randu.txt (which is available as dataset* `randu` *in the* R*-base distribution) has 3 columns and 400 rows. Each row is a consecutive triple of numbers generated by the old VAX FORTRAN function RANDU running under VMS 1.5. This random generator, which is discussed in* [**Knu98**, pp. 106/7]*, starts with an odd seed x_0, the $n+1$st seed is $x_{n+1} = (65539 x_n) \bmod 2^{31}$, and the data displayed are $x_n/2^{31}$ rounded to 6 digits. Load the data into* `xgobi` *and use the* `Rotation` *view to check whether you can see something suspicious.*

Answer. All data are concentrated in 15 parallel planes. All triplets of observations of `randu` fall into these planes; [**Knu98**, pp. ??] has a mathematical proof. VMS versions 2.0 and higher use a different random generator.

□

4.1. Alternatives to the Linear Congruential Random Generator

One of the common fallacies encountered in connection with random number generation is the idea that we can take a good generator and modify it a little in order to get an "even more random" sequence. This is often false.

Making the value dependent on the two preceding values increases the maximum possible period to μ^2. The simplest such generator, the Fibonacci sequence

$$x_{n+1} = (x_n + x_{n-1}) \bmod \mu \tag{4.1.1}$$

is definitely not satisfactorily random. But specific other combinations are good:

$$x_{n+1} = (x_{n-100} - x_{n-37}) \bmod 2^{30} \tag{4.1.2}$$

is one of the state of the art random generators used in R.

Using more work to get from one number to the next, not mere addition or multiplication:

$$x_{n+1} = (\alpha x_n^{-1} + \gamma) \bmod \mu \tag{4.1.3}$$

Efficient algorithms exist but are not in the repertoire of most computers. This generator is completely free of the lattice structure of multiplicative congruential generators.

Combine several random generators: If you have two random generators with modulus m, use

$$x_m - y_m \bmod \mu \tag{4.1.4}$$

The Wichmann-Hill portable random generator uses this trick.

Randomizing by shuffling: If you have x_n and y_n, put the first k observation of x_n into a buffer, call them $v_1, \ldots, v_k$ ($k = 100$ or so). Then construct x_{n+1} and y_{n+1}. Use y_{n+1} to generate a random integer j between 1 and k, use v_j as your next random observation, and put x_{n+1} in the buffer at place j. This still gives the same values as x_n but in a different order.

4.2. How to test random generators

Chi-Square Test: partition the outcomes into finitely many categories and test whether the relative frequencies are compatible with the probabilities.

Kolmogorov-Smirnoff test for continuous distributions: uses the maximum distance between the empirical distribution function and the theoretical distribution function.

Now there are 11 kinds of empirical tests, either on the original x_n which are supposedly uniform between 0 and 1, or on integer-valued y_n between 0 and d-1.

Equidistribution: either a Chi-Square test that the outcomes fall into d intervals, or a Kolmogoroff-Smirnov test.

Serial test: that all integer pairs in the integer-valued outcome are equally likely.

Gap test: for $0 \le \alpha < \beta \le 1$ a gap of length r is a sequence of $r+1$ consecutive numbers in which the last one is in the interval, and the others are not. Count the occurrence of such gaps, and make a Chi Squared test with the probabilities of such occurrences. For instance, if $\alpha = 0$ and $\beta = 1/2$ this computes the lengths of "runs above the mean."

Poker test: consider groups of 5 successive integers and classify them into the 7 categories: all different, one pair, two pairs, three of a kind, full house, four of a kind, five of a kind.

Coupon collectors test: observe the length of sequences required to get a full set of integers $0, \ldots, d-1$.

Permutation test: divide the input sequence of the continuous random variable into t-element groups and look at all possible relative orderings of these k-tuples. There are $t!$ different relative orderings, and each ordering has probability $1/t!$.

Run test: counts runs up, but don't use Chi Square test since subsquent runs are not independent; a long run up is likely to be followed by a short run up.

Maximum-of-t-Test: split the sample into batches of equal length and take the maximum of each batch. Taking these maxima to the tth power should again give an equidistributed sample.

Collision tests: 20 consecutive observations are all smaller than 1/2 with probability 2^{-20}; and every other partition defined by combinations of bigger or smaller than 1/2 has the same probability. If there are only 2^{14} observations, then on the average each of these partitions is populated only with probability 1/64. We count the number of "collisions", i.e., the number of partitions which have more than 1 observation in them, and compare this with the binomial distribution (the Chi Square cannot be applied here).

Birthday spacings test: lagged Fibonacci generators consistently fail it.

Serial correlation test: a statistic which looks like a sample correlation coefficient which can be easily computed with the Fast Fourier transformation.

Tests on subsequences: equally spaced subsequences are usually worse than the original sequence if it is a linear congruential generator.

4.3. The Wichmann Hill generator

The Wichmann Hill generator defined in [**WH82**] can be implemented in almost any high-level language. It used to be the default random number generator in R, but version 1.0 of R has different defaults.

Since even the largest allowable integers in ordinary programming languages are not large enough to make a good congruential random number generator, the

Wichmann Hill generator is the addition mod 1 of 3 different multiplicative congruential generators which can be computed using a high-level programming language. [**Zei86**] points out that due to the Chinese Remainder Theorem, see [**Knu81**, p. 286], this is equivalent to one single multiplicative congruential generator with $\alpha = 1655\,54252\,64690$ and $\mu = 2781\,71856\,04309$. Since such long integers cannot be used in ordinary computer programs, Wichmann-Hill's algorithm is an efficient method to compute a congruential generator with such large numbers.

PROBLEM 77. *Here is a more detailed description of the Wichmann-Hill generator: Its seed is a 3-vector* $\begin{bmatrix} x_1 & y_1 & z_1 \end{bmatrix}^\top$ *satisfying*

$$0 < x_1 \leq 30269 \tag{4.3.1}$$

$$0 < y_1 \leq 30307 \tag{4.3.2}$$

$$0 < z_1 \leq 30323 \tag{4.3.3}$$

A call to the random generator updates the seed as follows:

$$x_2 = 171 x_1 \bmod 30269 \tag{4.3.4}$$

$$y_2 = 172 y_1 \bmod 30307 \tag{4.3.5}$$

$$z_2 = 170 z_1 \bmod 30323 \tag{4.3.6}$$

and then it returns

$$\left(\frac{x_2}{30269} + \frac{y_2}{30307} + \frac{z_2}{30323}\right) \bmod 1 \tag{4.3.7}$$

as its latest drawing from a uniform distribution. If you have R *on your computer, do parts b and c, otherwise do a and b.*

• a. *4 points Program the Wichmann-Hill random generator in the programming language of your choice.*

ANSWER. A random generator does two things:

- It takes the current seed (or generates one if there is none), computes the next seed from it, and stores this next seed on disk as a side effect.
- Then it converts this next seed into a number between 0 and 1.

The `ecmet` package has two demonstration functions which perform these two tasks separately for the Wichmann-Hill generator, without side effects. The function `next.WHseed()` computes the next seed from its argument (which defaults to the seed stored in the official variable `.Random.seed`), and the function `WH.from.current.seed()` gets a number between 0 and 1 from its argument (which has the same default). Both functions are one-liners:

```
next.WHseed <- function(integer.seed = .Random.seed[-1])
  (c(  171,   172,   170) * integer.seed) %% c(30269, 30307, 30323)

WH.from.current.seed <- function(integer.seed = .Random.seed[-1])
  sum(integer.seed / c(30269, 30307, 30323)) %% 1
```


□

• b. *2 points Check that the 3 first numbers returned by the Wichmann-Hill random number generator after setting the seed to 1 10 2000 are 0.2759128 0.8713303 0.6150737. (one digit in those 3 numbers is wrong; which is it, and what is the right digit?)*

ANSWER. The R-code doing this is `ecmet.script(wichhill)`:

```
##This script generates 3 consecutive seeds, with the
##initial seed set as (1, 10, 2000), puts them into a matrix,
##and then generates the random numbers from the rows of
##this matrix:

  my.seeds <- matrix(nrow=3, ncol=3)

  my.seeds[1,] <- next.WHseed(c(1, 10, 2000))
  my.seeds[2,] <- next.WHseed(my.seeds[1,])
  my.seeds[3,] <- next.WHseed(my.seeds[2,])

  my.unif <- c(WH.from.current.seed(my.seeds[1,]),
               WH.from.current.seed(my.seeds[2,]),
               WH.from.current.seed(my.seeds[3,]))
```

□

• c. *4 points Check that the Wichmann-Hill random generator built into* R *is identical to the one described here.*

ANSWER. First make sure that R will actually use the Wichmann-Hill generator (since it is not the default): `RNGkind("Wichmann-Hill")`. Then call `runif(1)`. (This sets a seed if there was none, or uses the existing seed if there was one.) `.Random.seed[-1]` shows present value of the random seed associated with this last call, dropping 1st number which indicates which random generator this is for, which is not needed for our purposes. Therefore `WH.from.current.seed()`, which takes `.Random.seed[-1]` as default argument, should give the same result as the last call of the official random generator. And `WH.from.current.seed(next.WHseed())` takes the current seed, computes the next seed from it, and converts this next seed into a number between 0 and 1. It does not write the updated random seed back. Therefore if we issue now the official call `runif(1)` again, we should get the same result. □

4.4. Public Key Cryptology

The development of public key encryption is the greatest revolution so far in the history of cryptography. In ordinary encryption, the same key is used to encrypt and decrypt the message. In public-key encryption, there is a pair of keys, a public key and a private key. If the message is encrypted with the public key, then it must be decrypted with the private key, and vice versa. But knowledge of the public key will not allow you to determine the private key belonging to it.

This solves one of the most vexing problems for encryption, namely, the exchange of keys. In order to communicate privately with each other, the partners no longer have to exchange a secret key first. A broadcasts his public key to the world, and only has to safeguard his private key. Everyone who wants to send a secret message to A simply uses A's public key.

This same scheme can also be used for signing documents (i.e. ensuring that the author of a given document is the person who pretends to be the author): if A signs his document with his private key (i.e., attaches a checksum of the document encrypted with his private key), the recipient has to do two things: decrypt the checksum with A's public key, and compute the checksum of the document himself. If these two checksums agree, then B knows that the document indeed comes from A and that it has not been altered in transit.

The original computer program doing this kind of encryption was written by the programmer Phil Zimmerman. The program is called PGP, "Pretty Good Privacy," and manuals are [**Zim95**] and [**Sta95**]. More recently, a free version of this program has been written, called GNU Privacy Guard, textttwww.gnupg.org, which does not use the patented IDEA algorithm and is under the Gnu Public License.

Here is the mathematics of it. I am following [**Sch97**, p. 120, 130], a relevant and readable book. First some number-theoretic preliminaries.

Fermat's theorem: for a prime p and an integer b not divisible by p, $b^{p-1} \bmod p = 1$.

Euler's ϕ function or Euler's "totient" is the number of positive integers r smaller than m that are coprime to m, i.e., have no common divisors with m. Example: for $m = 10$, the coprime numbers are $r = 1, 3, 7, 9$, therefore $\phi(m) = 4$.

If m is prime, then $\phi(m) = m - 1$.

If m and n are coprime, then $\phi(mn) = \phi(m)\phi(n)$.

If p and q are to different prime numbers, then $\phi(pq) = (p-1)(q-1)$.

Euler's theorem extends Fermat's theorem: if b is coprime with e, then $b^{\phi(e)} \bmod e = 1$.

Application to digital encryption (RSA algorithm): r is a large "modulus" (indicating the largest message size which is encrypted in one step) and the plaintext message is a number M with $1 < M < r$ which must be coprime with r. (A real life text is first converted into a sequence of positive integers $M < r$ which are then encrypted individually. Indeed, since the RSA algorithm is rather slow, the message is encrypted with a temporary ordinary secret key, and only this key is encrypted with the RSA algorithm and attached to the conventionally encrypted message.) By applying Euler's theorem twice one can show that pairs of integers s and t exist such that encryption consists in raising to the sth power modulo r, and decryption in raising to the tth power modulo r. I.e., in order to encrypt one computes $E = M^s \bmod r$, and one can get M back by computing $M = E^t \bmod r$.

If s is any number coprime with $\phi(r)$, then $t = s^{\phi(\phi(r))-1} \bmod \phi(r)$ is the decryption key belonging to s. To prove this, we will first show that $E^t \bmod r = M^{st} \bmod r = M$. Now $st = s^{\phi(\phi(r))}$, and since s is coprime with $\phi(r)$, we can apply Euler's theorem to get $st \bmod \phi(r) = 1$, i.e., a k exists with $st = 1 + k\phi(r)$. Therefore $E^t \bmod r = M^{st} \bmod r = (MM^{k\phi(r)}) \bmod r = \big(M(M^{\phi(r)} \bmod r)^k\big) \bmod r$. A second application of Euler's theorem says that $M^{\phi(r)} \bmod r = 1$, therefore $M^{st} \bmod r = M \bmod r = M$. Finally, since $M^{\phi(r)} \bmod r = 1$, we get $M^{st} \bmod r = M^{st \bmod \phi(r)} \bmod r$.

If r is a prime and s is coprime with $r - 1$, then someone who has enough information to do the encryption, i.e., who knows s and r, can also easily compute t: $t = s^{\phi(r-1)-1}$.

But if r is the product of two different big primes, call them p and q, then someone who knows p and q can compute pairs s, t fairly easily, but it is computationally very expensive to get t from the knowledge of s and r alone, because no algorithm is known which easily determines the prime factors of huge integers.

Problem 78. [**Sta99**, p. 200] *As an example showing what is involved in the RSA algorithm, first generate the private and public keys as follows:*

• a. *2 points Select two primes, $p = 3$ and $q = 11$. The modulus of our encryption algorithm is their product $r = pq = 33$. Enumerate all numbers < 33 which are coprime to 33. You should come up with $\phi(r) = (3-1)(11-1) = 20$ numbers.*

Answer. 1, 2, 4, 5, 7, 8, 10, 13, 14, 16, 17, 19, 20, 23, 25, 26, 28, 29, 31, 32. □

• b. *2 points Now we have to select s such that s is relatively prime to $\phi(r) = 20$ and less than $\phi(r)$; a possible choice which we will use here is $s = 7$. To get a t such that $st \bmod 20 = 1$ we have to compute $t = s^{\phi(\phi(r))-1} \bmod \phi(r) = s^{\phi(20)-1} \bmod \phi(r)$. First compute $\phi(20)$ and then t.*

Answer. The numbers coprime with 20 are 1, 3, 7, 9, 11, 13, 17, 19. Therefore $\phi(20) = 8$. Therefore $t = 7^7 \bmod 20 = 823543 \bmod 20 = 3$. One easily verifies that $t = 3$ is correct because $st = 7 \cdot 3 = 20 + 1$. □

• c. *2 points Therefore the public key is $\{7, 33\}$ and the private key $\{t, 33\}$ with the t just computed. Now take a plaintext consisting of the number 5, use the public key to encrypt it. What is the encrypted text? Use the private key to decrypt it again.*

Answer. If the plaintext $= 5$, then encryption is the computation of $5^7 \bmod 33 = 78125 \bmod 33 = 14$. Decryption is the computation of $14^3 \bmod 33 = 2744 \bmod 33 = 5$. □

• d. *1 point This procedure is only valid if the plaintext is coprime with t. What should be done about this?*

Answer. Nothing. t is huge, and if it is selected in such a way that it does not have many different prime multipliers, the chance that a text happens to be not coprime with it is minuscule. □

• e. *2 points Now take the same plaintext and use the private key to encrypt it. What is the encrypted text? Then use the public key to decrypt it.*

ANSWER. If the plaintext $= 5$, then encryption is the computation of $5^3 \bmod 33 = 125 \bmod 33 = 26$. Decryption is the computation of $26^7 \bmod 33 = 8031810176 \bmod 33 = 5$. □

CHAPTER 5

Specific Random Variables

5.1. Binomial

We will begin with mean and variance of the binomial variable, i.e., the number of successes in n independent repetitions of a Bernoulli trial (3.7.1). The binomial variable has the two parameters n and p. Let us look first at the case $n = 1$, in which the binomial variable is also called *indicator variable:* If the event A has probability p, then its complement A' has the probability $q = 1 - p$. The indicator variable of A, which assumes the value 1 if A occurs, and 0 if it doesn't, has expected value p and variance pq. For the binomial variable with n observations, which is the sum of n independent indicator variables, the expected value (mean) is np and the variance is npq.

PROBLEM 79. *The random variable* x *assumes the value* a *with probability* p *and the value* b *with probability* $q = 1 - p$. *Show that* $\operatorname{var}[\mathsf{x}] = pq(a - b)^2$.

ANSWER. $\mathrm{E}[\mathsf{x}] = pa + qb$; $\operatorname{var}[\mathsf{x}] = \mathrm{E}[\mathsf{x}^2] - (\mathrm{E}[\mathsf{x}])^2 = pa^2 + qb^2 - (pa + qb)^2 = (p - p^2)a^2 - 2pqab + (q - q^2)b^2 = pq(a - b)^2$. For this last equality we need $p - p^2 = p(1 - p) = pq$. □

The *Negative Binomial Variable* is, like the binomial variable, derived from the Bernoulli experiment; but one reverses the question. Instead of asking how many successes one gets in a given number of trials, one asks, how many trials one must make to get a given number of successes, say, r successes.

First look at $r = 1$. Let t denote the number of the trial at which the first success occurs. Then

$$\Pr[\mathsf{t}{=}n] = pq^{n-1} \qquad (n = 1, 2, \ldots). \tag{5.1.1}$$

This is called the geometric probability.

Is the probability derived in this way σ-additive? The sum of a geometrically declining sequence is easily computed:

$$1 + q + q^2 + q^3 + \cdots = s \quad \text{Now multiply by } q: \tag{5.1.2}$$

$$q + q^2 + q^3 + \cdots = qs \quad \text{Now subtract and write } 1 - q = p: \tag{5.1.3}$$

$$1 = ps \tag{5.1.4}$$

Equation (5.1.4) means $1 = p + pq + pq^2 + \cdots$, i.e., the sum of all probabilities is indeed 1.

Now what is the expected value of a geometric variable? Use definition of expected value of a discrete variable: $\mathrm{E}[\mathsf{t}] = p\sum_{k=1}^{\infty} kq^{k-1}$. To evaluate the infinite sum, solve (5.1.4) for s:

$$s = \frac{1}{p} \quad \text{or} \quad 1 + q + q^2 + q^3 + q^4 \cdots = \sum_{k=0}^{\infty} q^k = \frac{1}{1-q} \tag{5.1.5}$$

and differentiate both sides with respect to q:

$$1 + 2q + 3q^2 + 4q^3 + \cdots = \sum_{k=1}^{\infty} kq^{k-1} = \frac{1}{(1-q)^2} = \frac{1}{p^2}. \tag{5.1.6}$$

The expected value of the geometric variable is therefore $\mathrm{E}[\mathsf{t}] = \frac{p}{p^2} = \frac{1}{p}$.

PROBLEM 80. *Assume* t *is a geometric random variable with parameter* p, *i.e., it has the values* $k = 1, 2, \ldots$ *with probabilities*

$$p_{\mathsf{t}}(k) = pq^{k-1}, \text{ where } q = 1 - p. \tag{5.1.7}$$

The geometric variable denotes the number of times one has to perform a Bernoulli experiment with success probability p *to get the first success.*

• a. *1 point Given a positive integer* n. *What is* $\Pr[\mathsf{t}{>}n]$? *(Easy with a simple trick!)*

ANSWER. $\mathsf{t}{>}n$ means, the first n trials must result in failures, i.e., $\Pr[\mathsf{t}{>}n] = q^n$. Since $\{\mathsf{t} > n\} = \{\mathsf{t} = n + 1\} \cup \{\mathsf{t} = n + 2\} \cup \cdots$, one can also get the same result in a more tedious way: It is $pq^n + pq^{n+1} + pq^{n+2} + \cdots = s$, say. Therefore $qs = pq^{n+1} + pq^{n+2} + \cdots$, and $(1 - q)s = pq^n$; since $p = 1 - q$, it follows $s = q^n$. □

• b. *2 points Let* m *and* n *be two positive integers with* $m < n$. *Show that* $\Pr[\mathsf{t}{=}n|\mathsf{t}{>}m] = \Pr[\mathsf{t}{=}n - m]$.

ANSWER. $\Pr[\mathsf{t}{=}n|\mathsf{t}{>}m] = \frac{\Pr[\mathsf{t}=n]}{\Pr[\mathsf{t}>m]} = \frac{pq^{n-1}}{q^m} = pq^{n-m-1} = \Pr[\mathsf{t}{=}n - m]$. □

• c. *1 point Why is this property called the memory-less property of the geometric random variable?*

ANSWER. If you have already waited for m periods without success, the probability that success will come in the nth period is the same as the probability that it comes in $n - m$ periods if you start now. Obvious if you remember that geometric random variable is time you have to wait until 1st success in Bernoulli trial.

□

PROBLEM 81. t *is a geometric random variable as in the preceding problem. In order to compute* $\mathrm{var}[\mathsf{t}]$ *it is most convenient to make a detour via* $\mathrm{E}[\mathsf{t}(\mathsf{t}-1)]$. *Here are the steps:*

- a. *Express* $\mathrm{E}[\mathsf{t}(\mathsf{t}-1)]$ *as an infinite sum.*

ANSWER. Just write it down according to the definition of expected values: $\sum_{k=0}^{\infty} k(k-1)pq^{k-1} = \sum_{k=2}^{\infty} k(k-1)pq^{k-1}$. □

- b. *Derive the formula*

$$\sum_{k=2}^{\infty} k(k-1)q^{k-2} = \frac{2}{(1-q)^3} \tag{5.1.8}$$

by the same trick by which we derived a similar formula in class. Note that the sum starts at $k = 2$.

ANSWER. This is just a second time differentiating the geometric series, i.e., first time differentiating (5.1.6). □

- c. *Use a. and b. to derive*

$$\mathrm{E}[\mathsf{t}(\mathsf{t}-1)] = \frac{2q}{p^2} \tag{5.1.9}$$

ANSWER.

$$\sum_{k=2}^{\infty} k(k-1)pq^{k-1} = pq\sum_{k=2}^{\infty} k(k-1)q^{k-2} = pq\frac{2}{(1-q)^3} = \frac{2q}{p^2}. \tag{5.1.10}$$

□

- d. *Use c. and the fact that* $\mathrm{E}[\mathsf{t}] = 1/p$ *to derive*

$$\mathrm{var}[\mathsf{t}] = \frac{q}{p^2}. \tag{5.1.11}$$

ANSWER.

$$\mathrm{var}[\mathsf{t}] = \mathrm{E}[\mathsf{t}^2] - (\mathrm{E}[\mathsf{t}])^2 = \mathrm{E}[\mathsf{t}(\mathsf{t}-1)] + \mathrm{E}[\mathsf{t}] - (\mathrm{E}[\mathsf{t}])^2 = \frac{2q}{p^2} + \frac{1}{p} - \frac{1}{p^2} = \frac{q}{p^2}. \tag{5.1.12}$$

□

Now let us look at the negative binomial with arbitrary r. What is the probability that it takes n trials to get r successes? (That means, with $n-1$ trials we did not yet have r successes.) The probability that the nth trial is a success is p. The probability that there are $r-1$ successes in the first $n-1$ trials is $\binom{n-1}{r-1}p^{r-1}q^{n-r}$. Multiply those to get:

$$\Pr[\mathsf{t}{=}n] = \binom{n-1}{r-1}p^r q^{n-r}. \tag{5.1.13}$$

This is the negative binomial, also called the Pascal probability distribution with parameters r and p.

One easily gets the mean and variance, because due to the memory-less property it is the sum of r independent geometric variables:

$$\mathrm{E}[\mathsf{t}] = \frac{r}{p} \qquad \mathrm{var}[\mathsf{t}] = \frac{rq}{p^2} \tag{5.1.14}$$

Some authors define the negative binomial as the number of failures before the rth success. Their formulas will look slightly different than ours.

PROBLEM 82. *3 points A fair coin is flipped until heads appear 10 times, and* x *is the number of times tails appear before the 10th appearance of heads. Show that the expected value* $\mathrm{E}[\mathsf{x}] = 10$.

ANSWER. Let t be the number of the throw which gives the 10th head. t is a negative binomial with $r = 10$ and $p = 1/2$, therefore $\mathrm{E}[\mathsf{t}] = 20$. Since $\mathsf{x} = \mathsf{t} - 10$, it follows $\mathrm{E}[\mathsf{x}] = 10$. □

PROBLEM 83. *(Banach's match-box problem) (Not eligible for in-class exams) There are two restaurants in town serving hamburgers. In the morning each of them obtains a shipment of* n *raw hamburgers. Every time someone in that town wants to eat a hamburger, he or she selects one of the two restaurants at random. What is the probability that the* $(n+k)$*th customer will have to be turned away because the restaurant selected has run out of hamburgers?*

ANSWER. For each restaurant it is the negative binomial probability distribution in disguise: if a restaurant runs out of hamburgers this is like having n successes in $n+k$ tries.

But one can also reason it out: Assume one of the restaurantes must turn customers away after the $n+k$th customer. Write down all the $n+k$ decisions made: write a 1 if the customer goes to the first restaurant, and a 2 if he goes to the second. I.e., write down $n+k$ ones and twos. Under what conditions will such a sequence result in the $n+k$th move eating the last hamburgerthe first restaurant? Exactly if it has n ones and k twos, a $n+k$th move is a one. As in the reasoning for the negative binomial probability distribution, there are $\binom{n+k-1}{n-1}$ possibilities, each of which has probability 2^{-n-k}. Emptying the second restaurant has the same probability. Together the probability is therefore $\binom{n+k-1}{n-1}2^{1-n-k}$. □

5.2. The Hypergeometric Probability Distribution

Until now we had independent events, such as, repeated throwing of coins or dice, sampling with replacement from finite populations, ar sampling from infinite populations. If we sample *without* replacement from a *finite* population, the probability of the second element of the sample depends on what the first element was. Here the hypergeometric probability distribution applies.

Assume we have an urn with w white and $n-w$ black balls in it, and we take a sample of m balls. What is the probability that y of them are white?

We are not interested in the order in which these balls are taken out; we may therefore assume that they are taken out simultaneously, therefore the set U of outcomes is the set of subsets containing m of the n balls. The total number of such subsets is $\binom{n}{m}$. How many of them have y white balls in them? Imagine you first pick y white balls from the set of all white balls (there are $\binom{w}{y}$ possibilities to do that), and then you pick $m-y$ black balls from the set of all black balls, which can be done in $\binom{n-w}{m-y}$ different ways. Every union of such a set of white balls with a set of black balls gives a set of m elements with exactly y white balls, as desired. There are therefore $\binom{w}{y}\binom{n-w}{m-y}$ different such sets, and the probability of picking such a set is

$$\Pr[\text{Sample of } m \text{ elements has exactly } y \text{ white balls}] = \frac{\binom{w}{y}\binom{n-w}{m-y}}{\binom{n}{m}}. \tag{5.2.1}$$

Problem 84. *You have an urn with w white and $n-w$ black balls in it, and you take a sample of m balls* with replacement*, i.e., after pulling each ball out you put it back in before you pull out the next ball. What is the probability that y of these balls are white? I.e., we are asking here for the counterpart of formula* (5.2.1) *if sampling is done with replacement.*

Answer.

$$\left(\frac{w}{n}\right)^y \left(\frac{n-w}{n}\right)^{m-y} \binom{m}{y} \tag{5.2.2}$$

□

Without proof we will state here that the expected value of y, the number of white balls in the sample, is $\mathrm{E}[\mathsf{y}] = m\frac{w}{n}$, which is the same as if one would select the balls with replacement.

Also without proof, the variance of y is

$$\mathrm{var}[\mathsf{y}] = m\frac{w}{n}\frac{(n-w)}{n}\frac{(n-m)}{(n-1)}. \tag{5.2.3}$$

This is smaller than the variance if one would choose with replacement, which is represented by the above formula without the last term $\frac{n-m}{n-1}$. This last term is called the finite population correction. More about all this is in [**Lar82**, p. 176–183].

5.3. The Poisson Distribution

The Poisson distribution counts the number of events in a given time interval. This number has the Poisson distribution if each event is the cumulative result of a large number of independent possibilities, each of which has only a small chance of occurring (law of rare events). The expected number of occurrences is proportional

to time with a proportionality factor λ, and in a short time span only zero or one event can occur, i.e., for infinitesimal time intervals it becomes a Bernoulli trial.

Approximate it by dividing the time from 0 to t into n intervals of length $\frac{t}{n}$; then the occurrences are approximately n independent Bernoulli trials with probability of success $\frac{\lambda t}{n}$. (This is an approximation since some of these intervals may have more than one occurrence; but if the intervals become very short the probability of having two occurrences in the same interval becomes negligible.)

In this discrete approximation, the probability to have k successes in time t is

$$\Pr[\mathsf{x}{=}k] = \binom{n}{k}\left(\frac{\lambda t}{n}\right)^k\left(1-\frac{\lambda t}{n}\right)^{(n-k)} \tag{5.3.1}$$

$$= \frac{1}{k!}\frac{n(n-1)\cdots(n-k+1)}{n^k}(\lambda t)^k\left(1-\frac{\lambda t}{n}\right)^n\left(1-\frac{\lambda t}{n}\right)^{-k} \tag{5.3.2}$$

$$\to \frac{(\lambda t)^k}{k!}e^{-\lambda t} \quad \text{for } n\to\infty \text{ while } k \text{ remains constant} \tag{5.3.3}$$

(5.3.3) is the limit because the second and the last term in (5.3.2) $\to 1$. The sum of all probabilities is 1 since $\sum_{k=0}^{\infty}\frac{(\lambda t)^k}{k!} = e^{\lambda t}$. The expected value is (note that we can have the sum start at $k=1$):

$$\mathrm{E}[\mathsf{x}] = e^{-\lambda t}\sum_{k=1}^{\infty} k\frac{(\lambda t)^k}{k!} = \lambda t e^{-\lambda t}\sum_{k=1}^{\infty}\frac{(\lambda t)^{k-1}}{(k-1)!} = \lambda t. \tag{5.3.4}$$

This is the same as the expected value of the discrete approximations.

Problem 85. x *follows a Poisson distribution, i.e.,*

$$\Pr[\mathsf{x}{=}k] = \frac{(\lambda t)^k}{k!}e^{-\lambda t} \quad \textit{for } k = 0, 1, \ldots. \tag{5.3.5}$$

- a. *2 points Show that* $\mathrm{E}[\mathsf{x}] = \lambda t$.

Answer. See (5.3.4). □

- b. *4 points Compute* $\mathrm{E}[\mathsf{x}(\mathsf{x}-1)]$ *and show that* $\mathrm{var}[\mathsf{x}] = \lambda t$.

Answer. For $\mathrm{E}[\mathsf{x}(\mathsf{x}-1)]$ we can have the sum start at $k=2$:

$$\mathrm{E}[\mathsf{x}(\mathsf{x}-1)] = e^{-\lambda t}\sum_{k=2}^{\infty} k(k-1)\frac{(\lambda t)^k}{k!} = (\lambda t)^2 e^{-\lambda t}\sum_{k=2}^{\infty}\frac{(\lambda t)^{k-2}}{(k-2)!} = (\lambda t)^2. \tag{5.3.6}$$

From this follows

$$\mathrm{var}[\mathsf{x}] = \mathrm{E}[\mathsf{x}^2] - (\mathrm{E}[\mathsf{x}])^2 = \mathrm{E}[\mathsf{x}(\mathsf{x}-1)] + \mathrm{E}[\mathsf{x}] - (\mathrm{E}[\mathsf{x}])^2 = (\lambda t)^2 + \lambda t - (\lambda t)^2 = \lambda t. \tag{5.3.7}$$

□

The Poisson distribution can be used as an approximation to the Binomial distribution when n large, p small, and np moderate.

PROBLEM 86. *Which value of λ would one need to approximate a given Binomial with n and p?*

ANSWER. That which gives the right expected value, i.e., $\lambda = np$. □

PROBLEM 87. *Two researchers counted cars coming down a road, which obey a Poisson distribution with unknown parameter λ. In other words, in an interval of length t one will have k cars with probability*

$$\frac{(\lambda t)^k}{k!}e^{-\lambda t}. \tag{5.3.8}$$

Their assignment was to count how many cars came in the first half hour, and how many cars came in the second half hour. However they forgot to keep track of the time when the first half hour was over, and therefore wound up only with one count, namely, they knew that 213 cars had come down the road during this hour. They were afraid they would get fired if they came back with one number only, so they applied the following remedy: they threw a coin 213 times and counted the number of heads. This number, they pretended, was the number of cars in the first half hour.

• a. *6 points Did the probability distribution of the number gained in this way differ from the distribution of actually counting the number of cars in the first half hour?*

ANSWER. First a few definitions: x is the total number of occurrences in the interval $[0,1]$. y is the number of occurrences in the interval $[0,t]$ (for a fixed t; in the problem it was $t=\frac{1}{2}$, but we will do it for general t, which will make the notation clearer and more compact. Then we want to compute $\Pr[\mathsf{y}{=}m|\mathsf{x}{=}n]$. By definition of conditional probability:

$$\Pr[\mathsf{y}{=}m|\mathsf{x}{=}n] = \frac{\Pr[\mathsf{y}{=}m \text{ and } \mathsf{x}{=}n]}{\Pr[\mathsf{x}{=}n]}. \tag{5.3.9}$$

How can we compute the probability of the intersection $\Pr[\mathsf{y}{=}m \text{ and } \mathsf{x}{=}n]$? Use a trick: express this intersection as the intersection of independent events. For this define z as the number of events in the interval $(t,1]$. Then $\{\mathsf{y}{=}m \text{ and } \mathsf{x}{=}n\} = \{\mathsf{y}{=}m \text{ and } \mathsf{z}{=}n-m\}$; therefore $\Pr[\mathsf{y}{=}m \text{ and } \mathsf{x}{=}n] = \Pr[\mathsf{y}{=}m]\Pr[\mathsf{z}{=}n-m]$; use this to get

(5.3.10)

$$\Pr[\mathsf{y}{=}m|\mathsf{x}{=}n] = \frac{\Pr[\mathsf{y}{=}m]\Pr[\mathsf{z}{=}n-m]}{\Pr[\mathsf{x}{=}n]} = \frac{\frac{\lambda^m t^m}{m!}e^{-\lambda t}\frac{\lambda^{n-m}(1-t)^{n-m}}{(n-m)!}e^{-\lambda(1-t)}}{\frac{\lambda^n}{n!}e^{-\lambda}} = \binom{n}{m}t^m(1-t)^{n-m},$$

Here we use the fact that $\Pr[\mathsf{x}{=}k] = \frac{t^k}{k!}e^{-t}$, $\Pr[\mathsf{y}{=}k] = \frac{(\lambda t)^k}{k!}e^{-\lambda t}$, $\Pr[\mathsf{z}{=}k] = \frac{(1-\lambda)^k t^k}{k!}e^{-(1-\lambda)t}$. One sees that a. $\Pr[\mathsf{y}{=}m|\mathsf{x}{=}n]$ does not depend on λ, and b. it is exactly the probability of having m successes and $n-m$ failures in a Bernoulli trial with success probability t. Therefore the procedure with the coins gave the two researchers a result which had the same probability distribution as if they had counted the number of cars in each half hour separately.

□

• b. *2 points Explain what it means that the probability distribution of the number for the first half hour gained by throwing the coins does not differ from the one gained by actually counting the cars. Which condition is absolutely necessary for this to hold?*

ANSWER. The supervisor would never be able to find out through statistical analysis of the data they delivered, even if they did it repeatedly. All estimation results based on the faked statistic would be as accurate regarding λ as the true statistics. All this is only true under the assumption that the cars really obey a Poisson distribution and that the coin is fair.

The fact that the Poisson as well as the binomial distributions are memoryless has nothing to do with them having a sufficient statistic.

□

PROBLEM 88. *8 points x is the number of customers arriving at a service counter in one hour. x follows a Poisson distribution with parameter $\lambda = 2$, i.e.,*

$$\Pr[\mathsf{x}{=}j] = \frac{2^j}{j!}e^{-2}. \tag{5.3.11}$$

• a. *Compute the probability that only one customer shows up at the service counter during the hour, the probability that two show up, and the probability that no one shows up.*

• b. *Despite the small number of customers, two employees are assigned to the service counter. They are hiding in the back, and whenever a customer steps up to the counter and rings the bell, they toss a coin. If the coin shows head, Herbert serves the customer, and if it shows tails, Karl does. Compute the probability that Herbert has to serve exactly one customer during the hour. Hint:*

$$e = 1 + 1 + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \cdots. \tag{5.3.12}$$

• c. *For any integer $k \geq 0$, compute the probability that Herbert has to serve exactly k customers during the hour.*

PROBLEM 89. *3 points Compute the moment generating function of a Poisson variable observed over a unit time interval, i.e., x satisfies $\Pr[\mathsf{x}{=}k] = \frac{\lambda^k}{k!}e^{-\lambda}$ and you want $\mathrm{E}[e^{t\mathsf{x}}]$ for all t.*

ANSWER. $\mathrm{E}[e^{t\mathsf{x}}] = \sum_{k=0}^{\infty} e^{tk}\frac{\lambda^k}{k!}e^{-\lambda} = \sum_{k=0}^{\infty}\frac{(\lambda e^t)^k}{k!}e^{-\lambda} = e^{\lambda e^t}e^{-\lambda} = e^{\lambda(e^t-1)}$. □

5.4. The Exponential Distribution

Now we will discuss random variables which are related to the Poisson distribution. At time $t = 0$ you start observing a Poisson process, and the random variable t denotes the time you have to wait until the first occurrence. t can have any nonnegative real number as value. One can derive its cumulative distribution

as follows. $\mathtt{t}>t$ if and only if there are no occurrences in the interval $[0,t]$. Therefore $\Pr[\mathtt{t}>t] = \frac{(\lambda t)^0}{0!}e^{-\lambda t} = e^{-\lambda t}$, and hence the cumulative distribution function $F_{\mathtt{t}}(t) = \Pr[\mathtt{t}\le t] = 1-e^{-\lambda t}$ when $t \ge 0$, and $F_{\mathtt{t}}(t) = 0$ for $t < 0$. The density function is therefore $f_{\mathtt{t}}(t) = \lambda e^{-\lambda t}$ for $t \ge 0$, and 0 otherwise. This is called the exponential density function (its discrete analog is the geometric random variable). It can also be called a Gamma variable with parameters $r = 1$ and λ.

Problem 90. *2 points An exponential random variable* $\mathtt{t}$ *with parameter* $\lambda > 0$ *has the density* $f_{\mathtt{t}}(t) = \lambda e^{-\lambda t}$ *for* $t \ge 0$*, and 0 for* $t < 0$*. Use this density to compute the expected value of* $\mathtt{t}$*.*

Answer. $\mathrm{E}[\mathtt{t}] = \int_0^\infty \lambda t e^{-\lambda t}dt = \int_0^\infty uv'dt = uv\Big|_0^\infty - \int_0^\infty u'vdt$, where $\begin{matrix} u=t & v'=\lambda e^{-\lambda t} \\ u'=1 & v=-e^{-\lambda t}\end{matrix}$. One can also use the more abbreviated notation $= \int_0^\infty u\,dv = uv\Big|_0^\infty - \int_0^\infty v\,du$, where $\begin{matrix} u=t & dv'=\lambda e^{-\lambda t}\,dt \\ du'=dt & v=-e^{-\lambda t}\end{matrix}$. Either way one obtains $\mathrm{E}[\mathtt{t}] = -te^{-\lambda t}\Big|_0^\infty + \int_0^\infty e^{-\lambda t}dt = 0 - \frac{1}{\lambda}e^{-\lambda t}\Big|_0^\infty = \frac{1}{\lambda}$. □

Problem 91. *4 points An exponential random variable* $\mathtt{t}$ *with parameter* $\lambda > 0$ *has the density* $f_{\mathtt{t}}(t) = \lambda e^{-\lambda t}$ *for* $t \ge 0$*, and 0 for* $t < 0$*. Use this density to compute the expected value of* $\mathtt{t}^2$*.*

Answer. One can use that $\Gamma(r) = \int_0^\infty \lambda^r t^{r-1}e^{-\lambda t}\,dt$ for $r = 3$ to get: $\mathrm{E}[\mathtt{t}^2] = (1/\lambda^2)\Gamma(3) = 2/\lambda^2$. Or all from scratch: $\mathrm{E}[\mathtt{t}^2] = \int_0^\infty \lambda t^2 e^{-\lambda t}dt = \int_0^\infty uv'dt = uv\Big|_0^\infty - \int_0^\infty u'vdt$, where $\begin{matrix} u=t^2 & v'=\lambda e^{-\lambda t} \\ u'=2t & v=-e^{-\lambda t}\end{matrix}$. Therefore $\mathrm{E}[\mathtt{t}^2] = -t^2e^{-\lambda t}\Big|_0^\infty + \int_0^\infty 2te^{-\lambda t}\,dt$. The first term vanishes, for the second do it again: $\int_0^\infty 2te^{-\lambda t}\,dt = \int_0^\infty uv'dt = uv\Big|_0^\infty - \int_0^\infty u'vdt$, where $\begin{matrix} u=t & v'=e^{-\lambda t} \\ u'=1 & v=-(1/\lambda)e^{-\lambda t}\end{matrix}$. Therefore the second term becomes $2(t/\lambda)e^{-\lambda t}\Big|_0^\infty + 2\int_0^\infty (1/\lambda)e^{-\lambda t}dt = 2/\lambda^2$. □

Problem 92. *2 points Does the exponential random variable with parameter* $\lambda > 0$*, whose cumulative distribution function is* $F_{\mathtt{t}}(t) = 1 - e^{-\lambda t}$ *for* $t \ge 0$*, and 0 otherwise, have a memory-less property? Compare Problem 80. Formulate this memory-less property and then verify whether it holds or not.*

Answer. Here is the formulation: for $\mathtt{s}<\mathtt{t}$ follows $\Pr[\mathtt{t}>t|\mathtt{t}>s] = \Pr[\mathtt{t}>t-s]$. This does indeed hold. Proof: lhs $= \frac{\Pr[\mathtt{t}>t \text{ and } \mathtt{t}>s]}{\Pr[\mathtt{t}>s]} = \frac{\Pr[\mathtt{t}>t]}{\Pr[\mathtt{t}>s]} = \frac{e^{-\lambda t}}{e^{-\lambda s}} = e^{-\lambda(t-s)}$. □

Problem 93. *The random variable* $\mathtt{t}$ *denotes the duration of an unemployment spell. It has the exponential distribution, which can be defined by:* $\Pr[\mathtt{t}>t] = e^{-\lambda t}$ *for* $t \ge 0$ *(*$\mathtt{t}$ *cannot assume negative values).*

• a. *1 point Use this formula to compute the cumulative distribution function* $F_{\mathtt{t}}(t)$ *and the density function* $f_{\mathtt{t}}(t)$

Answer. $F_{\mathtt{t}}(t) = \Pr[\mathtt{t}\le t] = 1 - \Pr[\mathtt{t}>t] = 1 - e^{-\lambda t}$ for $t \ge 0$, zero otherwise. Taking the derivative gives $f_{\mathtt{t}}(t) = \lambda e^{-\lambda t}$ for $t \ge 0$, zero otherwise. □

• b. *2 points What is the probability that an unemployment spell ends after time* $t+h$*, given that it has not yet ended at time* t*? Show that this is the same as the unconditional probability that an unemployment spell ends after time* h *(memory-less property).*

Answer.

$$\Pr[\mathtt{t}>t+h|\mathtt{t}>t] = \frac{\Pr[\mathtt{t}>t+h]}{\Pr[\mathtt{t}>t]} = \frac{e^{-\lambda(t+h)}}{e^{-\lambda t}} = e^{-\lambda h} \tag{5.4.1}$$

□

• c. *3 points Let* h *be a small number. What is the probability that an unemployment spell ends at or before* $t+h$*, given that it has not yet ended at time* t*? Hint: for small* h*, one can write approximately*

$$\Pr[t < \mathtt{t}\le t+h] = hf_{\mathtt{t}}(t). \tag{5.4.2}$$

Answer.

$$\begin{aligned}\Pr[\mathtt{t}\le t+h|\mathtt{t}>t] &= \frac{\Pr[\mathtt{t}\le t+h \text{ and } \mathtt{t}>t]}{\Pr[\mathtt{t}>t]} = \\ &= \frac{h\,f_{\mathtt{t}}(t)}{1-F_{\mathtt{t}}(t)} = \frac{h\,\lambda e^{-\lambda t}}{e^{-\lambda t}} = h\,\lambda.\end{aligned} \tag{5.4.3}$$

□

5.5. The Gamma Distribution

The time until the *second* occurrence of a Poisson event is a random variable which we will call $\mathtt{t}^{(2)}$. Its cumulative distribution function is $F_{\mathtt{t}^{(2)}}(t) = \Pr[\mathtt{t}^{(2)}\le t] = 1-\Pr[\mathtt{t}^{(2)}>t]$. But $\mathtt{t}^{(2)}>t$ means: there are either zero or one occurrences in the time between 0 and t; therefore $\Pr[\mathtt{t}^{(2)}>t] = \Pr[\mathtt{x}=0]+\Pr[\mathtt{x}=1] = e^{-\lambda t}+\lambda te^{-\lambda t}$. Putting it all together gives $F_{\mathtt{t}^{(2)}}(t) = 1-e^{-\lambda t}-\lambda te^{-\lambda t}$. In order to differentiate the cumulative distribution function we need the product rule of differentiation: $(uv)' = u'v + uv'$. This gives

$$f_{\mathtt{t}^{(2)}}(t) = \lambda e^{-\lambda t} - \lambda e^{-\lambda t} + \lambda^2 te^{-\lambda t} = \lambda^2 te^{-\lambda t}. \tag{5.5.1}$$

Problem 94. *3 points Compute the density function of* $\mathtt{t}^{(3)}$*, the time of the third occurrence of a Poisson variable.*

Answer.

$$\Pr[\mathtt{t}^{(3)}>t] = \Pr[\mathtt{x}=0] + \Pr[\mathtt{x}=1] + \Pr[\mathtt{x}=2] \tag{5.5.2}$$

$$F_{\mathtt{t}^{(3)}}(t) = \Pr[\mathtt{t}^{(3)}\le t] = 1 - (1+\lambda t+\frac{\lambda^2}{2}t^2)e^{-\lambda t} \tag{5.5.3}$$

$$f_{\mathtt{t}^{(3)}}(t) = \frac{\partial}{\partial t}F_{\mathtt{t}^{(3)}}(t) = -\Big(-\lambda(1+\lambda t+\frac{\lambda^2}{2}t^2) + (\lambda+\lambda^2 t)\Big)e^{-\lambda t} = \frac{\lambda^3}{2}t^2e^{-\lambda t}. \tag{5.5.4}$$

□

If one asks for the rth occurrence, again all but the last term cancel in the differentiation, and one gets

$$f_{\mathfrak{t}^{(r)}}(t) = \frac{\lambda^r}{(r-1)!} t^{r-1} e^{-\lambda t}. \tag{5.5.5}$$

This density is called the Gamma density with parameters λ and r.

The following definite integral, which is defined for all $r > 0$ and all $\lambda > 0$ is called the Gamma function:

$$\Gamma(r) = \int_0^\infty \lambda^r t^{r-1} e^{-\lambda t} dt. \tag{5.5.6}$$

Although this integral cannot be expressed in a closed form, it is an important function in mathematics. It is a well behaved function interpolating the factorials in the sense that $\Gamma(r) = (r-1)!$.

PROBLEM 95. *Show that $\Gamma(r)$ as defined in (5.5.6) is independent of λ, i.e., instead of (5.5.6) one can also use the simpler equation*

$$\Gamma(r) = \int_0^\infty t^{r-1} e^{-t} dt. \tag{5.5.7}$$

PROBLEM 96. *3 points Show by partial integration that the Gamma function satisfies $\Gamma(r+1) = r\Gamma(r)$.*

ANSWER. Start with

$$\Gamma(r+1) = \int_0^\infty \lambda^{r+1} t^r e^{-\lambda t} dt \tag{5.5.8}$$

and integrate by parts: $\int u'v dt = uv - \int uv' dt$ with $u' = \lambda e^{-\lambda t}$ and $v = \lambda^r t^r$, therefore $u = -e^{-\lambda t}$ and $v' = r\lambda^r t^{r-1}$:

$$\Gamma(r+1) = -\lambda^r t^r e^{-\lambda t}\Big|_0^\infty + \int_0^\infty r\lambda^r t^{r-1} e^{-\lambda t} dt = 0 + r\Gamma(r). \tag{5.5.9}$$

□

PROBLEM 97. *Show that $\Gamma(r) = (r-1)!$ for all natural numbers $r = 1, 2, \ldots$.*

ANSWER. Proof by induction. First verify that it holds for $r = 1$, i.e., that $\Gamma(1) = 1$:

$$\Gamma(1) = \int_0^\infty \lambda e^{-\lambda t} dt = -e^{-\lambda t}\Big|_0^\infty = 1 \tag{5.5.10}$$

and then, assuming that $\Gamma(r) = (r-1)!$ Problem 96 says that $\Gamma(r+1) = r\Gamma(r) = r(r-1)! = r!$. □

Without proof: $\Gamma(\frac{1}{2}) = \sqrt{\pi}$. This will be shown in Problem 161.

Therefore the following defines a density function, called the Gamma density with parameter r and λ, for all $r > 0$ and $\lambda > 0$:

$$f(x) = \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x} \quad \text{for} \quad x \geq 0, \quad 0 \text{ otherwise.} \tag{5.5.11}$$

The only application we have for it right now is: this is the distribution of the time one has to wait until the rth occurrence of a Poisson distribution with intensity λ. Later we will have other applications in which r is not an integer.

PROBLEM 98. *4 points Compute the moment generating function of the Gamma distribution.*

ANSWER.

$$m_{\mathfrak{x}}(t) = \mathrm{E}[e^{t\mathfrak{x}}] = \int_0^\infty e^{tx} \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x} dx \tag{5.5.12}$$

$$= \frac{\lambda^r}{(\lambda-t)^r} \int_0^\infty \frac{(\lambda-t)^r x^{r-1}}{\Gamma(r)} e^{-(\lambda-t)x}\, dx \tag{5.5.13}$$

$$= \left(\frac{\lambda}{\lambda-t}\right)^r \tag{5.5.14}$$

since the integrand in (5.5.12) is the density function of a Gamma distribution with parameters r and $\lambda - t$. □

PROBLEM 99. *2 points The density and moment generating functions of a Gamma variable $\mathfrak{x}$ with parameters $r > 0$ and $\lambda > 0$ are*

$$f_{\mathfrak{x}}(x) = \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x} \quad \textit{for} \quad x \geq 0, \quad 0 \textit{ otherwise.} \tag{5.5.15}$$

$$m_{\mathfrak{x}}(t) = \left(\frac{\lambda}{\lambda-t}\right)^r. \tag{5.5.16}$$

Show the following: If $\mathfrak{x}$ has a Gamma distribution with parameters r and 1, then $\mathfrak{v} = \mathfrak{x}/\lambda$ has a Gamma distribution with parameters r and λ. You can prove this either using the transformation theorem for densities, or the moment-generating function.

ANSWER. Solution using density function: The random variable whose density we know is $\mathfrak{x}$; its density is $\frac{1}{\Gamma(r)} x^{r-1} e^{-x}$. If $\mathfrak{x} = \lambda\mathfrak{v}$, then $\frac{dx}{dv} = \lambda$, and the absolute value is also λ. Therefore the density of $\mathfrak{v}$ is $\frac{\lambda^r}{\Gamma(r)} v^{r-1} e^{-\lambda v}$. Solution using the mgf:

$$m_{\mathfrak{x}}(t) = \mathrm{E}[e^{t\mathfrak{x}}] = \left(\frac{1}{1-t}\right)^r \tag{5.5.17}$$

$$m_{\mathfrak{v}}(t)\, \mathrm{E}[e^{t\mathfrak{v}}] = \mathrm{E}[e^{(t/\lambda)\mathfrak{x}}] = \left(\frac{1}{1-(t/\lambda)}\right)^r = \left(\frac{\lambda}{\lambda-t}\right)^r \tag{5.5.18}$$

but this last expression can be recognized to be the mgf of a Gamma with r and λ. □

PROBLEM 100. *2 points It* x *has a Gamma distribution with parameters* r *and* λ*, and* y *one with parameters* p *and* λ*, and both are independent, show that* $\mathsf{x}+\mathsf{y}$ *has a Gamma distribution with parameters* $r+p$ *and* λ *(reproductive property of the Gamma distribution.) You may use equation* (5.5.14) *without proof*

ANSWER.

$$\left(\frac{\lambda}{\lambda-t}\right)^r\left(\frac{\lambda}{\lambda-t}\right)^p=\left(\frac{\lambda}{\lambda-t}\right)^{r+p}. \tag{5.5.19}$$

□

PROBLEM 101. *Show that a Gamma variable* x *with parameters* r *and* λ *has expected value* $\mathrm{E}[\mathsf{x}]=r/\lambda$ *and variance* $\mathrm{var}[\mathsf{x}]=r/\lambda^2$.

ANSWER. Proof with moment generating function:

$$\frac{d}{dt}\left(\frac{\lambda}{\lambda-t}\right)^r=\frac{r}{\lambda}\left(\frac{\lambda}{\lambda-t}\right)^{r+1}, \tag{5.5.20}$$

therefore $\mathrm{E}[\mathsf{x}]=\frac{r}{\lambda}$, and by differentiating twice (apply the same formula again), $\mathrm{E}[\mathsf{x}^2]=\frac{r(r+1)}{\lambda^2}$, therefore $\mathrm{var}[\mathsf{x}]=\frac{r}{\lambda^2}$.

Proof using density function: For the expected value one gets $\mathrm{E}[\mathsf{t}]=\int_0^\infty t\cdot\frac{\lambda^r}{\Gamma(r)}t^{r-1}e^{-\lambda t}dt=\frac{r}{\lambda}\frac{1}{\Gamma(r+1)}\int_0^\infty t^r\lambda^{r+1}e^{-\lambda t}dt=\frac{r}{\lambda}\cdot\frac{\Gamma(r+1)}{\Gamma(r+1)}=\frac{r}{\lambda}$. Using the same tricks $\mathrm{E}[\mathsf{t}^2]=\int_0^\infty t^2\cdot\frac{\lambda^r}{\Gamma(r)}t^{r-1}e^{-\lambda t}dt=\frac{r(r+1)}{\lambda^2}\int_0^\infty\frac{\lambda^{r+2}}{\Gamma(r+2)}t^{r+1}e^{-\lambda t}dt=\frac{r(r+1)}{\lambda^2}$.

Therefore $\mathrm{var}[\mathsf{t}]=\mathrm{E}[\mathsf{t}^2]-(\mathrm{E}[\mathsf{t}])^2=r/\lambda^2$. □

5.6. The Uniform Distribution

PROBLEM 102. *Let* x *be uniformly distributed in the interval* $[a,b]$*, i.e., the density function of* x *is a constant for* $a\le x\le b$*, and zero otherwise.*

- a. *1 point What is the value of this constant?*

ANSWER. It is $\frac{1}{b-a}$ □

- b. *2 points Compute* $\mathrm{E}[\mathsf{x}]$

ANSWER. $\mathrm{E}[\mathsf{x}]=\int_a^b\frac{x}{b-a}\,dx=\frac{1}{b-a}\frac{b^2-a^2}{2}=\frac{a+b}{2}$ since $b^2-a^2=(b+a)(b-a)$. □

- c. *2 points Show that* $\mathrm{E}[\mathsf{x}^2]=\frac{a^2+ab+b^2}{3}$.

ANSWER. $\mathrm{E}[\mathsf{x}^2]=\int_a^b\frac{x^2}{b-a}\,dx=\frac{1}{b-a}\frac{b^3-a^3}{3}$. Now use the identity $b^3-a^3=(b-a)(b^2+ab+a^2)$ (check it by multiplying out). □

- d. *2 points Show that* $\mathrm{var}[\mathsf{x}]=\frac{(b-a)^2}{12}$.

ANSWER. $\mathrm{var}[\mathsf{x}]=\mathrm{E}[\mathsf{x}^2]-(\mathrm{E}[\mathsf{x}])^2=\frac{a^2+ab+b^2}{3}-\frac{(a+b)^2}{4}=\frac{4a^2+4ab+4b^2}{12}-\frac{3a^2+6ab+3b^2}{12}=\frac{(b-a)^2}{12}$. □

5.7. The Beta Distribution

Assume you have two independent variables, both distributed uniformly over the interval $[0,1]$, and you want to know the distribution of their maximum. Or of their minimum. Or you have three and you want the distribution of the one in the middle. Then the densities have their maximum to the right, or to the left, or in the middle. The distribution of the rth highest out of n independent uniform variables is an example of the Beta density function. Can also be done and is probability-theoretically meaningful for arbitrary real r and n.

PROBLEM 103. x *and* y *are two independent random variables distributed uniformly over the interval* $[0,1]$*. Let* u *be their minimum* $\mathsf{u}=\min(\mathsf{x},\mathsf{y})$ *(i.e.,* u *takes the value of* x *when* x *is smaller, and the value of* y *when* y *is smaller), and* $\mathsf{v}=\max(\mathsf{x},\mathsf{y})$.

- a. *2 points Given two numbers* q *and* r *between 0 and 1. Draw the events* $\mathsf{u}\le q$ *and* $\mathsf{v}\le r$ *into the unit square and compute their probabilities.*

- b. *2 points Compute the density functions* $f_{\mathsf{u}}(u)$ *and* $f_{\mathsf{v}}(v)$.

- c. *2 points Compute the expected values of* u *and* v.

ANSWER. For u: $\Pr[\mathsf{u}\le q]=1-\Pr[\mathsf{u}>q]=1-(1-q)^2=2q-q^2$. $f_{\mathsf{v}}(v)=2v$ Therefore $f_{\mathsf{u}}(u)=2-2u$

$$\mathrm{E}[\mathsf{u}]=\int_0^1(2-2u)u\,du=\left(u^2-\frac{2u^3}{3}\right)\Big|_0^1=\frac{1}{3}. \tag{5.7.1}$$

For v it is: $\Pr[\mathsf{v}\le r]=r^2$; this is at the same time the cumulative distribution function. Therefore the density function is $f_{\mathsf{v}}(v)=2v$ for $0\le v\le 1$ and 0 elsewhere.

$$\mathrm{E}[\mathsf{v}]=\int_0^1 v2v\,dv=\frac{2v^3}{3}\Big|_0^1=\frac{2}{3}. \tag{5.7.2}$$

□

5.8. The Normal Distribution

By definition, y is *normally distributed* with mean μ and variance σ^2, in symbols, $\mathsf{y}\sim\mathsf{N}(\mu,\sigma^2)$, if it has the density function

$$f_{\mathsf{y}}(y)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(y-\mu)^2}{2\sigma^2}}. \tag{5.8.1}$$

It will be shown a little later that this is indeed a density function. This distribution has the highest entropy among all distributions with a given mean and variance [**Kap89**, p. 47].

If $\mathsf{y}\sim\mathsf{N}(\mu,\sigma^2)$, then $\mathsf{z}=(\mathsf{y}-\mu)/\sigma\sim\mathsf{N}(0,1)$, which is called the standard Normal distribution.

PROBLEM 104. *2 points Compare* [**Gre97**, p. 68]*: Assume* $\mathsf{x} \sim \mathsf{N}(3,4)$ *(mean is* 3 *and variance* 4*). Determine with the help of a table of the Standard Normal Distribution function* $\Pr[2<\mathsf{x}\leq 5]$.

ANSWER. $\Pr[2<\mathsf{x}\leq 5] = \Pr[2-3<\mathsf{x}-3\leq 5-3] = \Pr[\frac{2-3}{2} < \frac{\mathsf{x}-3}{2} \leq \frac{5-3}{2}] = \Pr[-\frac{1}{2} < \frac{\mathsf{x}-3}{2} \leq 1] = \Phi(1) - \Phi(-\frac{1}{2}) = \Phi(1) - (1-\Phi(\frac{1}{2})) = \Phi(1) + \Phi(\frac{1}{2}) - 1 = 0.8413 + 0.6915 - 1 = 0.5328$. Some tables (Greene) give the area between 0 and all positive values; in this case it is $0.3413 + 0.1915$. □

The moment generating function of a standard normal $\mathsf{z} \sim \mathsf{N}(0,1)$ is the following integral:

$$m_{\mathsf{z}}(t) = \mathrm{E}[e^{t\mathsf{z}}] = \int_{-\infty}^{+\infty} e^{tz} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz. \tag{5.8.2}$$

To solve this integral, complete the square in the exponent:

$$tz - \frac{z^2}{2} = \frac{t^2}{2} - \frac{1}{2}(z-t)^2; \tag{5.8.3}$$

Note that the first summand, $\frac{t^2}{2}$, no longer depends on z; therefore the factor $e^{\frac{t^2}{2}}$ can be written in front of the integral:

$$m_{\mathsf{z}}(t) = e^{\frac{t^2}{2}} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z-t)^2} dz = e^{\frac{t^2}{2}}, \tag{5.8.4}$$

because now the integrand is simply the density function of a $\mathsf{N}(t,1)$.

A general univariate normal $\mathsf{x} \sim \mathsf{N}(\mu, \sigma^2)$ can be written as $\mathsf{x} = \mu + \sigma\mathsf{z}$ with $\mathsf{z} \sim \mathsf{N}(0,1)$, therefore

$$m_{\mathsf{x}}(t) = \mathrm{E}[e^{(\mu+\sigma z)t}] = e^{\mu t}\,\mathrm{E}[e^{\sigma z t}] = e^{(\mu t + \sigma^2 t^2/2)}. \tag{5.8.5}$$

PROBLEM 105. *Given two independent normal variables* $\mathsf{x} \sim \mathsf{N}(\mu_{\mathsf{x}}, \sigma_{\mathsf{x}}^2)$ *and* $\mathsf{y} \sim \mathsf{N}(\mu_{\mathsf{y}}, \sigma_{\mathsf{y}}^2)$*. Using the moment generating function, show that*

$$\alpha\mathsf{x} + \beta\mathsf{y} \sim \mathsf{N}(\alpha\mu_{\mathsf{x}} + \beta\mu_{\mathsf{y}}, \alpha^2\sigma_{\mathsf{x}}^2 + \beta^2\sigma_{\mathsf{y}}^2). \tag{5.8.6}$$

ANSWER. Because of independence, the moment generating function of $\alpha\mathsf{x}+\beta\mathsf{y}$ is the product of the m.g.f. of $\alpha\mathsf{x}$ and the one of $\beta\mathsf{y}$:

$$m_{\alpha\mathsf{x}+\beta\mathsf{y}}(t) = e^{\mu_{\mathsf{x}}\alpha t + \sigma_{\mathsf{x}}^2\alpha^2 t^2/2} e^{\mu_{\mathsf{y}}\beta t + \sigma_{\mathsf{y}}^2\beta^2 t^2/2} = e^{(\mu_{\mathsf{x}}\alpha + \mu_{\mathsf{y}}\beta)t + (\sigma_{\mathsf{x}}^2\alpha^2 + \sigma_{\mathsf{y}}^2\beta^2)t^2/2}, \tag{5.8.7}$$

which is the moment generating function of a $\mathsf{N}(\alpha\mu_{\mathsf{x}} + \beta\mu_{\mathsf{y}}, \alpha^2\sigma_{\mathsf{x}}^2 + \beta^2\sigma_{\mathsf{y}}^2)$. □

We will say more about the univariate normal later when we discuss the multivariate normal distribution.

Sometimes it is also necessary to use the truncated normal distributions. If z is standard normal, then

$$\mathrm{E}[\mathsf{z}|\mathsf{z}>z] = \frac{f_{\mathsf{z}}(z)}{1-F_{\mathsf{z}}(z)}, \qquad \mathrm{var}[\mathsf{z}|\mathsf{z}>z] = 1 - \mu(\mu - z), \text{ where } \mu = \mathrm{E}[\mathsf{z}|\mathsf{z}>z]. \tag{5.8.8}$$

This expected value is therefore the ordinate of the density function at point z divided by the tail area of the tail over which z is known to vary. (This rule is only valid for the normal density function, not in general!) These kinds of results can be found in [**JK70**, pp. 81–83] or in the original paper [**Coh50**]

PROBLEM 106. *Every customer entering a car dealership in a certain location can be thought of as having a reservation price* y *in his or her mind: if the car will be offered at or below this reservation price, then he or she will buy the car, otherwise there will be no sale. (Assume for the sake of the argument all cars are equal.) Assume this reservation price is Normally distributed with mean \$6000 and standard deviation \$1000 (if you randomly pick a customer and ask his or her reservation price). If a sale is made, a person's consumer surplus is the difference between the reservation price and the price actually paid, otherwise it is zero. For this question you will need the table for the standard normal cumulative distribution function.*

• a. *2 points A customer is offered a car at a price of \$5800. The probability that he or she will take the car is* $\boxed{}$.

ANSWER. We need $\Pr[\mathsf{y}\geq 5800]$. If $\mathsf{y}=5800$ then $\mathsf{z}=\frac{\mathsf{y}-6000}{1000} = -0.2$; $\Pr[\mathsf{z}\geq -0.2] = 1 - \Pr[\mathsf{z}\leq -0.2] = 1 - 0.4207 = 0.5793$. □

• b. *3 points Since it is the 63rd birthday of the owner of the dealership, all cars in the dealership are sold for the price of \$6300. You pick at random one of the people coming out of the dealership. The probability that this person bought a car and his or her consumer surplus was more than \$500 is* $\boxed{}$.

ANSWER. This is the unconditional probability that the reservation price was higher than $\$6300 + \$500 = \$6800$. i.e., $\Pr[\mathsf{y}\geq 6800]$. Define $\mathsf{z} = (\mathsf{y} - \$6000)/\$1000$. It is a standard normal, and $\mathsf{y}\leq\$6800 \iff \mathsf{z}\leq .8$, Therefore $p = 1 - \Pr[\mathsf{z}\leq .8] = .2119$. □

• c. *4 points Here is an alternative scenario: Since it is the 63rd birthday of the owner of the dealership, all cars in the dealership are sold for the "birthday special" price of \$6300. You pick at random one of the people who bought one of these "birthday specials" priced \$6300. The probability that this person's consumer surplus was more than \$500 is* $\boxed{}$.

The important part of this question is: it depends on the outcome of the experiment whether or not someone is included in the sample sample selection bias.

ANSWER. Here we need the conditional probability:

$$p = \Pr[\mathsf{y}>\$6800|\mathsf{y}>\$6300] = \frac{\Pr[\mathsf{y}>\$6800]}{\Pr[\mathsf{y}>\$6300]} = \frac{1-\Pr[\mathsf{y}\leq\$6800]}{1-\Pr[\mathsf{y}\leq\$6300]}. \tag{5.8.9}$$

Again use the standard normal $z = (\mathsf{y} - \$6000)/\$1000$. As before, $\mathsf{y} \leq \$6800 \iff z \leq .8$, and $\mathsf{y} \leq \$6300 \iff z \leq .3$. Therefore

$$p = \frac{1 - \Pr[z \leq .8]}{1 - \Pr[z \leq .3]} = \frac{.2119}{.3821} = .5546. \tag{5.8.10}$$

It depends on the layout of the normal distribution table how this should be looked up. □

• d. *5 points We are still picking out customers that have bought the birthday specials. Compute the median value m of such a customer's consumer surplus. It is defined by*

$$\Pr[\mathsf{y} > \$6300 + m | \mathsf{y} > \$6300] = \Pr[\mathsf{y} \leq \$6300 + m | \mathsf{y} > \$6300] = 1/2. \tag{5.8.11}$$

Answer. Obviously, $m \geq \$0$. Therefore

$$\Pr[\mathsf{y} > \$6300 + m | \mathsf{y} > \$6300] = \frac{\Pr[\mathsf{y} > \$6300 + m]}{\Pr[\mathsf{y} > \$6300]} = \frac{1}{2}, \tag{5.8.12}$$

or $\Pr[\mathsf{y} > \$6300 + m] = (1/2)\Pr[\mathsf{y} > \$6300] = (1/2).3821 = .1910$. I.e., $\Pr[\frac{\mathsf{y}-6000}{1000} > \frac{6300-6000+m}{1000} = \frac{300}{1000} + \frac{m}{1000}] = .1910$. For this we find in the table $\frac{300}{1000} + \frac{m}{1000} = 0.875$, therefore $300 + m = 875$, or $m = \$575$. □

• e. *3 points Is the expected value of the consumer surplus of all customers that have bought a birthday special larger or smaller than the median? Fill in your answer here:* $\boxed{}$. *Proof is not required, as long as the answer is correct.*

Answer. The mean is larger because it is more heavily influenced by outliers.

$$\begin{aligned}
\mathrm{E}[\mathsf{y} - 6300 | \mathsf{y} \geq 6300] &= \mathrm{E}[6000 + 1000\mathsf{z} - 6300 | 6000 + 1000\mathsf{z} \geq 6300] && (5.8.13)\\
&= \mathrm{E}[1000\mathsf{z} - 300 | 1000\mathsf{z} \geq 300] && (5.8.14)\\
&= \mathrm{E}[1000\mathsf{z} | \mathsf{z} \geq 0.3] - 300 && (5.8.15)\\
&= 1000\,\mathrm{E}[\mathsf{z} | \mathsf{z} \geq 0.3] - 300 && (5.8.16)\\
&= 1000\frac{f(0.3)}{1 - \Psi(0.3)} - 300 = 698 > 575. && (5.8.17)
\end{aligned}$$

□

5.9. The Chi-Square Distribution

A χ^2 with *one* degree of freedom is defined to be the distribution of the square $\mathsf{q} = \mathsf{z}^2$ of a univariate standard normal variable.

Call the cumulative distribution function of a standard normal $F_{\mathsf{z}}(z)$. Then the cumulative distribution function of the χ^2 variable $\mathsf{q} = \mathsf{z}^2$ is, according to Problem 47, $F_{\mathsf{q}}(q) = 2F_{\mathsf{z}}(\sqrt{q}) - 1$. To get the density of q take the derivative of $F_{\mathsf{q}}(q)$ with

respect to q. For this we need the chain rule, first taking the derivative with respect to $z = \sqrt{q}$ and multiply by $\frac{dz}{dq}$:

$$f_{\mathsf{q}}(q) = \frac{d}{dq}\big(2F_{\mathsf{z}}(\sqrt{q}) - 1\big) = \frac{d}{dq}\big(2F_{\mathsf{z}}(z) - 1\big) \tag{5.9.1}$$

$$= 2\frac{dF_{\mathsf{z}}}{dz}(z)\frac{dz}{dq} = \frac{2}{\sqrt{2\pi}}e^{-z^2/2}\frac{1}{2\sqrt{q}} \tag{5.9.2}$$

$$= \frac{1}{\sqrt{2\pi q}}e^{-q/2}. \tag{5.9.3}$$

Now remember the Gamma function. Since $\Gamma(1/2) = \sqrt{\pi}$ (Proof in Problem 161), one can rewrite (5.9.3) as

$$f_{\mathsf{q}}(q) = \frac{(1/2)^{1/2}q^{-1/2}e^{-q/2}}{\Gamma(1/2)}, \tag{5.9.4}$$

i.e., it is a Gamma density with parameters $r = 1/2$, $\lambda = 1/2$.

A χ^2 with p degrees of freedom is defined as the sum of p independent univariate χ^2 variables. By the reproductive property of the Gamma distribution (Problem 100) this gives a Gamma variable with parameters $r = p/2$ and $\lambda = 1/2$.

$$\text{If} \quad \mathsf{q} \sim \chi^2_p \qquad \text{then} \quad \mathrm{E}[\mathsf{q}] = p \qquad \text{and} \quad \mathrm{var}[\mathsf{q}] = 2p \tag{5.9.5}$$

We will say that a random variable q is distributed as a $\sigma^2\chi^2_p$ iff q/σ^2 is a χ^2_p. This is the distribution of a sum of p independent $\mathsf{N}(0, \sigma^2)$ variables.

5.10. The Lognormal Distribution

This is a random variable whose log has a normal distribution. See [**Gre97**, p. 71]. Parametrized by the μ and σ^2 of its log. Density is

$$\frac{1}{x\sqrt{2\pi\sigma^2}}e^{-(\ln x - \mu/\sigma^2)/2} \tag{5.10.1}$$

[**Cow77**, pp. 82–87] has an excellent discussion of the properties of the lognormal for income distributions.

5.11. The Cauchy Distribution

Problem 107. *6 points* [**JK70**, pp. 155/6] *An example of a distribution without mean and variance is the* Cauchy distribution, *whose density looks much like the normal density, but has much thicker tails. The density and characteristic functions are (I am not asking you to compute the characteristic function)*

$$f_{\mathsf{x}}(x) = \frac{1}{\pi(1 + x^2)} \qquad \mathrm{E}[e^{it\mathsf{x}}] = \exp(-|t|). \tag{5.11.1}$$

Here $i = \sqrt{-1}$, but you should not be afraid of it, in most respects, i behaves like any real number. The characteristic function has properties very similar to the moment generating function, with the added advantage that it always exists. Using the characteristic functions show that if $\mathtt{x}$ and $\mathtt{y}$ are independent Cauchy distributions, then $(\mathtt{x}+\mathtt{y})/2$ has the same distribution as $\mathtt{x}$ or $\mathtt{y}$.

Answer.

$$\mathrm{E}\Big[\exp\Big(it\frac{\mathtt{x}+\mathtt{y}}{2}\Big)\Big] = \mathrm{E}\Big[\exp\Big(i\frac{t}{2}\mathtt{x}\Big)\exp\Big(i\frac{t}{2}\mathtt{y}\Big)\Big] = \exp(-\Big|\frac{t}{2}\Big|)\exp(-\Big|\frac{t}{2}\Big|) = \exp(-|t|). \tag{5.11.2}$$

□

It has taken a historical learning process to distinguish significant from insignificant events. The order in which the birds sit down on a tree is insignificant, but the constellation of stars on the night sky is highly significant for the seasons etc. The confusion between significant and insignificant events can explain how astrology arose: after it was discovered that the constellation of stars was significant, but without knowledge of the mechanism through which the constellation of stars was significant, people experimented to find evidence of causality between those aspects of the night sky that were changing, like the locations of the planets, and events on earth, like the births of babies. Romans thought the constellation of birds in the sky was significant.

Freud discovered that human error may be significant. Modern political consciousness still underestimates the extent to which the actions of states are significant: If a welfare recipient is faced with an intractable labyrinth of regulations and a multitude of agencies, then this is not the unintended result of bureaucracy gone wild, but it is deliberate: this bureaucratic nightmare deters people from using welfare, but it creates the illusion that welfare exists and it does give relief in some blatant cases.

Also "mistakes" like the bombing of the Chinese embassy are not mistakes but are significant.

In statistics the common consensus is that the averages are significant and the deviations from the averages are insignificant. By taking averages one distills the significant, systematic part of the date from the insignificant part. Usually this is justified by the "law of large numbers." I.e., people think that this is something about reality which can be derived and proved mathematically. However this is an irrealist position: how can math tell us which events are significant?

Here the Cauchy distribution is an interesting counterexample: it is a probability distribution for which it does not make sense to take averages. If one takes the average of n observations, then this average does not have less randomness than each individual observation, but it has exactly the same distribution as one single

observation. (The law of large numbers does not apply here because the Cauchy distribution does not have an expected value.)

In a world in which random outcomes are Cauchy-distributed, taking averages is not be a good way to learn from one's experiences. People who try to keep track of things by taking averages (or by running regressions, which is a natural extension of taking averages) would have the same status in that world as astrologers have in our world. Taking medians and other quantiles would be considered scientific, but taking averages would be considered superstition.

The lesson of this is: even a scientific procedure as innocuous as that of taking averages cannot be justified on purely epistemological grounds. Although it is widely assumed that the law of large numbers is such a justification, it is not. The law of large numbers does not always hold; it only holds if the random variable under consideration has an expected value.

The transcendental realist can therefore say: since it apparently does make sense to take averages in our world, we can deduce transcendentally that many random variables which we are dealing with do have finite expected values.

This is perhaps the simplest case of a transcendental conclusion. But this simplest case also vindicates another one of Bhaskar's assumptions: these transcendental conclusions cannot be arrived at in a non-transcendental way, by staying in the science itself. It is impossible to decide, using statistical means alone, whether one's data come from a distribution which has finite expected values or not. The reason is that one always has only finite datasets, and the empirical distribution of a finite sample always has finite expected values, even if the sample comes from a population which does not have finite expected values.

CHAPTER 6

Sufficient Statistics and their Distributions

6.1. Factorization Theorem for Sufficient Statistics

Given a family of probability measures $\Pr_\theta$ defined on a sample set U, which depend on a parameter $\theta \in \Theta$. By definition, the scalar random variable $\mathsf{t} : \mathsf{U} \to \mathbb{R}$ is a "sufficient statistic" for parameter θ if and only if for all events $\mathsf{E} \subset \mathsf{U}$ and all t, the conditional probability $\Pr_\theta[\mathsf{E}|\mathsf{t}{=}t]$ does not involve θ. The factorization theorem for sufficient statistics allows you to tell, from the functional form of the probability mass function or density function, whether a given statistic t is sufficient or not. It says: t is sufficient if and only if there exists a function of two variables $g\colon \mathbb{R}\times\Theta \to \mathbb{R},\ (t,\theta) \mapsto g(t,\theta)$, and a scalar random variable $\mathsf{h}\colon \mathsf{U} \to \mathbb{R},\ \omega \mapsto \mathsf{h}(\omega)$ so that in the discrete case, the probability mass function $p_\theta(\omega) = \Pr_\theta[\{\omega\}]$ can be factorized as follows:

$$p_\theta(\omega) = g\big(\mathsf{t}(\omega),\theta\big)\cdot \mathsf{h}(\omega) \quad \text{for all } \omega \in \mathsf{U}.$$

If $\mathsf{U} \subset \mathbb{R}^n$, we can write $\omega = (y_1,\dots,y_n)$. If $\Pr_\theta$ is not discrete but generated by a family of probability densities $f(y_1,\dots,y_n;\theta)$, then the condition reads

$$f(y_1,\dots,y_n;\theta) = g\big(\mathsf{t}(y_1,\dots,y_n),\theta\big)\cdot \mathsf{h}(y_1,\dots,y_n).$$

Note what this means: the probability of an elementary event (or of an infinitesimal interval) is written as the product of two parts: one depends on ω through t, while the other depends on ω directly. Only that part of the probability that depends on ω through t is allowed to also depend on θ.

Proof in the discrete case: First let us show the necessity of this factorization. Assume that t is sufficient, i.e., that $\Pr_\theta[\omega|\mathsf{t}{=}t]$ does not involve θ. Then one possible factorization is

$$\Pr_\theta[\omega] = \Pr_\theta[\mathsf{t}{=}\mathsf{t}(\omega)]\cdot \Pr[\omega|\mathsf{t}{=}\mathsf{t}(\omega)] \tag{6.1.1}$$

$$= g(\mathsf{t}(\omega),\theta)\cdot \mathsf{h}(\omega). \tag{6.1.2}$$

Now let us prove that the factorization property implies sufficiency. Assume therefore (6.1) holds. We have to show that for all $\omega \in \mathsf{U}$ and $t \in \mathbb{R}$, the conditional probability $\Pr_\theta[\{\omega\}|\{\kappa \in \mathsf{U}\colon \mathsf{t}(\kappa) = t\}]$, which will in shorthand notation be written

as $\Pr_\theta[\omega|\mathsf{t}{=}t]$, does not depend on θ.

$$\Pr_\theta[\mathsf{t}{=}t] = \sum_{\omega\colon \mathsf{t}(\omega)=t} \Pr_\theta[\{\omega\}] = \sum_{\omega\colon \mathsf{t}(\omega)=t} g(\mathsf{t}(\omega),\theta)\cdot \mathsf{h}(\omega) \tag{6.1.3}$$

$$= g(t,\theta)\cdot \sum_{\omega\colon \mathsf{t}(\omega)=t} \mathsf{h}(\omega) = g(t,\theta)\cdot k(t), \text{ say.} \tag{6.1.4}$$

Here it is important that $k(t)$ does not depend on θ. Now

$$\Pr_\theta[\omega|\mathsf{t}{=}t] = \Pr_\theta[\{\omega\}\cap\{\mathsf{t}{=}t\}]\Pr_\theta[\mathsf{t}{=}t] \tag{6.1.5}$$

if $\mathsf{t}(\omega) \neq t$, this is zero, i.e., independent of θ. Now look at case $\mathsf{t}(\omega) = t$, i.e., $\{\omega\}\cap\{\mathsf{t}{=}t\} = \{\omega\}$. Then

$$\Pr_\theta[\omega|\mathsf{t}{=}t] = \frac{g(t,\theta)\mathsf{h}(\omega)}{g(t,\theta)k(t)} = \frac{\mathsf{h}(\omega)}{k(t)}, \text{ which is independent of } \theta. \tag{6.1.6}$$

Problem 108. *6 points Using the* factorization theorem for sufficient statistics, *show that in a n times repeated Bernoulli experiment (n is known), the number of successes is a sufficient statistic for the success probability p.*

• a. *Here is a formulation of the factorization theorem: Given a family of discrete probability measures $\Pr_\theta$ depending on a parameter θ. The statistic t is sufficient for parameter θ iff there exists a function of two variables $g\colon \mathbb{R}\times\Theta \to \mathbb{R},\ (t,\theta) \mapsto g(t;\theta)$, and a function of one variable $\mathsf{h}\colon \mathsf{U} \to \mathbb{R},\ \omega \mapsto \mathsf{h}(\omega)$ so that for all $\omega \in \mathsf{U}$*

$$\Pr_\theta[\{\omega\}] = g\big(\mathsf{t}(\omega),\theta\big)\cdot \mathsf{h}(\omega).$$

Before you apply this, ask yourself: what is ω?

Answer. This is very simple: the probability of every elementary event depends on this element only through the random variable $\mathsf{t}\colon \mathsf{U} \to N$, which is the number of successes. $\Pr[\{\omega\}] = p^{\mathsf{t}(\omega)}(1-p)^{n-\mathsf{t}(\omega)}$. Therefore $g(k;p) = p^k(1-p)^{n-k}$ and $\mathsf{h}(\omega) = 1$ does the trick. One can also say: the probability of one element ω is the probability of $\mathsf{t}(\omega)$ successes divided by $\binom{n}{\mathsf{t}(\omega)}$. This gives another easy-to-understand factorization. □

6.2. The Exponential Family of Probability Distributions

Assume the random variable x has values in $U \subset \mathbb{R}$. A family of density functions (or, in the discrete case, probability mass functions) $f_\mathsf{x}(x;\xi)$ that depends on the parameter $\xi \in \Xi$ is called a one-parameter exponential family if and only if there exist functions $s,u\colon \Xi \to \mathbb{R}$ and $r,t\colon U \to \mathbb{R}$ such that the density function can be written as

$$f_\mathsf{x}(x;\xi) = r(x)s(\xi)\exp\big(t(x)u(\xi)\big) \quad \text{if } x \in U, \text{ and } = 0 \text{ otherwise.} \tag{6.2.1}$$

For this definition it is important that $U \subset \mathbb{R}^n$ does not depend on ξ. Notice how symmetric this condition is between observations and parameters. If we put the factors $r(x)s(\xi)$ into the exponent we get

$$f_{\mathsf{x}}(x;\xi) = \exp\big(t(x)u(\xi) + v(x) + w(\xi)\big) \quad \text{if } x \in U, \text{ and } = 0 \text{ otherwise.} \tag{6.2.2}$$

If we plug the random variable x into the function t we get the transformed random variable $\mathsf{y} = t(\mathsf{x})$, and if we re-define the parameter $\theta = u(\xi)$, we get the density in its *canonical form*

$$f_{\mathsf{y}}(y;\theta) = \exp\big(y\theta - b(\theta) + c(y)\big) \quad \text{if } y \in U, \text{ and } = 0 \text{ otherwise.} \tag{6.2.3}$$

Note here the minus sign in front of b. We will see later that $\theta \mapsto b(\theta)$ is an important function; its derivatives yield the mean and variance functions used in the generalized linear model.

Problem 109. *3 points Show that the Binomial distribution* (3.7.1)

$$p_{\mathsf{x}}(k) = \Pr[\mathsf{x}{=}k] = \binom{n}{k} p^k (1-p)^{(n-k)} \quad k = 0, 1, 2, \ldots, n \tag{6.2.4}$$

is a member of the exponential family. Compute the canonical parameter θ and the function $b(\theta)$.

Answer. Rewrite (6.2.4) as

$$p_{\mathsf{x}}(k) = \tbinom{n}{k}\left(\frac{p}{1-p}\right)^k (1-p)^n = \exp\left(k\ln\left(\frac{p}{1-p}\right) + n\ln(1-p) + \ln\tbinom{n}{k}\right) \tag{6.2.5}$$

therefore $\theta = \ln\left(\frac{p}{1-p}\right)$. To compute $b(\theta)$ you have to express $n\ln(1-p)$ as a function of θ and then reverse the sign. The following steps are involved: $\exp\theta = \frac{p}{1-p} = \frac{1}{1-p} - 1$; $1 + \exp\theta = \frac{1}{1-p}$; $\ln(1+\exp\theta) = -\ln(1-p)$; therefore $b(\theta) = n\ln(1+\exp\theta)$. □

Problem 110. *2 points Show that the Poisson distribution* (5.3.5) *with $t = 1$, i.e.,*

$$\Pr[\mathsf{x}{=}k] = \frac{\lambda^k}{k!}e^{-\lambda} \quad \textit{for } k = 0, 1, \ldots \tag{6.2.6}$$

is a member of the exponential family. Compute the canonical parameter θ and the function $b(\theta)$.

Answer. The probability mass function can be written as

$$\Pr[\mathsf{x}{=}k] = \frac{e^{k\ln\lambda}}{k!}e^{-\lambda} = \exp(k\ln\lambda - \lambda - \ln k!) \quad \text{for } k = 0, 1, \ldots \tag{6.2.7}$$

This is (6.2.3) for the Poisson distribution, where the values of the random variable are called k instead of x, and $\theta = \ln\lambda$. Substituting $\lambda = \exp(\theta)$ in (6.2.7) gives

$$\Pr[\mathsf{x}{=}k] = \exp(k\theta - \exp(\theta) - \ln k!) \quad \text{for } k = 0, 1, \ldots \tag{6.2.8}$$

from which one sees $b(\theta) = \exp(\theta)$.

□

The one-parameter exponential family can be generalized by the inclusion of a scale parameter ϕ in the distribution. This gives the *exponential dispersion family*, see [**MN89**, p. 28]: Each observation has the density function

$$f_{\mathsf{y}}(y;\theta,\phi) = \exp\Big(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\Big). \tag{6.2.9}$$

Problem 111. [**MN89**, p. 28] *Show that the Normal distribution is a member of the exponential dispersion family.*

Answer.

$$f_{\mathsf{y}}(y) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(y-\mu)^2}{2\sigma^2}} = \exp\Big(\big(y\mu - \mu^2/2\big)/\sigma^2 - \frac{1}{2}\big(y^2/\sigma^2 + \log(2\pi\sigma^2)\big)\Big), \tag{6.2.10}$$

i.e., $\theta = \mu$, $\phi = \sigma^2$, $a(\phi) = \phi$, $b(\theta) = \theta^2/2$, $c(y,\phi) = -\frac{1}{2}\big(y^2/\sigma^2 + \log(2\pi\sigma^2)\big)$. □

Problem 112. *Show that the Gamma distribution is a member of the exponential dispersion family.*

Next observation: for the exponential and the exponential dispersion families, the expected value is the derivative of the function $b(\theta)$

$$\mathrm{E}[\mathsf{y}] = \frac{\partial b(\theta)}{\partial\theta}. \tag{6.2.11}$$

This follows from the basic theory associated with maximum likelihood estimation, see (13.4.12). $\mathrm{E}[\mathsf{y}]$ is therefore a function of the "canonical parameter" θ, and in the generalized linear model the assumption is made that this function has an inverse, i.e., the canonical parameter can be written $\theta = g(\mu)$ where g is called the "canonical link function."

Problem 113. *2 points In the case of the Binomial distribution (see Problem 109) compute $b'(\theta)$ and verify that it is the same as* $\mathrm{E}[\mathsf{x}]$.

Answer. $b(\theta) = n\ln(1+\exp\theta)$, therefore $b'(\theta) = n\frac{1}{1+\exp\theta}\exp(\theta)$. Now $\exp(\theta) = \frac{p}{1-p}$; plugging this in gives $b'(\theta) = np$, which is the same as $\mathrm{E}[\mathsf{x}]$. □

Problem 114. *1 point In the case of the Poisson distribution (see Problem 110) compute $b'(\theta)$ and verify that it is the same as* $\mathrm{E}[\mathsf{x}]$*, and compute $b''(\theta)$ and verify that it is the same as* $\mathrm{var}[\mathsf{x}]$*. You are allowed, without proof, that a Poisson distribution with parameter λ has expected value λ and variance λ.*

Answer. $b(\theta) = \exp\theta$, therefore $b'(\theta) = b''(\theta) = \exp(\theta) = \lambda$. □

From (13.4.20) follows furthermore that the variance is the second derivative of b, multiplied by $a(\phi)$:

$$\operatorname{var}[\mathsf{y}] = \frac{\partial^2 b(\theta)}{\partial \theta^2} a(\phi) \tag{6.2.12}$$

Since θ is a function of the mean, this means: the variance of each observation is the product of two factors, the first factor depends on the mean only, it is called the "variance function," and the other factor depends on ϕ. This is exactly the specification of the generalized linear model, see Section 69.3.

CHAPTER 7

Chebyshev Inequality, Weak Law of Large Numbers, and Central Limit Theorem

7.1. Chebyshev Inequality

If the random variable y has finite expected value μ and standard deviation σ, and k is some positive number, then the *Chebyshev Inequality* says

$$\Pr\big[|\mathsf{y}-\mu|\geq k\sigma\big] \leq \frac{1}{k^2}. \tag{7.1.1}$$

In words, the probability that a given random variable y differs from its expected value by more than k standard deviations is less than $1/k^2$. (Here "more than" and "less than" are short forms for "more than or equal to" and "less than or equal to.") One does not need to know the full distribution of y for that, only its expected value and standard deviation. We will give here a proof only if y has a discrete distribution, but the inequality is valid in general. Going over to the standardized variable $\mathsf{z}=\frac{\mathsf{y}-\mu}{\sigma}$ we have to show $\Pr[|\mathsf{z}|\geq k]\leq\frac{1}{k^2}$. Assuming z assumes the values z_1, z_2,... with probabilities $p(z_1)$, $p(z_2)$,..., then

$$\Pr[|\mathsf{z}|\geq k] = \sum_{i:\ |z_i|\geq k} p(z_i). \tag{7.1.2}$$

Now multiply by k^2:

$$k^2\Pr[|\mathsf{z}|\geq k] = \sum_{i:\ |z_i|\geq k} k^2 p(z_i) \tag{7.1.3}$$

$$\leq \sum_{i:\ |z_i|\geq k} z_i^2 p(z_i) \tag{7.1.4}$$

$$\leq \sum_{\text{all i}} z_i^2 p(z_i) = \operatorname{var}[\mathsf{z}] = 1. \tag{7.1.5}$$

The Chebyshev inequality is *sharp* for all $k\geq 1$. Proof: the random variable which takes the value $-k$ with probability $\frac{1}{2k^2}$ and the value $+k$ with probability

$\frac{1}{2k^2}$, and 0 with probability $1-\frac{1}{k^2}$, has expected value 0 and variance 1 and the $\leq$-sign in (7.1.1) becomes an equal sign.

Problem 115. [**HT83**, p. 316] *Let y be the number of successes in n trials of a Bernoulli experiment with success probability p. Show that*

$$\Pr\Big(\Big|\frac{\mathsf{y}}{n}-p\Big|<\varepsilon\Big)\geq 1-\frac{1}{4n\varepsilon^2}. \tag{7.1.6}$$

Hint: first compute what Chebyshev will tell you about the lefthand side, and then you will need still another inequality.

Answer. $\mathrm{E}[\mathsf{y}/n]=p$ and $\operatorname{var}[\mathsf{y}/n]=pq/n$ (where $q=1-p$). Chebyshev says therefore

$$\Pr\Big(\Big|\frac{\mathsf{y}}{n}-p\Big|\geq k\sqrt{\frac{pq}{n}}\Big)\leq\frac{1}{k^2}. \tag{7.1.7}$$

Setting $\varepsilon=k\sqrt{pq/n}$, therefore $1/k^2=pq/n\varepsilon^2$ one can rewerite (7.1.7) as

$$\Pr\Big(\Big|\frac{\mathsf{y}}{n}-p\Big|\geq\varepsilon\Big)\leq\frac{pq}{n\varepsilon^2}. \tag{7.1.8}$$

Now note that $pq\leq 1/4$ whatever their values are. □

Problem 116. *2 points For a standard normal variable, $\Pr[|\mathsf{z}|\geq 1]$ is approximately 1/3, please look up the precise value in a table. What does the Chebyshev inequality says about this probability? Also, $\Pr[|\mathsf{z}|\geq 2]$ is approximately 5%, again look up the precise value. What does Chebyshev say?*

Answer. $\Pr[|\mathsf{z}|\geq 1]=0.3174$, the Chebyshev inequality says that $\Pr[|\mathsf{z}|\geq 1]\leq 1$. Also, $\Pr[|\mathsf{z}|\geq 2]=0.0456$, while Chebyshev says it is ≤ 0.25. □

7.2. The Probability Limit and the Law of Large Numbers

Let $\mathsf{y}_1,\mathsf{y}_2,\mathsf{y}_3,\ldots$ be a sequence of independent random variables all of which have the same expected value μ and variance σ^2. Then $\bar{\mathsf{y}}_n=\frac{1}{n}\sum_{i=1}^n\mathsf{y}_i$ has expected value μ and variance $\frac{\sigma^2}{n}$. I.e., its probability mass is clustered much more closely around the value μ than the individual y_i. To make this statement more precise we need a concept of convergence of random variables. It is not possible to define it in the "obvious" way that the sequence of random variables y_n converges toward y if every realization of them converges, since it is possible, although extremely unlikely, that e.g. all throws of a coin show heads ad infinitum, or follow another sequence for which the average number of heads does not converge towards 1/2. Therefore we will use the following definition:

The sequence of random variables $\mathsf{y}_1,\mathsf{y}_2,\ldots$ *converges in probability* to another random variable y if and only if for every $\delta>0$

$$\lim_{n\to\infty}\Pr\big[|\mathsf{y}_n-\mathsf{y}|\geq\delta\big]=0. \tag{7.2.1}$$

One can also say that the probability limit of y_n is y, in formulas

$$\operatorname*{plim}_{n\to\infty} \mathsf{y}_n = \mathsf{y}. \tag{7.2.2}$$

In many applications, the limiting variable y is a degenerate random variable, i.e., it is a constant.

The *Weak Law of Large Numbers* says that, if the expected value exists, then the probability limit of the sample means of an ever increasing sample is the expected value, i.e., $\operatorname{plim}_{n\to\infty} \bar{\mathsf{y}}_n = \mu$.

PROBLEM 117. *5 points Assuming that not only the expected value but also the variance exists, derive the Weak Law of Large Numbers, which can be written as*

$$\lim_{n\to\infty} \Pr\big[|\bar{\mathsf{y}}_n - \mathrm{E}[\mathsf{y}]|\geq\delta\big] = 0 \textit{ for all } \delta > 0, \tag{7.2.3}$$

from the Chebyshev inequality

$$\Pr[|\mathsf{x} - \mu|\geq k\sigma] \leq \frac{1}{k^2} \qquad \textit{where } \mu = \mathrm{E}[\mathsf{x}] \textit{ and } \sigma^2 = \mathrm{var}[\mathsf{x}] \tag{7.2.4}$$

ANSWER. From nonnegativity of probability and the Chebyshev inequality for $\mathsf{x} = \bar{\mathsf{y}}$ follows $0 \leq \Pr[|\bar{\mathsf{y}} - \mu|\geq \frac{k\sigma}{\sqrt{n}}] \leq \frac{1}{k^2}$ for all k. Set $k = \frac{\delta\sqrt{n}}{\sigma}$ to get $0 \leq \Pr[|\bar{\mathsf{y}}_n - \mu|\geq\delta] \leq \frac{\sigma^2}{n\delta^2}$. For any fixed $\delta > 0$, the upper bound converges towards zero as $n \to \infty$, and the lower bound is zero, therefore the probability iself also converges towards zero. □

PROBLEM 118. *4 points Let $\mathsf{y}_1, \ldots, \mathsf{y}_n$ be a sample from some unknown probability distribution, with sample mean $\bar{\mathsf{y}} = \frac{1}{n}\sum_{i=1}^n \mathsf{y}_i$ and sample variance $\mathsf{s}^2 = \frac{1}{n}\sum_{i=1}^n(\mathsf{y}_i - \bar{\mathsf{y}})^2$. Show that the data satisfy the following "sample equivalent" of the Chebyshev inequality: if k is any fixed positive number, and m is the number of observations y_j which satisfy $|\mathsf{y}_j - \bar{\mathsf{y}}| \geq k\mathsf{s}$, then $\mathsf{m} \leq n/k^2$. In symbols,*

$$\#\{\mathsf{y}_i\colon\ |\mathsf{y}_i - \bar{\mathsf{y}}| \geq k\mathsf{s}\} \leq \frac{n}{k^2}. \tag{7.2.5}$$

Hint: apply the usual Chebyshev inequality to the so-called empirical distribution *of the sample. The empirical distribution is a discrete probability distribution defined by $\Pr[\mathsf{y}{=}\mathsf{y}_i] = k/n$, when the number y_i appears k times in the sample. (If all y_i are different, then all probabilities are $1/n$). The empirical distribution corresponds to the experiment of randomly picking one observation out of the given sample.*

ANSWER. The only thing to note is: the sample mean is the expected value in that empirical distribution, the sample variance is the variance, and the relative number m/n is the probability.

$$\#\{\mathsf{y}_i\colon\ \mathsf{y}_i \in S\} = n\Pr[S] \tag{7.2.6}$$

□

• a. *3 points What happens to this result when the distribution from which the y_i are taken does not have an expected value or a variance?*

ANSWER. The result still holds but $\bar{\mathsf{y}}$ and s^2 do not converge as the number of observations increases. □

7.3. Central Limit Theorem

Assume all y_i are independent and have the same distribution with mean μ, variance σ^2, and also a moment generating function. Again, let $\bar{\mathsf{y}}_n$ be the sample mean of the first n observations. The central limit theorem says that the probability distribution for

$$\frac{\bar{\mathsf{y}}_n - \mu}{\sigma/\sqrt{n}} \tag{7.3.1}$$

converges to a $\mathsf{N}(0, 1)$. This is a different concept of convergence than the probability limit, it is convergence in distribution.

PROBLEM 119. *1 point Construct a sequence of random variables $\mathsf{y}_1, \mathsf{y}_2 \ldots$ with the following property: their cumulative distribution functions converge to the cumulative distribution function of a standard normal, but the random variables themselves do not converge in probability. (This is easy!)*

ANSWER. One example would be: all y_i are independent standard normal variables. □

Why do we have the funny expression $\frac{\bar{\mathsf{y}}_n-\mu}{\sigma/\sqrt{n}}$? Because this is the standardized version of $\bar{\mathsf{y}}_n$. We know from the law of large numbers that the distribution of $\bar{\mathsf{y}}_n$ becomes more and more concentrated around μ. If we standardize the sample averages $\bar{\mathsf{y}}_n$, we compensate for this concentration. The central limit theorem tells us therefore what happens to the *shape* of the cumulative distribution function of $\bar{\mathsf{y}}_n$. If we disregard the fact that it becomes more and more concentrated (by multiplying it by a factor which is chosen such that the variance remains constant), then we see that its geometric shape comes closer and closer to a normal distribution.

Proof of the Central Limit Theorem: By Problem 120,

$$\frac{\bar{\mathsf{y}}_n - \mu}{\sigma/\sqrt{n}} = \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{\mathsf{y}_i - \mu}{\sigma} = \frac{1}{\sqrt{n}}\sum_{i=1}^n \mathsf{z}_i \quad \text{where } \mathsf{z}_i = \frac{\mathsf{y}_i - \mu}{\sigma}. \tag{7.3.2}$$

Let m_3, m_4, etc., be the third, fourth, etc., moments of z_i; then the m.g.f. of z_i is

$$m_{\mathsf{z}_i}(t) = 1 + \frac{t^2}{2!} + \frac{m_3 t^3}{3!} + \frac{m_4 t^4}{4!} + \cdots \tag{7.3.3}$$

Therefore the m.g.f. of $\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathsf{z}_i$ is (multiply and substitute $t/\sqrt{n}$ for t):

$$\left(1 + \frac{t^2}{2!n} + \frac{m_3 t^3}{3!\sqrt{n^3}} + \frac{m_4 t^4}{4!n^2} + \cdots\right)^n = \left(1 + \frac{w_n}{n}\right)^n \tag{7.3.4}$$

where

$$w_n = \frac{t^2}{2!} + \frac{m_3 t^3}{3!\sqrt{n}} + \frac{m_4 t^4}{4! n} + \cdots . \tag{7.3.5}$$

Now use Euler's limit, this time in the form: if $w_n \to w$ for $n \to \infty$, then $\left(1+\frac{w_n}{n}\right)^n \to e^w$. Since our $w_n \to \frac{t^2}{2}$, the m.g.f. of the standardized $\bar{\mathsf{y}}_n$ converges toward $e^{\frac{t^2}{2}}$, which is that of a standard normal distribution.

The Central Limit theorem is an example of emergence: independently of the distributions of the individual summands, the distribution of the sum has a very specific shape, the Gaussian bell curve. The signals turn into white noise. Here emergence is the emergence of homogenity and indeterminacy. In capitalism, much more specific outcomes emerge: whether one quits the job or not, whether one sells the stock or not, whether one gets a divorce or not, the outcome for society is to perpetuate the system. Not many activities don't have this outcome.

PROBLEM 120. *Show in detail that* $\frac{\bar{\mathsf{y}}_n-\mu}{\sigma/\sqrt{n}} = \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{\mathsf{y}_i-\mu}{\sigma}$.

ANSWER. Lhs $= \frac{\sqrt{n}}{\sigma}\left(\left(\frac{1}{n}\sum_{i=1}^n \mathsf{y}_i\right)-\mu\right) = \frac{\sqrt{n}}{\sigma}\left(\left(\frac{1}{n}\sum_{i=1}^n \mathsf{y}_i\right)-\left(\frac{1}{n}\sum_{i=1}^n \mu\right)\right) = \frac{\sqrt{n}}{\sigma}\frac{1}{n}\left(\sum_{i=1}^n \mathsf{y}_i - \mu\right)$ = rhs. □

PROBLEM 121. *3 points Explain verbally clearly what the law of large numbers means, what the Central Limit Theorem means, and what their difference is.*

PROBLEM 122. *(For this problem, a table is needed.)* [**Lar82**, exercise 5.6.1, p. 301] *If you roll a pair of dice 180 times, what is the approximate probability that the sum seven appears 25 or more times? Hint: use the Central Limit Theorem (but don't worry about the continuity correction, which is beyond the scope of this class).*

ANSWER. Let x_i be the random variable that equals one if the i-th roll is a seven, and zero otherwise. Since 7 can be obtained in six ways (1+6, 2+5, 3+4, 4+3, 5+2, 6+1), the probability to get a 7 (which is at the same time the expected value of x_i) is 6/36=1/6. Since $\mathsf{x}_i^2 = \mathsf{x}_i$, $\text{var}[\mathsf{x}_i] = \text{E}[\mathsf{x}_i] - (\text{E}[\mathsf{x}_i])^2 = \frac{1}{6} - \frac{1}{36} = \frac{5}{36}$. Define $\mathsf{x} = \sum_{i=1}^{180} \mathsf{x}_i$. We need $\Pr[\mathsf{x}\geq 25]$. Since x is the sum of many independent identically distributed random variables, the CLT says that x is asympotically normal. Which normal? That which has the same expected value and variance as x. $\text{E}[\mathsf{x}] = 180\cdot(1/6) = 30$ and $\text{var}[\mathsf{x}] = 180\cdot(5/36) = 25$. Therefore define $\mathsf{y} \sim \text{N}(30, 25)$. The CLT says that $\Pr[\mathsf{x}\geq 25] \approx \Pr[\mathsf{y}\geq 25]$. Now $\mathsf{y}\geq 25 \iff \mathsf{y}-30\geq -5 \iff \mathsf{y}-30\leq +5 \iff (\mathsf{y}-30)/5\leq 1$. But $z = (\mathsf{y}-30)/5$ is a standard Normal, therefore $\Pr[(\mathsf{y}-30)/5\leq 1] = F_z(1)$, i.e., the cumulative distribution of the standard Normal evaluated at +1. One can look this up in a table, the probability asked for is .8413. Larson uses the continuity correction: x is discrete, and $\Pr[\mathsf{x}\geq 25] = \Pr[\mathsf{x}>24]$. Therefore $\Pr[\mathsf{y}\geq 25]$ and $\Pr[\mathsf{y}>24]$ are two alternative good approximations; but the best is $\Pr[\mathsf{y}\geq 24.5] = .8643$. This is the continuity correction. □

CHAPTER 8

Vector Random Variables

In this chapter we will look at *two* random variables x and y defined on the same sample space U, i.e.,

$$\mathsf{x}\colon U \ni \omega \mapsto x(\omega) \in \mathbb{R} \qquad \text{and} \qquad \mathsf{y}\colon U \ni \omega \mapsto y(\omega) \in \mathbb{R}. \tag{8.0.6}$$

As we said before, x and y are called independent if all events of the form $\mathsf{x} \le x$ are independent of any event of the form $\mathsf{y} \le y$. But now let us assume they are *not* independent. In this case, we do not have all the information about them if we merely know the distribution of each.

The following example from [**Lar82**, example 5.1.7. on p. 233] illustrates the issues involved. This example involves two random variables that have only two possible outcomes each. Suppose you are told that a coin is to be flipped two times and that the probability of a head is .5 for each flip. This information is not enough to determine the probability of the second flip giving a head conditionally on the first flip giving a head.

For instance, the above two probabilities can be achieved by the following experimental setup: a person has one fair coin and flips it twice in a row. Then the two flips are independent.

But the probabilities of 1/2 for heads and 1/2 for tails can also be achieved as follows: The person has two coins in his or her pocket. One has two heads, and one has two tails. If at random one of these two coins is picked and flipped twice, then the second flip has the same outcome as the first flip.

What do we need to get the full picture? We must consider the two variables not separately but jointly, as a *totality*. In order to do this, we combine x and y into one entity, a vector $\begin{bmatrix}\mathsf{x}\\ \mathsf{y}\end{bmatrix} \in \mathbb{R}^2$. Consequently we need to know the probability measure induced by the mapping $\mathsf{U} \ni \omega \mapsto \begin{bmatrix}x(\omega)\\ y(\omega)\end{bmatrix} \in \mathbb{R}^2$.

It is not sufficient to look at random variables individually; one must look at them as a totality.

Therefore let us first get an overview over all possible probability measures on the plane $\mathbb{R}^2$. In strict analogy with the one-dimensional case, these probability measures

can be represented by the joint cumulative distribution function. It is defined as

$$F_{\mathsf{x},\mathsf{y}}(x,y) = \Pr[\begin{bmatrix}\mathsf{x}\\ \mathsf{y}\end{bmatrix} \le \begin{bmatrix}x\\ y\end{bmatrix}] = \Pr[\mathsf{x} \le x \text{ and } \mathsf{y} \le y]. \tag{8.0.7}$$

For discrete random variables, for which the cumulative distribution function is a step function, the joint probability mass function provides the same information:

$$p_{\mathsf{x},\mathsf{y}}(x,y) = \Pr[\begin{bmatrix}\mathsf{x}\\ \mathsf{y}\end{bmatrix} = \begin{bmatrix}x\\ y\end{bmatrix}] = \Pr[\mathsf{x}{=}x \text{ and } \mathsf{y}{=}y]. \tag{8.0.8}$$

PROBLEM 123. *Write down the joint probability mass functions for the two versions of the two coin flips discussed above.*

ANSWER. Here are the probability mass functions for these two cases:

(8.0.9)

		Second Flip		
		H	T	sum
First Flip	H	.25	.25	.50
	T	.25	.25	.50
	sum	.50	.50	1.00

		Second Flip		
		H	T	sum
First Flip	H	.50	.00	.50
	T	.00	.50	.50
	sum	.50	.50	1.00

□

The most important case is that with a differentiable cumulative distribution function. Then the joint density function $f_{\mathsf{x},\mathsf{y}}(x,y)$ can be used to define the probability measure. One obtains it from the cumulative distribution function by taking derivatives:

$$f_{\mathsf{x},\mathsf{y}}(x,y) = \frac{\partial^2}{\partial x\,\partial y} F_{\mathsf{x},\mathsf{y}}(x,y). \tag{8.0.10}$$

Probabilities can be obtained back from the density function either by the integral condition, or by the infinitesimal condition. I.e., either one says for a subset $B \subset \mathbb{R}^2$:

$$\Pr[\begin{bmatrix}\mathsf{x}\\ \mathsf{y}\end{bmatrix} \in B] = \int\int_B f(x,y)\,dx\,dy, \tag{8.0.11}$$

or one says, for a infinitesimal two-dimensional volume element $dV_{x,y}$ located at $\left[\begin{smallmatrix}x\\ y\end{smallmatrix}\right]$, which has the two-dimensional volume (i.e., area) $|dV|$,

$$\Pr[\begin{bmatrix}\mathsf{x}\\ \mathsf{y}\end{bmatrix} \in dV_{x,y}] = f(x,y)\,|dV|. \tag{8.0.12}$$

The vertical bars here do not mean the absolute value but the volume of the argument inside.

8.1. Expected Value, Variances, Covariances

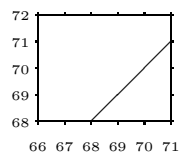
To get the expected value of a function of x and y, one simply has to put this function together with the density function into the integral, i.e., the formula is

$$\mathrm{E}[g(\mathsf{x},\mathsf{y})] = \int\int_{\mathbb{R}^2} g(x,y) f_{\mathsf{x},\mathsf{y}}(x,y)\,dx\,dy. \tag{8.1.1}$$

Problem 124. *Assume there are two transportation choices available: bus and car. If you pick at random a neoclassical individual ω and ask which utility this person derives from using bus or car, the answer will be two numbers that can be written as a vector $\begin{bmatrix} u(\omega) \\ v(\omega) \end{bmatrix}$ (u for bus and v for car).*

• a. *3 points Assuming $\begin{bmatrix} \mathsf{u} \\ \mathsf{v} \end{bmatrix}$ has a uniform density in the rectangle with corners $\begin{bmatrix} 66 \\ 68 \end{bmatrix}$, $\begin{bmatrix} 66 \\ 72 \end{bmatrix}$, $\begin{bmatrix} 71 \\ 68 \end{bmatrix}$, and $\begin{bmatrix} 71 \\ 72 \end{bmatrix}$, compute the probability that the bus will be preferred.*

Answer. The probability is 9/40. u and v have a joint density function that is uniform in the rectangle below and zero outside (u, the preference for buses, is on the horizontal, and v, the preference for cars, on the vertical axis). The probability is the fraction of this rectangle below the diagonal.



□

• b. *2 points How would you criticize an econometric study which argued along the above lines?*

Answer. The preferences are not for a bus or a car, but for a whole transportation systems. And these preferences are not formed independently and individualistically, but they depend on which other infrastructures are in place, whether there is suburban sprawl or concentrated walkable cities, etc. This is again the error of detotalization (which favors the status quo).

□

Jointly distributed random variables should be written as random *vectors*. Instead of $\begin{bmatrix} \mathsf{y} \\ \mathsf{z} \end{bmatrix}$ we will also write $\boldsymbol{\mathsf{x}}$ (bold face). Vectors are always considered to be column vectors. The expected value of a random vector is a vector of constants,

notation

$$\mathcal{E}[\boldsymbol{\mathsf{x}}] = \begin{bmatrix} \mathrm{E}[\mathsf{x}_1] \\ \vdots \\ \mathrm{E}[\mathsf{x}_n] \end{bmatrix} \tag{8.1.2}$$

For two random variables x and y, their *covariance* is defined as

$$\mathrm{cov}[\mathsf{x},\mathsf{y}] = \mathrm{E}\big[(\mathsf{x}-\mathrm{E}[\mathsf{x}])(\mathsf{y}-\mathrm{E}[\mathsf{y}])\big] \tag{8.1.3}$$

Computation rules with covariances are

$$\mathrm{cov}[\mathsf{x},\mathsf{z}] = \mathrm{cov}[\mathsf{z},\mathsf{x}] \quad \mathrm{cov}[\mathsf{x},\mathsf{x}] = \mathrm{var}[\mathsf{x}] \quad \mathrm{cov}[\mathsf{x},\alpha] = 0 \tag{8.1.4}$$

$$\mathrm{cov}[\mathsf{x}+\mathsf{y},\mathsf{z}] = \mathrm{cov}[\mathsf{x},\mathsf{z}] + \mathrm{cov}[\mathsf{y},\mathsf{z}] \quad \mathrm{cov}[\alpha\mathsf{x},\mathsf{y}] = \alpha\,\mathrm{cov}[\mathsf{x},\mathsf{y}] \tag{8.1.5}$$

Problem 125. *3 points Using definition* (8.1.3) *prove the following formula:*

$$\mathrm{cov}[\mathsf{x},\mathsf{y}] = \mathrm{E}[\mathsf{x}\mathsf{y}] - \mathrm{E}[\mathsf{x}]\,\mathrm{E}[\mathsf{y}]. \tag{8.1.6}$$

Write it down carefully, you will lose points for unbalanced or missing parantheses and brackets.

Answer. Here it is side by side with and without the notation $\mathrm{E}[\mathsf{x}] = \mu$ and $\mathrm{E}[\mathsf{y}] = \nu$:

$$\begin{aligned} \mathrm{cov}[\mathsf{x},\mathsf{y}] &= \mathrm{E}\big[(\mathsf{x}-\mathrm{E}[\mathsf{x}])(\mathsf{y}-\mathrm{E}[\mathsf{y}])\big] \\ &= \mathrm{E}\big[\mathsf{x}\mathsf{y} - \mathsf{x}\,\mathrm{E}[\mathsf{y}] - \mathrm{E}[\mathsf{x}]\mathsf{y} + \mathrm{E}[\mathsf{x}]\,\mathrm{E}[\mathsf{y}]\big] \\ &= \mathrm{E}[\mathsf{x}\mathsf{y}] - \mathrm{E}[\mathsf{x}]\,\mathrm{E}[\mathsf{y}] - \mathrm{E}[\mathsf{x}]\,\mathrm{E}[\mathsf{y}] + \mathrm{E}[\mathsf{x}]\,\mathrm{E}[\mathsf{y}] \\ &= \mathrm{E}[\mathsf{x}\mathsf{y}] - \mathrm{E}[\mathsf{x}]\,\mathrm{E}[\mathsf{y}]. \end{aligned} \qquad \begin{aligned} \mathrm{cov}[\mathsf{x},\mathsf{y}] &= \mathrm{E}[(\mathsf{x}-\mu)(\mathsf{y}-\nu)] \\ &= \mathrm{E}[\mathsf{x}\mathsf{y} - \mathsf{x}\nu - \mu\mathsf{y} + \mu\nu] \\ &= \mathrm{E}[\mathsf{x}\mathsf{y}] - \mu\nu - \mu\nu + \mu\nu \\ &= \mathrm{E}[\mathsf{x}\mathsf{y}] - \mu\nu. \end{aligned} \tag{8.1.7}$$

□

Problem 126. *1 point Using* (8.1.6) *prove the five computation rules with covariances* (8.1.4) *and* (8.1.5).

Problem 127. *Using the computation rules with covariances, show that*

$$\mathrm{var}[\mathsf{x}+\mathsf{y}] = \mathrm{var}[\mathsf{x}] + 2\,\mathrm{cov}[\mathsf{x},\mathsf{y}] + \mathrm{var}[\mathsf{y}]. \tag{8.1.8}$$

If one deals with random vectors, the expected value becomes a vector, and the variance becomes a matrix, which is called *dispersion matrix* or *variance-covariance matrix* or simply *covariance matrix*. We will write it $\mathcal{V}[\boldsymbol{\mathsf{x}}]$. Its formal definition is

$$\mathcal{V}[\boldsymbol{\mathsf{x}}] = \mathcal{E}\big[(\boldsymbol{\mathsf{x}}-\mathcal{E}[\boldsymbol{\mathsf{x}}])(\boldsymbol{\mathsf{x}}-\mathcal{E}[\boldsymbol{\mathsf{x}}])^\top\big], \tag{8.1.9}$$

but we can look at it simply as the matrix of all variances and covariances, for example

$$\mathcal{V}[\begin{bmatrix} \mathsf{x} \\ \mathsf{y} \end{bmatrix}] = \begin{bmatrix} \mathrm{var}[\mathsf{x}] & \mathrm{cov}[\mathsf{x},\mathsf{y}] \\ \mathrm{cov}[\mathsf{y},\mathsf{x}] & \mathrm{var}[\mathsf{y}] \end{bmatrix}. \tag{8.1.10}$$

An important computation rule for the covariance matrix is

$$\mathcal{V}[\mathbf{x}] = \boldsymbol{\Psi} \Rightarrow \mathcal{V}[\boldsymbol{A}\mathbf{x}] = \boldsymbol{A}\boldsymbol{\Psi}\boldsymbol{A}^\top. \tag{8.1.11}$$

PROBLEM 128. *4 points Let* $\mathbf{x} = \begin{bmatrix} \mathsf{y} \\ \mathsf{z} \end{bmatrix}$ *be a vector consisting of two random variables, with covariance matrix* $\mathcal{V}[\mathbf{x}] = \boldsymbol{\Psi}$*, and let* $\boldsymbol{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ *be an arbitrary* 2×2 *matrix. Prove that*

$$\mathcal{V}[\boldsymbol{A}\mathbf{x}] = \boldsymbol{A}\boldsymbol{\Psi}\boldsymbol{A}^\top. \tag{8.1.12}$$

Hint: You need to multiply matrices, and to use the following computation rules for covariances:

(8.1.13)

$$\operatorname{cov}[\mathsf{x}+\mathsf{y}, \mathsf{z}] = \operatorname{cov}[\mathsf{x}, \mathsf{z}] + \operatorname{cov}[\mathsf{y}, \mathsf{z}] \quad \operatorname{cov}[\alpha\mathsf{x}, \mathsf{y}] = \alpha \operatorname{cov}[\mathsf{x}, \mathsf{y}] \quad \operatorname{cov}[\mathsf{x}, \mathsf{x}] = \operatorname{var}[\mathsf{x}].$$

ANSWER. $\mathcal{V}[\boldsymbol{A}\mathbf{x}] =$

$$\mathcal{V}[\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} \mathsf{y} \\ \mathsf{z} \end{bmatrix}\right)] = \mathcal{V}[\begin{bmatrix} a\mathsf{y}+b\mathsf{z} \\ c\mathsf{y}+d\mathsf{z} \end{bmatrix}] = \begin{bmatrix} \operatorname{var}[a\mathsf{y}+b\mathsf{z}] & \operatorname{cov}[a\mathsf{y}+b\mathsf{z}, c\mathsf{y}+d\mathsf{z}] \\ \operatorname{cov}[c\mathsf{y}+d\mathsf{z}, a\mathsf{y}+b\mathsf{z}] & \operatorname{var}[c\mathsf{y}+d\mathsf{z}] \end{bmatrix}$$

On the other hand, $\boldsymbol{A}\boldsymbol{\Psi}\boldsymbol{A}^\top =$

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} \operatorname{var}[\mathsf{y}] & \operatorname{cov}[\mathsf{y}, \mathsf{z}] \\ \operatorname{cov}[\mathsf{y}, \mathsf{z}] & \operatorname{var}[\mathsf{z}] \end{bmatrix}\begin{bmatrix} a & c \\ b & d \end{bmatrix} = \begin{bmatrix} a\operatorname{var}[\mathsf{y}] + b\operatorname{cov}[\mathsf{y}, \mathsf{z}] & a\operatorname{cov}[\mathsf{y}, \mathsf{z}] + b\operatorname{var}[\mathsf{z}] \\ c\operatorname{var}[\mathsf{y}] + d\operatorname{cov}[\mathsf{y}, \mathsf{z}] & c\operatorname{cov}[\mathsf{y}, \mathsf{z}] + d\operatorname{var}[\mathsf{z}] \end{bmatrix}\begin{bmatrix} a & c \\ b & d \end{bmatrix}$$

Multiply out and show that it is the same thing. □

Since the variances are nonnegative, one can see from equation (8.1.11) that covariance matrices are *nonnegative definite* (which is in econometrics is often also called *positive semidefinite*). By definition, a symmetric matrix $\boldsymbol{\Sigma}$ is nonnegative definite if for all vectors $\boldsymbol{a}$ follows $\boldsymbol{a}^\top\boldsymbol{\Sigma}\boldsymbol{a} \geq 0$. It is *positive definite* if it is nonnegativbe definite, and $\boldsymbol{a}^\top\boldsymbol{\Sigma}\boldsymbol{a} = 0$ holds only if $\boldsymbol{a} = \boldsymbol{o}$.

PROBLEM 129. *1 point A symmetric matrix* $\boldsymbol{\Omega}$ *is nonnegative definite if and only if* $\boldsymbol{a}^\top\boldsymbol{\Omega}\boldsymbol{a} \geq 0$ *for every vector* $\boldsymbol{a}$*. Using this criterion, show that if* $\boldsymbol{\Sigma}$ *is symmetric and nonnegative definite, and if* $\boldsymbol{R}$ *is an arbitrary matrix, then* $\boldsymbol{R}^\top\boldsymbol{\Sigma}\boldsymbol{R}$ *is also nonnegative definite.*

One can also define a covariance matrix between *different* vectors, $\mathcal{C}[\mathbf{x}, \mathbf{y}]$; its i, j element is $\operatorname{cov}[\mathsf{x}_i, \mathsf{y}_j]$.

The *correlation coefficient* of two scalar random variables is defined as

$$\operatorname{corr}[\mathsf{x}, \mathsf{y}] = \frac{\operatorname{cov}[\mathsf{x}, \mathsf{y}]}{\sqrt{\operatorname{var}[\mathsf{x}]\operatorname{var}[\mathsf{y}]}}. \tag{8.1.14}$$

The advantage of the correlation coefficient over the covariance is that it is always between -1 and $+1$. This follows from the Cauchy-Schwartz inequality

$$(\operatorname{cov}[\mathsf{x}, \mathsf{y}])^2 \leq \operatorname{var}[\mathsf{x}]\operatorname{var}[\mathsf{y}]. \tag{8.1.15}$$

PROBLEM 130. *4 points Given two random variables* y *and* z *with* $\operatorname{var}[\mathsf{y}] \neq 0$*, compute that constant* a *for which* $\operatorname{var}[a\mathsf{y} - \mathsf{z}]$ *is the minimum. Then derive the Cauchy-Schwartz inequality from the fact that the minimum variance is nonnegative.*

ANSWER.

$$\operatorname{var}[a\mathsf{y} - \mathsf{z}] = a^2\operatorname{var}[\mathsf{y}] - 2a\operatorname{cov}[\mathsf{y}, \mathsf{z}] + \operatorname{var}[\mathsf{z}] \tag{8.1.16}$$

$$\text{First order condition:} \quad 0 = 2a\operatorname{var}[\mathsf{y}] - 2\operatorname{cov}[\mathsf{y}, \mathsf{z}] \tag{8.1.17}$$

Therefore the minimum value is $a^* = \operatorname{cov}[\mathsf{y}, \mathsf{z}]/\operatorname{var}[\mathsf{y}]$, for which the cross product term is -2 times the first item:

$$0 \leq \operatorname{var}[a^*\mathsf{y} - \mathsf{z}] = \frac{(\operatorname{cov}[\mathsf{y}, \mathsf{z}])^2}{\operatorname{var}[\mathsf{y}]} - \frac{2(\operatorname{cov}[\mathsf{y}, \mathsf{z}])^2}{\operatorname{var}[\mathsf{y}]} + \operatorname{var}[\mathsf{z}] \tag{8.1.18}$$

$$0 \leq -(\operatorname{cov}[\mathsf{y}, \mathsf{z}])^2 + \operatorname{var}[\mathsf{y}]\operatorname{var}[\mathsf{z}]. \tag{8.1.19}$$

This proves (8.1.15) for the case $\operatorname{var}[\mathsf{y}] \neq 0$. If $\operatorname{var}[\mathsf{y}] = 0$, then y is a constant, therefore $\operatorname{cov}[\mathsf{y}, \mathsf{z}] = 0$ and (8.1.15) holds trivially.

□

8.2. Marginal Probability Laws

The *marginal* probability distribution of x (or y) is simply the probability distribution of x (or y). The word "marginal" merely indicates that it is derived from the joint probability distribution of x and y.

If the probability distribution is characterized by a probability mass function, we can compute the marginal probability mass functions by writing down the joint probability mass function in a rectangular scheme and summing up the rows or columns:

$$p_{\mathsf{x}}(x) = \sum_{y:\, p(x,y) \neq 0} p_{\mathsf{x},\mathsf{y}}(x, y). \tag{8.2.1}$$

For density functions, the following argument can be given:

$$\Pr[\mathsf{x} \in dV_x] = \Pr[\begin{bmatrix} \mathsf{x} \\ \mathsf{y} \end{bmatrix} \in dV_x \times \mathbb{R}]. \tag{8.2.2}$$

By the definition of a product set: $\begin{bmatrix}\mathsf{x}\\ \mathsf{y}\end{bmatrix} \in A \times B \Leftrightarrow \mathsf{x} \in A$ and $\mathsf{y} \in B$. Split $\mathbb{R}$ into many small disjoint intervals, $\mathbb{R} = \bigcup_i dV_{y_i}$, then

$$\Pr[\mathsf{x} \in dV_x] = \sum_i \Pr\Big[\begin{bmatrix}\mathsf{x}\\ \mathsf{y}\end{bmatrix} \in dV_x \times dV_{y_i}\Big] \tag{8.2.3}$$

$$= \sum_i f_{\mathsf{x},\mathsf{y}}(x, y_i)|dV_x||dV_{y_i}| \tag{8.2.4}$$

$$= |dV_x| \sum_i f_{\mathsf{x},\mathsf{y}}(x, y_i)|dV_{y_i}|. \tag{8.2.5}$$

Therefore $\sum_i f_{\mathsf{x},\mathsf{y}}(x, y)|dV_{y_i}|$ is the density function we are looking for. Now the $|dV_{y_i}|$ are usually written as dy, and the sum is usually written as an integral (i.e., an infinite sum each summand of which is infinitesimal), therefore we get

$$f_{\mathsf{x}}(x) = \int_{y=-\infty}^{y=+\infty} f_{\mathsf{x},\mathsf{y}}(x, y)\, dy. \tag{8.2.6}$$

In other words, one has to "integrate out" the variable which one is not interested in.

8.3. Conditional Probability Distribution and Conditional Mean

The conditional probability distribution of y given $\mathsf{x}{=}x$ is the probability distribution of y if we count only those experiments in which the outcome of x is x. If the distribution is defined by a probability mass function, then this is no problem:

$$p_{\mathsf{y}|\mathsf{x}}(y, x) = \Pr[\mathsf{y}{=}y|\mathsf{x}{=}x] = \frac{\Pr[\mathsf{y}{=}y \text{ and } \mathsf{x}{=}x]}{\Pr[\mathsf{x}{=}x]} = \frac{p_{\mathsf{x},\mathsf{y}}(x, y)}{p_{\mathsf{x}}(x)}. \tag{8.3.1}$$

For a density function there is the problem that $\Pr[\mathsf{x}{=}x] = 0$, i.e., the conditional probability is strictly speaking not defined. Therefore take an infinitesimal volume element dV_x located at x and condition on $\mathsf{x} \in dV_x$:

$$\Pr[\mathsf{y} \in dV_y|\mathsf{x} \in dV_x] = \frac{\Pr[\mathsf{y} \in dV_y \text{ and } \mathsf{x} \in dV_x]}{\Pr[\mathsf{x} \in dV_x]} \tag{8.3.2}$$

$$= \frac{f_{\mathsf{x},\mathsf{y}}(x, y)|dV_x||dV_y|}{f_{\mathsf{x}}(x)|dV_x|} \tag{8.3.3}$$

$$= \frac{f_{\mathsf{x},\mathsf{y}}(x, y)}{f_{\mathsf{x}}(x)}|dV_y|. \tag{8.3.4}$$

This no longer depends on dV_x, only on its location x. The conditional density is therefore

$$f_{\mathsf{y}|\mathsf{x}}(y, x) = \frac{f_{\mathsf{x},\mathsf{y}}(x, y)}{f_{\mathsf{x}}(x)}. \tag{8.3.5}$$

As y varies, the conditional density is proportional to the joint density function, but for every given value of x the joint density is multiplied by an appropriate factor so that its integral with respect to y is 1. From (8.3.5) follows also that the joint density function is the product of the conditional times the marginal density functions.

Problem 131. *2 points The conditional density is the joint divided by the marginal:*

$$f_{\mathsf{y}|\mathsf{x}}(y, x) = \frac{f_{\mathsf{x},\mathsf{y}}(x, y)}{f_{\mathsf{x}}(x)}. \tag{8.3.6}$$

Show that this density integrates out to 1.

Answer. The conditional is a density in y with x as parameter. Therefore its integral with respect to y must be $= 1$. Indeed,

$$\int_{y=-\infty}^{+\infty} f_{\mathsf{y}|\mathsf{x}=x}(y, x)\, dy = \frac{\int_{y=-\infty}^{+\infty} f_{\mathsf{x},\mathsf{y}}(x, y)\, dy}{f_{\mathsf{x}}(x)} = \frac{f_{\mathsf{x}}(x)}{f_{\mathsf{x}}(x)} = 1 \tag{8.3.7}$$

because of the formula for the marginal:

$$f_{\mathsf{x}}(x) = \int_{y=-\infty}^{+\infty} f_{\mathsf{x},\mathsf{y}}(x, y)\, dy \tag{8.3.8}$$

You see that formula (8.3.6) divides the joint density exactly by the right number which makes the integral equal to 1. □

Problem 132. [**BD77**, example 1.1.4 on p. 7]. x *and* y *are two independent random variables uniformly distributed over* $[0, 1]$. *Define* $\mathsf{u} = \min(\mathsf{x}, \mathsf{y})$ *and* $\mathsf{v} = \max(\mathsf{x}, \mathsf{y})$.

• a. *Draw in the* x, y *plane the event* $\{\max(\mathsf{x}, \mathsf{y}) \leq 0.5$ *and* $\min(\mathsf{x}, \mathsf{y}) > 0.4\}$ *and compute its probability.*

Answer. The event is the square between 0.4 and 0.5, and its probability is 0.01. □

• b. *Compute the probability of the event* $\{\max(\mathsf{x}, \mathsf{y}) \leq 0.5$ *and* $\min(\mathsf{x}, \mathsf{y}) \leq 0.4\}$.

Answer. It is $\Pr[\max(\mathsf{x}, \mathsf{y}) \leq 0.5] - \Pr[\max(\mathsf{x}, \mathsf{y}) \leq 0.5 \text{ and } \min(\mathsf{x}, \mathsf{y}) > 0.4]$, i.e., the area of the square from 0 to 0.5 minus the square we just had, i.e., 0.24. □

• c. *Compute* $\Pr[\max(\mathsf{x}, \mathsf{y}) \leq 0.5| \min(\mathsf{x}, \mathsf{y}) \leq 0.4]$.

Answer.

$$\frac{\Pr[\max(\mathsf{x},\mathsf{y}) \leq 0.5 \text{ and } \min(\mathsf{x},\mathsf{y}) \leq 0.4]}{\Pr[\min(\mathsf{x},\mathsf{y}) \leq 0.4]} = \frac{0.24}{1-0.36} = \frac{0.24}{0.64} = \frac{3}{8}. \tag{8.3.9}$$

□

• d. *Compute the joint cumulative distribution function of* u *and* v.

Answer. One good way is to do it geometrically: for arbitrary $0 \leq u, v \leq 1$ draw the area $\{\mathsf{u} \leq u \text{ and } \mathsf{v} \leq v\}$ and then derive its size. If $u \leq v$ then $\Pr[\mathsf{u} \leq u \text{ and } \mathsf{v} \leq v] = \Pr[\mathsf{v} \leq v] - \Pr[\mathsf{u} \leq u \text{ and } \mathsf{v} > v] = v^2 - (v-u)^2 = 2uv - u^2$. If $u \geq v$ then $\Pr[\mathsf{u} \leq u \text{ and } \mathsf{v} \leq v] = \Pr[\mathsf{v} \leq v] = v^2$. □

• e. *Compute the joint density function of* u *and* v. *Note: this joint density is discontinuous. The values at the breakpoints themselves do not matter, but it is very important to give the limits within this is a nontrivial function and where it is zero.*

Answer. One can see from the way the cumulative distribution function was constructed that the density function must be

$$f_{\mathsf{u},\mathsf{v}}(u,v) = \begin{cases} 2 & \text{if } 0 \leq u \leq v \leq 1 \\ 0 & \text{otherwise} \end{cases} \tag{8.3.10}$$

I.e., it is uniform in the above-diagonal part of the square. This is also what one gets from differentiating $2vu - u^2$ once with respect to u and once with respect to v. □

• f. *Compute the marginal density function of* u.

Answer. Integrate v out: the marginal density of u is

$$f_{\mathsf{u}}(u) = \int_{v=u}^{1} 2\,dv = 2v\Big|_u^1 = 2 - 2u \quad \text{if } 0 \leq u \leq 1, \qquad \text{and 0 otherwise.} \tag{8.3.11}$$

□

• g. *Compute the conditional density of* v *given* $\mathsf{u} = u$.

Answer. Conditional density is easy to get too; it is the joint divided by the marginal, i.e., it is uniform:

$$f_{\mathsf{v}|\mathsf{u}=u}(v) = \begin{cases} \frac{1}{1-u} & \text{for } 0 \leq u \leq v \leq 1 \\ 0 & \text{otherwise.} \end{cases} \tag{8.3.12}$$

□

8.4. The Multinomial Distribution

Assume you have an experiment with r different possible outcomes, with outcome i having probability p_i $(i = 1, \ldots, r)$. You are repeating the experiment n different times, and you count how many times the ith outcome occurred. Therefore you get

a random vector with r different components x_i, indicating how often the ith event occurred. The probability to get the frequencies $x_1, \ldots, x_r$ is

$$\Pr[\mathsf{x}_1 = x_1, \ldots, \mathsf{x}_r = x_r] = \frac{m!}{x_1! \cdots x_r!} p_1^{x_1} p_2^{x_2} \cdots p_r^{x_r} \tag{8.4.1}$$

This can be explained as follows: The probability that the first x_1 experiments yield outcome 1, the next x_2 outcome 2, etc., is $p_1^{x_1} p_2^{x_2} \cdots p_r^{x_r}$. Now every other sequence of experiments which yields the same number of outcomes of the different categories is simply a permutation of this. But multiplying this probability by $n!$ may count certain sequences of outcomes more than once. Therefore we have to divide by the number of permutations of the whole n element set which yield the same original sequence. This is $x_1! \cdots x_r!$, because this must be a permutation which permutes the first x_1 elements amongst themselves, etc. Therefore the relevant count of permutations is $\frac{n!}{x_1! \cdots x_r!}$.

Problem 133. *You have an experiment with r different outcomes, the ith outcome occurring with probability p_i. You make n independent trials, and the ith outcome occurred x_i times. The joint distribution of the $\mathsf{x}_1, \ldots, \mathsf{x}_r$ is called a multinomial distribution with parameters n and $p_1, \ldots, p_r$.*

• a. *3 points Prove that their mean vector and covariance matrix are*

$$\boldsymbol{\mu} = \mathcal{E}\left[\begin{bmatrix} \mathsf{x}_1 \\ \vdots \\ \mathsf{x}_r \end{bmatrix}\right] = n \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_r \end{bmatrix} \quad \textit{and} \quad \boldsymbol{\Psi} = \mathcal{V}\left[\begin{bmatrix} \mathsf{x}_1 \\ \vdots \\ \mathsf{x}_r \end{bmatrix}\right] = n \begin{bmatrix} p_1 - p_1^2 & -p_1p_2 & \cdots & -p_1p_r \\ -p_2p_1 & p_2 - p_2^2 & \cdots & -p_2p_r \\ \vdots & \vdots & \ddots & \vdots \\ -p_rp_1 & -p_rp_2 & \cdots & p_r - p_r^2 \end{bmatrix}. \tag{8.4.2}$$

Hint: use the fact that the multinomial distribution with parameters n and $p_1, \ldots, p_r$ is the independent sum of n multinomial distributions with parameters 1 and $p_1, \ldots, p_r$.

Answer. In one trial, $\mathsf{x}_i^2 = \mathsf{x}_i$, from which follows the formula for the variance, and for $i \neq j$, $\mathsf{x}_i\mathsf{x}_j = 0$, since only one of them can occur. Therefore $\operatorname{cov}[\mathsf{x}_i, \mathsf{x}_j] = 0 - \operatorname{E}[\mathsf{x}_i]\operatorname{E}[\mathsf{x}_j]$. For several independent trials, just add this. □

• b. *1 point How can you show that this covariance matrix is singular?*

Answer. Since $\mathsf{x}_1 + \cdots + \mathsf{x}_r = n$ with zero variance, we should expect

$$n \begin{bmatrix} p_1 - p_1^2 & -p_1p_2 & \cdots & -p_1p_r \\ -p_2p_1 & p_2 - p_2^2 & \cdots & -p_2p_r \\ \vdots & \vdots & \ddots & \vdots \\ -p_rp_1 & -p_rp_2 & \cdots & p_r - p_r^2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{8.4.3}$$

□

8.5. Independent Random Vectors

The same definition of independence, which we already encountered with scalar random variables, also applies to vector random variables: the vector random variables $\mathbf{x} : U \to \mathbb{R}^m$ and $\mathbf{y} : U \to \mathbb{R}^n$ are called independent if all events that can be defined in terms of $\mathbf{x}$ are independent of all events that can be defined in terms of $\mathbf{y}$, i.e., all events of the form $\{\boldsymbol{x}(\omega) \in C\}$ are independent of all events of the form $\{\boldsymbol{y}(\omega) \in D\}$ with arbitrary (measurable) subsets $C \subset \mathbb{R}^m$ and $D \subset \mathbb{R}^n$.

For this it is sufficient that for all $\boldsymbol{x} \in \mathbb{R}^m$ and $\boldsymbol{y} \in \mathbb{R}^n$, the event $\{\mathbf{x} \leq \boldsymbol{x}\}$ is independent of the event $\{\mathbf{y} \leq \boldsymbol{y}\}$, i.e., that the joint cumulative distribution function is the product of the marginal ones.

Since the joint cumulative distribution function of independent variables is equal to the product of the univariate cumulative distribution functions, the same is true for the joint density function and the joint probability mass function.

Only under this strong definition of independence is it true that any functions of independent random variables are independent.

Problem 134. *4 points Prove that, if* x *and* y *are independent, then* $\mathrm{E}[\mathsf{xy}] = \mathrm{E}[\mathsf{x}]\,\mathrm{E}[\mathsf{y}]$ *and therefore* $\mathrm{cov}[\mathsf{x},\mathsf{y}] = 0$. *(You may assume* x *and* y *have density functions). Give a counterexample where the covariance is zero but the variables are nevertheless dependent.*

Answer. Just use that the joint density function is the product of the marginals. It can also be done as follows: $\mathrm{E}[\mathsf{xy}] = \mathrm{E}\big[\mathrm{E}[\mathsf{xy}|\mathsf{x}]\big] = \mathrm{E}\big[\mathsf{x}\,\mathrm{E}[\mathsf{y}|\mathsf{x}]\big] =$ now independence is needed $= \mathrm{E}\big[\mathsf{x}\,\mathrm{E}[\mathsf{y}]\big] = \mathrm{E}[\mathsf{x}]\,\mathrm{E}[\mathsf{y}]$. A counterexample is given in Problem 150. □

Problem 135. *3 points Prove the following: If the scalar random variables* x *and* y *are indicator variables (i.e., if each of them can only assume the values 0 and 1), and if* $\mathrm{cov}[\mathsf{x},\mathsf{y}] = 0$, *then* x *and* y *are independent. (I.e., in this respect indicator variables have similar properties as jointly normal random variables.)*

Answer. Define the events $\mathsf{A} = \{\omega \in U : x(\omega) = 1\}$ and $\mathsf{B} = \{\omega \in U : y(\omega) = 1\}$, i.e., $\mathsf{x} = \mathsf{i}_\mathsf{A}$ (the indicator variable of the event A) and $\mathsf{y} = \mathsf{i}_\mathsf{B}$. Then $\mathsf{xy} = \mathsf{i}_{\mathsf{A}\cap\mathsf{B}}$. If $\mathrm{cov}[\mathsf{x},\mathsf{y}] = \mathrm{E}[\mathsf{xy}] - \mathrm{E}[\mathsf{x}]\,\mathrm{E}[\mathsf{y}] = \Pr[\mathsf{A}\cap\mathsf{B}] - \Pr[\mathsf{A}]\Pr[\mathsf{B}] = 0$, then A and B are independent. □

Problem 136. *If the vector random variables* $\mathbf{x}$ *and* $\mathbf{y}$ *have the property that* x_i *is independent of every* y_j *for all* i *and* j, *does that make* $\mathbf{x}$ *and* $\mathbf{y}$ *independent random vectors? Interestingly, the answer is no. Give a counterexample that this fact does not even hold for indicator variables. I.e., construct two random vectors* $\mathbf{x}$ *and* $\mathbf{y}$, *consisting of indicator variables, with the property that each component of* $\mathbf{x}$ *is independent of each component of* $\mathbf{y}$, *but* $\mathbf{x}$ *and* $\mathbf{y}$ *are not independent as vector random variables. Hint: Such an example can be constructed in the simplest possible case that* $\mathbf{x}$ *has two components and* $\mathbf{y}$ *has one component; i.e., you merely have to find three indicator variables* x_1, x_2, *and* y *with the property that* x_1 *is independent of* y, *and* x_2 *is independent of* y, *but the vector* $\begin{bmatrix}\mathsf{x}_1\\ \mathsf{x}_2\end{bmatrix}$ *is not independent of* y. *For these three variables, you should use three events which are pairwise independent but not mutually independent.*

Answer. Go back to throwing a coin twice independently and define $\mathsf{A} = \{HH, HT\}$; $\mathsf{B} = \{TH, HH\}$, and $\mathsf{C} = \{HH, TT\}$, and $\mathsf{x}_1 = I_\mathsf{A}$, $\mathsf{x}_2 = I_\mathsf{B}$, and $\mathsf{y} = I_\mathsf{C}$. They are pairwise independent, but $\mathsf{A}\cap\mathsf{B}\cap\mathsf{C} = \mathsf{A}\cap\mathsf{B}$, i.e., $\mathsf{x}_1\mathsf{x}_2\mathsf{y} = \mathsf{x}_1\mathsf{x}_2$, therefore $\mathrm{E}[\mathsf{x}_1\mathsf{x}_2\mathsf{y}] \neq \mathrm{E}[\mathsf{x}_1\mathsf{x}_2]\,\mathrm{E}[\mathsf{y}]$ therefore they are not independent. □

Problem 137. *4 points Prove that, if* x *and* y *are independent, then* $\mathrm{var}[\mathsf{xy}] = (\mathrm{E}[\mathsf{x}])^2\,\mathrm{var}[\mathsf{y}] + (\mathrm{E}[\mathsf{y}])^2\,\mathrm{var}[\mathsf{x}] + \mathrm{var}[\mathsf{x}]\,\mathrm{var}[\mathsf{y}]$.

Answer. Start with result and replace all occurrences of $\mathrm{var}[\mathsf{z}]$ with $\mathrm{E}[\mathsf{z}^2] - \mathrm{E}[\mathsf{z}]^2$, then multiply out: $\mathrm{E}[\mathsf{x}]^2(\mathrm{E}[\mathsf{y}^2] - \mathrm{E}[\mathsf{y}]^2) + \mathrm{E}[\mathsf{y}]^2(\mathrm{E}[\mathsf{x}^2] - \mathrm{E}[\mathsf{x}]^2) + (\mathrm{E}[\mathsf{x}^2] - \mathrm{E}[\mathsf{x}]^2)(\mathrm{E}[\mathsf{y}^2] - \mathrm{E}[\mathsf{y}]^2) = \mathrm{E}[\mathsf{x}^2]\,\mathrm{E}[\mathsf{y}^2] - \mathrm{E}[\mathsf{x}]^2\,\mathrm{E}[\mathsf{y}]^2 = \mathrm{E}[(\mathsf{xy})^2] - \mathrm{E}[\mathsf{xy}]^2$. □

8.6. Conditional Expectation and Variance

The *conditional expectation* of y is the expected value of y under the conditional density. If joint densities exist, it follows

$$\mathrm{E}[\mathsf{y}|\mathsf{x}{=}x] = \frac{\int y\, f_{\mathsf{x},\mathsf{y}}(x,y)\,dy}{f_\mathsf{x}(x)} =: g(x). \tag{8.6.1}$$

This is not a random variable but a constant which depends on x, i.e., a function of x, which is called here $g(x)$. But often one uses the term $\mathrm{E}[\mathsf{y}|\mathsf{x}]$ without specifying x. This is, by definition, the random variable $g(\mathsf{x})$ which one gets by plugging x into g; it assigns to every outcome $\omega \in \mathsf{U}$ the conditional expectation of y given $\mathsf{x}{=}x(\omega)$.

Since $\mathrm{E}[\mathsf{y}|\mathsf{x}]$ is a random variable, it is possible to take its expected value. The law of iterated expectations is extremely important here. It says that you will get the same result as if you had taken the expected value of y:

$$\mathrm{E}\big[\mathrm{E}[\mathsf{y}|\mathsf{x}]\big] = \mathrm{E}[\mathsf{y}]. \tag{8.6.2}$$

Proof (for the case that the densities exist):

$$\begin{aligned}\mathrm{E}\big[\mathrm{E}[\mathsf{y}|\mathsf{x}]\big] = \mathrm{E}[g(\mathsf{x})] &= \int \frac{\int y\, f_{\mathsf{x},\mathsf{y}}(x,y)\,dy}{f_\mathsf{x}(x)} f_\mathsf{x}(x)\,dx \\ &= \int\int y\, f_{\mathsf{x},\mathsf{y}}(x,y)\,dy\,dx = \mathrm{E}[\mathsf{y}].\end{aligned} \tag{8.6.3}$$

Problem 138. *Let* x *and* y *be two jointly distributed variables. For every fixed value* x, $\mathrm{var}[\mathsf{y}|\mathsf{x} = x]$ *is the variance of* y *under the conditional distribution, and* $\mathrm{var}[\mathsf{y}|\mathsf{x}]$ *is this variance as a random variable, namely, as a function of* x.

• a. *1 point Prove that*

$$\operatorname{var}[\mathsf{y}|\mathsf{x}] = \operatorname{E}[\mathsf{y}^2|\mathsf{x}] - (\operatorname{E}[\mathsf{y}|\mathsf{x}])^2. \tag{8.6.4}$$

This is a very simple proof. Explain exactly what, if anything, needs to be done to prove it.

Answer. For every *fixed* value x, it is an instance of the law

$$\operatorname{var}[\mathsf{y}] = \operatorname{E}[\mathsf{y}^2] - (\operatorname{E}[\mathsf{y}])^2 \tag{8.6.5}$$

applied to the conditional density given $\mathsf{x} = x$. And since it is true for every fixed x, it is also true after plugging in the random variable x. □

• b. *3 points Prove that*

$$\operatorname{var}[\mathsf{y}] = \operatorname{var}\big[\operatorname{E}[\mathsf{y}|\mathsf{x}]\big] + \operatorname{E}\big[\operatorname{var}[\mathsf{y}|\mathsf{x}]\big], \tag{8.6.6}$$

i.e., the variance consists of two components: the variance of the conditional mean and the mean of the conditional variances. This decomposition of the variance is given e.g. in [**Rao73**, p. 97] *or* [**Ame94**, theorem 4.4.2 on p. 78].

Answer. The first term on the rhs is $\operatorname{E}[(\operatorname{E}[\mathsf{y}|\mathsf{x}])^2] - (\operatorname{E}[\operatorname{E}[\mathsf{y}|\mathsf{x}]])^2$, and the second term, due to (8.6.4), becomes $\operatorname{E}[\operatorname{E}[\mathsf{y}^2|\mathsf{x}]] - \operatorname{E}[(\operatorname{E}[\mathsf{y}|\mathsf{x}])^2]$. If one adds, the two $\operatorname{E}[(\operatorname{E}[\mathsf{y}|\mathsf{x}])^2]$ cancel out, and the other two terms can be simplified by the law of iterated expectations to give $\operatorname{E}[\mathsf{y}^2] - (\operatorname{E}[\mathsf{y}])^2$. □

• c. *2 points* [**Coo98**, p. 23] *The conditional expected value is sometimes called the population regression function. In graphical data analysis, the sample equivalent of the variance ratio*

$$\frac{\operatorname{E}\big[\operatorname{var}[\mathsf{y}|\mathsf{x}]\big]}{\operatorname{var}\big[\operatorname{E}[\mathsf{y}|\mathsf{x}]\big]} \tag{8.6.7}$$

can be used to determine whether the regression function $\operatorname{E}[\mathsf{y}|\mathsf{x}]$ *appears to be visually well-determined or not. Does a small or a big variance ratio indicate a well-determined regression function?*

Answer. For a well-determined regression function the variance ratio should be *small*. [**Coo98**, p. 23] writes: "This ratio is reminiscent of a one-way analysis of variance, with the numerator representing the average within group (slice) variance, and the denominator representing the varince between group (slice) means." □

Now some general questions:

Problem 139. *The figure on page 116 shows 250 independent observations of the random vector* $\left[\begin{smallmatrix}\mathsf{x}\\ \mathsf{y}\end{smallmatrix}\right]$.

• a. *2 points Draw in by hand the approximate location of* $\mathcal{E}[\left[\begin{smallmatrix}\mathsf{x}\\ \mathsf{y}\end{smallmatrix}\right]]$ *and the graph of* $\operatorname{E}[\mathsf{y}|\mathsf{x}]$. *Draw into the second diagram the approximate marginal density of* x.

• b. *2 points Is there a law that the graph of the conditional expectation* $\operatorname{E}[\mathsf{y}|\mathsf{x}]$ *always goes through the point* $\mathcal{E}[\left[\begin{smallmatrix}\mathsf{x}\\ \mathsf{y}\end{smallmatrix}\right]]$*—for arbitrary probability distributions for which these expectations exist, or perhaps for an important special case? Indicate how this could be proved or otherwise give (maybe geometrically) a simple counterexample.*

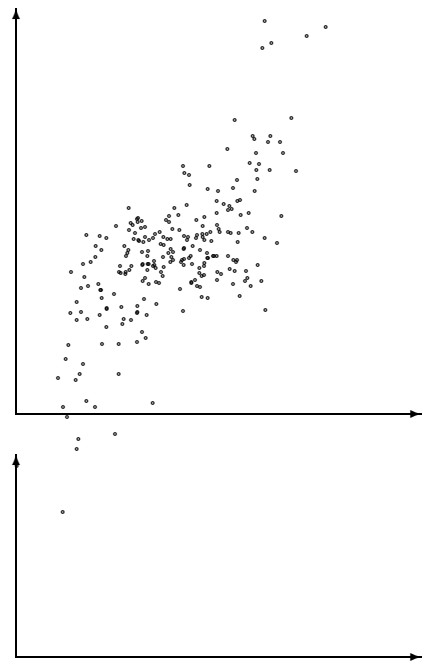
Answer. This is *not* the law of iterated expectations. It is true for jointly normal variables, not in general. It is also true if $\mathbf{x}$ and $\mathbf{y}$ are independent; then the graph of $\operatorname{E}[\mathsf{y}|\mathsf{x}]$ is a horizontal line at the height of the unconditional expectation $\operatorname{E}[\mathsf{y}]$. A distribution with U-shaped unconditional distribution has the unconditional mean in the center of the U, i.e., here the unconditional mean does not lie on the curve drawn out by the conditional mean. □

• c. *2 points Do you have any ideas how the strange-looking cluster of points in the figure on page 116 was generated?*

Problem 140. *2 points Given two independent random variables* x *and* y *with density functions* $f_{\mathsf{x}}(x)$ *and* $g_{\mathsf{y}}(y)$. *Write down their joint, marginal, and conditional densities.*

Answer. Joint density: $f_{\mathsf{x},\mathsf{y}}(x,(y) = f_{\mathsf{x}}(x)g_{\mathsf{y}}(y)$.

Marginal density of x is $\int_{-\infty}^{\infty} f_{\mathsf{x}}(x)g_{\mathsf{y}}(y)\,dy = f_{\mathsf{x}}(x)\int_{-\infty}^{\infty} g_{\mathsf{y}}(y)\,dy = f_{\mathsf{x}}(x)$, and that of y is $g_{\mathsf{y}}(y)$. The text of the question should have been: "Given two independent random variables x and y with *marginal* density functions $f_{\mathsf{x}}(x)$ and $g_{\mathsf{y}}(y)$"; by just calling them "density functions" without specifying "marginal" it committed the error of de-totalization, i.e., it treated elements of



a totality, i.e., of an ensemble in which each depends on everything else, as if they could be defined independently of each other.

Conditional density functions: $f_{\mathsf{x}|\mathsf{y}=y}(x;y) = f_{\mathsf{x}}(x)$ (i.e., it does not depend on y); and $g_{\mathsf{y}|\mathsf{x}=x}(y;x) = g_{\mathsf{y}}(y)$. You can see this by dividing the joint by the marginal. □

8.7. Expected Values as Predictors

Expected values and conditional expected values have optimal properties as predictors.

PROBLEM 141. *3 points What is the best predictor of a random variable* y *by a constant* a*, if the loss function is the "mean squared error" (MSE)* $\mathrm{E}[(\mathsf{y}-a)^2]$*?*

ANSWER. Write $\mathrm{E}[\mathsf{y}] = \mu$; then

$$\begin{aligned}(\mathsf{y}-a)^2 &= \big((\mathsf{y}-\mu)-(a-\mu)\big)^2 \\ &= (\mathsf{y}-\mu)^2 - 2(\mathsf{y}-\mu)(a-\mu) + (a-\mu)^2; \\ \text{therefore}\quad \mathrm{E}[(\mathsf{y}-a)^2] &= \mathrm{E}[(\mathsf{y}-\mu)^2] - 0 + (a-\mu)^2\end{aligned} \tag{8.7.1}$$

This is minimized by $a = \mu$. □

The expected value of y is therefore that constant which, as predictor of y, has smallest MSE.

What if we want to predict y not by a constant but by a function of the random vector $\mathbf{x}$, call it $h(\mathbf{x})$?

PROBLEM 142. *2 points Assume the vector* $\mathbf{x} = [\mathsf{x}_1, \dots \mathsf{x}_j]^\top$ *and the scalar* y *are jointly distributed random variables, and assume conditional means exist.* $\mathbf{x}$ *is observed, but* y *is not observed. The joint distribution of* $\mathbf{x}$ *and* y *is known. Show that the conditional expectation* $\mathrm{E}[\mathsf{y}|\mathbf{x}]$ *is the minimum* MSE *predictor of* y *given* $\mathbf{x}$*, i.e., show that for any other function of* $\mathbf{x}$*, call it* $h(\mathbf{x})$*, the following inequality holds:*

$$\mathrm{E}[\big(\mathsf{y}-h(\mathbf{x})\big)^2] \geq \mathrm{E}[\big(\mathsf{y}-\mathrm{E}[\mathsf{y}|\mathbf{x}]\big)^2]. \tag{8.7.2}$$

For this proof and the proofs required in Problems 143 and 144, you may use (1) the theorem of iterated expectations $\mathrm{E}\big[\mathrm{E}[\mathsf{y}|\mathbf{x}]\big] = \mathrm{E}[\mathsf{y}]$*, (2) the additivity* $\mathrm{E}[g(\mathsf{y}) + h(\mathsf{y})|\mathbf{x}] = \mathrm{E}[g(\mathsf{y})|\mathbf{x}] + \mathrm{E}[h(\mathsf{y})|\mathbf{x}]$*, and (3) the fact that* $\mathrm{E}[g(\mathbf{x})h(\mathsf{y})|\mathbf{x}] = g(\mathbf{x})\mathrm{E}[h(\mathsf{y})|\mathbf{x}]$*. Be very specific about which rules you are applying at every step. You must show that you understand what you are writing down.*

ANSWER.

$$\begin{aligned}\mathrm{E}[\big(\mathsf{y}-h(\mathbf{x})\big)^2] &= \mathrm{E}\Big[\big(\mathsf{y}-\mathrm{E}[\mathsf{y}|\mathbf{x}]-(h(\mathbf{x})-\mathrm{E}[\mathsf{y}|\mathbf{x}])\big)^2\Big] \\ &= \mathrm{E}[(\mathsf{y}-\mathrm{E}[\mathsf{y}|\mathbf{x}])^2] - 2\,\mathrm{E}[(\mathsf{y}-\mathrm{E}[\mathsf{y}|\mathbf{x}])(h(\mathbf{x})-\mathrm{E}[\mathsf{y}|\mathbf{x}])] + \mathrm{E}[(h(\mathbf{x})-\mathrm{E}[\mathsf{y}|\mathbf{x}])^2].\end{aligned} \tag{8.7.3}$$

Here the cross product term $\mathrm{E}[(\mathsf{y}-\mathrm{E}[\mathsf{y}|\mathbf{x}])(h(\mathbf{x})-\mathrm{E}[\mathsf{y}|\mathbf{x}])]$ is zero. In order to see this, first use the law of iterated expectations

$$\mathrm{E}[(\mathsf{y}-\mathrm{E}[\mathsf{y}|\mathbf{x}])(h(\mathbf{x})-\mathrm{E}[\mathsf{y}|\mathbf{x}])] = \mathrm{E}\Big[\mathrm{E}[(\mathsf{y}-\mathrm{E}[\mathsf{y}|\mathbf{x}])(h(\mathbf{x})-\mathrm{E}[\mathsf{y}|\mathbf{x}])|\mathbf{x}]\Big] \tag{8.7.4}$$

and then look at the inner term, not yet doing the outer expectation:

$$\mathrm{E}[(\mathsf{y}-\mathrm{E}[\mathsf{y}|\mathbf{x}])(h(\mathbf{x})-\mathrm{E}[\mathsf{y}|\mathbf{x}])|\mathbf{x}] = (h(\mathbf{x})-\mathrm{E}[\mathsf{y}|\mathbf{x}])\,\mathrm{E}[(\mathsf{y}-\mathrm{E}[\mathsf{y}|\mathbf{x}])|\mathbf{x}] = (h(\mathbf{x})-\mathrm{E}[\mathsf{y}|\mathbf{x}])(\mathrm{E}[\mathsf{y}|\mathbf{x}]-\mathrm{E}[\mathsf{y}|\mathbf{x}]) == (h(\mathbf{x})-\mathrm{E}[\mathsf{y}|\mathbf{x}])\cdot 0 = 0$$

Plugging this into (8.7.4) gives $\mathrm{E}[(\mathsf{y}-\mathrm{E}[\mathsf{y}|\mathbf{x}])(h(\mathbf{x})-\mathrm{E}[\mathsf{y}|\mathbf{x}])] = \mathrm{E}\Big[0\Big] = 0$.

□

This is one of the few clear cut results in probability theory where a best estimator/predictor exists. In this case, however, all parameters of the distribution are known, the only uncertainty comes from the fact that some random variables are unobserved.

PROBLEM 143. *Assume the vector* $\mathbf{x} = [\mathsf{x}_1, \dots \mathsf{x}_j]^\top$ *and the scalar* y *are jointly distributed random variables, and assume conditional means exist. Define* $\varepsilon = \mathsf{y} - \mathrm{E}[\mathsf{y}|\mathbf{x}]$.

• a. *5 points Demonstrate the following identities:*

$$\mathrm{E}[\varepsilon|\mathbf{x}] = 0 \tag{8.7.5}$$

$$\mathrm{E}[\varepsilon] = 0 \tag{8.7.6}$$

$$\mathrm{E}[\mathsf{x}_i\varepsilon|\mathbf{x}] = 0 \quad \textit{for all } i,\ 1 \leq i \leq j \tag{8.7.7}$$

$$\mathrm{E}[\mathsf{x}_i\varepsilon] = 0 \quad \textit{for all } i,\ 1 \leq i \leq j \tag{8.7.8}$$

$$\mathrm{cov}[\mathsf{x}_i, \varepsilon] = 0 \quad \textit{for all } i,\ 1 \leq i \leq j. \tag{8.7.9}$$

Interpretation of (8.7.9)*:* ε *is the error in the best prediction of* y *based on* $\mathbf{x}$*. If this error were correlated with one of the components* x_i*, then this correlation could be used to construct a better prediction of* y.

ANSWER. (8.7.5): $\mathrm{E}[\varepsilon|\mathbf{x}] = \mathrm{E}[\mathsf{y}|\mathbf{x}] - \mathrm{E}\Big[\mathrm{E}[\mathsf{y}|\mathbf{x}]|\mathbf{x}\Big] = 0$ since $\mathrm{E}[\mathsf{y}|\mathbf{x}]$ is a function of $\mathbf{x}$ and therefore equal to its own expectation conditionally on $\mathbf{x}$. (This is *not* the law of iterated expectations but the law that the expected value of a constant is a constant.)

(8.7.6) follows from (8.7.5) (i.e., (8.7.5) is stronger than (8.7.6)): if an expectation is zero conditionally on every possible outcome of $\mathbf{x}$ then it is zero altogether. In formulas, $\mathrm{E}[\varepsilon] = \mathrm{E}\Big[\mathrm{E}[\varepsilon|\mathbf{x}]\Big] = \mathrm{E}[0] = 0$. It is also easy to show it in one swoop, without using (8.7.5): $\mathrm{E}[\varepsilon] = \mathrm{E}[\mathsf{y} - \mathrm{E}[\mathsf{y}|\mathbf{x}]] = 0$. Either way you need the law of iterated expectations for this.

(8.7.7): $\mathrm{E}[\mathsf{x}_i\varepsilon|\mathbf{x}] = \mathsf{x}_i\mathrm{E}[\varepsilon|\mathbf{x}] = 0$.

(8.7.8): $\mathrm{E}[\mathsf{x}_i\varepsilon] = \mathrm{E}\Big[\mathrm{E}[\mathsf{x}_i\varepsilon|\mathbf{x}]\Big] = \mathrm{E}[0] = 0$; or in one swoop: $\mathrm{E}[\mathsf{x}_i\varepsilon] = \mathrm{E}\Big[\mathsf{x}_i\mathsf{y} - \mathsf{x}_i\mathrm{E}[\mathsf{y}|\mathbf{x}]\Big] = \mathrm{E}\Big[\mathsf{x}_i\mathsf{y} - \mathrm{E}[\mathsf{x}_i\mathsf{y}|\mathbf{x}]\Big] = \mathrm{E}[\mathsf{x}_i\mathsf{y}] - \mathrm{E}[\mathsf{x}_i\mathsf{y}] = 0$. The following "proof" is not correct: $\mathrm{E}[\mathsf{x}_i\varepsilon] = \mathrm{E}[\mathsf{x}_i]\,\mathrm{E}[\varepsilon] = \mathrm{E}[\mathsf{x}_i]\cdot 0 = 0$. x_i and ε are generally not independent, therefore the multiplication rule $\mathrm{E}[\mathsf{x}_i\varepsilon] = \mathrm{E}[\mathsf{x}_i]\,\mathrm{E}[\varepsilon]$ cannot be used. Of course, the following "proof" does not work either: $\mathrm{E}[\mathsf{x}_i\varepsilon] = \mathsf{x}_i\,\mathrm{E}[\varepsilon] = \mathsf{x}_i \cdot 0 = 0$. x_i is a random variable and $\mathrm{E}[\mathsf{x}_i\varepsilon]$ is a constant; therefore $\mathrm{E}[\mathsf{x}_i\varepsilon] = \mathsf{x}_i\,\mathrm{E}[\varepsilon]$ cannot hold.

(8.7.9): $\mathrm{cov}[\mathsf{x}_i, \varepsilon] = \mathrm{E}[\mathsf{x}_i\varepsilon] - \mathrm{E}[\mathsf{x}_i]\,\mathrm{E}[\varepsilon] = 0 - \mathrm{E}[\mathsf{x}_i]\cdot 0 = 0$. □

• b. *2 points This part can only be done after discussing the multivariate normal distribution:If* $\mathbf{x}$ *and* y *are jointly normal, show that* $\mathbf{x}$ *and* ε *are independent, and*

that the variance of ε does not depend on $\mathbf{x}$. (This is why one can consider it an error term.)

Answer. If $\mathbf{x}$ and y are jointly normal, then $\mathbf{x}$ and ε are jointly normal as well, and independence follows from the fact that their covariance is zero. The variance is constant because in the Normal case, the conditional variance is constant, i.e., $\mathrm{E}[\varepsilon^2] = \mathrm{E}\big[\mathrm{E}[\varepsilon^2|\mathbf{x}]\big] = \text{constant}$ (does not depend on $\mathbf{x}$). □

Problem 144. *5 points Under the permanent income hypothesis, the assumption is made that consumers' lifetime utility is highest if the same amount is consumed every year. The utility-maximizing level of consumption c for a given consumer depends on the actual state of the economy in each of the n years of the consumer's life $\mathsf{c} = \mathsf{f}(\mathsf{y}_1, \ldots, \mathsf{y}_n)$. Since c depends on future states of the economy, which are not known, it is impossible for the consumer to know this optimal c in advance; but it is assumed that the function f and the joint distribution of $\mathsf{y}_1, \ldots, \mathsf{y}_n$ are known to him. Therefore in period t, when he only knows the values of $\mathsf{y}_1, \ldots, \mathsf{y}_t$, but not yet the future values, the consumer decides to consume the amount $\mathsf{c}_t = \mathrm{E}[\mathsf{c}|\mathsf{y}_1, \ldots, \mathsf{y}_t]$, which is the best possible prediction of c given the information available to him. Show that in this situation, $\mathsf{c}_{t+1} - \mathsf{c}_t$ is uncorrelated with all $\mathsf{y}_1, \ldots, \mathsf{y}_t$. This implication of the permanent income hypothesis can be tested empirically, see* [**Hal78**]. *Hint: you are allowed to use without proof the following extension of the theorem of iterated expectations:*

$$\mathrm{E}\big[\mathrm{E}[\mathsf{x}|\mathbf{y}, \mathbf{z}]\big|\mathbf{y}\big] = \mathrm{E}[\mathsf{x}|\mathbf{y}]. \tag{8.7.10}$$

Here is an explanation of (8.7.10)*: $\mathrm{E}[\mathsf{x}|\mathbf{y}]$ is the best predictor of x based on information set $\mathbf{y}$. $\mathrm{E}[\mathsf{x}|\mathbf{y}, \mathbf{z}]$ is the best predictor of x based on the extended information set consisting of $\mathbf{y}$ and $\mathbf{z}$. $\mathrm{E}\big[\mathrm{E}[\mathsf{x}|\mathbf{y}, \mathbf{z}]\big|\mathbf{y}\big]$ is therefore my prediction, based on $\mathbf{y}$ only, how I will refine my prediction when $\mathbf{z}$ becomes available as well. Its equality with $\mathrm{E}[\mathsf{x}|\mathbf{y}]$, i.e.,* (8.7.10) *says therefore that I cannot predict how I will change my mind after better information becomes available.*

Answer. In (8.7.10) set $\mathsf{x} = \mathsf{c} = f(\mathsf{y}_1, \ldots, \mathsf{y}_t, \mathsf{y}_{t+1}, \ldots, \mathsf{y}_n)$, $\mathbf{y} = [\mathsf{y}_1, \ldots, \mathsf{y}_t]^\top$, and $\mathbf{z} = \mathsf{y}_{t+1}$ to get

$$\mathrm{E}\big[\mathrm{E}[\mathsf{c}|\mathsf{y}_1, \ldots, \mathsf{y}_{t+1}]\big|\mathsf{y}_1, \ldots, \mathsf{y}_t\big] = \mathrm{E}[\mathsf{c}|\mathsf{y}_1, \ldots, \mathsf{y}_t]. \tag{8.7.11}$$

Writing c_t for $\mathrm{E}[\mathsf{c}|\mathsf{y}_1, \ldots, \mathsf{y}_t]$, this becomes $\mathrm{E}[\mathsf{c}_{t+1}|\mathsf{y}_1, \ldots, \mathsf{y}_t] = \mathsf{c}_t$, i.e., c_t is not only the best predictor of c, but also that of c_{t+1}. The change in consumption $\mathsf{c}_{t+1} - \mathsf{c}_t$ is therefore the prediction error, which is uncorrelated with the conditioning variables, as shown in Problem 143. □

Problem 145. *3 points Show that for any two random variables x and y whose covariance exists, the following equation holds:*

$$\mathrm{cov}[\mathsf{x}, \mathsf{y}] = \mathrm{cov}\big[\mathsf{x}, \mathrm{E}[\mathsf{y}|\mathsf{x}]\big] \tag{8.7.12}$$

Note: Since $\mathrm{E}[\mathsf{y}|\mathsf{x}]$ is the best predictor of y based on the observation of x, (8.7.12) *can also be written as*

$$\mathrm{cov}\big[\mathsf{x}, (\mathsf{y} - \mathrm{E}[\mathsf{y}|\mathsf{x}])\big] = 0, \tag{8.7.13}$$

i.e., x is uncorrelated with the prediction error of the best prediction of y given x. (Nothing to prove for this Note.)

Answer. Apply (8.1.6) to the righthand side of (8.7.12):

$$\mathrm{cov}\big[\mathsf{x}, \mathrm{E}[\mathsf{y}|\mathsf{x}]\big] = \mathrm{E}\big[\mathsf{x}\mathrm{E}[\mathsf{y}|\mathsf{x}]\big] - \mathrm{E}[\mathsf{x}]\,\mathrm{E}\big[\mathrm{E}[\mathsf{y}|\mathsf{x}]\big] = \mathrm{E}\big[\mathrm{E}[\mathsf{x}\mathsf{y}|\mathsf{x}]\big] - \mathrm{E}[\mathsf{x}]\,\mathrm{E}[\mathsf{y}] = \mathrm{E}[\mathsf{x}\mathsf{y}] - \mathrm{E}[\mathsf{x}]\,\mathrm{E}[\mathsf{y}] = \mathrm{cov}[\mathsf{x}, \mathsf{y}]. \tag{8.7.14}$$

The tricky part here is to see that $\mathsf{x}\mathrm{E}[\mathsf{y}|\mathsf{x}] = \mathrm{E}[\mathsf{x}\mathsf{y}|\mathsf{x}]$.

□

Problem 146. *Assume x and y have a joint density function $f_{\mathsf{x},\mathsf{y}}(x, y)$ which is symmetric about the x-axis, i.e.,*

$$f_{\mathsf{x},\mathsf{y}}(x, y) = f_{\mathsf{x},\mathsf{y}}(x, -y).$$

Also assume that variances and covariances exist. Show that $\mathrm{cov}[\mathsf{x}, \mathsf{y}] = 0$. Hint: one way to do it is to look at $\mathrm{E}[\mathsf{y}|\mathsf{x}]$.

Answer. We know that $\mathrm{cov}[\mathsf{x}, \mathsf{y}] = \mathrm{cov}\big[\mathsf{x}, \mathrm{E}[\mathsf{y}|\mathsf{x}]\big]$. Furthermore, from symmetry follows $\mathrm{E}[\mathsf{y}|\mathsf{x}] = 0$. Therefore $\mathrm{cov}[\mathsf{x}, \mathsf{y}] = \mathrm{cov}[\mathsf{x}, 0] = 0$. Here is a detailed proof of $\mathrm{E}[\mathsf{y}|\mathsf{x}] = 0$: $\mathrm{E}[\mathsf{y}|\mathsf{x}{=}x] = \int_{-\infty}^{\infty} y \frac{f_{\mathsf{x},\mathsf{y}}(x,y)}{f_{\mathsf{x}}(x)} dy$. Now substitute $z = -y$, then also $dz = -dy$, and the boundaries of integration are reversed:

$$\mathrm{E}[\mathsf{y}|\mathsf{x}{=}x] = \int_{\infty}^{-\infty} z \frac{f_{\mathsf{x},\mathsf{y}}(x, -z)}{f_{\mathsf{x}}(x)} dz = \int_{\infty}^{-\infty} z \frac{f_{\mathsf{x},\mathsf{y}}(x, z)}{f_{\mathsf{x}}(x)} dz = -\,\mathrm{E}[\mathsf{y}|\mathsf{x}{=}x]. \tag{8.7.15}$$

One can also prove directly under this presupposition $\mathrm{cov}[\mathsf{x}, \mathsf{y}] = \mathrm{cov}[\mathsf{x}, -\mathsf{y}]$ and therefore it must be zero.

□

Problem 147. [**Wit85**, footnote on p. 241] *Let p be the logarithm of the price level, m the logarithm of the money supply, and x a variable representing real influences on the price level (for instance productivity). We will work in a model of the economy in which $\mathsf{p} = \mathsf{m} + \gamma\mathsf{x}$, where γ is a nonrandom parameter, and m and x are independent normal with expected values μ_{m}, μ_{x}, and variances σ^2_{m}, σ^2_{x}. According to the rational expectations assumption, the economic agents know the probability distribution of the economy they live in, i.e., they know the expected values and variances of m and x and the value of γ. But they are unable to observe m and x, they can only observe p. Then the best predictor of x using p is the conditional expectation $\mathrm{E}[\mathsf{x}|\mathsf{p}]$.*

• a. *Assume you are one of these agents and you observe $\mathsf{p} = p$. How great would you predict x to be, i.e., what is the value of $\mathrm{E}[\mathsf{x}|\mathsf{p} = p]$?*

Answer. It is, according to formula (10.3.18), $\mathrm{E}[\mathsf{x}|\mathsf{p}=p]=\mu_{\mathsf{x}}+\frac{\mathrm{cov}(\mathsf{x},\mathsf{p})}{\mathrm{var}(\mathsf{p})}(p-\mathrm{E}[\mathsf{p}])$. Now $\mathrm{E}[\mathsf{p}]=\mu_{\mathsf{m}}+\gamma\mu_{\mathsf{x}}$, $\mathrm{cov}[\mathsf{x},\mathsf{p}]=\mathrm{cov}[\mathsf{x},\mathsf{m}]+\gamma\,\mathrm{cov}[\mathsf{x},\mathsf{x}]=\gamma\sigma_{\mathsf{x}}^2$, and $\mathrm{var}(\mathsf{p})=\sigma_{\mathsf{m}}^2+\gamma^2\sigma_{\mathsf{x}}^2$. Therefore

$$\mathrm{E}[\mathsf{x}|\mathsf{p}=p]=\mu_{\mathsf{x}}+\frac{\gamma\sigma_{\mathsf{x}}^2}{\sigma_{\mathsf{m}}^2+\gamma^2\sigma_{\mathsf{x}}^2}(p-\mu_{\mathsf{m}}-\gamma\mu_{\mathsf{x}}). \tag{8.7.16}$$

□

• b. *Define the prediction error* $\varepsilon=\mathsf{x}-\mathrm{E}[\mathsf{x}|\mathsf{p}]$. *Compute expected value and variance of* ε.

Answer.

$$\varepsilon=\mathsf{x}-\mu_{\mathsf{x}}-\frac{\gamma\sigma_{\mathsf{x}}^2}{\sigma_{\mathsf{m}}^2+\gamma^2\sigma_{\mathsf{x}}^2}(\mathsf{p}-\mu_{\mathsf{m}}-\gamma\mu_{\mathsf{x}}). \tag{8.7.17}$$

This has zero expected value, and its variance is

$$\mathrm{var}[\varepsilon]=\mathrm{var}[\mathsf{x}]+\Big(\frac{\gamma\sigma_{\mathsf{x}}^2}{\sigma_{\mathsf{m}}^2+\gamma^2\sigma_{\mathsf{x}}^2}\Big)^2\mathrm{var}[\mathsf{p}]-2\Big(\frac{\gamma\sigma_{\mathsf{x}}^2}{\sigma_{\mathsf{m}}^2+\gamma^2\sigma_{\mathsf{x}}^2}\Big)\mathrm{cov}[\mathsf{x},\mathsf{p}]= \tag{8.7.18}$$

$$=\sigma_{\mathsf{x}}^2+\frac{\gamma^2(\sigma_{\mathsf{x}}^2)^2}{\sigma_{\mathsf{m}}^2+\gamma^2\sigma_{\mathsf{x}}^2}-2\frac{\gamma^2(\sigma_{\mathsf{x}}^2)^2}{\sigma_{\mathsf{m}}^2+\gamma^2\sigma_{\mathsf{x}}^2} \tag{8.7.19}$$

$$=\frac{\sigma_{\mathsf{x}}^2\sigma_{\mathsf{m}}^2}{\sigma_{\mathsf{m}}^2+\gamma^2\sigma_{\mathsf{x}}^2}=\frac{\sigma_{\mathsf{x}}^2}{1+\gamma^2\sigma_{\mathsf{x}}^2/\sigma_{\mathsf{m}}^2}. \tag{8.7.20}$$

□

• c. *In an attempt to fine tune the economy, the central bank increases* σ_{m}^2. *Does that increase or decrease* $\mathrm{var}(\varepsilon)$*?*

Answer. From (8.7.20) follows that it increases the variance. □

8.8. Transformation of Vector Random Variables

In order to obtain the density or probability mass function of a one-to-one transformation of random variables, we have to follow the same 4 steps described in Section 3.6 for a scalar random variable. (1) Determine A, the range of the new variable, whose density we want to compute; (2) express the old variable, the one whose density/mass function is known, in terms of the new variable, the one whose density or mass function is needed. If that of $\begin{bmatrix}\mathsf{x}\\ \mathsf{y}\end{bmatrix}$ is known, set $\begin{bmatrix}\mathsf{x}\\ \mathsf{y}\end{bmatrix}=\boldsymbol{t}(\mathsf{u},\mathsf{v})$. Here $\boldsymbol{t}$ is a vector-valued function, (i.e., it could be written $\boldsymbol{t}(\mathsf{u},\mathsf{v})=\begin{bmatrix}\mathsf{q}(\mathsf{u},\mathsf{v})\\ \mathsf{r}(\mathsf{u},\mathsf{v})\end{bmatrix}$, but we will use one symbol $\boldsymbol{t}$ for this whole transformation), and you have to check that it is one-to-one on A, i.e., $\boldsymbol{t}(u,v)=\boldsymbol{t}(u_1,v_1)$ implies $u=u_1$ and $v=v_1$ for all (u,v) and $u_1,v_1)$ in A. (A function for which two different arguments (u,v) and $u_1,v_1)$ give the same function value is called many-to-one.)

If the joint probability distribution of x and y is described by a probability mass function, then the joint probability mass function of u and v can simply be obtained

by substituting $\boldsymbol{t}$ into the joint probability mass function of x and y (and it is zero for any values which are not in A):

$$p_{\mathsf{u},\mathsf{v}}(u,v)=\Pr\Big[\begin{bmatrix}\mathsf{u}\\ \mathsf{v}\end{bmatrix}=\begin{bmatrix}u\\ v\end{bmatrix}\Big]=\Pr\big[\boldsymbol{t}(\mathsf{u},\mathsf{v})=\boldsymbol{t}(u,v)\big]=\Pr\Big[\begin{bmatrix}\mathsf{x}\\ \mathsf{y}\end{bmatrix}=\boldsymbol{t}(u,v)\Big]=p_{\mathsf{x},\mathsf{y}}\big(\boldsymbol{t}(u,v)\big). \tag{8.8.1}$$

The second equal sign is where the condition enters that $\boldsymbol{t}:\mathbb{R}^2\to\mathbb{R}^2$ is one-to-one.

If one works with the density function instead of a mass function, one must perform an additional step besides substituting $\boldsymbol{t}$. Since $\boldsymbol{t}$ is one-to-one, it follows

$$\Big\{\begin{bmatrix}\mathsf{u}\\ \mathsf{v}\end{bmatrix}\in dV_{u,v}\Big\}=\{\boldsymbol{t}(\mathsf{u},\mathsf{v})\in\boldsymbol{t}(dV)_{x,y}\}. \tag{8.8.2}$$

Therefore

$$f_{\mathsf{u},\mathsf{v}}(u,v)|dV_{u,v}|=\Pr\Big[\begin{bmatrix}\mathsf{u}\\ \mathsf{v}\end{bmatrix}\in dV_{u,v}\Big]=\Pr[\boldsymbol{t}(\mathsf{u},\mathsf{v})\in\boldsymbol{t}(dV)_{x,y}]=f_{\mathsf{x},\mathsf{y}}(\boldsymbol{t}(u,v))|\boldsymbol{t}(dV)_{x,y}|= \tag{8.8.3}$$

$$=f_{\mathsf{x},\mathsf{y}}(\boldsymbol{t}(u,v))\frac{|\boldsymbol{t}(dV)_{x,y}|}{|dV_{u,v}|}|dV_{u,v}|. \tag{8.8.4}$$

The term $\frac{|\boldsymbol{t}(dV)_{x,y}|}{|dV_{u,v}|}$ is the local magnification factor of the transformation $\boldsymbol{t}$; analytically it is the absolute value $|J|$ of the Jacobian determinant

$$J=\begin{vmatrix}\frac{\partial\mathsf{x}}{\partial\mathsf{u}}&\frac{\partial\mathsf{x}}{\partial\mathsf{v}}\\ \frac{\partial\mathsf{y}}{\partial\mathsf{u}}&\frac{\partial\mathsf{y}}{\partial\mathsf{v}}\end{vmatrix}=\begin{vmatrix}\frac{\partial\mathsf{q}}{\partial u}(u,v)&\frac{\partial\mathsf{q}}{\partial v}(u,v)\\ \frac{\partial\mathsf{r}}{\partial u}(u,v)&\frac{\partial\mathsf{r}}{\partial v}(u,v)\end{vmatrix}. \tag{8.8.5}$$

Remember, u,v are the new and x,y the old variables. To compute J one has to express the old in terms of the new variables. If one expresses the new in terms of the old, one has to take the inverse of the corresponding determinant! The transformation rule for density functions can therefore be summarized as:

$$(\mathsf{x},\mathsf{y})=\boldsymbol{t}(\mathsf{u},\mathsf{v})\quad\text{one-to-one}\quad\Rightarrow\quad f_{\mathsf{u},\mathsf{v}}(u,v)=f_{\mathsf{x},\mathsf{y}}\big(\boldsymbol{t}(u,v)\big)\,|J|\quad\text{where}\quad J=\begin{vmatrix}\frac{\partial\mathsf{x}}{\partial\mathsf{u}}&\frac{\partial\mathsf{x}}{\partial\mathsf{v}}\\ \frac{\partial\mathsf{y}}{\partial\mathsf{u}}&\frac{\partial\mathsf{y}}{\partial\mathsf{v}}\end{vmatrix}.$$

Problem 148. *Let* x *and* y *be two random variables with joint density function* $f_{\mathsf{x},\mathsf{y}}(x,y)$.

• a. *3 points Define* $\mathsf{u}=\mathsf{x}+\mathsf{y}$. *Derive the joint density function of* u *and* y.

Answer. You have to express the "old" x and y as functions of the "new" u and y:

$$\begin{aligned}\mathsf{x}&=\mathsf{u}-\mathsf{y}\\ \mathsf{y}&=\mathsf{y}\end{aligned}\quad\text{or}\quad\begin{bmatrix}\mathsf{x}\\ \mathsf{y}\end{bmatrix}=\begin{bmatrix}1&-1\\ 0&1\end{bmatrix}\begin{bmatrix}\mathsf{u}\\ \mathsf{y}\end{bmatrix}\quad\text{therefore}\quad J=\begin{vmatrix}\frac{\partial\mathsf{x}}{\partial\mathsf{u}}&\frac{\partial\mathsf{x}}{\partial\mathsf{y}}\\ \frac{\partial\mathsf{y}}{\partial\mathsf{u}}&\frac{\partial\mathsf{y}}{\partial\mathsf{y}}\end{vmatrix}=\begin{vmatrix}1&-1\\ 0&1\end{vmatrix}=1.$$

Therefore

$$f_{\mathsf{u},\mathsf{y}}(u,y)=f_{\mathsf{x},\mathsf{y}}(u-y,y). \tag{8.8.6}$$

□

• b. *1 point Derive from this the following formula computing the density function $f_{\mathfrak{u}}(u)$ of the sum $\mathfrak{u} = \mathfrak{x} + \mathfrak{y}$ from the joint density function $f_{\mathfrak{x},\mathfrak{y}}(x, y)$ of $\mathfrak{x}$ and $\mathfrak{y}$.*

$$f_{\mathfrak{u}}(u) = \int_{y=-\infty}^{y=\infty} f_{\mathfrak{x},\mathfrak{y}}(u-y, y)dy. \tag{8.8.7}$$

Answer. Write down the joint density of $\mathfrak{u}$ and $\mathfrak{y}$ and then integrate y out, i.e., take its integral over y from $-\infty$ to $+\infty$:

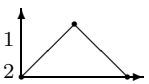
$$f_{\mathfrak{u}}(u) = \int_{y=-\infty}^{y=\infty} f_{\mathfrak{u},\mathfrak{y}}(u, y)dy = \int_{y=-\infty}^{y=\infty} f_{\mathfrak{x},\mathfrak{y}}(u-y, y)dy. \tag{8.8.8}$$

i.e., one integrates over all $\begin{bmatrix} x \\ y \end{bmatrix}$ with $x + y = u$. □

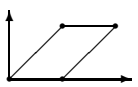
Problem 149. *6 points Let $\mathfrak{x}$ and $\mathfrak{y}$ be independent and uniformly distributed over the interval $[0, 1]$. Compute the density function of $\mathfrak{u} = \mathfrak{x} + \mathfrak{y}$ and draw its graph. Hint: you may use formula (8.8.7) for the density of the sum of two jointly distributed random variables. An alternative approach would be to first compute the cumulative distribution function $\Pr[\mathfrak{x} + \mathfrak{y} \leq u]$ for all u.*

Answer. Using equation (8.8.7):

$$f_{\mathfrak{x}+\mathfrak{y}}(u) = \int_{-\infty}^{\infty} f_{\mathfrak{x},\mathfrak{y}}(u-y, y)\, dy = \begin{cases} u & \text{for } 0 \leq u \leq 1 \\ 2-u & \text{for } 1 \leq u \leq 2 \\ 0 & \text{otherwise.} \end{cases} \tag{8.8.9}$$



To help evaluate this integral, here is the area in u, y-plane ($u = x + y$ on the horizontal and y on the vertical axis) in which $f_{\mathfrak{x},\mathfrak{y}}(u - v, v)$ has the value 1:



This is the area between (0,0), (1,1), (2,1), and (1,0).

One can also show it this way: $f_{\mathfrak{x},\mathfrak{y}}(x, y) = 1$ iff $0 \leq x \leq 1$ and $0 \leq y \leq 1$. Now take any fixed u. It must be between 0 and 2. First assume $0 \leq u \leq 1$: then $f_{\mathfrak{x},\mathfrak{y}}(u - y, y) = 1$ iff $0 \leq u - y \leq 1$ and $0 \leq y \leq 1$ iff $0 \leq y \leq u$. Now assume $1 \leq u \leq 2$: then $f_{\mathfrak{x},\mathfrak{y}}(u - y, y) = 1$ iff $u - 1 \leq y \leq 1$. □

Problem 150. *Assume $\left[\begin{smallmatrix} \mathfrak{x} \\ \mathfrak{y} \end{smallmatrix}\right]$ is uniformly distributed on a round disk around the origin with radius 10.*

• a. *4 points Derive the joint density, the marginal density of $\mathfrak{x}$, and the conditional density of $\mathfrak{y}$ given $\mathfrak{x}$=x.*

• b. *3 points Now let us go over to polar coordinates $\mathfrak{r}$ and ϕ, which satisfy*

$$\begin{aligned} \mathfrak{x} &= \mathfrak{r}\cos\phi \\ \mathfrak{y} &= \mathfrak{r}\sin\phi \end{aligned}, \quad \textit{i.e., the vector transformation } \boldsymbol{t} \textit{ is } \quad \boldsymbol{t}\left(\begin{bmatrix} \mathfrak{r} \\ \phi \end{bmatrix}\right) = \begin{bmatrix} \mathfrak{r}\cos\phi \\ \mathfrak{r}\sin\phi \end{bmatrix}. \tag{8.8.10}$$

Which region in $\left(\begin{smallmatrix} \mathfrak{r} \\ \phi \end{smallmatrix}\right)$-space is necessary to cover $\left(\begin{smallmatrix} \mathfrak{x} \\ \mathfrak{y} \end{smallmatrix}\right)$-space? Compute the Jacobian determinant of this transformation. Give an intuitive explanation in terms of local magnification factor of the formula you get. Finally compute the transformed density function.

• c. *1 point Compute* $\operatorname{cov}[\mathfrak{x}, \mathfrak{y}]$.

• d. *2 points Compute the conditional variance* $\operatorname{var}[\mathfrak{y}|\mathfrak{x}{=}x]$.

• e. *2 points Are $\mathfrak{x}$ and $\mathfrak{y}$ independent?*

Problem 151. [**Ame85**, pp. 296–7] *Assume three transportation choices are available: bus, train, and car. If you pick at random a neoclassical individual ω and ask him or her which utility this person derives from using bus, train, and car, the answer will be three numbers $\mathfrak{u}_1(\omega), \mathfrak{u}_2(\omega), \mathfrak{u}_3(\omega)$. Here $\mathfrak{u}_1$, $\mathfrak{u}_2$, and $\mathfrak{u}_3$ are assumed to be* independent *random variables with the following cumulative distribution functions:*

$$\Pr[\mathfrak{u}_i \leq u] = F_i(u) = \exp\bigl(-\exp(\mu_i - u)\bigr), \qquad i = 1, 2, 3. \tag{8.8.11}$$

*I.e., the functional form is the same for all three transportation choices (*exp *indicates the exponential function); the F_i only differ by the parameters μ_i. These probability distributions are called Type I extreme value distributions, or log Weibull distributions.*

Often these kinds of models are set up in such a way that these μ_i to depend on the income etc. of the individual, but we assume for this exercise that this distribution applies to the population as a whole.

• a. *1 point Show that the F_i are indeed cumulative distribution functions, and derive the density functions $f_i(u)$.*

Individual ω likes cars best if and only if his utilities satisfy $\mathfrak{u}_3(\omega) \geq \mathfrak{u}_1(\omega)$ and $\mathfrak{u}_3(\omega) \geq \mathfrak{u}_2(\omega)$. Let I be a function of three arguments such that $I(\mathfrak{u}_1, \mathfrak{u}_2, \mathfrak{u}_3)$ is the indicator function of the event that one randomly chooses an individual ω who likes cars best, i.e.,

$$I(u_1, u_2, u_3) = \begin{cases} 1 & \textit{if } u_1 \leq u_3 \textit{ and } u_2 \leq u_3 \\ 0 & \textit{otherwise.} \end{cases} \tag{8.8.12}$$

Then $\Pr[car] = \mathrm{E}[I(\mathfrak{u}_1, \mathfrak{u}_2, \mathfrak{u}_3)]$. *The following steps have the purpose to compute this probability:*

• b. *2 points For any fixed number u, define $g(u) = \mathrm{E}[I(\mathfrak{u}_1, \mathfrak{u}_2, \mathfrak{u}_3)|\mathfrak{u}_3 = u]$. Show that*

$$g(u) = \exp\bigl(-\exp(\mu_1 - u) - \exp(\mu_2 - u)\bigr). \tag{8.8.13}$$

• c. *2 points This here is merely the evaluation of an integral. Show that*

$$\int_{-\infty}^{+\infty} \exp\bigl(-\exp(\mu_1 - u) - \exp(\mu_2 - u) - \exp(\mu_3 - u)\bigr) \exp(\mu_3 - u)\, du =$$
$$= \frac{\exp \mu_3}{\exp \mu_1 + \exp \mu_2 + \exp \mu_3}.$$

Hint: use substitution rule with $y = -\exp(\mu_1 - u) - \exp(\mu_2 - u) - \exp(\mu_3 - u)$.

• d. *1 point Use b and c to show that*

$$\Pr[car] = \frac{\exp \mu_3}{\exp \mu_1 + \exp \mu_2 + \exp \mu_3}. \tag{8.8.14}$$

CHAPTER 9

Random Matrices

The step from random vectors to random matrices (and higher order random arrays) is not as big as the step from individual random variables to random vectors. We will first give a few quite trivial verifications that the expected value operator is indeed a linear operator, and them make some not quite as trivial observations about the expected values and higher moments of quadratic forms.

9.1. Linearity of Expected Values

Definition 9.1.1. Let $\mathbf{Z}$ be a random matrix with elements z_{ij}. Then $\mathcal{E}[\mathbf{Z}]$ is the matrix with elements $\mathrm{E}[\mathrm{z}_{ij}]$.

Theorem 9.1.2. *If $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$ are constant matrices, then $\mathcal{E}[\boldsymbol{A}\mathbf{Z}\boldsymbol{B}+\boldsymbol{C}] = \boldsymbol{A}\,\mathcal{E}[\mathbf{Z}]\boldsymbol{B}+\boldsymbol{C}$.*

Proof by multiplying out.

Theorem 9.1.3. $\mathcal{E}[\mathbf{Z}^\top] = (\mathcal{E}[\mathbf{Z}])^\top$; $\mathcal{E}[\operatorname{tr}\mathbf{Z}] = \operatorname{tr}\mathcal{E}[\mathbf{Z}]$.

Theorem 9.1.4. *For partitioned matrices* $\mathcal{E}[\begin{bmatrix}\mathbf{X}\\ \mathbf{Y}\end{bmatrix}] = \begin{bmatrix}\mathcal{E}[\mathbf{X}]\\ \mathcal{E}[\mathbf{Y}]\end{bmatrix}$.

Special cases: If $\boldsymbol{C}$ is a constant, then $\mathcal{E}[\boldsymbol{C}] = \boldsymbol{C}$, $\mathcal{E}[\boldsymbol{A}\mathbf{X}+\boldsymbol{B}\mathbf{Y}] = \boldsymbol{A}\,\mathcal{E}[\mathbf{X}]+\boldsymbol{B}\,\mathcal{E}[\mathbf{Y}]$, and $\mathcal{E}[a\cdot\mathbf{X}+b\cdot\mathbf{Y}] = a\cdot\mathcal{E}[\mathbf{X}]+b\cdot\mathcal{E}[\mathbf{Y}]$.

If $\mathbf{X}$ and $\mathbf{Y}$ are random matrices, then the covariance of these two matrices is a four-way array containing the covariances of all elements of $\mathbf{X}$ with all elements of $\mathbf{Y}$. Certain conventions are necessary to arrange this four-way array in a two-dimensional scheme that can be written on a sheet of paper. Before we develop those, we will first define the covariance matrix for two random vectors.

Definition 9.1.5. The covariance matrix of two random vectors is defined as

$$\mathcal{C}[\mathbf{x},\mathbf{y}] = \mathcal{E}[(\mathbf{x}-\mathcal{E}[\mathbf{x}])(\mathbf{y}-\mathcal{E}[\mathbf{y}])^\top]. \tag{9.1.1}$$

Theorem 9.1.6. $\mathcal{C}[\mathbf{x},\mathbf{y}] = \mathcal{E}[\mathbf{x}\mathbf{y}^\top]-(\mathcal{E}[\mathbf{x}])(\mathcal{E}[\mathbf{y}])^\top$.

Theorem 9.1.7. $\mathcal{C}[\boldsymbol{A}\mathbf{x}+\boldsymbol{b},\boldsymbol{C}\mathbf{y}+\boldsymbol{d}] = \boldsymbol{A}\,\mathcal{C}[\mathbf{x},\mathbf{y}]\boldsymbol{C}^\top$.

Problem 152. *Prove theorem 9.1.7.*

Theorem 9.1.8. $\mathcal{C}[\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix},\begin{bmatrix}\mathbf{u}\\ \boldsymbol{v}\end{bmatrix}] = \begin{bmatrix}\mathcal{C}[\mathbf{x},\mathbf{u}] & \mathcal{C}[\mathbf{x},\boldsymbol{v}]\\ \mathcal{C}[\mathbf{y},\mathbf{u}] & \mathcal{C}[\mathbf{y},\boldsymbol{v}]\end{bmatrix}$.

Special case: $\mathcal{C}[\boldsymbol{A}\mathbf{x}+\boldsymbol{B}\mathbf{y},\boldsymbol{C}\mathbf{u}+\boldsymbol{D}\boldsymbol{v}] = \boldsymbol{A}\,\mathcal{C}[\mathbf{x},\mathbf{u}]\boldsymbol{C}^\top+\boldsymbol{A}\,\mathcal{C}[\mathbf{x},\boldsymbol{v}]\boldsymbol{D}^\top+\boldsymbol{B}\,\mathcal{C}[\mathbf{y},\mathbf{u}]\boldsymbol{C}^\top+\boldsymbol{B}\,\mathcal{C}[\mathbf{y},\boldsymbol{v}]\boldsymbol{D}^\top$. To show this, express each of the arguments as a partitioned matrix, then use theorem 9.1.7.

Definition 9.1.9. $\mathcal{V}[\mathbf{x}] = \mathcal{C}[\mathbf{x},\mathbf{x}]$ is called the dispersion matrix.

It follows from theorem 9.1.8 that

$$\mathcal{V}[\mathbf{x}] = \begin{bmatrix}\operatorname{var}[\mathrm{x}_1] & \operatorname{cov}[\mathrm{x}_1,\mathrm{x}_2] & \cdots & \operatorname{cov}[\mathrm{x}_1,\mathrm{x}_n]\\ \operatorname{cov}[\mathrm{x}_2,\mathrm{x}_1] & \operatorname{var}[\mathrm{x}_2] & \cdots & \operatorname{cov}[\mathrm{x}_2,\mathrm{x}_n]\\ \vdots & \vdots & \ddots & \vdots\\ \operatorname{cov}[\mathrm{x}_n,\mathrm{x}_1] & \operatorname{cov}[\mathrm{x}_n,\mathrm{x}_2] & \cdots & \operatorname{var}[\mathrm{x}_n]\end{bmatrix} \tag{9.1.2}$$

Theorem 9.1.10. $\mathcal{V}[\boldsymbol{A}\mathbf{x}] = \boldsymbol{A}\,\mathcal{V}[\mathbf{x}]\boldsymbol{A}^\top$.

From this follows that $\mathcal{V}[\mathbf{x}]$ is nonnegative definite (or, as it is also called, positive semidefinite).

Problem 153. *Assume $\mathbf{y}$ is a random vector, and $\operatorname{var}[\mathrm{y}_i]$ exists for every component y_i. Then the whole dispersion matrix $\mathcal{V}[\mathbf{y}]$ exists.*

Theorem 9.1.11. *$\mathcal{V}[\mathbf{x}]$ is singular if and only if a vector $\boldsymbol{a}$ exists so that $\boldsymbol{a}^\top\mathbf{x}$ is almost surely a constant.*

Proof: Call $\mathcal{V}[\mathbf{x}] = \boldsymbol{\Sigma}$. Then $\boldsymbol{\Sigma}$ singular iff an $\boldsymbol{a}$ exists with $\boldsymbol{\Sigma}\boldsymbol{a} = \boldsymbol{o}$ iff an $\boldsymbol{a}$ exists with $\boldsymbol{a}^\top\boldsymbol{\Sigma}\boldsymbol{a} = \operatorname{var}[\boldsymbol{a}^\top\mathbf{x}] = 0$ iff an $\boldsymbol{a}$ exists so that $\boldsymbol{a}^\top\mathbf{x}$ is almost surely a constant.

This means, singular random variables have a restricted range, their values are contained in a linear subspace. This has relevance for estimators involving singular random variables: two such estimators (i.e., functions of a singular random variable) should still be considered the same if their values coincide in that subspace in which the values of the random variable is concentrated—even if elsewhere their values differ.

Problem 154. [**Seb77**, exercise 1a–3 on p. 13] *Let $\mathbf{x} = [\mathrm{x}_1,\ldots,\mathrm{x}_n]^\top$ be a vector of random variables, and let $\mathrm{y}_1 = \mathrm{x}_1$ and $\mathrm{y}_i = \mathrm{x}_i - \mathrm{x}_{i-1}$ for $i = 2,3,\ldots,n$. What must the dispersion matrix $\mathcal{V}[\mathbf{x}]$ be so that the y_i are uncorrelated with each other and each have unit variance?*

ANSWER. $\operatorname{cov}[\mathsf{x}_i, \mathsf{x}_j] = \min(i, j)$.

$$\mathsf{y} = \boldsymbol{A}\mathsf{x} \qquad \text{with} \qquad \boldsymbol{A} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & -1 & 1 \end{bmatrix}$$

$$\boldsymbol{A}^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \qquad \text{and} \qquad \boldsymbol{A}^{-1}(\boldsymbol{A}^{-1})^\top = (\boldsymbol{A}^\top\boldsymbol{A})^{-1} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 & 2 \\ 1 & 2 & 3 & 3 & 3 \\ 1 & 2 & 3 & 4 & 4 \\ 1 & 2 & 3 & 4 & 5 \end{bmatrix}$$

□

9.2. Means and Variances of Quadratic Forms in Random Matrices

9.2.1. Expected Value of Quadratic Form.

THEOREM 9.2.1. *Assume $\mathcal{E}[\mathsf{y}] = \boldsymbol{\eta}$, $\mathcal{V}[\mathsf{y}] = \sigma^2\boldsymbol{\Psi}$, and $\boldsymbol{A}$ is a matrix of constants. Then*

$$\mathrm{E}[\mathsf{y}^\top\boldsymbol{A}\mathsf{y}] = \sigma^2\operatorname{tr}(\boldsymbol{A}\boldsymbol{\Psi}) + \boldsymbol{\eta}^\top\boldsymbol{A}\boldsymbol{\eta}. \tag{9.2.1}$$

PROOF. Write y as the sum of $\boldsymbol{\eta}$ and $\boldsymbol{\varepsilon} = \mathsf{y} - \boldsymbol{\eta}$; then

$$\mathsf{y}^\top\boldsymbol{A}\mathsf{y} = (\boldsymbol{\varepsilon} + \boldsymbol{\eta})^\top\boldsymbol{A}(\boldsymbol{\varepsilon} + \boldsymbol{\eta}) \tag{9.2.2}$$

$$= \boldsymbol{\varepsilon}^\top\boldsymbol{A}\boldsymbol{\varepsilon} + \boldsymbol{\varepsilon}^\top\boldsymbol{A}\boldsymbol{\eta} + \boldsymbol{\eta}^\top\boldsymbol{A}\boldsymbol{\varepsilon} + \boldsymbol{\eta}^\top\boldsymbol{A}\boldsymbol{\eta} \tag{9.2.3}$$

$\boldsymbol{\eta}^\top\boldsymbol{A}\boldsymbol{\eta}$ is nonstochastic, and since $\mathcal{E}[\boldsymbol{\varepsilon}] = \boldsymbol{o}$ it follows

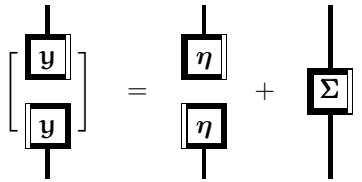
$$\mathrm{E}[\mathsf{y}^\top\boldsymbol{A}\mathsf{y}] - \boldsymbol{\eta}^\top\boldsymbol{A}\boldsymbol{\eta} = \mathrm{E}[\boldsymbol{\varepsilon}^\top\boldsymbol{A}\boldsymbol{\varepsilon}] \tag{9.2.4}$$

$$= \mathrm{E}[\operatorname{tr}(\boldsymbol{\varepsilon}^\top\boldsymbol{A}\boldsymbol{\varepsilon})] = \mathrm{E}[\operatorname{tr}(\boldsymbol{A}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top)] = \operatorname{tr}(\boldsymbol{A}\,\mathcal{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top]) \tag{9.2.5}$$

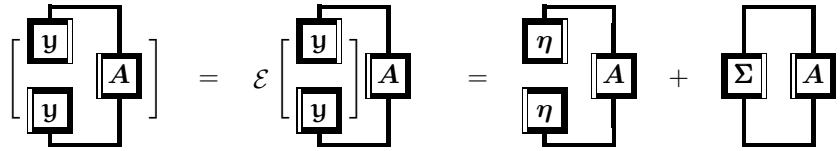
$$= \sigma^2\operatorname{tr}(\boldsymbol{A}\boldsymbol{\Psi}). \tag{9.2.6}$$

Here we used that $\operatorname{tr}(\boldsymbol{AB}) = \operatorname{tr}(\boldsymbol{BA})$ and, if c is a scalar, i.e., a 1×1 matrix, then $\operatorname{tr}(c) = c$. □

In tile notation (see Appendix B), the proof of theorem 9.2.1 is much more straightforward and no longer seems to rely on "tricks." From $\mathsf{y} \sim (\boldsymbol{\eta}, \boldsymbol{\Sigma})$, i.e., we are writing now $\sigma^2\boldsymbol{\Psi} = \boldsymbol{\Sigma}$, follows $\mathcal{E}[\mathsf{y}\mathsf{y}^\top] = \boldsymbol{\eta}\boldsymbol{\eta}^\top + \boldsymbol{\Sigma}$, therefore

(9.2.7) 

; therefore

(9.2.8) 

.

PROBLEM 155. [**Seb77**, Exercise 1b–2 on p. 16] *If $\mathsf{y}_1, \mathsf{y}_2, \dots, \mathsf{y}_n$ are mutually independent random variables with commom mean η, and with variances $\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2$, respectively, prove that*

$$\frac{1}{n(n-1)}\sum_i(\mathsf{y}_i - \bar{\mathsf{y}})^2 \tag{9.2.9}$$

is an unbiased estimator of $\operatorname{var}[\bar{\mathsf{y}}]$. *It is recommended to use theorem 9.2.1 for this.*

ANSWER. Write $\mathsf{y} = \begin{bmatrix}\mathsf{y}_1 & \mathsf{y}_2 & \dots & \mathsf{y}_n\end{bmatrix}^\top$ and $\boldsymbol{\Sigma} = \operatorname{diag}(\begin{bmatrix}\sigma_1^2 & \sigma_2^2 & \dots & \sigma_n^2\end{bmatrix})$. Then the vector $\begin{bmatrix}\mathsf{y}_1 - \bar{\mathsf{y}} & \mathsf{y}_2 - \bar{\mathsf{y}} & \dots & \mathsf{y}_n - \bar{\mathsf{y}}\end{bmatrix}^\top$ can be written as $(\boldsymbol{I} - \frac{1}{n}\boldsymbol{\iota}\boldsymbol{\iota}^\top)\mathsf{y}$. $\frac{1}{n}\boldsymbol{\iota}\boldsymbol{\iota}^\top$ is idempotent, therefore $\boldsymbol{D} = \boldsymbol{I} - \frac{1}{n}\boldsymbol{\iota}\boldsymbol{\iota}^\top$ is idempotent too. Our estimator is $\frac{1}{n(n-1)}\mathsf{y}^\top\boldsymbol{D}\mathsf{y}$, and since the mean vector $\boldsymbol{\eta} = \boldsymbol{\iota}\eta$ satisfies $\boldsymbol{D}\boldsymbol{\eta} = \boldsymbol{o}$, theorem 9.2.1 gives

$$\mathrm{E}[\mathsf{y}^\top\boldsymbol{D}\mathsf{y}] = \operatorname{tr}[\boldsymbol{D}\boldsymbol{\Sigma}] = \operatorname{tr}[\boldsymbol{\Sigma}] - \frac{1}{n}\operatorname{tr}[\boldsymbol{\iota}\boldsymbol{\iota}^\top\boldsymbol{\Sigma}] \tag{9.2.10}$$

$$= (\sigma_1^2 + \dots + \sigma_n^2) - \frac{1}{n}\operatorname{tr}[\boldsymbol{\iota}^\top\boldsymbol{\Sigma}\boldsymbol{\iota}] \tag{9.2.11}$$

$$= \frac{n-1}{n}(\sigma_1^2 + \dots + \sigma_n^2). \tag{9.2.12}$$

Divide this by $n(n-1)$ to get $(\sigma_1^2 + \dots + \sigma_n^2)/n^2$, which is $\operatorname{var}[\bar{\mathsf{y}}]$, as claimed. □

For the variances of quadratic forms we need the third and fourth moments of the underlying random variables.

PROBLEM 156. *Let $\mu_i = \mathrm{E}[(\mathsf{y} - \mathrm{E}[\mathsf{y}])^i]$ be the ith centered moment of y, and let $\sigma = \sqrt{\mu_2}$ be its standard deviation. Then the skewness is defined as $\gamma_1 = \mu_3/\sigma^3$, and kurtosis is $\gamma_2 = (\mu_4/\sigma^4) - 3$. Show that skewness and kurtosis of $a\mathsf{y} + b$ are equal to those of y if $a > 0$; for $a < 0$ the skewness changes its sign. Show that skewness γ_1 and kurtosis γ_2 always satisfy*

$$\gamma_1^2 \le \gamma_2 + 2. \tag{9.2.13}$$

ANSWER. Define $\varepsilon = \mathsf{y} - \mu$, and apply Cauchy-Schwartz for the variables ε and ε^2:

$$(\sigma^3\gamma_1)^2 = (\mathrm{E}[\varepsilon^3])^2 = \left(\operatorname{cov}[\varepsilon, \varepsilon^2]\right)^2 \le \operatorname{var}[\varepsilon]\operatorname{var}[\varepsilon^2] = \sigma^6(\gamma_2 + 2) \tag{9.2.14}$$

□

PROBLEM 157. *Show that any real numbers γ_1 and γ_2 satisfying* (9.2.13) *can be the skewness and kurtosis of a random variable.*

Answer. To show that all combinations satisfying this inequality are possible, define

$$r = \sqrt{\gamma_2 + 3 - 3\gamma_1^2/4} \qquad a = r + \gamma_1/2 \qquad b = r - \gamma_1/2$$

and construct a random variable x which assumes the following three values:

$$\mathrm{x} = \begin{cases} a & \text{with probability } 1/2ar \\ 0 & \text{with probability } 1/(\gamma_2 + 3 - \gamma_1^2), \\ -b & \text{with probability } 1/2br \end{cases} \tag{9.2.15}$$

This variable has expected value zero, variance 1, its third moment is γ_1, and its fourth moment $\gamma_2 + 3$.

□

Theorem 9.2.2. *Given a random vector $\boldsymbol{\varepsilon}$ of independent variables ε_i with zero expected value $\mathrm{E}[\varepsilon_i] = 0$, and whose second and third moments are identical. Call $\mathrm{var}[\varepsilon_i] = \sigma^2$, and $\mathrm{E}[\varepsilon_i^3] = \sigma^3\gamma_1$ (where σ is the positive square root of σ^2). Here γ_1 is called the skewness of these variables. Then the following holds for the third mixed moments:*

$$\mathrm{E}[\varepsilon_i\varepsilon_j\varepsilon_k] = \begin{cases} \sigma^3\gamma_1 & \text{if } i = j = k \\ 0 & \text{otherwise} \end{cases} \tag{9.2.16}$$

and from (9.2.16) *follows that for any $n \times 1$ vector $\boldsymbol{a}$ and symmetric $n \times n$ matrices $\boldsymbol{C}$ whose vector of diagonal elements is $\boldsymbol{c}$,*

$$\mathrm{E}[(\boldsymbol{a}^\top\boldsymbol{\varepsilon})(\boldsymbol{\varepsilon}^\top\boldsymbol{C}\boldsymbol{\varepsilon})] = \sigma^3\gamma_1\boldsymbol{a}^\top\boldsymbol{c}. \tag{9.2.17}$$

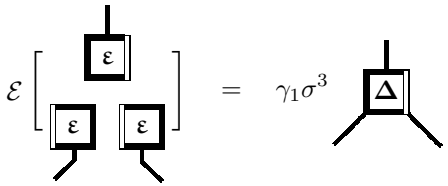
Proof. If $i \neq j \neq k \neq i$, then $\mathrm{E}[\varepsilon_i\varepsilon_j\varepsilon_k] = 0 \cdot 0 \cdot 0 = 0$; if $i = j \neq k$ then $\mathrm{E}[\varepsilon_i\varepsilon_j\varepsilon_k] = \sigma^2 \cdot 0 = 0$, same for $i \neq j = k$ and $j \neq i = k$. Therefore only $\mathrm{E}[\varepsilon_i^3]$ remains, which proves (9.2.16). Now

$$(\boldsymbol{a}^\top\boldsymbol{\varepsilon})(\boldsymbol{\varepsilon}^\top\boldsymbol{C}\boldsymbol{\varepsilon}) = \sum_{i,j,k} a_i c_{jk}\varepsilon_i\varepsilon_j\varepsilon_k \tag{9.2.18}$$

$$\mathrm{E}[\boldsymbol{a}^\top\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\boldsymbol{C}\boldsymbol{\varepsilon}] = \sigma^3\gamma_1\sum_i a_i c_{ii} = \sigma^3\gamma_1\boldsymbol{a}^\top\boldsymbol{c}. \tag{9.2.19}$$

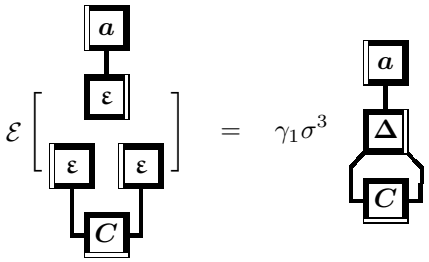
One would like to have a matrix notation for (9.2.16) from which (9.2.17) follows by a trivial operation. This is not easily possible in the usual notation, but it is possible in tile notation:

(9.2.20)



Therefore

(9.2.21)



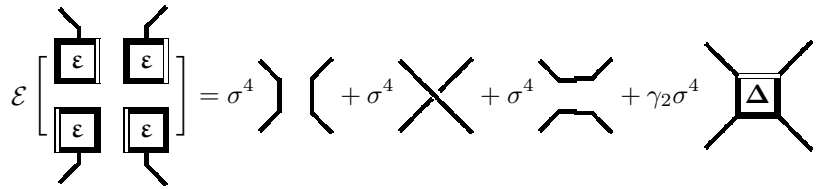
Since n is the vector of diagonal elements of $\boldsymbol{C}$, called $\boldsymbol{c}$, the last term in equation (9.2.21) is the scalar product $\boldsymbol{a}^\top\boldsymbol{c}$. □

Given a random vector $\boldsymbol{\varepsilon}$ of independent variables ε_i with zero expected value $\mathrm{E}[\varepsilon_i] = 0$ and identical second and fourth moments. Call $\mathrm{var}[\varepsilon_i] = \sigma^2$ and $\mathrm{E}[\varepsilon_i^4] = \sigma^4(\gamma_2+3)$, where γ_2 is the kurtosis. Then the following holds for the fourth moments:

$$\mathrm{E}[\varepsilon_i\varepsilon_j\varepsilon_k\varepsilon_l] = \begin{cases} \sigma^4(\gamma_2 + 3) & \text{if } i = j = k = l \\ \sigma^4 & \text{if } i = j \neq k = l \text{ or } i = k \neq j = l \\ & \quad \text{or } i = l \neq j = k \\ 0 & \text{otherwise.} \end{cases} \tag{9.2.22}$$

It is not an accident that (9.2.22) is given element by element and not in matrix notation. It is not possible to do this, not even with the Kronecker product. But it is easy in tile notation:

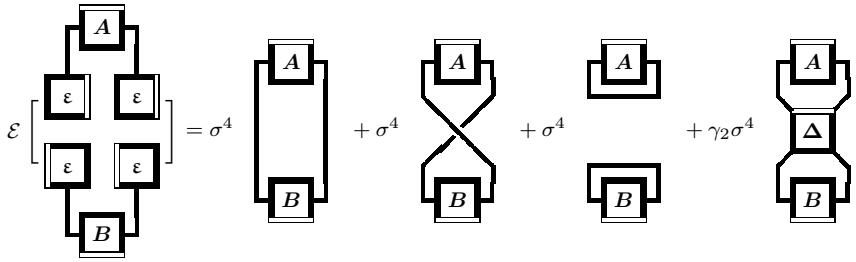
(9.2.23)



Problem 158. [**Seb77**, pp. 14–16 and 52] *Show that for any symmetric $n \times n$ matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, whose vectors of diagonal elements are $\boldsymbol{a}$ and $\boldsymbol{b}$,*

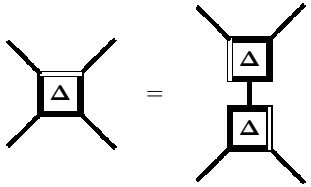
$$\mathrm{E}[(\boldsymbol{\varepsilon}^\top\boldsymbol{A}\boldsymbol{\varepsilon})(\boldsymbol{\varepsilon}^\top\boldsymbol{B}\boldsymbol{\varepsilon})] = \sigma^4\Big(\operatorname{tr}\boldsymbol{A}\operatorname{tr}\boldsymbol{B} + 2\operatorname{tr}(\boldsymbol{A}\boldsymbol{B}) + \gamma_2\boldsymbol{a}^\top\boldsymbol{b}\Big). \tag{9.2.24}$$

Answer. (9.2.24) is an immediate consequence of (9.2.23); this step is now trivial due to linearity of the expected value:



The first term is $\operatorname{tr} \boldsymbol{AB}$. The second is $\operatorname{tr} \boldsymbol{AB}^\top$, but since $\boldsymbol{A}$ and $\boldsymbol{B}$ are symmetric, this is equal to $\operatorname{tr} \boldsymbol{AB}$. The third term is $\operatorname{tr} \boldsymbol{A} \operatorname{tr} \boldsymbol{B}$. What is the fourth term? Diagonal arrays exist with any number of arms, and any connected concatenation of diagonal arrays is again a diagonal array, see (B.2.1). For instance,

(9.2.25)



From this together with (B.1.4) one can see that the fourth term is the scalar product of the diagonal vectors of $\boldsymbol{A}$ and $\boldsymbol{B}$. □

Problem 159. *Under the conditions of equation* (9.2.23) *show that*

$$\operatorname{cov}[\boldsymbol{\varepsilon}^\top \boldsymbol{C}\boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}^\top \boldsymbol{D}\boldsymbol{\varepsilon}] = \sigma^4 \gamma_2 \boldsymbol{c}^\top \boldsymbol{d} + 2\sigma^4 \operatorname{tr}(\boldsymbol{CD}). \tag{9.2.26}$$

Answer. Use $\operatorname{cov}[\boldsymbol{\varepsilon}^\top \boldsymbol{C}\boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}^\top \boldsymbol{D}\boldsymbol{\varepsilon}] = \operatorname{E}[(\boldsymbol{\varepsilon}^\top \boldsymbol{C}\boldsymbol{\varepsilon})(\boldsymbol{\varepsilon}^\top \boldsymbol{D}\boldsymbol{\varepsilon})] - \operatorname{E}[\boldsymbol{\varepsilon}^\top \boldsymbol{C}\boldsymbol{\varepsilon}] \operatorname{E}[\boldsymbol{\varepsilon}^\top \boldsymbol{D}\boldsymbol{\varepsilon}]$. The first term is given by (9.2.24). The second term is $\sigma^4 \operatorname{tr} \boldsymbol{C} \operatorname{tr} \boldsymbol{D}$, according to (9.2.1). □

Problem 160. *(Not eligible for in-class exams) Take any symmetric matrix* $\boldsymbol{A}$ *and denote the vector of diagonal elements by* $\boldsymbol{a}$. *Let* $\boldsymbol{x} = \boldsymbol{\theta} + \boldsymbol{\varepsilon}$ *where* $\boldsymbol{\varepsilon}$ *satisfies the conditions of theorem 9.2.2 and equation* (9.2.23). *Then*

$$\operatorname{var}[\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}] = 4\sigma^2 \boldsymbol{\theta}^\top \boldsymbol{A}^2 \boldsymbol{\theta} + 4\sigma^3 \gamma_1 \boldsymbol{\theta}^\top \boldsymbol{A}\boldsymbol{a} + \sigma^4 \Big(\gamma_2 \boldsymbol{a}^\top \boldsymbol{a} + 2 \operatorname{tr}(\boldsymbol{A}^2)\Big). \tag{9.2.27}$$

Answer. Proof: $\operatorname{var}[\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}] = \operatorname{E}[(\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x})^2] - (\operatorname{E}[\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}])^2$. Since by assumption $\mathcal{V}[\boldsymbol{x}] = \sigma^2 \boldsymbol{I}$, the second term is, by theorem 9.2.1, $(\sigma^2 \operatorname{tr} \boldsymbol{A} + \boldsymbol{\theta}^\top \boldsymbol{A}\boldsymbol{\theta})^2$. Now look at first term. Again using the notation $\boldsymbol{\varepsilon} = \boldsymbol{x} - \boldsymbol{\theta}$ it follows from (9.2.3) that

$$(\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x})^2 = (\boldsymbol{\varepsilon}^\top \boldsymbol{A}\boldsymbol{\varepsilon})^2 + 4(\boldsymbol{\theta}^\top \boldsymbol{A}\boldsymbol{\varepsilon})^2 + (\boldsymbol{\theta}^\top \boldsymbol{A}\boldsymbol{\theta})^2 \tag{9.2.28}$$

$$+ 4\, \boldsymbol{\varepsilon}^\top \boldsymbol{A}\boldsymbol{\varepsilon}\, \boldsymbol{\theta}^\top \boldsymbol{A}\boldsymbol{\varepsilon} + 2\, \boldsymbol{\varepsilon}^\top \boldsymbol{A}\boldsymbol{\varepsilon}\, \boldsymbol{\theta}^\top \boldsymbol{A}\boldsymbol{\theta} + 4\, \boldsymbol{\theta}^\top \boldsymbol{A}\boldsymbol{\varepsilon}\, \boldsymbol{\theta}^\top \boldsymbol{A}\boldsymbol{\theta}. \tag{9.2.29}$$

We will take expectations of these terms one by one. Use (9.2.24) for first term:

$$(\boldsymbol{\varepsilon}^\top \boldsymbol{A}\boldsymbol{\varepsilon})^2 = \sigma^4 \Big(\gamma_2 \boldsymbol{a}^\top \boldsymbol{a} + (\operatorname{tr} \boldsymbol{A})^2 + 2 \operatorname{tr}(\boldsymbol{A}^2)\Big). \tag{9.2.30}$$

To deal with the second term in (9.2.29) define $\boldsymbol{b} = \boldsymbol{A}\boldsymbol{\theta}$; then

$$(\boldsymbol{\theta}^\top \boldsymbol{A}\boldsymbol{\varepsilon})^2 = (\boldsymbol{b}^\top \boldsymbol{\varepsilon})^2 = \boldsymbol{b}^\top \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top \boldsymbol{b} = \operatorname{tr}(\boldsymbol{b}^\top \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top \boldsymbol{b}) = \operatorname{tr}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top \boldsymbol{b}\boldsymbol{b}^\top) \tag{9.2.31}$$

$$\operatorname{E}[(\boldsymbol{\theta}^\top \boldsymbol{A}\boldsymbol{\varepsilon})^2] = \sigma^2 \operatorname{tr}(\boldsymbol{b}\boldsymbol{b}^\top) = \sigma^2 \boldsymbol{b}^\top \boldsymbol{b} = \sigma^2 \boldsymbol{\theta}^\top \boldsymbol{A}^2 \boldsymbol{\theta} \tag{9.2.32}$$

The third term is a constant which remains as it is; for the fourth term use (9.2.17)

$$\boldsymbol{\varepsilon}^\top \boldsymbol{A}\boldsymbol{\varepsilon}\, \boldsymbol{\theta}^\top \boldsymbol{A}\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}^\top \boldsymbol{A}\boldsymbol{\varepsilon}\, \boldsymbol{b}^\top \boldsymbol{\varepsilon} \tag{9.2.33}$$

$$\operatorname{E}[\boldsymbol{\varepsilon}^\top \boldsymbol{A}\boldsymbol{\varepsilon}\, \boldsymbol{\theta}^\top \boldsymbol{A}\boldsymbol{\varepsilon}] = \sigma^3 \gamma_1 \boldsymbol{a}^\top \boldsymbol{b} = \sigma^3 \gamma_1 \boldsymbol{a}^\top \boldsymbol{A}\boldsymbol{\theta} \tag{9.2.34}$$

If one takes expected values, the fifth term becomes $2\sigma^2 \operatorname{tr}(\boldsymbol{A})\, \boldsymbol{\theta}^\top \boldsymbol{A}\boldsymbol{\theta}$, and the last term falls away. Putting the pieces together the statement follows. □

CHAPTER 10

The Multivariate Normal Probability Distribution

10.1. More About the Univariate Case

By definition, z is a *standard* normal variable, in symbols, $\mathsf{z} \sim \mathsf{N}(0,1)$, if it has the density function

$$f_{\mathsf{z}}(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}. \tag{10.1.1}$$

To verify that this is a density function we have to check two conditions. (1) It is everywhere nonnegative. (2) Its integral from $-\infty$ to ∞ is 1. In order to evaluate this integral, it is easier to work with the independent product of two standard normal variables x and y; their joint density function is $f_{\mathsf{x},\mathsf{y}}(x,y) = \frac{1}{2\pi} e^{-\frac{x^2+y^2}{2}}$. In order to see that this joint density integrates to 1, go over to polar coordinates $x = r\cos\phi$, $y = r\sin\phi$, i.e., compute the joint distribution of r and ϕ from that of x and y: the absolute value of the Jacobian determinant is r, i.e., $dx\,dy = r\,dr\,d\phi$, therefore

$$\int_{y=-\infty}^{y=\infty} \int_{x=-\infty}^{x=\infty} \frac{1}{2\pi} e^{-\frac{x^2+y^2}{2}}\,dx\,dy = \int_{\phi=0}^{2\pi} \int_{r=0}^{\infty} \frac{1}{2\pi} e^{-\frac{r^2}{2}} r\,dr\,d\phi. \tag{10.1.2}$$

By substituting $t = r^2/2$, therefore $dt = r\,dr$, the inner integral becomes $-\frac{1}{2\pi} e^{-t}\Big|_0^\infty = \frac{1}{2\pi}$; therefore the whole integral is 1. Therefore the product of the integrals of the marginal densities is 1, and since each such marginal integral is positive and they are equal, each of the marginal integrals is 1 too.

Problem 161. *6 points The Gamma function can be defined as* $\Gamma(r) = \int_0^\infty x^{r-1} e^{-x}\,dx$ *Show that* $\Gamma(\frac{1}{2}) = \sqrt{\pi}$. *(Hint: after substituting* $r = 1/2$*, apply the variable transformation* $x = z^2/2$ *for nonnegative* x *and* z *only, and then reduce the resulting integral to the integral over the normal density function.)*

Answer. Then $dx = z\,dz$, $\frac{dx}{\sqrt{x}} = dz\sqrt{2}$. Therefore one can reduce it to the integral over the normal density:

$$\int_0^\infty \frac{1}{\sqrt{x}} e^{-x}\,dx = \sqrt{2}\int_0^\infty e^{-z^2/2}\,dz = \frac{1}{\sqrt{2}}\int_{-\infty}^{\infty} e^{-z^2/2}\,dz = \frac{\sqrt{2\pi}}{\sqrt{2}} = \sqrt{\pi}. \tag{10.1.3}$$

□

A univariate normal variable with mean μ and variance σ^2 is a variable x whose standardized version $\mathsf{z} = \frac{\mathsf{x}-\mu}{\sigma} \sim \mathsf{N}(0,1)$. In this transformation from x to z, the Jacobian determinant is $\frac{dz}{dx} = \frac{1}{\sigma}$; therefore the density function of $\mathsf{x} \sim \mathsf{N}(\mu,\sigma^2)$ is (two notations, the second is perhaps more modern:)

$$f_{\mathsf{x}}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = (2\pi\sigma^2)^{-1/2} \exp\bigl(-(x-\mu)^2/2\sigma^2\bigr). \tag{10.1.4}$$

Problem 162. *3 points Given* n *independent observations of a Normally distributed variable* $\mathsf{y} \sim \mathsf{N}(\mu,1)$. *Show that the sample mean* $\bar{\mathsf{y}}$ *is a sufficient statistic for* μ. *Here is a formulation of the factorization theorem for sufficient statistics, which you will need for this question: Given a family of probability densities* $f_{\mathsf{y}}(y_1,\ldots,y_n;\theta)$ *defined on* $\mathbb{R}^n$*, which depend on a parameter* $\theta \in \Theta$. *The statistic* $T\colon \mathbb{R}^n \to \mathbb{R}$, $y_1,\ldots,y_n \mapsto T(y_1,\ldots,y_n)$ *is sufficient for parameter* θ *if and only if there exists a function of two variables* $g\colon \mathbb{R}\times\Theta \to \mathbb{R}$, $t,\theta \mapsto g(t;\theta)$*, and a function of* n *variables* $h\colon \mathbb{R}^n \to \mathbb{R}$, $y_1,\ldots,y_n \mapsto h(y_1,\ldots,y_n)$ *so that*

$$f_{\mathsf{y}}(y_1,\ldots,y_n;\theta) = g\bigl(T(y_1,\ldots,y_n);\theta\bigr)\cdot h(y_1,\ldots,y_n). \tag{10.1.5}$$

Answer. The joint density function can be written (factorization indicated by $\cdot$):

$$(2\pi)^{-n/2}\exp\Bigl(-\frac{1}{2}\sum_{i=1}^n (y_i-\mu)^2\Bigr) = (2\pi)^{-n/2}\exp\Bigl(-\frac{1}{2}\sum_{i=1}^n (y_i-\bar{y})^2\Bigr)\cdot\exp\Bigl(-\frac{n}{2}(\bar{y}-\mu)^2\Bigr) = h(y_1,\ldots,y_n)\cdot g(\bar{y};\mu). \tag{10.1.6}$$

□

10.2. Definition of Multivariate Normal

The multivariate normal distribution is an important family of distributions with very nice properties. But one must be a little careful how to define it. One might naively think a multivariate Normal is a vector random variable each component of which is univariate Normal. But this is not the right definition. Normality of the components is a necessary but not sufficient condition for a multivariate normal vector. If $\mathbf{u} = \begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}$ with both $\mathbf{x}$ and $\mathbf{y}$ multivariate normal, $\mathbf{u}$ is not necessarily multivariate normal.

Here is a recursive definition from which one gets all multivariate normal distributions:

(1) The univariate standard normal z, considered as a vector with one component, is multivariate normal.

(2) If $\mathbf{x}$ and $\mathbf{y}$ are multivariate normal *and they are independent*, then $\mathbf{u} = \begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}$ is multivariate normal.

(3) If y is multivariate normal, and $\boldsymbol{A}$ a matrix of constants (which need not be square and is allowed to be singular), and $\boldsymbol{b}$ a vector of constants, then $\boldsymbol{A}\mathsf{y}+\boldsymbol{b}$ is multivariate normal. In words: A vector consisting of linear combinations of the same set of multivariate normal variables is again multivariate normal.

For simplicity we will go over now to the bivariate Normal distribution.

10.3. Special Case: Bivariate Normal

The following two simple rules allow to obtain all bivariate Normal random variables:

(1) If x and y are independent and each of them has a (univariate) normal distribution with mean 0 and the same variance σ^2, then they are bivariate normal. (They would be bivariate normal even if their variances were different and their means not zero, but for the calculations below we will use only this special case, which together with principle (2) is sufficient to get all bivariate normal distributions.)

(2) If $\boldsymbol{\mathsf{x}} = \begin{bmatrix}\mathsf{x}\\ \mathsf{y}\end{bmatrix}$ is bivariate normal and $\boldsymbol{P}$ is a 2×2 nonrandom matrix and $\boldsymbol{\mu}$ a nonrandom column vector with two elements, then $\boldsymbol{P}\boldsymbol{\mathsf{x}}+\boldsymbol{\mu}$ is bivariate normal as well.

All other properties of bivariate Normal variables can be derived from this.

First let us derive the density function of a bivariate Normal distribution. Write $\boldsymbol{\mathsf{x}} = \begin{bmatrix}\mathsf{x}\\ \mathsf{y}\end{bmatrix}$. x and y are independent $\mathsf{N}(0,\sigma^2)$. Therefore by principle (1) above the vector $\boldsymbol{\mathsf{x}}$ is bivariate normal. Take any *nonsingular* 2×2 matrix $\boldsymbol{P}$ and a 2 vector $\boldsymbol{\mu} = \begin{bmatrix}\mu\\ \nu\end{bmatrix}$, and define $\begin{bmatrix}\mathsf{u}\\ \mathsf{v}\end{bmatrix} = \boldsymbol{\mathsf{u}} = \boldsymbol{P}\boldsymbol{\mathsf{x}}+\boldsymbol{\mu}$. We need nonsingularity because otherwise the resulting variable would not have a bivariate density; its probability mass would be concentrated on one straight line in the two-dimensional plane. What is the joint density function of $\boldsymbol{\mathsf{u}}$? Since $\boldsymbol{P}$ is nonsingular, the transformation is on-to-one, therefore we can apply the transformation theorem for densities. Let us first write down the density function of $\boldsymbol{\mathsf{x}}$ which we know:

$$f_{\mathsf{x},\mathsf{y}}(x,y) = \frac{1}{2\pi\sigma^2}\exp\Big(-\frac{1}{2\sigma^2}(x^2+y^2)\Big). \tag{10.3.1}$$

For the next step, remember that we have to express the old variable in terms of the new one: $\boldsymbol{\mathsf{x}} = \boldsymbol{P}^{-1}(\boldsymbol{\mathsf{u}}-\boldsymbol{\mu})$. The Jacobian determinant is therefore $J = \det(\boldsymbol{P}^{-1})$. Also notice that, after the substitution $\begin{bmatrix}x\\ y\end{bmatrix} = \boldsymbol{P}^{-1}\begin{bmatrix}u-\mu\\ v-\nu\end{bmatrix}$, the exponent in the joint density function of x and y is $-\frac{1}{2\sigma^2}(x^2+y^2) = -\frac{1}{2\sigma^2}\begin{bmatrix}x\\ y\end{bmatrix}^\top\begin{bmatrix}x\\ y\end{bmatrix} =$

$-\frac{1}{2\sigma^2}\begin{bmatrix}u-\mu\\ v-\nu\end{bmatrix}^\top \boldsymbol{P}^{-1\top}\boldsymbol{P}^{-1}\begin{bmatrix}u-\mu\\ v-\nu\end{bmatrix}$. Therefore the transformation theorem of density functions gives

$$f_{\mathsf{u},\mathsf{v}}(u,v) = \frac{1}{2\pi\sigma^2}\left|\det(\boldsymbol{P}^{-1})\right|\exp\Big(-\frac{1}{2\sigma^2}\begin{bmatrix}u-\mu\\ v-\nu\end{bmatrix}^\top \boldsymbol{P}^{-1\top}\boldsymbol{P}^{-1}\begin{bmatrix}u-\mu\\ v-\nu\end{bmatrix}\Big). \tag{10.3.2}$$

This expression can be made nicer. Note that the covariance matrix of the transformed variables is $\mathcal{V}[\begin{bmatrix}\mathsf{u}\\ \mathsf{v}\end{bmatrix}] = \sigma^2\boldsymbol{P}\boldsymbol{P}^\top = \sigma^2\boldsymbol{\Psi}$, say. Since $\boldsymbol{P}^{-1\top}\boldsymbol{P}^{-1}\boldsymbol{P}\boldsymbol{P}^\top = \boldsymbol{I}$, it follows $\boldsymbol{P}^{-1\top}\boldsymbol{P}^{-1} = \boldsymbol{\Psi}^{-1}$ and $\left|\det(\boldsymbol{P}^{-1})\right| = 1/\sqrt{\det(\boldsymbol{\Psi})}$, therefore

$$f_{\mathsf{u},\mathsf{v}}(u,v) = \frac{1}{2\pi\sigma^2}\frac{1}{\sqrt{\det(\boldsymbol{\Psi})}}\exp\Big(-\frac{1}{2\sigma^2}\begin{bmatrix}u-\mu\\ v-\nu\end{bmatrix}^\top \boldsymbol{\Psi}^{-1}\begin{bmatrix}u-\mu\\ v-\nu\end{bmatrix}\Big). \tag{10.3.3}$$

This is the general formula for the density function of a bivariate normal with nonsingular covariance matrix $\sigma^2\boldsymbol{\Psi}$ and mean vector $\boldsymbol{\mu}$. One can also use the following notation which is valid for the multivariate Normal variable with n dimensions, with mean vector $\boldsymbol{\mu}$ and nonsingular covariance matrix $\sigma^2\boldsymbol{\Psi}$:

$$f_{\boldsymbol{\mathsf{x}}}(\boldsymbol{x}) = (2\pi\sigma^2)^{-n/2}(\det\boldsymbol{\Psi})^{-1/2}\exp\Big(-\frac{1}{2\sigma^2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Psi}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\Big). \tag{10.3.4}$$

Problem 163. *1 point Show that the matrix product of* $(\boldsymbol{P}^{-1})^\top\boldsymbol{P}^{-1}$ *and* $\boldsymbol{P}\boldsymbol{P}^\top$ *is the identity matrix.*

Problem 164. *3 points All vectors in this question are* $n\times 1$ *column vectors. Let* $\mathsf{y} = \boldsymbol{\alpha}+\boldsymbol{\varepsilon}$, *where* $\boldsymbol{\alpha}$ *is a vector of constants and* $\boldsymbol{\varepsilon}$ *is jointly normal with* $\mathcal{E}[\boldsymbol{\varepsilon}] = \boldsymbol{o}$. *Often, the covariance matrix* $\mathcal{V}[\boldsymbol{\varepsilon}]$ *is not given directly, but a* $n\times n$ *nonsingular matrix* $\boldsymbol{T}$ *is known which has the property that the covariance matrix of* $\boldsymbol{T}\boldsymbol{\varepsilon}$ *is* σ^2 *times the* $n\times n$ *unit matrix, i.e.,*

$$\mathcal{V}[\boldsymbol{T}\boldsymbol{\varepsilon}] = \sigma^2\boldsymbol{I}_n. \tag{10.3.5}$$

Show that in this case the density function of y *is*

$$f_{\mathsf{y}}(\boldsymbol{y}) = (2\pi\sigma^2)^{-n/2}\left|\det(\boldsymbol{T})\right|\exp\Big(-\frac{1}{2\sigma^2}\big(\boldsymbol{T}(\boldsymbol{y}-\boldsymbol{\alpha})\big)^\top\boldsymbol{T}(\boldsymbol{y}-\boldsymbol{\alpha})\Big). \tag{10.3.6}$$

Hint: define $\boldsymbol{z} = \boldsymbol{T}\boldsymbol{\varepsilon}$, *write down the density function of* $\boldsymbol{z}$, *and make a transformation between* $\boldsymbol{z}$ *and* y.

Answer. Since $\mathcal{E}[\boldsymbol{z}] = \boldsymbol{o}$ and $\mathcal{V}[\boldsymbol{z}] = \sigma^2\boldsymbol{I}_n$, its density function is $(2\pi\sigma^2)^{-n/2}\exp(-z^\top z/2\sigma^2)$. Now express $\boldsymbol{z}$, whose density we know, as a function of y, whose density function we want to know.

$\boldsymbol{z} = \boldsymbol{T}(\mathfrak{y} - \boldsymbol{\alpha})$ or

$$z_1 = t_{11}(y_1 - \alpha_1) + t_{12}(y_2 - \alpha_2) + \cdots + t_{1n}(y_n - \alpha_n) \tag{10.3.7}$$

$$\vdots \tag{10.3.8}$$

$$z_n = t_{n1}(y_1 - \alpha_1) + t_{n2}(y_2 - \alpha_2) + \cdots + t_{nn}(y_n - \alpha_n) \tag{10.3.9}$$

therefore the Jacobian determinant is $\det(\boldsymbol{T})$. This gives the result.

□

10.3.1. Most Natural Form of Bivariate Normal Density.

PROBLEM 165. *In this exercise we will write the bivariate normal density in its most natural form. For this we set the multiplicative "nuisance parameter" $\sigma^2 = 1$, i.e., write the covariance matrix as $\boldsymbol{\Psi}$ instead of $\sigma^2\boldsymbol{\Psi}$.*

• a. *1 point Write the covariance matrix $\boldsymbol{\Psi} = \mathcal{V}[\begin{bmatrix}\mathfrak{u}\\ \mathfrak{v}\end{bmatrix}]$ in terms of the standard deviations $\sigma_{\mathfrak{u}}$ and $\sigma_{\mathfrak{v}}$ and the correlation coefficient ρ.*

• b. *1 point Show that the inverse of a 2×2 matrix has the following form:*

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}. \tag{10.3.10}$$

• c. *2 points Show that*

$$\mathfrak{q}^2 = \begin{bmatrix}\mathfrak{u} - \mu & \mathfrak{v} - \nu\end{bmatrix} \boldsymbol{\Psi}^{-1} \begin{bmatrix}\mathfrak{u} - \mu \\ \mathfrak{v} - \nu\end{bmatrix} \tag{10.3.11}$$

$$= \frac{1}{1 - \rho^2}\Big(\frac{(\mathfrak{u} - \mu)^2}{\sigma_{\mathfrak{u}}^2} - 2\rho\frac{\mathfrak{u} - \mu}{\sigma_{\mathfrak{u}}}\frac{\mathfrak{v} - \nu}{\sigma_{\mathfrak{v}}} + \frac{(\mathfrak{v} - \nu)^2}{\sigma_{\mathfrak{v}}^2}\Big). \tag{10.3.12}$$

• d. *2 points Show the following quadratic decomposition:*

$$\mathfrak{q}^2 = \frac{(\mathfrak{u} - \mu)^2}{\sigma_{\mathfrak{u}}^2} + \frac{1}{(1 - \rho^2)\sigma_{\mathfrak{v}}^2}\Big(\mathfrak{v} - \nu - \rho\frac{\sigma_{\mathfrak{v}}}{\sigma_{\mathfrak{u}}}(\mathfrak{u} - \mu)\Big)^2. \tag{10.3.13}$$

• e. *1 point Show that* (10.3.13) *can also be written in the form*

$$\mathfrak{q}^2 = \frac{(\mathfrak{u} - \mu)^2}{\sigma_{\mathfrak{u}}^2} + \frac{\sigma_{\mathfrak{u}}^2}{\sigma_{\mathfrak{u}}^2\sigma_{\mathfrak{v}}^2 - (\sigma_{\mathfrak{u}\mathfrak{v}})^2}\Big(\mathfrak{v} - \nu - \frac{\sigma_{\mathfrak{u}\mathfrak{v}}}{\sigma_{\mathfrak{u}}^2}(\mathfrak{u} - \mu)\Big)^2. \tag{10.3.14}$$

• f. *1 point Show that $d = \sqrt{\det \boldsymbol{\Psi}}$ can be split up, not additively but multiplicatively, as follows: $d = \sigma_{\mathfrak{u}} \cdot \sigma_{\mathfrak{v}}\sqrt{1 - \rho^2}$.*

• g. *1 point Using these decompositions of d and $\mathfrak{q}^2$, show that the density function $f_{\mathfrak{u},\mathfrak{v}}(u, v)$ reads*

$$\frac{1}{\sqrt{2\pi\sigma_{\mathfrak{u}}^2}}\exp\Big(-\frac{(u - \mu)^2}{2\sigma_{\mathfrak{u}}^2}\Big) \cdot \frac{1}{\sqrt{2\pi\sigma_{\mathfrak{v}}^2}\sqrt{1 - \rho^2}}\exp\Big(-\frac{\big((v - \nu) - \rho\frac{\sigma_{\mathfrak{v}}}{\sigma_{\mathfrak{u}}}(u - \mu)\big)^2}{2(1 - \rho^2)\sigma_{\mathfrak{v}}^2}\Big). \tag{10.3.15}$$

The second factor in (10.3.15) is the density of a $\mathsf{N}(\rho\frac{\sigma_{\mathfrak{v}}}{\sigma_{\mathfrak{u}}}u, (1 - \rho^2)\sigma_{\mathfrak{v}}^2)$ evaluated at v, and the first factor does not depend on v. Therefore if I integrate v out to get the marginal density of $\mathfrak{u}$, this simply gives me the first factor. The conditional density of $\mathfrak{v}$ given $\mathfrak{u} = u$ is the joint divided by the marginal, i.e., it is the second factor. In other words, by completing the square we wrote the joint density function in its natural form as the product of a marginal and a conditional density function: $f_{\mathfrak{u},\mathfrak{v}}(u, v) = f_{\mathfrak{u}}(u) \cdot f_{\mathfrak{v}|\mathfrak{u}}(v; u)$.

From this decomposition one can draw the following conclusions:

- $\mathfrak{u} \sim \mathsf{N}(0, \sigma_{\mathfrak{u}}^2)$ is normal and, by symmetry, $\mathfrak{v}$ is normal as well. Note that $\mathfrak{u}$ (or $\mathfrak{v}$) can be chosen to be any nonzero linear combination of $\mathfrak{x}$ and $\mathfrak{y}$. Any nonzero linear transformation of independent standard normal variables is therefore univariate normal.
- If $\rho = 0$ then the joint density function is the product of two independent univariate normal density functions. In other words, if the variables are normal, then they are independent whenever they are uncorrelated. For general distributions only the reverse is true.
- The conditional density of $\mathfrak{v}$ conditionally on $\mathfrak{u} = u$ is the second term on the rhs of (10.3.15), i.e., it is normal too.
- The conditional mean is

$$\mathrm{E}[\mathfrak{v}|\mathfrak{u} = u] = \rho\frac{\sigma_{\mathfrak{v}}}{\sigma_{\mathfrak{u}}}u, \tag{10.3.16}$$

  i.e., it is a linear function of u. If the (unconditional) means are not zero, then the conditional mean is

$$\mathrm{E}[\mathfrak{v}|\mathfrak{u} = u] = \mu_{\mathfrak{v}} + \rho\frac{\sigma_{\mathfrak{v}}}{\sigma_{\mathfrak{u}}}(u - \mu_{\mathfrak{u}}). \tag{10.3.17}$$

  Since $\rho = \frac{\mathrm{cov}[\mathfrak{u},\mathfrak{v}]}{\sigma_{\mathfrak{u}}\sigma_{\mathfrak{v}}}$, (10.3.17) can als be written as follows:

$$\mathrm{E}[\mathfrak{v}|\mathfrak{u} = u] = \mathrm{E}[\mathfrak{v}] + \frac{\mathrm{cov}[\mathfrak{u},\mathfrak{v}]}{\mathrm{var}[\mathfrak{u}]}(u - \mathrm{E}[\mathfrak{u}]) \tag{10.3.18}$$

- The conditional variance is the same whatever value of u was chosen: its value is

$$\mathrm{var}[\mathfrak{v}|\mathfrak{u} = u] = \sigma_{\mathfrak{v}}^2(1 - \rho^2), \tag{10.3.19}$$

which can also be written as

$$\operatorname{var}[\mathsf{v}|\mathsf{u}=u] = \operatorname{var}[\mathsf{v}] - \frac{(\operatorname{cov}[\mathsf{u},\mathsf{v}])^2}{\operatorname{var}[\mathsf{u}]}. \tag{10.3.20}$$

We did this in such detail because any bivariate normal with zero mean has this form. A multivariate normal distribution is determined by its means and variances and covariances (or correlations coefficients). If the means are not zero, then the densities merely differ from the above by an additive constant in the arguments, i.e., if one needs formulas for nonzero mean, one has to replace u and v in the above equations by $u-\mu_{\mathsf{u}}$ and $v-\mu_{\mathsf{v}}$. du and dv remain the same, because the Jacobian of the translation $u \mapsto u-\mu_{\mathsf{u}}$, $v \mapsto v-\mu_{\mathsf{v}}$ is 1. While the univariate normal was determined by mean and standard deviation, the bivariate normal is determined by the two means μ_{u} and μ_{v}, the two standard deviations σ_{u} and σ_{v}, and the correlation coefficient ρ.

10.3.2. Level Lines of the Normal Density.

Problem 166. *8 points Define the angle $\delta = \arccos(\rho)$, i.e, $\rho = \cos\delta$. In terms of δ, the covariance matrix (**??**) has the form*

$$\boldsymbol{\Psi} = \begin{bmatrix} \sigma_{\mathsf{u}}^2 & \sigma_{\mathsf{u}}\sigma_{\mathsf{v}}\cos\delta \\ \sigma_{\mathsf{u}}\sigma_{\mathsf{v}}\cos\delta & \sigma_{\mathsf{v}}^2 \end{bmatrix} \tag{10.3.21}$$

Show that for all ϕ, the vector

$$\boldsymbol{x} = \begin{bmatrix} r\,\sigma_{\mathsf{u}}\cos\phi \\ r\,\sigma_{\mathsf{v}}\cos(\phi+\delta) \end{bmatrix} \tag{10.3.22}$$

*satisfies $\boldsymbol{x}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{x} = r^2$. The opposite holds too, all vectors $\boldsymbol{x}$ satisfying $\boldsymbol{x}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{x} = r^2$ can be written in the form (10.3.22) for some ϕ, but I am not asking to prove this. This formula can be used to draw level lines of the bivariate Normal density and confidence ellipses, more details in (**??**).*

Problem 167. *The ellipse in Figure 1 contains all the points x, y for which*

$$\begin{bmatrix} x-1 & y-1 \end{bmatrix} \begin{bmatrix} 0.5 & -0.25 \\ -0.25 & 1 \end{bmatrix}^{-1} \begin{bmatrix} x-1 \\ y-1 \end{bmatrix} \leq 6 \tag{10.3.23}$$

• a. *3 points Compute the probability that a random variable*

$$\begin{bmatrix} \mathsf{x} \\ \mathsf{y} \end{bmatrix} \sim N\left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 0.5 & -0.25 \\ -0.25 & 1 \end{bmatrix}\right) \tag{10.3.24}$$

falls into this ellipse. Hint: you should apply equation (10.4.9). Then you will have to look up the values of a χ^2 distribution in a table, or use your statistics software to get it.

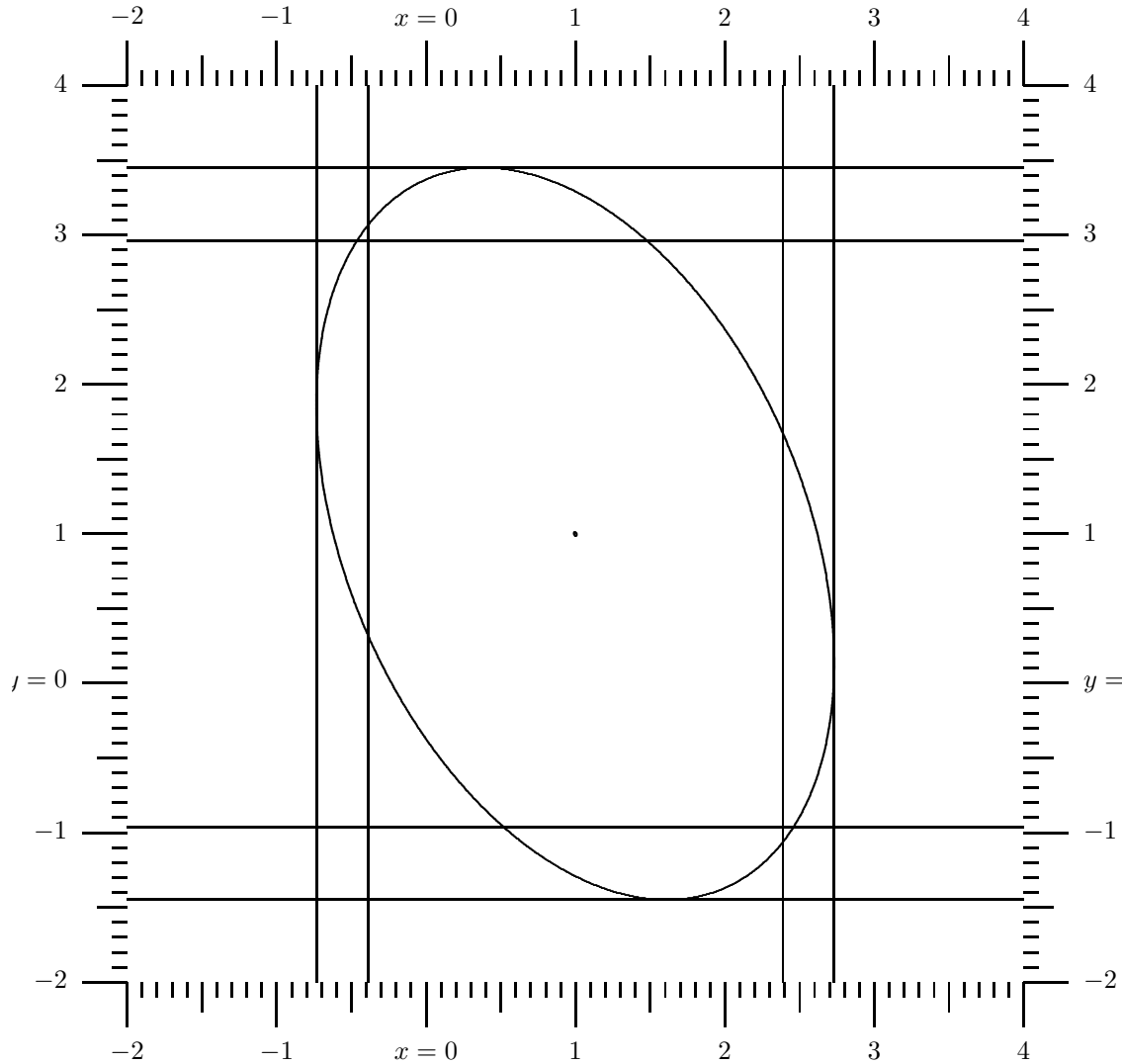


Figure 1. Level Line for Normal Density

• b. *1 point Compute the standard deviations of x and y, and the correlation coefficient* $\operatorname{corr}(\mathsf{x},\mathsf{y})$

• c. *2 points The vertical tangents to the ellipse in Figure 1 are at the locations* $x = 1 \pm \sqrt{3}$. *What is the probability that* $\left[\begin{smallmatrix} \mathsf{x} \\ \mathsf{y} \end{smallmatrix}\right]$ *falls between these two vertical tangents?*

• d. *1 point The* horizontal *tangents are at the locations* $y = 1 \pm \sqrt{6}$. *What is the probability that* $\left[\begin{smallmatrix} \mathsf{x} \\ \mathsf{y} \end{smallmatrix}\right]$ *falls between the horizontal tangents?*

• e. *1 point Now take an arbitrary linear combination* $u = ax + by$. *Write down its mean and its standard deviation.*

• f. *1 point Show that the set of realizations* x, y *for which* u *lies less than* $\sqrt{6}$ *standard deviation away from its mean is*

$$|a(x-1) + b(y-1)| \leq \sqrt{6}\,\sqrt{a^2 \operatorname{var}[x] + 2ab \operatorname{cov}[x,y] + b^2 \operatorname{var}[y]}. \tag{10.3.25}$$

The set of all these points forms a band limited by two parallel lines. What is the probability that $\left[\begin{smallmatrix} \mathsf{x} \\ \mathsf{y} \end{smallmatrix}\right]$ *falls between these two lines?*

• g. *1 point It is our purpose to show that this band is again tangent to the ellipse. This is easiest if we use matrix notation. Define*

$$\boldsymbol{x} = \begin{bmatrix} \mathsf{x} \\ \mathsf{y} \end{bmatrix} \quad \boldsymbol{\mu} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \boldsymbol{\Psi} = \begin{bmatrix} 0.5 & -0.25 \\ -0.25 & 1 \end{bmatrix} \quad \boldsymbol{a} = \begin{bmatrix} a \\ b \end{bmatrix} \tag{10.3.26}$$

Equation (10.3.23) *in matrix notation says: the ellipse contains all the points for which*

$$(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Psi}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \leq 6. \tag{10.3.27}$$

Show that the band defined by inequality (10.3.25) *contains all the points for which*

$$\frac{\left(\boldsymbol{a}^\top (\boldsymbol{x} - \boldsymbol{\mu})\right)^2}{\boldsymbol{a}^\top \boldsymbol{\Psi} \boldsymbol{a}} \leq 6. \tag{10.3.28}$$

• h. *2 points Inequality* (10.3.28) *can also be written as:*

$$(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{a} (\boldsymbol{a}^\top \boldsymbol{\Psi} \boldsymbol{a})^{-1} \boldsymbol{a}^\top (\boldsymbol{x} - \boldsymbol{\mu}) \leq 6 \tag{10.3.29}$$

or alternatively

$$\begin{bmatrix} x-1 & y-1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} \left(\begin{bmatrix} a & b \end{bmatrix} \boldsymbol{\Psi}^{-1} \begin{bmatrix} a \\ b \end{bmatrix} \right)^{-1} \begin{bmatrix} x-1 \\ y-1 \end{bmatrix} \begin{bmatrix} a & b \end{bmatrix} \leq 6. \tag{10.3.30}$$

Show that the matrix

$$\boldsymbol{\Omega} = \boldsymbol{\Psi}^{-1} - \boldsymbol{a}(\boldsymbol{a}^\top \boldsymbol{\Psi} \boldsymbol{a})^{-1} \boldsymbol{a}^\top \tag{10.3.31}$$

satisfies $\boldsymbol{\Omega}\boldsymbol{\Psi}\boldsymbol{\Omega} = \boldsymbol{\Omega}$. *Derive from this that* $\boldsymbol{\Omega}$ *is nonnegative definite. Hint: you may use, without proof, that any symmetric matrix is nonnegative definite if and only if it can be written in the form* $\boldsymbol{R}\boldsymbol{R}^\top$.

• i. *1 point As an aside: Show that* $\boldsymbol{\Omega}\boldsymbol{\Psi}\boldsymbol{a} = \boldsymbol{o}$ *and derive from this that* $\boldsymbol{\Omega}$ *is not* positive *definite but only nonnegative definite.*

• j. *1 point Show that the following inequality holds for all* $\boldsymbol{x} - \boldsymbol{\mu}$,

$$(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Psi}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \geq (\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{a} (\boldsymbol{a}^\top \boldsymbol{\Psi} \boldsymbol{a})^{-1} \boldsymbol{a}^\top (\boldsymbol{x} - \boldsymbol{\mu}). \tag{10.3.32}$$

In other words, if $\boldsymbol{x}$ *lies in the ellipse then it also lies in each band. I.e., the ellipse is contained in the intersection of all the bands.*

• k. *1 point Show: If* $\boldsymbol{x} - \boldsymbol{\mu} = \boldsymbol{\Psi}\boldsymbol{a}\alpha$ *with some arbitrary scalar* α, *then* (10.3.32) *is an equality, and if* $\alpha = \pm\sqrt{6/\boldsymbol{a}^\top \boldsymbol{\Psi} \boldsymbol{a}}$, *then both sides in* (10.3.32) *have the value 6. I.e., the boundary of the ellipse and the boundary lines of the band intersect. Since the ellipse is completely inside the band, this can only be the case if the boundary lines of the band are tangent to the ellipse.*

• l. *2 points The vertical lines in Figure 1 which are* not *tangent to the ellipse delimit a band which, if extended to infinity, has as much probability mass as the ellipse itself. Compute the* x*-coordinates of these two lines.*

10.3.3. Miscellaneous Exercises.

PROBLEM 168. *Figure 2 shows the level line for a bivariate Normal density which contains 95% of the probability mass.*

• a. *3 points One of the following matrices is the covariance matrix of* $\begin{bmatrix} \mathsf{x} \\ \mathsf{y} \end{bmatrix}$. $\boldsymbol{\Psi}_1 = \begin{bmatrix} 0.62 & -0.56 \\ -0.56 & 1.04 \end{bmatrix}$, $\boldsymbol{\Psi}_2 = \begin{bmatrix} 1.85 & 1.67 \\ 1.67 & 3.12 \end{bmatrix}$, $\boldsymbol{\Psi}_3 = \begin{bmatrix} 0.62 & 0.56 \\ 0.56 & 1.04 \end{bmatrix}$, $\boldsymbol{\Psi}_4 = \begin{bmatrix} 1.85 & -1.67 \\ 1.67 & 3.12 \end{bmatrix}$, $\boldsymbol{\Psi}_5 = \begin{bmatrix} 3.12 & -1.67 \\ -1.67 & 1.85 \end{bmatrix}$, $\boldsymbol{\Psi}_6 = \begin{bmatrix} 1.04 & 0.56 \\ 0.56 & 0.62 \end{bmatrix}$, $\boldsymbol{\Psi}_7 = \begin{bmatrix} 3.12 & 1.67 \\ 1.67 & 1.85 \end{bmatrix}$, $\boldsymbol{\Psi}_8 = \begin{bmatrix} 0.62 & 0.81 \\ 0.81 & 1.04 \end{bmatrix}$, $\boldsymbol{\Psi}_9 = \begin{bmatrix} 3.12 & 1.67 \\ 2.67 & 1.85 \end{bmatrix}$, $\boldsymbol{\Psi}_{10} = \begin{bmatrix} 0.56 & 0.62 \\ 0.62 & -1.04 \end{bmatrix}$. *Which is it? Remember that for a univariate Normal, 95% of the probability mass lie within* ± 2 *standard deviations from the mean. If you are not sure, cross out as many of these covariance matrices as possible and write down why you think they should be crossed out.*

ANSWER. Covariance matrix must be symmetric, therefore we can cross out 4 and 9. It must also be nonnegative definite (i.e., it must have nonnegative elements in the diagonal), therefore cross out 10, and a nonnegative determinant, therefore cross out 8. Covariance must be positive, so cross out 1 and 5. Variance in x-direction is smaller than in y-direction, therefore cross out 6 and 7. Remains 2 and 3.

Of these it is number 3. By comparison with Figure 1 one can say that the vertical band between 0.4 and 2.6 and the horizontal band between 3 and -1 roughly have the same probability as the ellipse, namely 95%. Since a univariate Normal has 95% of its probability mass in an interval centered around the mean which is 4 standard deviations long, standard deviations must be approximately 0.8 in the horizontal and 1 in the vertical directions.

$\boldsymbol{\Psi}_1$ is negatively correlated; $\boldsymbol{\Psi}_2$ has the right correlation but is scaled too big; $\boldsymbol{\Psi}_3$ this is it; $\boldsymbol{\Psi}_4$ not symmetric; $\boldsymbol{\Psi}_5$ negatively correlated, and x has larger variance than y; $\boldsymbol{\Psi}_6$ x has larger variance

than $\mathtt{y}$; $\mathbf{\Psi}_7$ too large, $\mathtt{x}$ has larger variance than $\mathtt{y}$; $\mathbf{\Psi}_8$ not positive definite; $\mathbf{\Psi}_9$ not symmetric; $\mathbf{\Psi}_{10}$ not positive definite.

□

The next Problem constructs a counterexample which shows that a bivariate distribution, which is not bivariate Normal, can nevertheless have two marginal densities which are univariate Normal.

PROBLEM 169. *Let* $\mathtt{x}$ *and* $\mathtt{y}$ *be two independent standard normal random variables, and let* $\mathtt{u}$ *and* $\mathtt{v}$ *be bivariate normal with mean zero, variances* $\sigma_{\mathtt{u}}^2 = \sigma_{\mathtt{v}}^2 = 1$, *and correlation coefficient* $\rho \neq 0$. *Let* $f_{\mathtt{x},\mathtt{y}}$ *and* $f_{\mathtt{u},\mathtt{v}}$ *be the corresponding density functions, i.e.,*

$$f_{\mathtt{x},\mathtt{y}}(a,b) = \frac{1}{2\pi}\exp(-\frac{a^2+b^2}{2}) \quad f_{\mathtt{u},\mathtt{v}}(a,b) = \frac{1}{2\pi\sqrt{1-\rho^2}}\exp(-a^2+b^2-2\rho a\frac{b}{2(1-\rho^2)}).$$

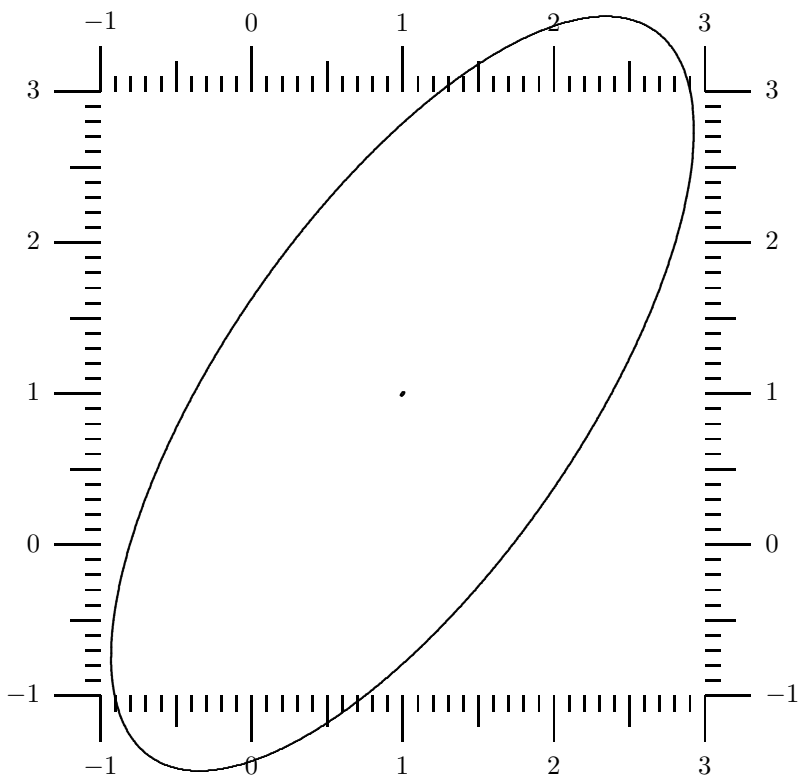


FIGURE 2. Level Line of Bivariate Normal Density, see Problem 168

Assume the random variables $\mathtt{a}$ *and* $\mathtt{b}$ *are defined by the following experiment: You flip a fair coin; if it shows head, then you observe* $\mathtt{x}$ *and* $\mathtt{y}$ *and give* $\mathtt{a}$ *the value observed on* $\mathtt{x}$, *and* $\mathtt{b}$ *the value observed of* $\mathtt{y}$. *If the coin shows tails, then you observe* $\mathtt{u}$ *and* $\mathtt{v}$ *and give* $\mathtt{a}$ *the value of* $\mathtt{u}$, *and* $\mathtt{b}$ *the value of* $\mathtt{v}$.

• a. Prove *that the joint density of* $\mathtt{a}$ *and* $\mathtt{b}$ *is*

$$f_{\mathtt{a},\mathtt{b}}(a,b) = \frac{1}{2}f_{\mathtt{x},\mathtt{y}}(a,b) + \frac{1}{2}f_{\mathtt{u},\mathtt{v}}(a,b). \tag{10.3.33}$$

Hint: first show the corresponding equation for the cumulative distribution functions.

ANSWER. Following this hint:

$$F_{\mathtt{a},\mathtt{b}}(a,b) = \Pr[\mathtt{a} \leq a \text{ and } \mathtt{b} \leq b] = \tag{10.3.34}$$

$$= \Pr[\mathtt{a} \leq a \text{ and } \mathtt{b} \leq b|\text{head}]\Pr[\text{head}] + \Pr[\mathtt{a} \leq a \text{ and } \mathtt{b} \leq b|\text{tail}]\Pr[\text{tail}] \tag{10.3.35}$$

$$= F_{\mathtt{x},\mathtt{y}}(a,b)\frac{1}{2} + F_{\mathtt{u},\mathtt{v}}(a,b)\frac{1}{2}. \tag{10.3.36}$$

The density function is the function which, if integrated, gives the above cumulative distribution function. □

• b. *Show that the marginal distribution of* $\mathtt{a}$ *and* $\mathtt{b}$ *each is normal.*

ANSWER. You can either argue it out: each of the above marginal distributions is standard normal, but you can also say integrate $\mathtt{b}$ out; for this it is better to use form (10.3.15) for $f_{\mathtt{u},\mathtt{v}}$, i.e., write

$$f_{\mathtt{u},\mathtt{v}}(a,b) = \frac{1}{\sqrt{2\pi}}\exp\Big(-\frac{a^2}{2}\Big)\cdot\frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}}\exp\Big(-\frac{(b-\rho a)^2}{2(1-\rho^2)}\Big). \tag{10.3.37}$$

Then you can see that the marginal is standard normal. Therefore you get a mixture of two distributions each of which is standard normal, therefore it is not really a mixture any more. □

• c. *Compute the density of* $\mathtt{b}$ *conditionally on* $\mathtt{a} = 0$. *What are its mean and variance? Is it a normal density?*

ANSWER. $F_{\mathtt{b}|\mathtt{a}}(b;a) = \frac{f_{\mathtt{a},\mathtt{b}}(a,b)}{f_{\mathtt{a}}(a)}$. We don't need it for every a, only for $a = 0$. Since $f_{\mathtt{a}}(0) = 1/\sqrt{2\pi}$, therefore

$$f_{\mathtt{b}|\mathtt{a}=0}(b) = \sqrt{2\pi}f_{\mathtt{a},\mathtt{b}}(0,b) = \frac{1}{2}\frac{1}{\sqrt{2\pi}}\exp\frac{-b^2}{2} + \frac{1}{2}\frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}}\exp\frac{-b^2}{2(1-\rho^2)}. \tag{10.3.38}$$

It is not normal, it is a mixture of normals with different variances. This has mean zero and variance $\frac{1}{2}(1+(1-\rho^2)) = 1 - \frac{1}{2}\rho^2$. □

• d. *Are* $\mathtt{a}$ *and* $\mathtt{b}$ *jointly normal?*

ANSWER. Since the conditional distribution is not normal, they cannot be jointly normal. □

PROBLEM 170. *This is* [**HT83**, 4.8-6 on p. 263] *with variance* σ^2 *instead of* 1*: Let* x *and* y *be independent normal with mean 0 and variance* σ^2*. Go over to polar coordinates* r *and* ϕ*, which satisfy*

$$\begin{aligned} \mathrm{x} &= \mathrm{r}\cos\phi \\ \mathrm{y} &= \mathrm{r}\sin\phi. \end{aligned} \tag{10.3.39}$$

• a. *1 point Compute the Jacobian determinant.*

ANSWER. Express the variables whose density you know in terms of those whose density you want to know. The Jacobian determinant is

$$J = \begin{vmatrix} \frac{\partial \mathrm{x}}{\partial \mathrm{r}} & \frac{\partial \mathrm{x}}{\partial \phi} \\ \frac{\partial \mathrm{y}}{\partial \mathrm{r}} & \frac{\partial \mathrm{y}}{\partial \phi} \end{vmatrix} = \begin{vmatrix} \cos\phi & -\mathrm{r}\sin\phi \\ \sin\phi & \mathrm{r}\cos\phi \end{vmatrix} = \big((\cos\phi)^2 + (\sin\phi)^2\big)\mathrm{r} = \mathrm{r}. \tag{10.3.40}$$

□

• b. *2 points Find the joint probability density function of* r *and* ϕ*. Also indicate the area in* (r, ϕ) *space in which it is nonzero.*

ANSWER. $f_{\mathrm{x},\mathrm{y}}(x, y) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2}$; therefore $f_{\mathrm{r},\phi}(r, \phi) = \frac{1}{2\pi\sigma^2} r e^{-r^2/2\sigma^2}$ for $0 \le r < \infty$ and $0 \le \phi < 2\pi$. □

• c. *3 points Find the marginal distributions of* r *and* ϕ*. Hint: for one of the integrals it is convenient to make the substitution* $q = r^2/2\sigma^2$*.*

ANSWER. $f_{\mathrm{r}}(r) = \frac{1}{\sigma^2} r e^{-r^2/2\sigma^2}$ for $0 \le r < \infty$, and $f_\phi(\phi) = \frac{1}{2\pi}$ for $0 \le \phi < 2\pi$. For the latter we need $\frac{1}{2\pi\sigma^2}\int_0^\infty r e^{-r^2/2\sigma^2}\, dr = \frac{1}{2\pi}$, set $q = r^2/2\sigma^2$, then $dq = \frac{1}{\sigma^2} r\, dr$, and the integral becomes $\frac{1}{2\pi}\int_0^\infty e^{-q}\, dq$. □

• d. *1 point Are* r *and* ϕ *independent?*

ANSWER. Yes, because joint density function is the product of the marginals. □

10.4. Multivariate Standard Normal in Higher Dimensions

Here is an important fact about the multivariate normal, which one cannot see in two dimensions: if the partitioned vector $\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}$ is jointly normal, and every component of $\mathbf{x}$ is independent of every component of $\mathbf{y}$, then the vectors $\mathbf{x}$ and $\mathbf{y}$ are already independent. Not surprised? You should be, see Problem 136.

Let's go back to the construction scheme at the beginning of this chapter. First we will introduce the multivariate *standard* normal, which one obtains by applying only operations (1) and (2), i.e., it is a vector composed of independent univariate standard normals, and give some properties of it. Then we will go over to the multivariate normal with arbitrary covariance matrix, which is simply an arbitrary linear transformation of the multivariate standard normal. We will always carry the "nuisance parameter" σ^2 along.

DEFINITION 10.4.1. The random vector $\boldsymbol{z}$ is said to have a multivariate standard normal distribution with variance σ^2, written as $\boldsymbol{z} \sim \mathsf{N}(\boldsymbol{o}, \sigma^2\boldsymbol{I})$, if each element z_i is a standard normal with same variance σ^2, and all elements are mutually independent of each other. (Note that this definition of the standard normal is a little broader than the usual one; the usual one requires that $\sigma^2 = 1$.)

The density function of a multivariate standard normal $\boldsymbol{z}$ is therefore the product of the univariate densities, which gives $f_{\mathbf{x}}(\boldsymbol{z}) = (2\pi\sigma^2)^{-n/2}\exp(-\boldsymbol{z}^\top\boldsymbol{z}/2\sigma^2)$.

The following property of the multivariate standard normal distributions is basic:

THEOREM 10.4.2. *Let* $\boldsymbol{z}$ *be multivariate standard normal* p*-vector with variance* σ^2*, and let* $\boldsymbol{P}$ *be a* $m \times p$ *matrix with* $\boldsymbol{P}\boldsymbol{P}^\top = \boldsymbol{I}$*. Then* $\mathbf{x} = \boldsymbol{P}\boldsymbol{z}$ *is a multivariate standard normal* m*-vector with the same variance* σ^2*, and* $\boldsymbol{z}^\top\boldsymbol{z} - \mathbf{x}^\top\mathbf{x} \sim \sigma^2\chi^2_{p-m}$ *independent of* $\mathbf{x}$*.*

PROOF. $\boldsymbol{P}\boldsymbol{P}^\top = \boldsymbol{I}$ means all rows are orthonormal. If $\boldsymbol{P}$ is not square, it must therefore have more columns than rows, and one can add more rows to get an orthogonal square matrix, call it $\boldsymbol{T} = \begin{bmatrix} \boldsymbol{P} \\ \boldsymbol{Q} \end{bmatrix}$. Define $\mathbf{y} = \boldsymbol{T}\boldsymbol{z}$, i.e., $\boldsymbol{z} = \boldsymbol{T}^\top\mathbf{y}$. Then $\boldsymbol{z}^\top\boldsymbol{z} = \mathbf{y}^\top\boldsymbol{T}\boldsymbol{T}^\top\mathbf{y} = \mathbf{y}^\top\mathbf{y}$, and the Jacobian of the transformation from $\mathbf{y}$ to $\boldsymbol{z}$ has absolute value one. Therefore the density function of $\mathbf{y}$ is $(2\pi\sigma^2)^{-n/2}\exp(-\boldsymbol{y}^\top\boldsymbol{y}/2\sigma^2)$, which means $\mathbf{y}$ is standard normal as well. In other words, every y_i is univariate standard normal with same variance σ^2 and y_i is independent of y_j for $i \ne j$. Therefore also any subvector of $\mathbf{y}$, such as $\mathbf{x}$, is standard normal. Since $\boldsymbol{z}^\top\boldsymbol{z} - \mathbf{x}^\top\mathbf{x} = \mathbf{y}^\top\mathbf{y} - \mathbf{x}^\top\mathbf{x}$ is the sum of the squares of those elements of $\mathbf{y}$ which are not in $\mathbf{x}$, it follows that it is an independent $\sigma^2\chi^2_{p-m}$. □

PROBLEM 171. *Show that the moment generating function of a multivariate standard normal with variance* σ^2 *is* $m_{\boldsymbol{z}}(\boldsymbol{t}) = \mathcal{E}[\exp(\boldsymbol{t}^\top\boldsymbol{z})] = \exp(\sigma^2\boldsymbol{t}^\top\boldsymbol{t}/2)$*.*

ANSWER. *Proof:* The moment generating function is defined as

$$m_{\boldsymbol{z}}(\boldsymbol{t}) = \mathrm{E}[\exp(\boldsymbol{t}^\top\boldsymbol{z})] \tag{10.4.1}$$

$$= (2\pi\sigma^2)^{n/2}\int\cdots\int \exp(-\frac{1}{2\sigma^2}\boldsymbol{z}^\top\boldsymbol{z})\,\exp(\boldsymbol{t}^\top\boldsymbol{z})\, dz_1\cdots dz_n \tag{10.4.2}$$

$$= (2\pi\sigma^2)^{n/2}\int\cdots\int \exp(-\frac{1}{2\sigma^2}(\boldsymbol{z}-\sigma^2\boldsymbol{t})^\top(\boldsymbol{z}-\sigma^2\boldsymbol{t}) + \frac{\sigma^2}{2}\boldsymbol{t}^\top\boldsymbol{t})\, dz_1\cdots dz_n \tag{10.4.3}$$

$$= \exp(\frac{\sigma^2}{2}\boldsymbol{t}^\top\boldsymbol{t}) \quad \text{since first part of integrand is density function.} \tag{10.4.4}$$

□

THEOREM 10.4.3. *Let* $\boldsymbol{z} \sim \mathsf{N}(\boldsymbol{o}, \sigma^2\boldsymbol{I})$*, and* $\boldsymbol{P}$ *symmetric and of rank* r*. A necessary and sufficient condition for* $\mathsf{q} = \boldsymbol{z}^\top\boldsymbol{P}\boldsymbol{z}$ *to have a* $\sigma^2\chi^2$ *distribution is* $\boldsymbol{P}^2 = \boldsymbol{P}$*. In this case, the* χ^2 *has* r *degrees of freedom.*

Proof of sufficiency: If $\boldsymbol{P}^2 = \boldsymbol{P}$ with rank r, then a matrix $\boldsymbol{T}$ exists with $\boldsymbol{P} = \boldsymbol{T}^\top\boldsymbol{T}$ and $\boldsymbol{T}\boldsymbol{T}^\top = \boldsymbol{I}$. Define $\mathsf{x} = \boldsymbol{T}\boldsymbol{z}$; it is standard normal by theorem 10.4.2. Therefore $\mathsf{q} = \boldsymbol{z}^\top\boldsymbol{T}^\top\boldsymbol{T}\boldsymbol{z} = \sum_{i=1}^r \mathsf{x}^2$.

Proof of necessity by construction of the moment generating function of $\mathsf{q} = \boldsymbol{z}^\top\boldsymbol{P}\boldsymbol{z}$ for arbitrary symmetric $\boldsymbol{P}$ with rank r. Since $\boldsymbol{P}$ is symmetric, there exists a $\boldsymbol{T}$ with $\boldsymbol{T}\boldsymbol{T}^\top = \boldsymbol{I}_r$ and $\boldsymbol{P} = \boldsymbol{T}^\top\boldsymbol{\Lambda}\boldsymbol{T}$ where $\boldsymbol{\Lambda}$ is a nonsingular diagonal matrix, write it $\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_r)$. Therefore $\mathsf{q} = \boldsymbol{z}^\top\boldsymbol{T}^\top\boldsymbol{\Lambda}\boldsymbol{T}\boldsymbol{z} = \boldsymbol{\mathsf{x}}^\top\boldsymbol{\Lambda}\boldsymbol{\mathsf{x}} = \sum_{i=1}^r \lambda_i \mathsf{x}_i^2$ where $\boldsymbol{\mathsf{x}} = \boldsymbol{T}\boldsymbol{z} \sim \mathsf{N}(\boldsymbol{o}, \sigma^2\boldsymbol{I}_r)$. Therefore the moment generating function

$$\mathrm{E}[\exp(\mathsf{q}t)] = \mathrm{E}[\exp(t\sum_{i=1}^r \lambda_i \mathsf{x}_i^2)] \tag{10.4.5}$$

$$= \mathrm{E}[\exp(t\lambda_1\mathsf{x}_1^2)]\cdots\mathrm{E}[\exp(t\lambda_r\mathsf{x}_r^2)] \tag{10.4.6}$$

$$= (1-2\lambda_1\sigma^2 t)^{-1/2}\cdots(1-2\lambda_r\sigma^2 t)^{-1/2}. \tag{10.4.7}$$

By assumption this is equal to $(1-2\sigma^2 t)^{-k/2}$ with some integer $k \geq 1$. Taking squares and inverses one obtains

$$(1-2\lambda_1\sigma^2 t)\cdots(1-2\lambda_r\sigma^2 t) = (1-2\sigma^2 t)^k. \tag{10.4.8}$$

Since the $\lambda_i \neq 0$, one obtains $\lambda_i = 1$ by uniqueness of the polynomial roots. Furthermore, this also implies $r = k$.

From Theorem 10.4.3 one can derive a characterization of all the quadratic forms of multivariate normal variables with arbitrary covariance matrices that are χ^2's. Assume $\mathfrak{y}$ is a multivariate normal vector random variable with mean vector $\boldsymbol{\mu}$ and covariance matrix $\sigma^2\boldsymbol{\Psi}$, and $\boldsymbol{\Omega}$ is a symmetric nonnegative definite matrix. Then $(\mathfrak{y}-\boldsymbol{\mu})^\top\boldsymbol{\Omega}(\mathfrak{y}-\boldsymbol{\mu}) \sim \sigma^2\chi_k^2$ iff

$$\boldsymbol{\Psi}\boldsymbol{\Omega}\boldsymbol{\Psi}\boldsymbol{\Omega}\boldsymbol{\Psi} = \boldsymbol{\Psi}\boldsymbol{\Omega}\boldsymbol{\Psi}, \tag{10.4.9}$$

and k is the rank of $\boldsymbol{\Psi}\boldsymbol{\Omega}$.

Here are the three best known special cases (with examples):

- $\boldsymbol{\Psi} = \boldsymbol{I}$ (the identity matrix) and $\boldsymbol{\Omega}^2 = \boldsymbol{\Omega}$, i.e., the case of theorem 10.4.3. This is the reason why the minimum value of the SSE has a $\sigma^2\chi^2$ distribution, see (35.0.10).
- $\boldsymbol{\Psi}$ nonsingular and $\boldsymbol{\Omega} = \boldsymbol{\Psi}^{-1}$. The quadratic form in the exponent of the normal density function is therefore a χ^2; one needs therefore the χ^2 to compute the probability that the realization of a Normal is in a given equidensity-ellipse (Problem 167).
- $\boldsymbol{\Psi}$ singular and $\boldsymbol{\Omega} = \boldsymbol{\Psi}^-$, its g-inverse. The multinomial distribution has a singular covariance matrix, and equation (15.4.2) gives a convenient g-inverse which enters the equation for Pearson's goodness of fit test.

Here are, without proof, two more useful theorems about the standard normal:

THEOREM 10.4.4. *Let $\boldsymbol{\mathsf{x}}$ a multivariate standard normal. Then $\boldsymbol{\mathsf{x}}^\top\boldsymbol{P}\boldsymbol{\mathsf{x}}$ is independent of $\boldsymbol{\mathsf{x}}^\top\boldsymbol{Q}\boldsymbol{\mathsf{x}}$ if and only if $\boldsymbol{P}\boldsymbol{Q} = \boldsymbol{O}$.*

This is called Craig's theorem, although Craig's proof in [**Cra43**] is incorrect. Kshirsagar [**Ksh19**, p. 41] describes the correct proof; he and Seber [**Seb77**] give Lancaster's book [**Lan69**] as basic reference. Seber [**Seb77**] gives a proof which is only valid if the two quadratic forms are χ^2.

The next theorem is known as James's theorem, it is a stronger version of Cochrane's theorem. It is from Kshirsagar [**Ksh19**, p. 41].

THEOREM 10.4.5. *Let $\boldsymbol{\mathsf{x}}$ be p-variate standard normal with variance σ^2, and $\boldsymbol{\mathsf{x}}^\top\boldsymbol{\mathsf{x}} = \sum_{i=1}^k \boldsymbol{\mathsf{x}}^\top\boldsymbol{P}_i\boldsymbol{\mathsf{x}}$. Then for the quadratic forms $\boldsymbol{\mathsf{x}}^\top\boldsymbol{P}_i\boldsymbol{\mathsf{x}}$ to be independently distributed as $\sigma^2\chi^2$, any one of the following three equivalent conditions is necessary and sufficient:*

$$\boldsymbol{P}_i^2 = \boldsymbol{P}_i \qquad \text{for all } i \tag{10.4.10}$$

$$\boldsymbol{P}_i\boldsymbol{P}_j = \boldsymbol{O} \qquad i \neq j \tag{10.4.11}$$

$$\sum_{i=1}^k \operatorname{rank}(\boldsymbol{P}_i) = p \tag{10.4.12}$$

10.5. Higher Moments of the Multivariate Standard Normal

For the definition of skewness and kurtosis see question 156

PROBLEM 172. *Show that, if z is a standard normal scalar variable, then skewness and kurtosis are both zero: $\gamma_1 = \gamma_2 = 0$.*

ANSWER. To compute the kurtosis of a standard Normal $\mathsf{z} \sim \mathsf{N}(0,1)$, define $u = z^3$ and $v = \exp(-z^2/2)$, therefore $v' = -z\exp(-z^2/2)$. Then

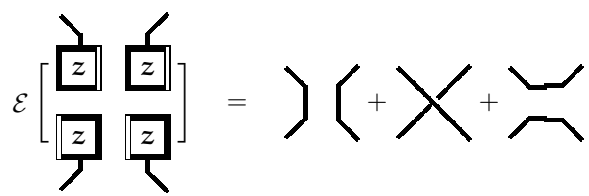
$$\mathrm{E}[z^4] = \frac{1}{\sqrt{2\pi}}\int z^4\exp-\frac{z^2}{2}\,dz = \frac{1}{\sqrt{2\pi}}\int uv'\,dz = \frac{1}{\sqrt{2\pi}}uv\Big|_{-\infty}^{\infty} - \frac{1}{\sqrt{2\pi}}\int u'v\,dz = 0 + \frac{3}{\sqrt{2\pi}}\int z^2\exp(-z^2/2) \tag{10.5.1}$$

since the last integral is just the variance of the standard normal.

□

I will give a brief overview in tile notation of the higher moments of the multivariate standard normal $\boldsymbol{z}$. All odd moments disappear, and the fourth moments are

(10.5.2)

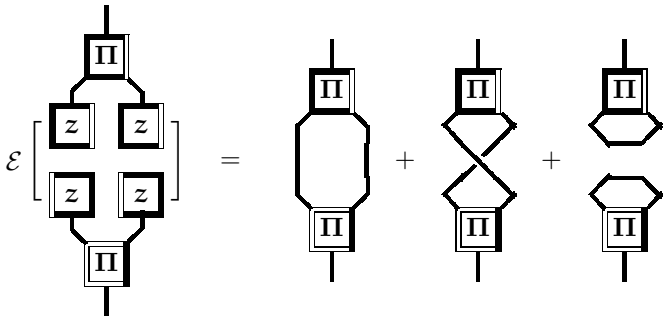


Compared with (9.2.23), the last term, which depends on the kurtosis, is missing. What remains is a sum of outer products of unit matrices, with every possibility appearing exactly once. In the present case, it happens to be possible to write down the four-way arrays in (10.5.2) in terms of Kronecker products and the commutation matrix $\boldsymbol{K}^{(n,n)}$ introduced in (B.5.21): It is

$$\mathcal{E}[(\boldsymbol{z}\boldsymbol{z}^\top)\otimes(\boldsymbol{z}\boldsymbol{z}^\top)] = \boldsymbol{I}_{n^2} + \boldsymbol{K}^{(n,n)} + (\operatorname{vec}[\boldsymbol{I}_n])(\operatorname{vec}[\boldsymbol{I}_n])^\top \tag{10.5.3}$$

Compare [**Gra83**, 10.9.2 on p. 361]. Here is a proof of (10.5.3) in tile notation:

(10.5.4)

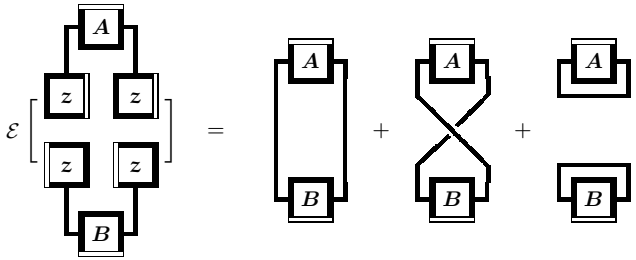


The first term is $\boldsymbol{I}_{n^2}$ due to (B.5.26), the second is $\boldsymbol{K}^{(n,n)}$ due to (B.5.35), and the third is $(\operatorname{vec}[\boldsymbol{I}_n])(\operatorname{vec}[\boldsymbol{I}_n])^\top$ because of (B.5.24). Graybill [**Gra83**, p. 312] considers it a justification of the interest of the commutation matrix that it appears in the higher moments of the standard normal. In my view, the commutation matrix is ubiquitous only because the Kronecker-notation blows up something as trivial as the crossing of two arms into a mysterious-sounding special matrix.

It is much easier to work with (10.5.2) without the detour over Kronecker products:

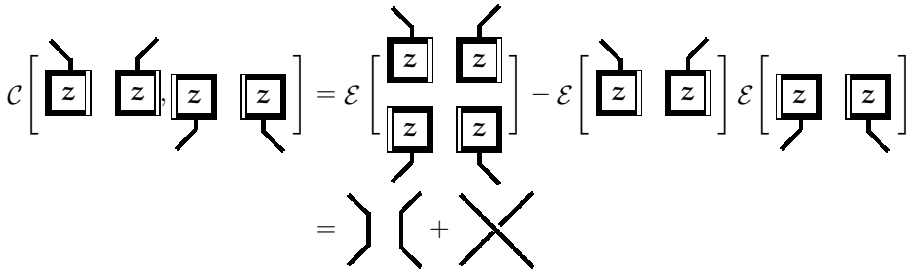
Problem 173. [**Gra83**, 10.9.10 (1) on p. 366] *Show that for symmetric* $\boldsymbol{A}$ *and* $\boldsymbol{B}$ $\mathrm{E}[\boldsymbol{z}^\top\boldsymbol{A}\boldsymbol{z}\,\boldsymbol{z}^\top\boldsymbol{B}\boldsymbol{z}] = 2\operatorname{tr}(\boldsymbol{A}\boldsymbol{B}) + \operatorname{tr}(\boldsymbol{A})\operatorname{tr}(\boldsymbol{B})$.

Answer. This is (9.2.24) in the case of zero kurtosis, but here is a direct proof based on (10.5.2):



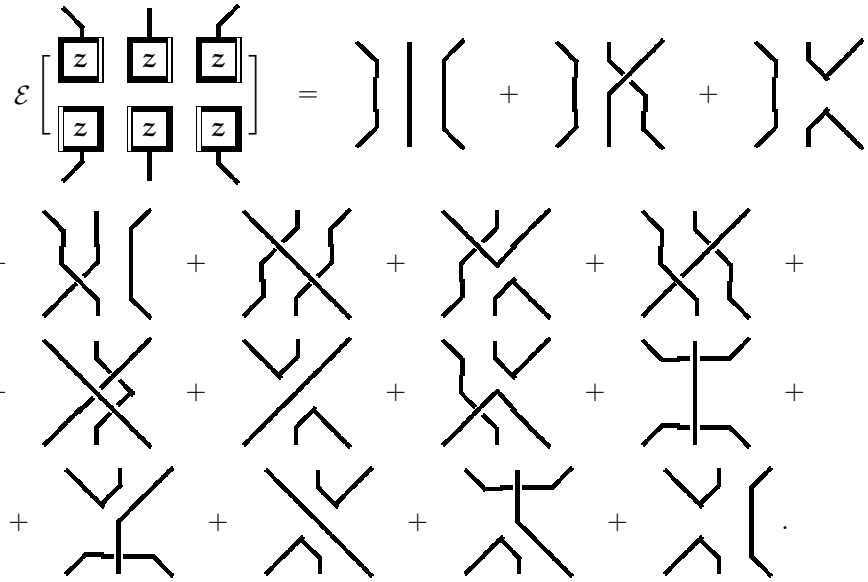
□

If one takes the variance-covariance matrix, which should in tile notation always be written with a $\mathcal{C}$, so that one knows which arms stick out in which direction, then the third term in (10.5.2) falls away:



The sixth moments of the standard normal, in analogy to the fourth, are the sum of all the different possible outer products of unit matrices:

(10.5.5)



Here is the principle how these were written down: Fix one branch, here the Southwest branch. First combine the Southwest branch with the Northwest one, and then you have three possibilities to pair up the others as in (10.5.2). Next combine the Southwest branch with the North branch, and you again have three possibilities for the others. Etc. This gives 15 possibilities altogether.

This can no longer be written as a Kronecker product, see [**Gra83**, 10.9.4 (3) on p. 363]. However (10.5.5) can be applied directly, for instance in order to show

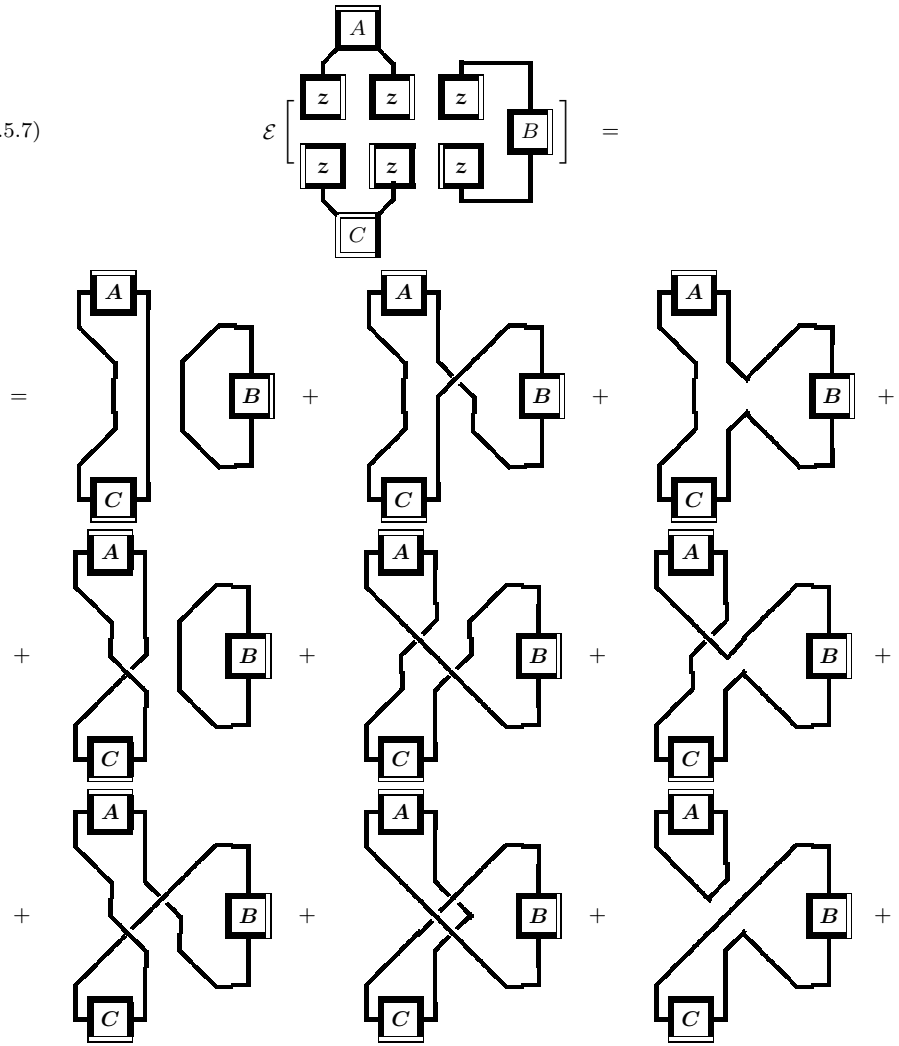
(10.5.6), which is [**Gra83**, 10.9.12 (1) on p. 368]:

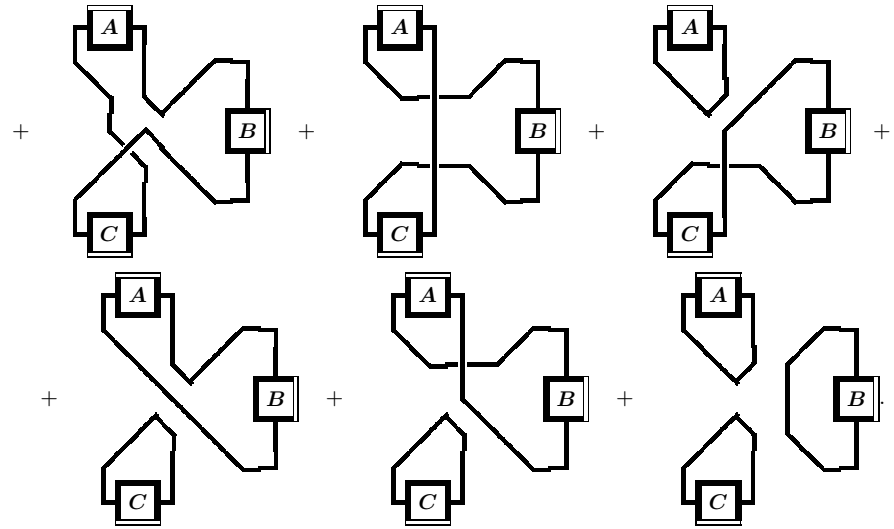
$$\mathrm{E}[(\boldsymbol{z}^\top \boldsymbol{A}\boldsymbol{z})(\boldsymbol{z}^\top \boldsymbol{B}\boldsymbol{z})(\boldsymbol{z}^\top \boldsymbol{C}\boldsymbol{z})] = \operatorname{tr}(\boldsymbol{A})\operatorname{tr}(\boldsymbol{B})\operatorname{tr}(\boldsymbol{C}) + 2\operatorname{tr}(\boldsymbol{A})\operatorname{tr}(\boldsymbol{B}\boldsymbol{C}) + 2\operatorname{tr}(\boldsymbol{B})\operatorname{tr}(\boldsymbol{A}\boldsymbol{C}) + 2\operatorname{tr}(\boldsymbol{C})\operatorname{tr}(\boldsymbol{A}\boldsymbol{B}) + 8\operatorname{tr}(\boldsymbol{A}\boldsymbol{B}\boldsymbol{C}). \tag{10.5.6}$$

Problem 174. *Assuming that $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$ are symmetric matrices, prove* (10.5.6) *in tile notation.*

Answer.

(10.5.7)





These 15 summands are, in order, $\operatorname{tr}(\boldsymbol{B})\operatorname{tr}(\boldsymbol{A}\boldsymbol{C})$, $\operatorname{tr}(\boldsymbol{A}\boldsymbol{B}\boldsymbol{C})$ twice, $\operatorname{tr}(\boldsymbol{B})\operatorname{tr}(\boldsymbol{A}\boldsymbol{C})$, $\operatorname{tr}(\boldsymbol{A}\boldsymbol{B}\boldsymbol{C})$ four times, $\operatorname{tr}(\boldsymbol{A})\operatorname{tr}(\boldsymbol{B}\boldsymbol{C})$, $\operatorname{tr}(\boldsymbol{A}\boldsymbol{B}\boldsymbol{C})$ twice, $\operatorname{tr}(\boldsymbol{A})\operatorname{tr}(\boldsymbol{B}\boldsymbol{C})$, $\operatorname{tr}(\boldsymbol{C})\operatorname{tr}(\boldsymbol{A}\boldsymbol{B})$ twice, and $\operatorname{tr}(\boldsymbol{A})\operatorname{tr}(\boldsymbol{B})\operatorname{tr}(\boldsymbol{C})$. □

10.6. The General Multivariate Normal

Definition 10.6.1. The random vector $\mathbf{y}$ is multivariate normal if and only if there is a multivariate standard normal variable $\boldsymbol{z}$, a nonstochastic matrix $\boldsymbol{C}$, and a nonstochastic vector $\boldsymbol{c}$ with $\mathbf{y} = \boldsymbol{C}\boldsymbol{z} + \boldsymbol{c}$.

In this case, clearly, $\mathcal{E}[\mathbf{y}] = \boldsymbol{c}$ and $\mathcal{V}[\mathbf{y}] = \sigma^2\boldsymbol{C}\boldsymbol{C}^\top$, where σ^2 is the variance of the standard normal.

We will say: the vector is multivariate normal, and its elements or subvectors are jointly normal, i.e., $\mathbf{x}$ and $\mathbf{y}$ are jointly normal if and only if $\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}$ is multivariate normal. This is not transitive. If $\mathbf{x}$ and $\mathbf{y}$ are jointly normal and $\mathbf{y}$ and $\boldsymbol{z}$ are, then $\mathbf{x}$ and $\boldsymbol{z}$ need not be. And even if all three pairs are jointly normal, this does not mean they are multivariate normal.

Theorem 10.6.2. *The distribution of a multivariate normal variable is fully determined by its expected value and dispersion matrix. Therefore the notation* $\mathbf{y} \sim \mathrm{N}(\boldsymbol{\theta}, \sigma^2\boldsymbol{\Sigma})$.

For the proof we will use the following theorem: The distribution of a random variable $\mathbf{y}$ is fully characterized by the univariate distributions of all $\boldsymbol{a}^\top\mathbf{y}$ for all vectors $\boldsymbol{a}$. A proof can be found in [**Rao73**, p. 517].

Assume $\mathbf{u} = \boldsymbol{C}\mathbf{x} + \boldsymbol{c}$ and $\boldsymbol{\nu} = \boldsymbol{D}\boldsymbol{z} + \boldsymbol{d}$ where $\mathbf{x}$ and $\boldsymbol{z}$ are standard normal, and $\mathbf{u}$ and $\boldsymbol{\nu}$ have equal mean and variances, i.e., $\boldsymbol{c} = \boldsymbol{d}$ and $\boldsymbol{C}\boldsymbol{C}^\top = \boldsymbol{D}\boldsymbol{D}^\top$. We will show that $\mathbf{u}$ and $\boldsymbol{\nu}$ or, equivalently, $\boldsymbol{C}\mathbf{x}$ and $\boldsymbol{D}\boldsymbol{z}$ are identically distributed, by verifying that for every vector $\boldsymbol{a}$, the distribution of $\boldsymbol{a}^\top\boldsymbol{C}\mathbf{x}$ is the same as the distribution of $\boldsymbol{a}^\top\boldsymbol{D}\boldsymbol{z}$. There are two cases: either $\boldsymbol{a}^\top\boldsymbol{C} = \boldsymbol{o}^\top \Rightarrow \boldsymbol{a}^\top\boldsymbol{C}\boldsymbol{C}^\top = \boldsymbol{o}^\top \Rightarrow \boldsymbol{a}^\top\boldsymbol{D}\boldsymbol{D}^\top = \boldsymbol{o}^\top \Rightarrow \boldsymbol{a}^\top\boldsymbol{D} = \boldsymbol{o}^\top$, therefore $\boldsymbol{a}^\top\boldsymbol{C}\mathbf{x}$ and $\boldsymbol{a}^\top\boldsymbol{D}\boldsymbol{z}$ have equal distributions degenerate at zero. Now if $\boldsymbol{a}^\top\boldsymbol{C} \neq \boldsymbol{o}^\top$, then without loss of generality one can restrict oneself to the $\boldsymbol{a}$ with $\boldsymbol{a}^\top\boldsymbol{C}\boldsymbol{C}^\top\boldsymbol{a} = 1$, therefore also $\boldsymbol{a}^\top\boldsymbol{D}\boldsymbol{D}^\top\boldsymbol{a} = 1$. By theorem 10.4.2, both $\boldsymbol{a}^\top\boldsymbol{C}\mathbf{x}$ and $\boldsymbol{a}^\top\boldsymbol{D}\mathbf{y}$ are standard normal.

THEOREM 10.6.3. *If* $\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}$ *is multivariate normal and* $\mathcal{C}[\mathbf{x}, \mathbf{y}] = \boldsymbol{O}$, *then* $\mathbf{x}$ *and* $\mathbf{y}$ *are independent.*

PROOF. Let $\boldsymbol{\mu} = \mathcal{E}[\mathbf{x}]$ and $\boldsymbol{\nu} = \mathcal{E}[\mathbf{y}]$, and $\boldsymbol{A}$ and $\boldsymbol{B}$ two matrices with $\boldsymbol{A}\boldsymbol{A}^\top = \mathcal{V}[\mathbf{x}]$ and $\boldsymbol{B}\boldsymbol{B}^\top = \mathcal{V}[\mathbf{y}]$, and let $\mathbf{u}$ and $\boldsymbol{\nu}$ independent standard normal variables. Then $\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}$ has the same distribution as

$$\begin{bmatrix}\boldsymbol{A} & \boldsymbol{O}\\ \boldsymbol{O} & \boldsymbol{B}\end{bmatrix}\begin{bmatrix}\mathbf{u}\\ \boldsymbol{\nu}\end{bmatrix} + \begin{bmatrix}\boldsymbol{\mu}\\ \boldsymbol{\nu}\end{bmatrix}. \tag{10.6.1}$$

Since $\mathbf{u}$ and $\boldsymbol{\nu}$ are independent, $\mathbf{x}$ and $\mathbf{y}$ are also independent. □

PROBLEM 175. *Show that, if* $\mathbf{y} \sim N_n(\boldsymbol{\theta}, \sigma^2\boldsymbol{\Sigma})$, *then*

$$\boldsymbol{D}\mathbf{y} + \boldsymbol{d} \sim N_k(\boldsymbol{D}\boldsymbol{\theta} + \boldsymbol{d}, \sigma^2\boldsymbol{D}\boldsymbol{\Sigma}\boldsymbol{D}^\top) \tag{10.6.2}$$

ANSWER. Follows immediately from our definition of a multivariate normal. □

THEOREM 10.6.4. *Let* $\mathbf{y} \sim \mathsf{N}(\boldsymbol{\theta}, \sigma^2\boldsymbol{\Sigma})$. *Then one can find two matrices* $\boldsymbol{B}$ *and* $\boldsymbol{D}$ *so that* $\boldsymbol{z} = \boldsymbol{B}(\mathbf{y} - \boldsymbol{\theta})$ *is standard normal, and* $\mathbf{y} = \boldsymbol{D}\boldsymbol{z} + \boldsymbol{\theta}$.

PROOF. According to theorem A.9.1, a $\boldsymbol{T}$ exists with $\boldsymbol{\Sigma} = \boldsymbol{T}^\top\boldsymbol{\Lambda}\boldsymbol{T}$, where $\boldsymbol{\Lambda}$ is a positive definite diagonal matrix, and $\boldsymbol{T}\boldsymbol{T}^\top = \boldsymbol{I}_k$, where k is the rank of $\boldsymbol{\Sigma}$. Define $\boldsymbol{D} = \boldsymbol{T}^\top\boldsymbol{\Lambda}^{1/2}$ and $\boldsymbol{B} = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{T}$. Since $\mathbf{y}$ is multivariate normal, it can be written in the form $\mathbf{y} = \boldsymbol{C}\mathbf{x} + \boldsymbol{\theta}$ for some standard normal $\mathbf{x}$, where $\boldsymbol{C}\boldsymbol{C}^\top = \boldsymbol{\Sigma}$. Therefore $\boldsymbol{z} = \boldsymbol{B}(\mathbf{y} - \boldsymbol{\theta}) = \boldsymbol{B}\boldsymbol{C}\mathbf{x}$; this is standard normal because $\mathbf{x}$ is and

$$\boldsymbol{B}\boldsymbol{C}\boldsymbol{C}^\top\boldsymbol{B}^\top = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{T}\boldsymbol{C}\boldsymbol{C}^\top\boldsymbol{T}^\top\boldsymbol{\Lambda}^{-1/2} = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{T}\boldsymbol{\Sigma}\boldsymbol{T}^\top\boldsymbol{\Lambda}^{-1/2} = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{T}\boldsymbol{T}^\top\boldsymbol{\Lambda}\boldsymbol{T}\boldsymbol{T}^\top\boldsymbol{\Lambda}^{-1/2}; \tag{10.6.3}$$

We still have to show that $\boldsymbol{D}\boldsymbol{z}+\boldsymbol{\theta} = \mathbf{y}$. Plugging in gives $\boldsymbol{D}\boldsymbol{z}+\boldsymbol{\theta} = \boldsymbol{D}\boldsymbol{B}(\mathbf{y}-\boldsymbol{\theta})+\boldsymbol{\theta} = \boldsymbol{T}^\top\boldsymbol{\Lambda}^{1/2}\boldsymbol{\Lambda}^{-1/2}\boldsymbol{T}(\mathbf{y}-\boldsymbol{\theta})+\boldsymbol{\theta} = \boldsymbol{T}^\top\boldsymbol{T}(\mathbf{y}-\boldsymbol{\theta})+\boldsymbol{\theta}$. Now we have to use the fact that with probability 1, $\mathbf{y}-\boldsymbol{\theta}$ lies in the range space of $\boldsymbol{\Sigma}$, i.e., $\mathbf{y}-\boldsymbol{\theta} = \boldsymbol{\Sigma}\boldsymbol{a} = \boldsymbol{T}^\top\boldsymbol{\Lambda}\boldsymbol{T}\boldsymbol{a}$ for some $\boldsymbol{a}$. This makes $\boldsymbol{D}\boldsymbol{z} + \boldsymbol{\theta} = \boldsymbol{T}^\top\boldsymbol{T}\boldsymbol{T}^\top\boldsymbol{\Lambda}\boldsymbol{T}\boldsymbol{a} + \boldsymbol{\theta} = \boldsymbol{T}^\top\boldsymbol{T}\boldsymbol{T}^\top\boldsymbol{\Lambda}\boldsymbol{T}\boldsymbol{a} + \boldsymbol{\theta} = \boldsymbol{T}^\top\boldsymbol{\Lambda}\boldsymbol{T}\boldsymbol{a} + \boldsymbol{\theta} = \boldsymbol{\Sigma}\boldsymbol{a} + \boldsymbol{\theta} = \mathbf{y}$. □

PROBLEM 176. *Show that a random variable* $\mathbf{y}$ *with expected value* $\boldsymbol{\theta}$ *and nonsingular covariance matrix* $\sigma^2\boldsymbol{\Sigma}$ *is multivariate normal iff its density function is*

$$f_{\mathbf{y}}(\boldsymbol{y}) = (2\pi\sigma^2)^{-n/2}(\det\boldsymbol{\Sigma})^{-1/2}\exp\Big(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\theta})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\theta})\Big). \tag{10.6.4}$$

Hint: use the matrices $\boldsymbol{B}$ *and* $\boldsymbol{D}$ *from theorem 10.6.4.*

ANSWER. First assume $\mathbf{y}$ is multivariate normal. Then by theorem 10.6.4, $\boldsymbol{z} = \boldsymbol{B}(\mathbf{y} - \boldsymbol{\theta})$ is standard normal, i.e., its density function is the product of n univariate standard normal densities:

$$f_{\boldsymbol{z}}(\boldsymbol{z}) = (2\pi\sigma^2)^{-n/2}\exp(-\frac{1}{2\sigma^2}\boldsymbol{z}^\top\boldsymbol{z}). \tag{10.6.5}$$

From this we get the one of $\mathbf{y}$. Since $\boldsymbol{I} = \boldsymbol{B}^\top\boldsymbol{\Sigma}\boldsymbol{B}$, it follows $1 = \det(\boldsymbol{B}^\top\boldsymbol{\Sigma}\boldsymbol{B}) = (\det\boldsymbol{B})^2\det\boldsymbol{\Sigma}$, therefore $J = \det\boldsymbol{B} = \pm\sqrt{\det\boldsymbol{\Sigma}}$, and $|J| = |\det\boldsymbol{B}| = \sqrt{\det\boldsymbol{\Sigma}}$. Since $\boldsymbol{z}^\top\boldsymbol{z} = (\mathbf{y}-\boldsymbol{\theta})^\top\boldsymbol{B}^\top\boldsymbol{B}(\mathbf{y}-\boldsymbol{\theta}) = (\mathbf{y}-\boldsymbol{\theta})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\theta})$, $\mathbf{y}$ has the density function (10.6.4).

Conversely, assume we know that $\mathbf{y}$ has the density function (10.6.4). Then let us derive from this the density function of $\boldsymbol{z} = \boldsymbol{B}(\mathbf{y} - \boldsymbol{\theta})$. Since $\boldsymbol{\Sigma}$ is nonsingular, one can solve $\mathbf{y} = \boldsymbol{D}\boldsymbol{z} + \boldsymbol{\theta}$. Since $\boldsymbol{D}\boldsymbol{D}^\top = \boldsymbol{\Sigma}$, it follows $J = \det\boldsymbol{D} = \pm\sqrt{\det\boldsymbol{\Sigma}}$, and therefore $|J| = \sqrt{\det\boldsymbol{\Sigma}}$. Furthermore, $(\mathbf{y}-\boldsymbol{\theta})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\theta}) = \boldsymbol{z}^\top\boldsymbol{z}$, i.e., the density of $\boldsymbol{z}$ is that of a standard normal. Since $\mathbf{y}$ is a linear transformation of $\boldsymbol{z}$, it is multivariate normal. □

PROBLEM 177. *Show that the moment generating function of a multivariate normal* $\mathbf{y} \sim \mathsf{N}(\boldsymbol{\theta}, \sigma^2\boldsymbol{\Sigma})$ *is*

$$m_{\mathbf{y}}(\boldsymbol{t}) = \exp(\boldsymbol{t}^\top\boldsymbol{\theta} + \sigma^2\boldsymbol{t}^\top\boldsymbol{\Sigma}\boldsymbol{t}/2). \tag{10.6.6}$$

Give a proof which is valid for singular as well as nonsingular $\boldsymbol{\Sigma}$. *You may use the formula for the moment generating function of a multivariate Standard normal for this proof.*

ANSWER.

$$m_{\mathbf{y}}(\boldsymbol{t}) = \mathrm{E}\Big[\exp(\boldsymbol{t}^\top\mathbf{y})\Big] = \mathrm{E}\Big[\exp\Big(\boldsymbol{t}^\top(\boldsymbol{D}\boldsymbol{z}+\boldsymbol{\theta})\Big)\Big] = \exp(\boldsymbol{t}^\top\boldsymbol{\theta})\,\mathrm{E}\Big[\exp\Big(\boldsymbol{t}^\top\boldsymbol{D}\boldsymbol{z}\Big)\Big] = \exp(\boldsymbol{t}^\top\boldsymbol{\theta})\exp\Big(\sigma^2\boldsymbol{t}^\top\boldsymbol{D}\boldsymbol{D}^\top\boldsymbol{t}/2\Big) = \tag{10.6.7}$$

□

THEOREM 10.6.5. $\mathbf{y}$ *is multivariate normal if and only if* $\boldsymbol{a}^\top\mathbf{y}$ *is univariate normal for all* $\boldsymbol{a}$.

PROOF. Necessity by (10.6.2). Sufficiency: If $\boldsymbol{a}^\top\mathbf{y}$ is normal, its first and second moments exist for every $\boldsymbol{a}$; by one of the early homework problems this means the whole dispersion matrix of $\mathbf{y}$ exists. Say $\mathcal{E}[\mathbf{y}] = \boldsymbol{\theta}$ and $\mathcal{V}[\mathbf{y}] = \sigma^2\boldsymbol{\Sigma}$. Then one sees that $\boldsymbol{a}^\top\mathbf{y}$ has the same distribution as $\boldsymbol{a}^\top\mathbf{u}$ where $\mathbf{u} \sim \mathsf{N}(\boldsymbol{\theta}, \sigma^2\boldsymbol{\Sigma})$. This is enough to establish that $\mathbf{y}$ and $\mathbf{u}$ have identical joint distributions. □

THEOREM 10.6.6. *Let* $\mathsf{y} \sim \mathsf{N}(\boldsymbol{\theta}, \sigma^2\boldsymbol{\Sigma})$ *with possibly singular* $\boldsymbol{\Sigma}$. *A necessary and sufficient condition for*

$$\mathsf{q} = (\mathsf{y} - \boldsymbol{\theta})^\top \boldsymbol{P}(\mathsf{y} - \boldsymbol{\theta}) \tag{10.6.8}$$

to be distributed like a $\sigma^2\chi^2$ *is that*

$$\boldsymbol{\Sigma}\boldsymbol{P}\boldsymbol{\Sigma}\boldsymbol{P}\boldsymbol{\Sigma} = \boldsymbol{\Sigma}\boldsymbol{P}\boldsymbol{\Sigma}. \tag{10.6.9}$$

In this case, the number of degrees of freedom of the $\sigma^2\chi^2$ *is* $\operatorname{rank}(\boldsymbol{P}\boldsymbol{\Sigma})$.

Proof: By Theorem 10.6.4, $\mathsf{y} = \boldsymbol{C}\boldsymbol{z} + \boldsymbol{\theta}$ for a standard normal $\boldsymbol{z}$ and a $\boldsymbol{C}$ with $\boldsymbol{C}\boldsymbol{C}^\top = \boldsymbol{\Sigma}$. Therefore $\mathsf{q} = \boldsymbol{z}^\top\boldsymbol{C}^\top\boldsymbol{P}\boldsymbol{C}\boldsymbol{z}$. By theorem 10.4.3 this is a $\sigma^2\chi^2$ iff

$$\boldsymbol{C}^\top\boldsymbol{P}\boldsymbol{C} = \boldsymbol{C}^\top\boldsymbol{P}\boldsymbol{C}\boldsymbol{C}^\top\boldsymbol{P}\boldsymbol{C} \tag{10.6.10}$$

Premultiply by $\boldsymbol{C}$ and postmultiply by $\boldsymbol{C}^\top$ to get (10.6.9). On the other hand, premultipy (10.6.9) by $\boldsymbol{C}^\top\boldsymbol{\Sigma}^-$ and postmultiply by the transpose to get (10.6.10). The number of degrees of freedom is the rank of $\boldsymbol{C}^\top\boldsymbol{P}\boldsymbol{C}$, which is that of $\boldsymbol{\Sigma}\boldsymbol{P}$.

PROBLEM 178. *Assume* $\mathsf{x}_1, \mathsf{x}_2, \ldots, \mathsf{x}_n$ *are independently and identically distributed as* $\mathsf{N}(\theta, \sigma^2)$. *The usual unbiased estimator of* σ^2 *is*

$$\mathsf{s}^2 = \frac{1}{n-1}\sum_i(\mathsf{x}_i - \bar{\mathsf{x}})^2. \tag{10.6.11}$$

Look at the alternative estimator

$$\mathsf{t}^2 = \frac{1}{n+1}\sum_i(\mathsf{x}_i - \bar{\mathsf{x}})^2. \tag{10.6.12}$$

Show that the "mean squared error"

$$\mathrm{E}[(\mathsf{s}^2 - \sigma^2)^2] > \mathrm{E}[(\mathsf{t}^2 - \sigma^2)^2]. \tag{10.6.13}$$

ANSWER. For every estimator $\hat{\theta}$ of a constant parameter θ, $\mathrm{MSE}[\hat{\theta};\theta] = \mathrm{var}[\hat{\theta}] + (\mathrm{E}[\hat{\theta} - \theta])^2$, i.e., it is variance plus squared bias. The MSE of s^2 is therefore equal to its variance, which is $\frac{2\sigma^4}{n-1}$. The alternative $\mathsf{t}^2 = \frac{n-1}{n+1}\mathsf{s}^2$; therefore its bias is $-\frac{2\sigma^2}{n+1}$ and its variance is $\frac{2(n-1)\sigma^4}{(n+1)^2}$, and the MSE is $\frac{2\sigma^4}{n+1}$. □

PROBLEM 179. *The* $n \times 1$ *vector* y *and distribution* $\mathsf{y} \sim \mathsf{N}(\boldsymbol{\iota}\theta, \sigma^2\boldsymbol{I})$. *Show that* $\bar{\mathsf{y}}$ *is independent of* $\mathsf{q} = \sum(\mathsf{y}_i - \bar{\mathsf{y}})^2$, *and that* $\mathsf{q} \sim \sigma^2\chi^2_{n-1}$.

ANSWER. Set $\boldsymbol{z} = \mathsf{y} - \boldsymbol{\iota}\theta$. Then $\mathsf{q} = \sum(z_i - \bar{z})^2$ and $\bar{\mathsf{y}} = \bar{z} + \theta$, and the statement follows from theorem 10.4.2 with $\boldsymbol{P} = \frac{1}{\sqrt{n}}\boldsymbol{\iota}^\top$. □

CHAPTER 11

The Regression Fallacy

Only for the sake of this exercise we will assume that "intelligence" is an innate property of individuals and can be represented by a real number z. If one picks at random a student entering the U of U, the intelligence of this student is a random variable which we assume to be normally distributed with mean μ and standard deviation σ. Also assume every student has to take two intelligence tests, the first at the beginning of his or her studies, the other half a year later. The outcomes of these tests are x and y. x and y measure the intelligence z (which is assumed to be the same in both tests) plus a random error ε and δ, i.e.,

$$\mathsf{x} = \mathsf{z} + \varepsilon \tag{11.0.14}$$

$$\mathsf{y} = \mathsf{z} + \delta \tag{11.0.15}$$

Here $\mathsf{z} \sim \mathsf{N}(\mu, \tau^2)$, $\varepsilon \sim \mathsf{N}(0, \sigma^2)$, and $\delta \sim \mathsf{N}(0, \sigma^2)$ (i.e., we assume that both errors have the same variance). The three variables ε, δ, and z are independent of each other. Therefore x and y are jointly normal. $\operatorname{var}[\mathsf{x}] = \tau^2 + \sigma^2$, $\operatorname{var}[\mathsf{y}] = \tau^2 + \sigma^2$, $\operatorname{cov}[\mathsf{x}, \mathsf{y}] = \operatorname{cov}[\mathsf{z} + \varepsilon, \mathsf{z} + \delta] = \tau^2 + 0 + 0 + 0 = \tau^2$. Therefore $\rho = \frac{\tau^2}{\tau^2 + \sigma^2}$. The contour lines of the joint density are ellipses with center (μ, μ) whose main axes are the lines $y = x$ and $y = -x$ in the x, y-plane.

Now what is the conditional mean? Since $\operatorname{var}[\mathsf{x}] = \operatorname{var}[\mathsf{y}]$, (10.3.17) gives the line $\operatorname{E}[\mathsf{y}|\mathsf{x}{=}x] = \mu + \rho(x - \mu)$, i.e., it is a line which goes through the center of the ellipses but which is flatter than the line $x = y$ representing the real underlying linear relationship if there are no errors. Geometrically one can get it as the line which intersects each ellipse exactly where the ellipse is vertical.

Therefore, the parameters of the best prediction of y on the basis of x are *not* the parameters of the underlying relationship. Why not? Because not only y but also x is subject to errors. Assume you pick an individual by random, and it turns out that his or her first test result is very much higher than the average. Then it is more likely that this is an individual which was lucky in the first exam, and his or her true IQ is lower than the one measured, than that the individual is an Einstein who had a bad day. This is simply because z is normally distributed, i.e., among the students entering a given University, there are more individuals with lower IQ's than

Einsteins. In order to make a good prediction of the result of the second test one must make allowance for the fact that the individual's IQ is most likely lower than his first score indicated, therefore one will predict the second score to be lower than the first score. The converse is true for individuals who scored lower than average, i.e., in your prediction you will do as if a "regression towards the mean" had taken place.

The next important point to note here is: the "true regression line," i.e., the prediction line, is uniquely determined by the joint distribution of x and y. However the line representing the underlying relationship can only be determined if one has information in addition to the joint density, i.e., in addition to the observations. E.g., assume the two tests have different standard deviations, which may be the case simply because the second test has more questions and is therefore more accurate. Then the underlying 45° line is no longer one of the main axes of the ellipse! To be more precise, the underlying line can only be identified if one knows the ratio of the variances, or if one knows one of the two variances. Without any knowledge of the variances, the only thing one can say about the underlying line is that it lies between the line predicting y on the basis of x and the line predicting x on the basis of y.

The name "regression" stems from a confusion between the prediction line and the real underlying relationship. Francis Galton, the cousin of the famous Darwin, measured the height of fathers and sons, and concluded from his evidence that the heights of sons tended to be closer to the average height than the height of the fathers, a purported law of "regression towards the mean." Problem 180 illustrates this:

Problem 180. *The evaluation of two intelligence tests, one at the beginning of the semester, one at the end, gives the following disturbing outcome: While the underlying intelligence during the first test was* $\mathsf{z} \sim \mathsf{N}(100, 20)$*, it changed between the first and second test due to the learning experience at the university. If* w *is the intelligence of each student at the second test, it is connected to his intelligence* z *at the first test by the formula* $\mathsf{w} = 0.5\mathsf{z} + 50$*, i.e., those students with intelligence below 100 gained, but those students with intelligence above 100 lost. (The errors of both intelligence tests are normally distributed with expected value zero, and the variance of the first intelligence test was 5, and that of the second test, which had more questions, was 4. As usual, the errors are independent of each other and of the actual intelligence.)*

• a. *3 points If* x *and* y *are the outcomes of the first and second intelligence test, compute* $\operatorname{E}[\mathsf{x}]$, $\operatorname{E}[\mathsf{y}]$, $\operatorname{var}[\mathsf{x}]$, $\operatorname{var}[\mathsf{y}]$, *and the correlation coefficient* $\rho = \operatorname{corr}[\mathsf{x}, \mathsf{y}]$. *Figure 1 shows an equi-density line of their joint distribution; 95% of the probability mass of the test results are inside this ellipse. Draw the line* $\mathsf{w} = 0.5\mathsf{z} + 50$ *into Figure 1.*

Answer. We know $z \sim \mathsf{N}(100, 20)$; $w = 0.5z + 50$; $\mathsf{x} = z + \varepsilon$; $\varepsilon \sim \mathsf{N}(0, 4)$; $\mathsf{y} = w + \delta$; $\delta \sim \mathsf{N}(0, 5)$; therefore $\mathrm{E}[\mathsf{x}] = 100$; $\mathrm{E}[\mathsf{y}] = 100$; $\mathrm{var}[\mathsf{x}] = 20 + 5 = 25$; $\mathrm{var}[\mathsf{y}] = 5 + 4 = 9$; $\mathrm{cov}[\mathsf{x}, \mathsf{y}] = 10$; $\mathrm{corr}[\mathsf{x}, \mathsf{y}] = 10/15 = 2/3$. In matrix notation

$$\begin{bmatrix} \mathsf{x} \\ \mathsf{y} \end{bmatrix} \sim \mathsf{N}\Big[\begin{bmatrix} 100 \\ 100 \end{bmatrix}, \begin{bmatrix} 25 & 10 \\ 10 & 9 \end{bmatrix}\Big] \tag{11.0.16}$$

The line $\mathsf{y} = 50 + 0.5\mathsf{x}$ goes through the points $(80, 90)$ and $(120, 110)$. □

• b. *4 points Compute* $\mathrm{E}[\mathsf{y}|\mathsf{x}{=}x]$ *and* $\mathrm{E}[\mathsf{x}|\mathsf{y}{=}y]$. *The first is a linear function of* x *and the second a linear function of* y. *Draw the two lines representing these linear functions into Figure 1. Use* (10.3.18) *for this.*

Answer.

$$\mathrm{E}[\mathsf{y}|\mathsf{x}{=}x] = 100 + \frac{10}{25}(x - 100) = 60 + \frac{2}{5}x \tag{11.0.17}$$

$$\mathrm{E}[x|\mathsf{y}{=}y] = 100 + \frac{10}{9}(y - 100) = -\frac{100}{9} + \frac{10}{9}y. \tag{11.0.18}$$

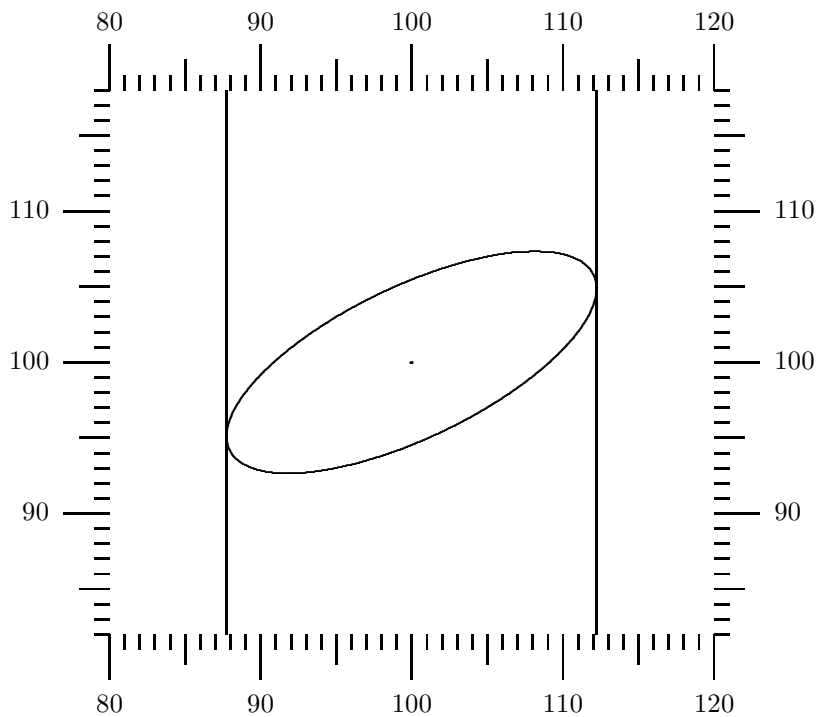


Figure 1. Ellipse containing 95% of the probability mass of test results x and y

The line $y = \mathrm{E}[\mathsf{y}|\mathsf{x}{=}x]$ goes through the points $(80, 92)$ and $(120, 108)$ at the edge of Figure 1; it intersects the ellipse where it is vertical. The line $x = \mathrm{E}[\mathsf{x}|\mathsf{y}{=}y]$ goes through the points $(80, 82)$ and $(120, 118)$, which are the corner points of Figure 1; it intersects the ellipse where it is horizontal. The two lines intersect in the center of the ellipse, i.e., at the point $(100, 100)$. □

• c. *2 points Another researcher says that* $w = \frac{6}{10}z + 40$, $z \sim \mathsf{N}(100, \frac{100}{6})$, $\varepsilon \sim \mathsf{N}(0, \frac{50}{6})$, $\delta \sim \mathsf{N}(0, 3)$. *Is this compatible with the data?*

Answer. Yes, it is compatible: $\mathrm{E}[\mathsf{x}] = \mathrm{E}[z]+\mathrm{E}[\varepsilon] = 100$; $\mathrm{E}[\mathsf{y}] = \mathrm{E}[w]+\mathrm{E}[\delta] = \frac{6}{10}100+40 = 100$; $\mathrm{var}[\mathsf{x}] = \frac{100}{6} + \frac{50}{6} = 25$; $\mathrm{var}[\mathsf{y}] = \left(\frac{6}{10}\right)^2 \mathrm{var}[z] + \mathrm{var}[\delta] = \frac{63}{100}\frac{100}{6} + 3 = 9$; $\mathrm{cov}[\mathsf{x}, \mathsf{y}] = \frac{6}{10}\mathrm{var}[z] = 10$. □

• d. *4 points A third researcher asserts that the IQ of the students really did not change. He says* $\mathsf{w} = \mathsf{z}$, $\mathsf{z} \sim \mathsf{N}(100, 5)$, $\varepsilon \sim \mathsf{N}(0, 20)$, $\delta \sim \mathsf{N}(0, 4)$. *Is this compatible with the data? Is there unambiguous evidence in the data that the IQ declined?*

Answer. This is not compatible. This scenario gets everything right except the covariance: $\mathrm{E}[\mathsf{x}] = \mathrm{E}[\mathsf{z}] + \mathrm{E}[\varepsilon] = 100$; $\mathrm{E}[\mathsf{y}] = \mathrm{E}[\mathsf{z}] + \mathrm{E}[\delta] = 100$; $\mathrm{var}[\mathsf{x}] = 5 + 20 = 25$; $\mathrm{var}[\mathsf{y}] = 5 + 4 = 9$; $\mathrm{cov}[\mathsf{x}, \mathsf{y}] = 5$. A scenario in which both tests have same underlying intelligence cannot be found. Since the two conditional expectations are on the same side of the diagonal, the hypothesis that the intelligence did not change between the two tests is not consistent with the joint distribution of x and y. The diagonal goes through the points $(82, 82)$ and $(118, 118)$, i.e., it intersects the two horizontal boundaries of Figure 1. □

We just showed that the parameters of the true underlying relationship cannot be inferred from the data alone if there are errors in both variables. We also showed that this lack of identification is not complete, because one can specify an interval which in the plim contains the true parameter value.

Chapter 53 has a much more detailed discussion of all this. There we will see that this lack of identification can be removed if more information is available, i.e., if one knows that the two error variances are equal, or if one knows that the regression has zero intercept, etc. Question 181 shows that in this latter case, the OLS estimate is not consistent, but other estimates exist that are consistent.

Problem 181. [**Fri57**, chapter 3] *According to Friedman's permanent income hypothesis, drawing at random families in a given country and asking them about their income* y *and consumption* c *can be modeled as the independent observations of two random variables which satisfy*

$$\mathsf{y} = \mathsf{y}^{\mathsf{p}} + \mathsf{y}^{\mathsf{t}}, \tag{11.0.19}$$

$$\mathsf{c} = \mathsf{c}^{\mathsf{p}} + \mathsf{c}^{\mathsf{t}}, \tag{11.0.20}$$

$$\mathsf{c}^{\mathsf{p}} = \beta\mathsf{y}^{\mathsf{p}}. \tag{11.0.21}$$

Here y^{p} *and* c^{p} *are the permanent and* y^{t} *and* c^{t} *the transitory components of income and consumption. These components are not observed separately, only their*

sums y *and* c *are observed. We assume that the permanent income* y^{p} *is random, with* $\mathrm{E}[\mathsf{y}^{\mathsf{p}}] = \mu \neq 0$ *and* $\mathrm{var}[\mathsf{y}^{\mathsf{p}}] = \tau_{\mathsf{y}}^2$. *The transitory components* y^{t} *and* c^{t} *are assumed to be independent of each other and of* y^{p}, *and* $\mathrm{E}[\mathsf{y}^{\mathsf{t}}] = 0$, $\mathrm{var}[\mathsf{y}^{\mathsf{t}}] = \sigma_{\mathsf{y}}^2$, $\mathrm{E}[\mathsf{c}^{\mathsf{t}}] = 0$, *and* $\mathrm{var}[\mathsf{c}^{\mathsf{t}}] = \sigma_{\mathsf{c}}^2$. *Finally, it is assumed that all variables are normally distributed.*

• a. *2 points Given the above information, write down the vector of expected values* $\mathcal{E}\big[\big[\begin{smallmatrix}\mathsf{y}\\ \mathsf{c}\end{smallmatrix}\big]\big]$ *and the covariance matrix* $\mathcal{V}\big[\big[\begin{smallmatrix}\mathsf{y}\\ \mathsf{c}\end{smallmatrix}\big]\big]$ *in terms of the five unknown parameters of the model* μ, β, τ_{y}^2, σ_{y}^2, *and* σ_{c}^2.

Answer.

$$\mathcal{E}\Big[\begin{bmatrix}\mathsf{y}\\ \mathsf{c}\end{bmatrix}\Big] = \begin{bmatrix}\mu\\ \beta\mu\end{bmatrix} \quad \text{and} \quad \mathcal{V}\Big[\begin{bmatrix}\mathsf{y}\\ \mathsf{c}\end{bmatrix}\Big] = \begin{bmatrix}\tau_{\mathsf{y}}^2+\sigma_{\mathsf{y}}^2 & \beta\tau_{\mathsf{y}}^2\\ \beta\tau_{\mathsf{y}}^2 & \beta^2\tau_{\mathsf{y}}^2+\sigma_{\mathsf{c}}^2\end{bmatrix}. \tag{11.0.22}$$

□

• b. *3 points Assume that you know the true parameter values and you observe a family's actual income* y. *Show that your best guess (minimum mean squared error) of this family's permanent income* y^{p} *is*

$$\mathsf{y}^{\mathsf{p}*} = \frac{\sigma_{\mathsf{y}}^2}{\tau_{\mathsf{y}}^2+\sigma_{\mathsf{y}}^2}\mu + \frac{\tau_{\mathsf{y}}^2}{\tau_{\mathsf{y}}^2+\sigma_{\mathsf{y}}^2}\mathsf{y}. \tag{11.0.23}$$

Note: here we are guessing income, not yet consumption! Use (10.3.17) *for this!*

Answer. This answer also does the math for part *c*. The best guess is the conditional mean

$$\begin{aligned}\mathrm{E}[\mathsf{y}^{\mathsf{p}}|y = 22{,}000] &= \mathrm{E}[\mathsf{y}^{\mathsf{p}}] + \frac{\mathrm{cov}[\mathsf{y}^{\mathsf{p}},\mathsf{y}]}{\mathrm{var}[\mathsf{y}]}(22{,}000 - \mathrm{E}[\mathsf{y}])\\ &= 12{,}000 + \frac{16{,}000{,}000}{20{,}000{,}000}(22{,}000 - 12{,}000) = 20{,}000\end{aligned}$$

or equivalently

$$\begin{aligned}\mathrm{E}[\mathsf{y}^{\mathsf{p}}|y = 22{,}000] &= \mu + \frac{\tau_{\mathsf{y}}^2}{\tau_{\mathsf{y}}^2+\sigma_{\mathsf{y}}^2}(22{,}000 - \mu)\\ &= \frac{\sigma_{\mathsf{y}}^2}{\tau_{\mathsf{y}}^2+\sigma_{\mathsf{y}}^2}\mu + \frac{\tau_{\mathsf{y}}^2}{\tau_{\mathsf{y}}^2+\sigma_{\mathsf{y}}^2}22{,}000\\ &= (0.2)(12{,}000) + (0.8)(22{,}000) = 20{,}000.\end{aligned}$$

□

• c. *3 points To make things more concrete, assume the parameters are*

$$\beta = 0.7 \tag{11.0.24}$$
$$\sigma_{\mathsf{y}} = 2{,}000 \tag{11.0.25}$$
$$\sigma_{\mathsf{c}} = 1{,}000 \tag{11.0.26}$$
$$\mu = 12{,}000 \tag{11.0.27}$$
$$\tau_{\mathsf{y}} = 4{,}000. \tag{11.0.28}$$

If a family's income is $y = 22{,}000$, *what is your best guess of this family's permanent income* y^{p}? *Give an intuitive explanation why this best guess is smaller than* 22,000.

Answer. Since the observed income of 22,000 is above the average of 12,000, chances are greater that it is someone with a positive transitory income than someone with a negative one. □

• d. *2 points If a family's income is* y, *show that your best guess about this family's* consumption *is*

$$\mathsf{c}^* = \beta\Big(\frac{\sigma_{\mathsf{y}}^2}{\tau_{\mathsf{y}}^2+\sigma_{\mathsf{y}}^2}\mu + \frac{\tau_{\mathsf{y}}^2}{\tau_{\mathsf{y}}^2+\sigma_{\mathsf{y}}^2}\mathsf{y}\Big). \tag{11.0.29}$$

Instead of an exact mathematical proof you may also reason out how it can be obtained from (11.0.23). *Give the numbers for a family whose actual income is 22,000.*

Answer. This is 0.7 times the best guess about the family's permanent income, since the transitory consumption is uncorrelated with everything else and therefore must be predicted by 0. This is an acceptable answer, but one can also derive it from scratch:

$$\mathrm{E}[\mathsf{c}|y = 22{,}000] = \mathrm{E}[\mathsf{c}] + \frac{\mathrm{cov}[\mathsf{c},\mathsf{y}]}{\mathrm{var}[\mathsf{y}]}(22{,}000 - \mathrm{E}[\mathsf{y}]) \tag{11.0.30}$$

$$= \beta\mu + \frac{\beta\tau_{\mathsf{y}}^2}{\tau_{\mathsf{y}}^2+\sigma_{\mathsf{y}}^2}(22{,}000 - \mu) = 8{,}400 + 0.7\frac{16{,}000{,}000}{20{,}000{,}000}(22{,}000 - 12{,}000) = 14{,}000 \tag{11.0.31}$$

$$\text{or} \quad = \beta\Big(\frac{\sigma_{\mathsf{y}}^2}{\tau_{\mathsf{y}}^2+\sigma_{\mathsf{y}}^2}\mu + \frac{\tau_{\mathsf{y}}^2}{\tau_{\mathsf{y}}^2+\sigma_{\mathsf{y}}^2}22{,}000\Big) \tag{11.0.32}$$

$$= 0.7\Big((0.2)(12{,}000) + (0.8)(22{,}000)\Big) = (0.7)(20{,}000) = 14{,}000. \tag{11.0.33}$$

□

The remainder of this Problem uses material that comes later in these Notes:

• e. *4 points From now on we will assume that the true values of the parameters are not known, but two vectors* $\mathbf{y}$ *and* $\mathbf{c}$ *of independent observations are available. We will show that it is not correct in this situation to estimate* β *by regressing* $\mathbf{c}$ *on*

y *with the intercept suppressed. This would give the estimator*

$$\hat{\beta} = \frac{\sum \mathsf{c}_i \mathsf{y}_i}{\sum \mathsf{y}_i^2} \tag{11.0.34}$$

Show that the plim *of this estimator is*

$$\operatorname{plim}[\hat{\beta}] = \frac{\mathrm{E}[\mathsf{c}\mathsf{y}]}{\mathrm{E}[\mathsf{y}^2]} \tag{11.0.35}$$

Which theorems do you need for this proof? Show that $\hat{\beta}$ *is an inconsistent estimator of* β*, which yields too small values for* β*.*

Answer. First rewrite the formula for $\hat{\beta}$ in such a way that numerator and denominator each has a plim: by the weak law of large numbers the plim of the average is the expected value, therefore we have to divide both numerator and denominator by n. Then we can use the Slutsky theorem that the plim of the fraction is the fraction of the plims.

$$\hat{\beta} = \frac{\frac{1}{n}\sum \mathsf{c}_i \mathsf{y}_i}{\frac{1}{n}\sum \mathsf{y}_i^2}; \quad \operatorname{plim}[\hat{\beta}] = \frac{\mathrm{E}[\mathsf{c}\mathsf{y}]}{\mathrm{E}[\mathsf{y}^2]} = \frac{\mathrm{E}[\mathsf{c}]\,\mathrm{E}[\mathsf{y}] + \operatorname{cov}[\mathsf{c},\mathsf{y}]}{(\mathrm{E}[\mathsf{y}])^2 + \operatorname{var}[\mathsf{y}]} = \frac{\mu\beta\mu + \beta\tau_{\mathsf{y}}^2}{\mu^2 + \tau_{\mathsf{y}}^2 + \sigma_{\mathsf{y}}^2} = \beta\frac{\mu^2 + \tau_{\mathsf{y}}^2}{\mu^2 + \tau_{\mathsf{y}}^2 + \sigma_{\mathsf{y}}^2}.$$

□

• f. *4 points Give the formulas of the method of moments estimators of the five paramaters of this model:* μ*,* β*,* τ_{y}^2*,* σ_{y}^2*, and* σ_{p}^2*. (For this you have to express these five parameters in terms of the five moments* $\mathrm{E}[\mathsf{y}]$*,* $\mathrm{E}[\mathsf{c}]$*,* $\operatorname{var}[\mathsf{y}]$*,* $\operatorname{var}[\mathsf{c}]$*, and* $\operatorname{cov}[\mathsf{y},\mathsf{c}]$*, and then simply replace the population moments by the sample moments.) Are these consistent estimators?*

Answer. From (11.0.22) follows $\mathrm{E}[\mathsf{c}] = \beta\,\mathrm{E}[\mathsf{y}]$, therefore $\beta = \frac{\mathrm{E}[\mathsf{c}]}{\mathrm{E}[\mathsf{y}]}$. This together with $\operatorname{cov}[\mathsf{y},\mathsf{c}] = \beta\tau_{\mathsf{y}}^2$ gives $\tau_{\mathsf{y}}^2 = \frac{\operatorname{cov}[\mathsf{y},\mathsf{c}]}{\beta} = \frac{\operatorname{cov}[\mathsf{y},\mathsf{c}]\,\mathrm{E}[\mathsf{y}]}{\mathrm{E}[\mathsf{c}]}$. This together with $\operatorname{var}[\mathsf{y}] = \tau_{\mathsf{y}}^2 + \sigma_{\mathsf{y}}^2$ gives $\sigma_{\mathsf{y}}^2 = \operatorname{var}[\mathsf{y}] - \tau_{\mathsf{y}}^2 = \operatorname{var}[\mathsf{y}] - \frac{\operatorname{cov}[\mathsf{y},\mathsf{c}]\,\mathrm{E}[\mathsf{y}]}{\mathrm{E}[\mathsf{c}]}$. And from the last equation $\operatorname{var}[\mathsf{c}] = \beta^2\tau_{\mathsf{y}}^2 + \sigma_{\mathsf{c}}^2$ one get $\sigma_{\mathsf{c}}^2 = \operatorname{var}[\mathsf{c}] - \frac{\operatorname{cov}[\mathsf{y},\mathsf{c}]\,\mathrm{E}[\mathsf{c}]}{\mathrm{E}[\mathsf{y}]}$. All these are consistent estimators, as long as $\mathrm{E}[\mathsf{y}] \neq 0$ and $\beta \neq 0$. □

• g. *4 points Now assume you are not interested in estimating* β *itself, but in addition to the two* n*-vectors* $\mathbf{y}$ *and* $\mathbf{c}$ *you have an observation of* y_{n+1} *and you want to predict the corresponding* c_{n+1}*. One obvious way to do this would be to plug the method-of moments estimators of the unknown parameters into formula* (11.0.29) *for the best linear predictor. Show that this is equivalent to using the ordinary least squares predictor* $\mathsf{c}^* = \hat{\alpha} + \hat{\beta}\mathsf{y}_{n+1}$ *where* $\hat{\alpha}$ *and* $\hat{\beta}$ *are intercept and slope in the simple regression of* $\mathbf{c}$ *on* $\mathbf{y}$*, i.e.,*

$$\hat{\beta} = \frac{\sum(\mathsf{y}_i - \bar{\mathsf{y}})(\mathsf{c}_i - \bar{\mathsf{c}})}{\sum(\mathsf{y}_i - \bar{\mathsf{y}})^2} \tag{11.0.36}$$

$$\hat{\alpha} = \bar{\mathsf{c}} - \hat{\beta}\bar{\mathsf{y}} \tag{11.0.37}$$

Note that we are regressing $\mathbf{c}$ *on* $\mathbf{y}$ with *an intercept, although the original model does not have an intercept.*

Answer. Here I am writing population moments where I should be writing sample moments. First substitute the method of moments estimators in the denominator in (11.0.29): $\tau_{\mathsf{y}}^2 + \sigma_{\mathsf{y}}^2 = \operatorname{var}[\mathsf{y}]$. Therefore the first summand becomes

$$\beta\sigma_{\mathsf{y}}^2\mu\frac{1}{\operatorname{var}[\mathsf{y}]} = \frac{\mathrm{E}[\mathsf{c}]}{\mathrm{E}[\mathsf{y}]}\left(\operatorname{var}[\mathsf{y}] - \frac{\operatorname{cov}[\mathsf{y},\mathsf{c}]\,\mathrm{E}[\mathsf{y}]}{\mathrm{E}[\mathsf{c}]}\right)\mathrm{E}[\mathsf{y}]\frac{1}{\operatorname{var}[\mathsf{y}]} = \mathrm{E}[\mathsf{c}]\left(1 - \frac{\operatorname{cov}[\mathsf{y},\mathsf{c}]\,\mathrm{E}[\mathsf{y}]}{\operatorname{var}[\mathsf{y}]\,\mathrm{E}[\mathsf{c}]}\right) = \mathrm{E}[\mathsf{c}] - \frac{\operatorname{cov}[\mathsf{y},\mathsf{c}]\,\mathrm{E}[\mathsf{y}]}{\operatorname{var}[\mathsf{y}]}$$

But since $\frac{\operatorname{cov}[\mathsf{y},\mathsf{c}]}{\operatorname{var}[\mathsf{y}]} = \hat{\beta}$ and $\hat{\alpha} + \hat{\beta}\,\mathrm{E}[\mathsf{y}] = \mathrm{E}[\mathsf{c}]$ this expression is simply $\hat{\alpha}$. The second term is easier to show:

$$\beta\frac{\tau_{\mathsf{y}}^2}{\operatorname{var}[\mathsf{y}]}\mathsf{y} = \frac{\operatorname{cov}[\mathsf{y},\mathsf{c}]}{\operatorname{var}[\mathsf{y}]}\mathsf{y} = \hat{\beta}\mathsf{y}$$

□

• h. *2 points What is the "Iron Law of Econometrics," and how does the above relate to it?*

Answer. The Iron Law says that all effects are underestimated because of errors in the independent variable. Friedman says Keynesians obtain their low marginal propensity to consume due to the "Iron Law of Econometrics": they ignore that actual income is a measurement with error of the true underlying variable, permanent income. □

Problem 182. *This question follows the original article* [**SW76**] *much more closely than* [**HVdP02**] *does. Sargent and Wallace first reproduce the usual argument why "activist" policy rules, in which the Fed "looks at many things" and "leans against the wind," are superior to policy rules without feedback as promoted by the monetarists.*

They work with a very stylized model in which national income is represented by the following time series:

$$\mathsf{y}_t = \alpha + \lambda\mathsf{y}_{t-1} + \beta\mathsf{m}_t + \mathsf{u}_t \tag{11.0.38}$$

Here y_t *is GNP, measured as its deviation from "potential" GNP or as unemployment rate, and* m_t *is the rate of growth of the money supply. The random disturbance* u_t *is assumed independent of* y_{t-1}*, it has zero expected value, and its variance* $\operatorname{var}[\mathsf{u}_t]$ *is constant over time, we will call it* $\operatorname{var}[\mathsf{u}]$ *(no time subscript).*

• a. *4 points First assume that the Fed tries to maintain a constant money supply, i.e.,* $\mathsf{m}_t = g_0 + \varepsilon_t$ *where* g_0 *is a constant, and* ε_t *is a random disturbance since the Fed does not have full control over the money supply. The* ε_t *have zero expected value; they are serially uncorrelated, and they are independent of the* u_t*. This constant money supply rule does not necessarily make* y_t *a stationary time series (i.e., a time series where mean, variance, and covariances do not depend on* t*), but if* $|\lambda| < 1$ *then* y_t converges *towards a stationary time series, i.e., any initial deviations from the "steady state" die out over time. You are not required here to prove that the time series converges towards a stationary time series, but you are asked to compute* $\mathrm{E}[\mathsf{y}_t]$ *in this stationary time series.*

• b. *8 points Now assume the policy makers want to steer the economy towards a desired steady state, call it y^*, which they think makes the best tradeoff between unemployment and inflation, by setting $\mathfrak{m}_t$ according to a rule with feedback:*

$$\mathfrak{m}_t = g_0 + g_1\mathfrak{y}_{t-1} + \varepsilon_t \tag{11.0.39}$$

Show that the following values of g_0 and g_1

$$g_0 = (y^* - \alpha)/\beta \qquad\qquad g_1 = -\lambda/\beta \tag{11.0.40}$$

represent an optimal monetary policy, since they bring the expected value of the steady state $\mathrm{E}[\mathfrak{y}_t]$ to y^* and minimize the steady state variance $\mathrm{var}[\mathfrak{y}_t]$.

• c. *3 points This is the conventional reasoning which comes to the result that a policy rule with feedback, i.e., a policy rule in which $g_1 \neq 0$, is better than a policy rule without feedback. Sargent and Wallace argue that there is a flaw in this reasoning. Which flaw?*

• d. *5 points A possible system of structural equations from which* (11.0.38) *can be derived are equations* (11.0.41)–(11.0.43) *below. Equation* (11.0.41) *indicates that unanticipated increases in the growth rate of the money supply increase output, while anticipated ones do not. This is a typical assumption of the rational expectations school (Lucas supply curve).*

$$\mathfrak{y}_t = \xi_0 + \xi_1(\mathfrak{m}_t - \mathrm{E}_{t-1}\,\mathfrak{m}_t) + \xi_2\mathfrak{y}_{t-1} + \mathfrak{u}_t \tag{11.0.41}$$

The Fed uses the policy rule

$$\mathfrak{m}_t = g_0 + g_1\mathfrak{y}_{t-1} + \varepsilon_t \tag{11.0.42}$$

and the agents know this policy rule, therefore

$$\mathrm{E}_{t-1}\,\mathfrak{m}_t = g_0 + g_1\mathfrak{y}_{t-1}. \tag{11.0.43}$$

Show that in this system, the parameters g_0 and g_1 have no influence on the time path of $\mathfrak{y}$.

• e. *4 points On the other hand, the econometric estimations which the policy makers are running seem to show that these coefficients have an impact. During a certain period during which a constant policy rule g_0, g_1 is followed, the econometricians regress $\mathfrak{y}_t$ on $\mathfrak{y}_{t-1}$ and $\mathfrak{m}_t$ in order to estimate the coefficients in* (11.0.38). *Which values of α, λ, and β will such a regression yield?*

CHAPTER 12

A Simple Example of Estimation

We will discuss here a simple estimation problem, which can be considered the prototype of all least squares estimation. Assume we have n independent observations $y_1, \ldots, y_n$ of a Normally distributed random variable $\mathsf{y} \sim \mathsf{N}(\mu, \sigma^2)$ with unknown location parameter μ and dispersion parameter σ^2. Our goal is to estimate the *location parameter* and also estimate some measure of the precision of this estimator.

12.1. Sample Mean as Estimator of the Location Parameter

The obvious (and in many cases also the best) estimate of the location parameter of a distribution is the sample mean $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$. Why is this a reasonable estimate?

1. The location parameter of the Normal distribution is its expected value, and by the weak law of large numbers, the probability limit for $n \to \infty$ of the sample mean is the expected value.

2. The expected value μ is sometimes called the "population mean," while $\bar{y}$ is the sample mean. This terminology indicates that there is a correspondence between population quantities and sample quantities, which is often used for estimation. This is the principle of estimating the unknown distribution of the population by the empirical distribution of the sample. Compare Problem 63.

3. This estimator is also unbiased. By definition, an estimator t of the parameter θ is unbiased if $\mathrm{E}[\mathsf{t}] = \theta$. $\bar{y}$ is an unbiased estimator of μ, since $\mathrm{E}[\bar{y}] = \mu$.

4. Given n observations $y_1, \ldots, y_n$, the sample mean is the number $a = \bar{y}$ which minimizes $(y_1 - a)^2 + (y_2 - a)^2 + \cdots + (y_n - a)^2$. One can say it is the number whose squared distance to the given sample numbers is smallest. This idea is generalized in the least squares principle of estimation. It follows from the following frequently used fact:

5. In the case of normality the sample mean is also the maximum likelihood estimate.

Problem 183. *4 points Let $y_1, \ldots, y_n$ be an arbitrary vector and α an arbitrary number. As usual, $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$. Show that*

$$\sum_{i=1}^n (y_i - \alpha)^2 = \sum_{i=1}^n (y_i - \bar{y})^2 + n(\bar{y} - \alpha)^2 \tag{12.1.1}$$

Answer.

$$\sum_{i=1}^n (y_i - \alpha)^2 = \sum_{i=1}^n \big((y_i - \bar{y}) + (\bar{y} - \alpha)\big)^2 \tag{12.1.2}$$

$$= \sum_{i=1}^n (y_i - \bar{y})^2 + 2\sum_{i=1}^n \big((y_i - \bar{y})(\bar{y} - \alpha)\big) + \sum_{i=1}^n (\bar{y} - \alpha)^2 \tag{12.1.3}$$

$$= \sum_{i=1}^n (y_i - \bar{y})^2 + 2(\bar{y} - \alpha)\sum_{i=1}^n (y_i - \bar{y}) + n(\bar{y} - \alpha)^2 \tag{12.1.4}$$

Since the middle term is zero, (12.1.1) follows.

□

Problem 184. *2 points Let $\boldsymbol{y}$ be a n-vector. (It may be a vector of observations of a random variable y, but it does not matter how the y_i were obtained.) Prove that the scalar α which minimizes the sum*

$$(y_1 - \alpha)^2 + (y_2 - \alpha)^2 + \cdots + (y_n - \alpha)^2 = \sum (y_i - \alpha)^2 \tag{12.1.5}$$

is the arithmetic mean $\alpha = \bar{y}$.

Answer. Use (12.1.1). □

Problem 185. *Give an example of a distribution in which the sample mean is not a good estimate of the location parameter. Which other estimate (or estimates) would be preferable in that situation?*

12.2. Intuition of the Maximum Likelihood Estimator

In order to make intuitively clear what is involved in maximum likelihood estimation, look at the simplest case $\mathsf{y} = \mu + \varepsilon$, $\varepsilon \sim \mathsf{N}(0, 1)$, where μ is an unknown parameter. In other words: we know that one of the functions shown in Figure 1 is the density function of y, but we do not know which:

Assume we have only one observation y. What is then the MLE of μ? It is that $\tilde{\mu}$ for which the value of the likelihood function, evaluated at y, is greatest. I.e., you look at all possible density functions and pick the one which is highest at point y, and use the μ which belongs this density as your estimate.

2) Now assume two independent observations of y are given, y_1 and y_2. The family of density functions is still the same. Which of these density functions do we choose now? The one for which the *product* of the ordinates over y_1 and y_2 gives

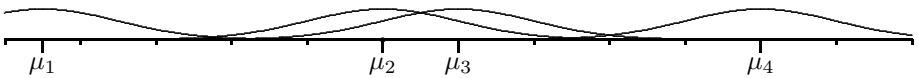


FIGURE 1. Possible Density Functions for $\mathfrak{y}$

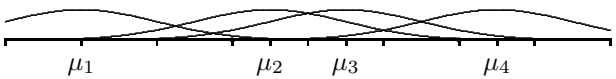


FIGURE 2. Two observations, $\sigma^2 = 1$

FIGURE 3. Two observations, σ^2 unknown

the highest value. For this the peak of the density function must be exactly in the middle between the two observations.

3) Assume again that we made two independent observations y_1 and y_2 of $\mathfrak{y}$, but this time not only the expected value but also the variance of $\mathfrak{y}$ is unknown, call it σ^2. This gives a larger family of density functions to choose from: they do not only differ by location, but some are low and fat and others tall and skinny.

For which density function is the product of the ordinates over y_1 and y_2 the largest again? Before even knowing our estimate of σ^2 we can already tell what $\tilde{\mu}$ is: it must again be $(y_1+y_2)/2$. Then among those density functions which are centered over $(y_1 + y_2)/2$, there is one which is highest over y_1 and y_2. Figure 4 shows the densities for standard deviations 0.01, 0.05, 0.1, 0.5, 1, and 5. All curves, except the last one, are truncated at the point where the resolution of TEX can no longer distinguish between their level and zero. For the last curve this point would only be reached at the coordinates ± 25.

4) If we have many observations, then the density pattern of the observations, as indicated by the histogram below, approximates the actual density function of $\mathfrak{y}$ itself. That likelihood function must be chosen which has a high value where the points are dense, and which has a low value where the points are not so dense.

12.2.1. Precision of the Estimator. How good is $\bar{\mathfrak{y}}$ as estimate of μ? To answer this question we need some criterion how to measure "goodness." Assume your

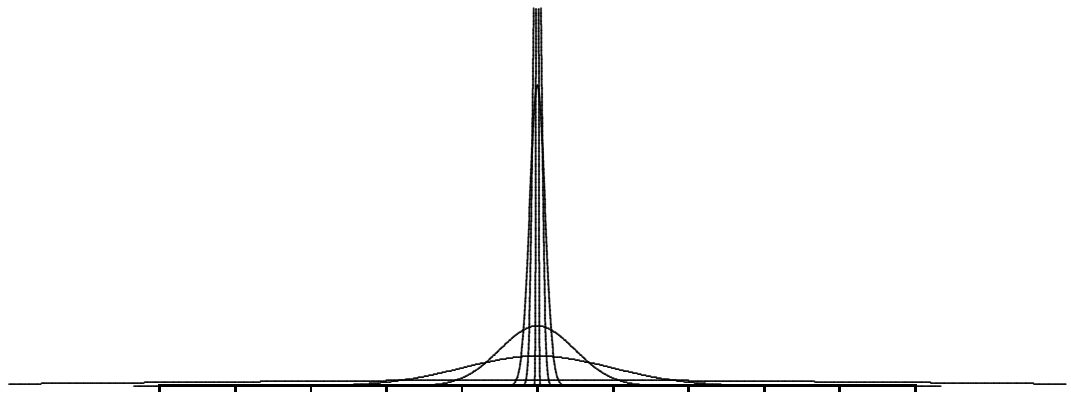


FIGURE 4. Only those centered over the two observations need to be considered

FIGURE 5. Many Observations

business depends on the precision of the estimate $\hat{\mu}$ of μ. It incurs a penalty (extra cost) amounting to $(\hat{\mu} - \mu)^2$. You don't know what this error will be beforehand, but the expected value of this "loss function" may be an indication how good the estimate is. Generally, the expected value of a loss function is called the "risk," and for the quadratic loss function $\mathrm{E}[(\hat{\mu} - \mu)^2]$ it has the name "mean squared error of $\hat{\mu}$ as an estimate of μ," write it $\mathrm{MSE}[\hat{\mu}; \mu]$. What is the mean squared error of $\bar{\mathfrak{y}}$? Since $\mathrm{E}[\bar{\mathfrak{y}}] = \mu$, it is $\mathrm{E}[(\bar{\mathfrak{y}} - \mathrm{E}[\bar{\mathfrak{y}}])^2] = \mathrm{var}[\bar{\mathfrak{y}}] = \frac{\sigma^2}{n}$.

Note that the MSE of $\bar{\mathfrak{y}}$ as an estimate of μ does not depend on μ. This is convenient, since usually the MSE depends on unknown parameters, and therefore one usually does not know how good the estimator is. But it has more important advantages. For any estimator $\tilde{\mathfrak{y}}$ of μ follows $\mathrm{MSE}[\tilde{\mathfrak{y}}; \mu] = \mathrm{var}[\tilde{\mathfrak{y}}] + (\mathrm{E}[\tilde{\mathfrak{y}}] - \mu)^2$. If $\tilde{\mathfrak{y}}$ is linear (perhaps with a constant term), then $\mathrm{var}[\tilde{\mathfrak{y}}]$ is a constant which does not depend on μ, therefore the MSE is a constant if $\tilde{\mathfrak{y}}$ is unbiased and a quadratic function of μ (parabola) if $\tilde{\mathfrak{y}}$ is biased. Since a parabola is an unbounded function, a biased linear estimator has therefore the disadvantage that for certain values of μ its MSE may be very high. Some estimators are very good when μ is in one area, and very bad when μ is in another area. Since our unbiased estimator $\bar{\mathfrak{y}}$ has *bounded* MSE, it will not let us down, wherever nature has hidden the μ.

On the other hand, the MSE does depend on the unknown σ^2. So we have to estimate σ^2.

12.3. Variance Estimation and Degrees of Freedom

It is not so clear what the best estimator of σ^2 is. At least two possibilities are in common use:

$$\mathsf{s}_m^2 = \frac{1}{n}\sum(\mathsf{y}_i - \bar{\mathsf{y}})^2 \tag{12.3.1}$$

or

$$\mathsf{s}_u^2 = \frac{1}{n-1}\sum(\mathsf{y}_i - \bar{\mathsf{y}})^2. \tag{12.3.2}$$

Let us compute the expected value of our two estimators. Equation (12.1.1) with $\alpha = \mathrm{E}[\mathsf{y}]$ allows us to simplify the sum of squared errors so that it becomes easy to take expected values:

$$\mathrm{E}[\sum_{i=1}^n(\mathsf{y}_i - \bar{\mathsf{y}})^2] = \sum_{i=1}^n \mathrm{E}[(\mathsf{y}_i - \mu)^2] - n\,\mathrm{E}[(\bar{\mathsf{y}} - \mu)^2] \tag{12.3.3}$$

$$= \sum_{i=1}^n \sigma^2 - n\frac{\sigma^2}{n} = (n-1)\sigma^2. \tag{12.3.4}$$

because $\mathrm{E}[(\mathsf{y}_i - \mu)^2] = \mathrm{var}[\mathsf{y}_i] = \sigma^2$ and $\mathrm{E}[(\bar{\mathsf{y}} - \mu)^2] = var[\bar{\mathsf{y}}] = \frac{\sigma^2}{n}$. Therefore, if we use as estimator of σ^2 the quantity

$$\mathsf{s}_u^2 = \frac{1}{n-1}\sum_{i=1}^n(\mathsf{y}_i - \bar{\mathsf{y}})^2 \tag{12.3.5}$$

then this is an unbiased estimate.

Problem 186. *4 points Show that*

$$\mathsf{s}_u^2 = \frac{1}{n-1}\sum_{i=1}^n(\mathsf{y}_i - \bar{\mathsf{y}})^2 \tag{12.3.6}$$

is an unbiased estimator of the variance. List the assumptions which have to be made about y_i so that this proof goes through. Do you need Normality of the individual observations y_i to prove this?

Answer. Use equation (12.1.1) with $\alpha = \mathrm{E}[\mathsf{y}]$:

$$\mathrm{E}[\sum_{i=1}^n(\mathsf{y}_i - \bar{\mathsf{y}})^2] = \sum_{i=1}^n \mathrm{E}[(\mathsf{y}_i - \mu)^2] - n\,\mathrm{E}[(\bar{\mathsf{y}} - \mu)^2] \tag{12.3.7}$$

$$= \sum_{i=1}^n \sigma^2 - n\frac{\sigma^2}{n} = (n-1)\sigma^2. \tag{12.3.8}$$

You do not need Normality for this. □

For testing, confidence intervals, etc., one also needs to know the probability distribution of s_u^2. For this look up once more Section 5.9 about the Chi-Square distribution. There we introduced the terminology that a random variable q is distributed as a $\sigma^2\chi^2$ iff q/σ^2 is a χ^2. In our model with n independent normal variables y_i with same mean and variance, the variable $\sum(\mathsf{y}_i - \bar{\mathsf{y}})^2$ is a $\sigma^2\chi^2_{n-1}$. Problem 187 gives a proof of this in the simplest case $n = 2$, and Problem 188 looks at the case $n = 3$. But it is valid for higher n too. Therefore s_u^2 is a $\frac{\sigma^2}{n-1}\chi^2_{n-1}$. This is remarkable: the distribution of s_u^2 does not depend on μ. Now use (5.9.5) to get the variance of s_u^2: it is $\frac{2\sigma^4}{n-1}$.

Problem 187. *Let y_1 and y_2 be two independent Normally distributed variables with mean μ and variance σ^2, and let $\bar{\mathsf{y}}$ be their arithmetic mean.*

- a. *2 points Show that*

$$\mathsf{SSE} = \sum_{i-1}^2(\mathsf{y}_i - \bar{\mathsf{y}})^2 \sim \sigma^2\chi^2_1 \tag{12.3.9}$$

Hint: Find a Normally distributed random variable z with expected value 0 and variance 1 such that $\mathsf{SSE} = \sigma^2\mathsf{z}^2$.

Answer.

$$\bar{\mathsf{y}} = \frac{\mathsf{y}_1 + \mathsf{y}_2}{2} \tag{12.3.10}$$

$$\mathsf{y}_1 - \bar{\mathsf{y}} = \frac{\mathsf{y}_1 - \mathsf{y}_2}{2} \tag{12.3.11}$$

$$\mathsf{y}_2 - \bar{\mathsf{y}} = -\frac{\mathsf{y}_1 - \mathsf{y}_2}{2} \tag{12.3.12}$$

$$(\mathsf{y}_1 - \bar{\mathsf{y}})^2 + (\mathsf{y}_2 - \bar{\mathsf{y}})^2 = \frac{(\mathsf{y}_1 - \mathsf{y}_2)^2}{4} + \frac{(\mathsf{y}_1 - \mathsf{y}_2)^2}{4} \tag{12.3.13}$$

$$= \frac{(\mathsf{y}_1 - \mathsf{y}_2)^2}{2} = \sigma^2\left(\frac{\mathsf{y}_1 - \mathsf{y}_2}{\sqrt{2\sigma^2}}\right)^2, \tag{12.3.14}$$

and since $\mathsf{z} = (\mathsf{y}_1 - \mathsf{y}_2)/\sqrt{2\sigma^2} \sim \mathrm{N}(0,1)$, its square is a χ^2_1. □

- b. *4 points Write down the covariance matrix of the vector*

$$\begin{bmatrix}\mathsf{y}_1 - \bar{\mathsf{y}} \\ \mathsf{y}_2 - \bar{\mathsf{y}}\end{bmatrix} \tag{12.3.15}$$

and show that it is singular.

Answer. (12.3.11) and (12.3.12) give

$$\begin{bmatrix}\mathsf{y}_1 - \bar{\mathsf{y}} \\ \mathsf{y}_2 - \bar{\mathsf{y}}\end{bmatrix} = \begin{bmatrix}\frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2}\end{bmatrix}\begin{bmatrix}\mathsf{y}_1 \\ \mathsf{y}_2\end{bmatrix} = \boldsymbol{D}\boldsymbol{\mathsf{y}} \tag{12.3.16}$$

and $\mathcal{V}[\boldsymbol{D}\mathsf{y}] = \boldsymbol{D}\mathcal{V}[\mathsf{y}]\boldsymbol{D}^\top = \sigma^2\boldsymbol{D}$ because $\mathcal{V}[\mathsf{y}] = \sigma^2\boldsymbol{I}$ and $\boldsymbol{D} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{bmatrix}$ is symmetric and idempotent. $\boldsymbol{D}$ is singular because its determinant is zero. □

• c. *1 point The joint distribution of* y_1 *and* y_2 *is bivariate normal, why did we then get a* χ^2 *with one, instead of two, degrees of freedom?*

Answer. Because $\mathsf{y}_1 - \bar{\mathsf{y}}$ and $\mathsf{y}_2 - \bar{\mathsf{y}}$ are not independent; one is exactly the negative of the other; therefore summing their squares is really only the square of one univariate normal. □

Problem 188. *Assume* y_1, y_2, *and* y_3 *are independent* $\mathsf{N}(\mu, \sigma^2)$. *Define three new variables* z_1, z_2, *and* z_3 *as follows:* z_1 *is that multiple of* $\bar{\mathsf{y}}$ *which has variance* σ^2. z_2 *is that linear combination of* z_1 *and* y_2 *which has zero covariance with* z_1 *and has variance* σ^2. z_3 *is that linear combination of* z_1, z_2, *and* y_3 *which has zero covariance with both* z_1 *and* z_2 *and has again variance* σ^2. *These properties define* z_1, z_2, *and* z_3 *uniquely up factors* ± 1, *i.e., if* z_1 *satisfies the above conditions, then* $-\mathsf{z}_1$ *does too, and these are the only two solutions.*

• a. *2 points Write* z_1 *and* z_2 *(not yet* z_3*) as linear combinations of* y_1, y_2, *and* y_3.

• b. *1 point To make the computation of* z_3 *less tedious, first show the following: if* z_3 *has zero covariance with* z_1 *and* z_2, *it also has zero covariance with* y_2.

• c. *1 point Therefore* z_3 *is a linear combination of* y_1 *and* y_3 *only. Compute its coefficients.*

• d. *1 point How does the joint distribution of* z_1, z_2, *and* z_3 *differ from that of* y_1, y_2, *and* y_3*? Since they are jointly normal, you merely have to look at the expected values, variances, and covariances.*

• e. *2 points Show that* $\mathsf{z}_1^2 + \mathsf{z}_2^2 + \mathsf{z}_3^2 = \mathsf{y}_1^2 + \mathsf{y}_2^2 + \mathsf{y}_3^2$. *Is this a surprise?*

• f. *1 point Show further that* $\mathsf{s}_u^2 = \frac{1}{2}\sum_{i=1}^3(\mathsf{y}_i - \bar{\mathsf{y}})^2 = \frac{1}{2}(\mathsf{z}_2^2 + \mathsf{z}_3^2)$. *(There is a simple trick!) Conclude from this that* $\mathsf{s}_u^2 \sim \frac{\sigma^2}{2}\chi_2^2$, *independent of* $\bar{\mathsf{y}}$.

For a matrix-interpretation of what is happening, see equation (10.4.9) together with Problem 189.

Problem 189. *3 points Verify that the matrix* $\boldsymbol{D} = \boldsymbol{I} - \frac{1}{n}\boldsymbol{\iota}\boldsymbol{\iota}^\top$ *is symmetric and idempotent, and that the sample covariance of two vectors of observations* $\boldsymbol{x}$ *and* $\boldsymbol{y}$ *can be written in matrix notation as*

$$\textit{sample covariance}(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\boldsymbol{x}^\top\boldsymbol{D}\boldsymbol{y} \tag{12.3.17}$$

In general, one can always find $n-1$ normal variables with variance σ^2, independent of each other and of $\bar{\mathsf{y}}$, whose sum of squares is equal to $\sum(\mathsf{y}_i - \bar{\mathsf{y}})^2$. Simply

start with $\bar{\mathsf{y}}\sqrt{n}$ and generate $n-1$ linear combinations of the y_i which are pairwise uncorrelated and have variances σ^2. You are simply building an orthonormal coordinate system with $\bar{\mathsf{y}}\sqrt{n}$ as its first vector; there are many different ways to do this.

Next let us show that $\bar{\mathsf{y}}$ and s_u^2 are statistically independent. This is an advantage. Assume, hypothetically, $\bar{\mathsf{y}}$ and s_u^2 were negatively correlated. Then, if the observed value of $\bar{\mathsf{y}}$ is too high, chances are that the one of s_u^2 is too low, and a look at s_u^2 will not reveal how far off the mark $\bar{\mathsf{y}}$ may be. To prove independence, we will first show that $\bar{\mathsf{y}}$ and $\mathsf{y}_i - \bar{\mathsf{y}}$ are uncorrelated:

$$\operatorname{cov}[\bar{\mathsf{y}}, \mathsf{y}_i - \bar{\mathsf{y}}] = \operatorname{cov}[\bar{\mathsf{y}}, \mathsf{y}_i] - \operatorname{var}[\bar{\mathsf{y}}] \tag{12.3.18}$$

$$= \operatorname{cov}[\frac{1}{n}(\mathsf{y}_1 + \cdots + \mathsf{y}_i + \cdots + \mathsf{y}_n), \mathsf{y}_i] - \frac{\sigma^2}{n} = 0 \tag{12.3.19}$$

By normality, $\bar{\mathsf{y}}$ is therefore *independent* of $\mathsf{y}_i - \bar{\mathsf{y}}$ for all i. Since all variables involved are jointly normal, it follows from this that $\bar{\mathsf{y}}$ is independent of the vector $\begin{bmatrix} \mathsf{y}_1 - \bar{\mathsf{y}} & \cdots & \mathsf{y}_n - \bar{\mathsf{y}} \end{bmatrix}^\top$; therefore it is also independent of any function of this vector, such as s_u^2.

The above calculations explain why the parameter of the χ^2 distribution has the colorful name "degrees of freedom." This term is sometimes used in a very broad sense, referring to estimation in general, and sometimes in a narrower sense, in conjunction with the linear model. Here is first an interpretation of the general use of the term. A "statistic" is defined to be a function of the observations and of other known parameters of the problem, but not of the unknown parameters. Estimators are statistics. If one has n observations, then one can find at most n mathematically independent statistics; any other statistic is then a function of these n. If therefore a model has k independent unknown parameters, then one must have at least k observations to be able to estimate all parameters of the model. The number $n-k$, i.e., the number of observations not "used up" for estimation, is called the number of "degrees of freedom."

There are at least three reasons why one does not want to make the model such that it uses up too many degrees of freedom. (1) the estimators become too inaccurate if one does; (2) if there are no degrees of freedom left, it is no longer possible to make any "diagnostic" tests whether the model really fits the data, because it always gives a perfect fit whatever the given set of data; (3) if there are no degrees of freedom left, then one can usually also no longer make estimates of the precision of the estimates.

Specifically in our linear estimation problem, the number of degrees of freedom is $n-1$, since one observation has been used up for estimating the mean. If one runs a regression, the number of degrees of freedom is $n-k$, where k is the number of regression coefficients. In the linear model, the number of degrees of freedom becomes immediately relevant for the estimation of σ^2. If k observations are used

up for estimating the slope parameters, then the other $n-k$ observations can be combined into a $n-k$-variate Normal whose expected value does not depend on the slope parameter at all but is zero, which allows one to estimate the variance.

If we assume that the original observations are normally distributed, i.e., $\mathsf{y}_i \sim \mathsf{NID}(\mu, \sigma^2)$, then we know that $\mathsf{s}_u^2 \sim \frac{\sigma^2}{n-1}\chi^2_{n-1}$. Therefore $\mathrm{E}[\mathsf{s}_u^2] = \sigma^2$ and $\mathrm{var}[\mathsf{s}_u^2] = 2\sigma^4/(n-1)$. This estimate of σ^2 therefore not only gives us an estimate of the precision of $\bar{\mathsf{y}}$, but it has an estimate of its own precision built in.

Interestingly, the MSE of the alternative estimator $\mathsf{s}_m^2 = \frac{\sum(\mathsf{y}_i - \bar{\mathsf{y}})^2}{n}$ is smaller than that of s_u^2, although s_m^2 is a biased estimator and s_u^2 an unbiased estimator of σ^2. For every estimator t, $\mathrm{MSE}[\mathsf{t};\theta] = \mathrm{var}[\mathsf{t}] + (\mathrm{E}[\mathsf{t}-\theta])^2$, i.e., it is variance plus squared bias. The MSE of s_u^2 is therefore equal to its variance, which is $\frac{2\sigma^4}{n-1}$. The alternative $\mathsf{s}_m^2 = \frac{n-1}{n}\mathsf{s}_u^2$ has bias $-\frac{\sigma^2}{n}$ and variance $\frac{2\sigma^4(n-1)}{n^2}$. Its MSE is $\frac{(2-1/n)\sigma^4}{n}$. Comparing that with the formula for the MSE of s_u^2 one sees that the numerator is smaller and the denominator is bigger, therefore s_m^2 has smaller MSE.

Problem 190. *4 points Assume $\mathsf{y}_i \sim \mathsf{NID}(\mu, \sigma^2)$. Show that the so-called Theil Schweitzer estimator* [**TS61**]

$$\mathsf{s}_t^2 = \frac{1}{n+1}\sum(\mathsf{y}_i - \bar{\mathsf{y}})^2 \tag{12.3.20}$$

has even smaller MSE *than s_u^2 and s_m^2 as an estimator of σ^2.*

Answer. $\mathsf{s}_t^2 = \frac{n-1}{n+1}\mathsf{s}_u^2$; therefore its bias is $-\frac{2\sigma^2}{n+1}$ and its variance is $\frac{2(n-1)\sigma^4}{(n+1)^2}$, and the MSE is $\frac{2\sigma^4}{n+1}$. That this is smaller than the MSE of s_m^2 means $\frac{2n-1}{n^2} \geq \frac{2}{n+1}$, which follows from $(2n-1)(n+1) = 2n^2 + n - 1 > 2n^2$ for $n > 1$. □

Problem 191. *3 points Computer assignment: Given 20 independent observations of a random variable $\mathsf{y} \sim \mathsf{N}(\mu, \sigma^2)$. Assume you know that $\sigma^2 = 2$. Plot the density function of s_u^2. Hint: In* R, *the command* `dchisq(x,df=25)` *returns the density of a Chi-square distribution with 25 degrees of freedom evaluated at* `x`. *But the number 25 was only taken as an example, this is not the number of degrees of freedom you need here. You also do not need the density of a Chi-Square but that of a certain multiple of a Chi-square. (Use the transformation theorem for density functions!)*

Answer. $\mathsf{s}_u^2 \sim \frac{2}{19}\chi^2_{19}$. To express the density of the variable whose density is known by that whose density one wants to know, say $\frac{19}{2}\mathsf{s}_u^2 \sim \chi^2_{19}$. Therefore

$$f_{\mathsf{s}_u^2}(x) = \frac{19}{2} f_{\chi^2_{19}}(\frac{19}{2}x). \tag{12.3.21}$$

□

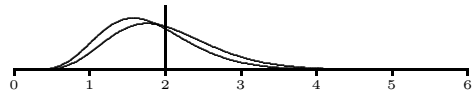


Figure 6. Densities of Unbiased and Theil Schweitzer Estimators

• a. *2 points In the same plot, plot the density function of the Theil-Schweitzer estimate s_t^2 defined in equation* (12.3.20). *This gives a plot as in Figure 6. Can one see from the comparison of these density functions that the Theil-Schweitzer estimator has a better* MSE*?*

Answer. Start with plotting the Theil-Schweitzer plot, because it is higher, and therefore it will give the right dimensions of the plot. You can run this by giving the command `ecmetscript(theilsch)`. The two areas between the densities have equal size, but the area where the Theil-Schweitzer density is higher is overall closer to the true value than the area where the unbiased density is higher. □

Problem 192. *4 points The following problem illustrates the general fact that if one starts with an unbiased estimator and "shrinks" it a little, one will end up with a better* MSE. *Assume $\mathrm{E}[\mathsf{y}] = \mu$, $\mathrm{var}(\mathsf{y}) = \sigma^2$, and you make n independent observations y_i. The best linear unbiased estimator of μ on the basis of these observations is the sample mean $\bar{\mathsf{y}}$. Show that, whenever α satisfies*

$$\frac{n\mu^2 - \sigma^2}{n\mu^2 + \sigma^2} < \alpha < 1 \tag{12.3.22}$$

then $\mathrm{MSE}[\alpha\bar{\mathsf{y}};\mu] < \mathrm{MSE}[\bar{\mathsf{y}};\mu]$. Unfortunately, this condition depends on μ and σ^2 and can therefore not be used to improve the estimate.

Answer. Here is the mathematical relationship:

$$\mathrm{MSE}[\alpha\bar{\mathsf{y}};\mu] = \mathrm{E}\Big[(\alpha\bar{\mathsf{y}} - \mu)^2\Big] = \mathrm{E}\Big[(\alpha\bar{\mathsf{y}} - \alpha\mu + \alpha\mu - \mu)^2\Big] < \mathrm{MSE}[\bar{\mathsf{y}};\mu] = \mathrm{var}[\bar{\mathsf{y}}] \tag{12.3.23}$$

$$\alpha^2\sigma^2/n + (1-\alpha)^2\mu^2 < \sigma^2/n \tag{12.3.24}$$

Now simplify it:

$$(1-\alpha)^2\mu^2 < (1-\alpha^2)\sigma^2/n = (1-\alpha)(1+\alpha)\sigma^2/n \tag{12.3.25}$$

This cannot be true for $\alpha \geq 1$, because for $\alpha = 1$ one has equality, and for $\alpha > 1$, the righthand side is negative. Therefore we are allowed to assume $\alpha < 1$, and can divide by $1-\alpha$ without disturbing the inequality:

$$(1-\alpha)\mu^2 < (1+\alpha)\sigma^2/n \tag{12.3.26}$$

$$\mu^2 - \sigma^2/n < \alpha(\mu^2 + \sigma^2/n) \tag{12.3.27}$$

The answer is therefore

$$\frac{n\mu^2 - \sigma^2}{n\mu^2 + \sigma^2} < \alpha < 1. \tag{12.3.28}$$

This the range. Note that $n\mu^2 - \sigma^2 < 0$ may be negative. The best value is in the middle of this range, see Problem 193. □

PROBLEM 193. [**KS79**, example 17.14 on p. 22] *The mathematics in the following problem is easier than it looks. If you can't prove a., assume it and derive b. from it, etc.*

• a. *2 points Let* t *be an estimator of the nonrandom scalar parameter* θ. $\mathrm{E}[\mathsf{t}-\theta]$ *is called the bias of* t, *and* $\mathrm{E}\big[(\mathsf{t}-\theta)^2\big]$ *is called the mean squared error of* t *as an estimator of* θ, *written* $\mathrm{MSE}[\mathsf{t};\theta]$. *Show that the* MSE *is the variance plus the squared bias, i.e., that*

$$\mathrm{MSE}[\mathsf{t};\theta] = \mathrm{var}[\mathsf{t}] + \big(\mathrm{E}[\mathsf{t}-\theta]\big)^2. \tag{12.3.29}$$

ANSWER. The most elegant proof, which also indicates what to do when θ is random, is:

$$\mathrm{MSE}[\mathsf{t};\theta] = \mathrm{E}\big[(\mathsf{t}-\theta)^2\big] = \mathrm{var}[\mathsf{t}-\theta] + (\mathrm{E}[\mathsf{t}-\theta])^2 = \mathrm{var}[\mathsf{t}] + (\mathrm{E}[\mathsf{t}-\theta])^2. \tag{12.3.30}$$

□

• b. *2 points For the rest of this problem assume that* t *is an unbiased estimator of* θ *with* $\mathrm{var}[\mathsf{t}] > 0$. *We will investigate whether one can get a better* MSE *if one estimates* θ *by a constant multiple* $a\mathsf{t}$ *instead of* t. *Show that*

$$\mathrm{MSE}[a\mathsf{t};\theta] = a^2\,\mathrm{var}[\mathsf{t}] + (a-1)^2\theta^2. \tag{12.3.31}$$

ANSWER. $\mathrm{var}[a\mathsf{t}] = a^2\,\mathrm{var}[\mathsf{t}]$ and the bias of $a\mathsf{t}$ is $\mathrm{E}[a\mathsf{t}-\theta] = (a-1)\theta$. Now apply (12.3.30). □

• c. *1 point Show that, whenever* $a > 1$, *then* $\mathrm{MSE}[a\mathsf{t};\theta] > \mathrm{MSE}[\mathsf{t};\theta]$. *If one wants to decrease the* MSE, *one should therefore not choose* $a > 1$.

ANSWER. $\mathrm{MSE}[a\mathsf{t};\theta]-\mathrm{MSE}[\mathsf{t};\theta] = (a^2-1)\,\mathrm{var}[\mathsf{t}]+(a-1)^2\theta^2 > 0$ since $a > 1$ and $\mathrm{var}[\mathsf{t}] > 0$. □

• d. *2 points Show that*

$$\frac{d}{da}\,\mathrm{MSE}[a\mathsf{t};\theta]\bigg|_{a=1} > 0. \tag{12.3.32}$$

From this follows that the MSE *of* $a\mathsf{t}$ *is smaller than the* MSE *of* t, *as long as* $a < 1$ *and close enough to* 1.

ANSWER. The derivative of (12.3.31) is

$$\frac{d}{da}\,\mathrm{MSE}[a\mathsf{t};\theta] = 2a\,\mathrm{var}[\mathsf{t}] + 2(a-1)\theta^2 \tag{12.3.33}$$

Plug $a = 1$ into this to get $2\,\mathrm{var}[\mathsf{t}] > 0$. □

• e. *2 points By solving the first order condition show that the factor* a *which gives smallest* MSE *is*

$$a = \frac{\theta^2}{\mathrm{var}[\mathsf{t}] + \theta^2}. \tag{12.3.34}$$

ANSWER. Rewrite (12.3.33) as $2a(\mathrm{var}[\mathsf{t}] + \theta^2) - 2\theta^2$ and set it zero. □

• f. *1 point Assume* t *has an exponential distribution with parameter* $\lambda > 0$, *i.e.,*

$$f_{\mathsf{t}}(t) = \lambda\exp(-\lambda t), \qquad t \geq 0 \quad \textit{and} \quad f_{\mathsf{t}}(t) = 0 \quad \textit{otherwise.} \tag{12.3.35}$$

Check that $f_{\mathsf{t}}(t)$ *is indeed a density function.*

ANSWER. Since $\lambda > 0$, $f_{\mathsf{t}}(t) > 0$ for all $t \geq 0$. To evaluate $\int_0^\infty \lambda\exp(-\lambda t)\,dt$, substitute $s = -\lambda t$, therefore $ds = -\lambda dt$, and the upper integration limit changes from $+\infty$ to $-\infty$, therefore the integral is $-\int_0^{-\infty}\exp(s)\,ds = 1$. □

• g. *4 points Using this density function (and no other knowledge about the exponential distribution)* prove *that* t *is an unbiased estimator of* $1/\lambda$, *with* $\mathrm{var}[\mathsf{t}] = 1/\lambda^2$.

ANSWER. To evaluate $\int_0^\infty \lambda t\exp(-\lambda t)\,dt$, use partial integration $\int uv'\,dt = uv - \int u'v\,dt$ with $u = t$, $u' = 1$, $v = -\exp(-\lambda t)$, $v' = \lambda\exp(-\lambda t)$. Therefore the integral is $-t\exp(-\lambda t)\big|_0^\infty + \int_0^\infty \exp(-\lambda t)\,dt = 1/\lambda$, since we just saw that $\int_0^\infty \lambda\exp(-\lambda t)\,dt = 1$.

To evaluate $\int_0^\infty \lambda t^2\exp(-\lambda t)\,dt$, use partial integration with $u = t^2$, $u' = 2t$, $v = -\exp(-\lambda t)$, $v' = \lambda\exp(-\lambda t)$. Therefore the integral is $-t^2\exp(-\lambda t)\big|_0^\infty + 2\int_0^\infty t\exp(-\lambda t)\,dt = \frac{2}{\lambda}\int_0^\infty \lambda t\exp(-\lambda t)\,dt = 2/\lambda^2$. Therefore $\mathrm{var}[\mathsf{t}] = \mathrm{E}[\mathsf{t}^2] - (\mathrm{E}[\mathsf{t}])^2 = 2/\lambda^2 - 1/\lambda^2 = 1/\lambda^2$. □

• h. *2 points Which multiple of* t *has the lowest* MSE *as an estimator of* $1/\lambda$?

ANSWER. It is $\mathsf{t}/2$. Just plug $\theta = 1/\lambda$ into (12.3.34).

$$a = \frac{1/\lambda^2}{\mathrm{var}[\mathsf{t}] + 1/\lambda^2} = \frac{1/\lambda^2}{1/\lambda^2 + 1/\lambda^2} = \frac{1}{2}. \tag{12.3.36}$$

□

• i. *2 points Assume* $\mathsf{t}_1, \ldots, \mathsf{t}_n$ *are independently distributed, and each of them has the exponential distribution with the same parameter* λ. *Which multiple of the sample mean* $\bar{\mathsf{t}} = \frac{1}{n}\sum_{i=1}^n \mathsf{t}_i$ *has best* MSE *as estimator of* $1/\lambda$?

ANSWER. $\bar{\mathsf{t}}$ has expected value $1/\lambda$ and variance $1/n\lambda^2$. Therefore

$$a = \frac{1/\lambda^2}{\mathrm{var}[\mathsf{t}] + 1/\lambda^2} = \frac{1/\lambda^2}{1/n\lambda^2 + 1/\lambda^2} = \frac{n}{n+1}, \tag{12.3.37}$$

i.e., for the best estimator $\tilde{\mathsf{t}} = \frac{1}{n+1}\sum \mathsf{t}_i$ divide the sum by $n+1$ instead of n. □

• j. *3 points Assume* $\mathsf{q} \sim \sigma^2\chi^2_m$ *(in other words,* $\frac{1}{\sigma^2}\mathsf{q} \sim \chi^2_m$, *a Chi-square distribution with* m *degrees of freedom). Using the fact that* $\mathrm{E}[\chi^2_m] = m$ *and* $\mathrm{var}[\chi^2_m] = 2m$, *compute that multiple of* q *that has minimum* MSE *as estimator of* σ^2.

ANSWER. This is a trick question since q itself is not an unbiased estimator of σ^2. $\mathrm{E}[\mathsf{q}] = m\sigma^2$, therefore q/m is the unbiased estimator. Since $\mathrm{var}[\mathsf{q}/m] = 2\sigma^4/m$, it follows from (12.3.34) that $a = m/(m+2)$, therefore the minimum MSE multiple of q is $\frac{\mathsf{q}}{m}\frac{m}{m+2} = \frac{\mathsf{q}}{m+2}$. I.e., divide q by $m+2$ instead of m. □

• k. *3 points Assume you have n independent observations of a Normally distributed random variable* $\mathtt{y}$ *with unknown mean μ and standard deviation σ^2. The best unbiased estimator of σ^2 is $\frac{1}{n-1}\sum(\mathtt{y}_i-\bar{\mathtt{y}})^2$, and the maximum likelihood extimator is $\frac{1}{n}\sum(\mathtt{y}_i-\bar{\mathtt{y}})^2$. What are the implications of the above for the question whether one should use the first or the second or still some other multiple of $\sum(\mathtt{y}_i-\bar{\mathtt{y}})^2$?*

Answer. Taking that multiple of the sum of squared errors which makes the estimator unbiased is not necessarily a good choice. In terms of MSE, the best multiple of $\sum(\mathtt{y}_i-\bar{\mathtt{y}})^2$ is $\frac{1}{n+1}\sum(\mathtt{y}_i-\bar{\mathtt{y}})^2$.

□

• l. *3 points We are still in the model defined in* k. *Which multiple of the sample mean $\bar{\mathtt{y}}$ has smallest* MSE *as estimator of μ? How does this example differ from the ones given above? Can this formula have practical significance?*

Answer. Here the optimal $a=\frac{\mu^2}{\mu^2+(\sigma^2/n)}$. Unlike in the earlier examples, this a depends on the unknown parameters. One can "operationalize" it by estimating the parameters from the data, but the noise introduced by this estimation can easily make the estimator worse than the simple $\bar{\mathtt{y}}$. Indeed, $\bar{\mathtt{y}}$ is admissible, i.e., it cannot be uniformly improved upon. On the other hand, the Stein rule, which can be considered an operationalization of a very similar formula (the only difference being that one estimates the mean vector of a vector with at least 3 elements), by estimating μ^2 and $\mu^2+\frac{1}{n}\sigma^2$ from the data, shows that such an operationalization is sometimes successful. □

We will discuss here one more property of $\bar{\mathtt{y}}$ and $\mathtt{s}_u^2$: They together form sufficient statistics for μ and σ^2. I.e., any estimator of μ and σ^2 which is not a function of $\bar{\mathtt{y}}$ and $\mathtt{s}_u^2$ is less efficient than it could be. Since the factorization theorem for sufficient statistics holds even if the parameter θ and its estimate $\mathtt{t}$ are vectors, we have to write the joint density of the observation vector $\mathbf{y}$ as a product of two functions, one depending on the parameters and the sufficient statistics, and the other depending on the value taken by $\mathbf{y}$, but not on the parameters. Indeed, it will turn out that this second function can just be taken to be $h(\boldsymbol{y})=1$, since the density function can be rearranged as

$$f_{\mathbf{y}}(y_1,\ldots,y_n;\mu,\sigma^2)=(2\pi\sigma^2)^{-n/2}\exp\Big(-\sum_{i=1}^n(y_i-\mu)^2/2\sigma^2\Big)= \tag{12.3.38}$$

$$=(2\pi\sigma^2)^{-n/2}\exp\Big(-\big(\sum_{i=1}^n(y_i-\bar{y})^2-n(\bar{y}-\mu)^2\big)/2\sigma^2\Big)= \tag{12.3.39}$$

$$=(2\pi\sigma^2)^{-n/2}\exp\Big(-\frac{(n-1)\mathtt{s}_u^2-n(\bar{y}+\mu)^2}{2\sigma^2}\Big). \tag{12.3.40}$$

CHAPTER 13

Estimation Principles and Classification of Estimators

13.1. Asymptotic or Large-Sample Properties of Estimators

We will discuss asymptotic properties first, because the idea of estimation is to get more certainty by increasing the sample size.

Strictly speaking, asymptotic properties do not refer to individual estimators but to sequences of estimators, one for each sample size n. And strictly speaking, if one alters the first 10 estimators or the first million estimators and leaves the others unchanged, one still gets a sequence with the same asymptotic properties. The results that follow should therefore be used with caution. The asymptotic properties may say very little about the concrete estimator at hand.

The most basic asymptotic property is (weak) consistency. An estimator t_n (where n is the sample size) of the parameter θ is consistent iff

$$\operatorname*{plim}_{n\to\infty} \mathsf{t}_n = \theta. \tag{13.1.1}$$

Roughly, a consistent estimation procedure is one which gives the correct parameter values if the sample is large enough. There are only very few exceptional situations in which an estimator is acceptable which is not *consistent*, i.e., which does not converge in the plim to the true parameter value.

Problem 194. *Can you think of a situation where an estimator which is not consistent is acceptable?*

Answer. If additional data no longer give information, like when estimating the initial state of a timeseries, or in prediction. And if there is no identification but the value can be confined to an interval. This is also inconsistency. □

The following is an important property of consistent estimators:

Slutsky theorem: If t is a consistent estimator for θ, and the function g is continuous at the true value of θ, then $g(\mathsf{t})$ is consistent for $g(\theta)$.

For the proof of the Slutsky theorem remember the definition of a continuous function. g is continuous at θ iff for all $\varepsilon > 0$ there exists a $\delta > 0$ with the property

that for all θ_1 with $|\theta_1 - \theta| < \delta$ follows $|g(\theta_1) - g(\theta)| < \varepsilon$. To prove consistency of $g(\mathsf{t})$ we have to show that for all $\varepsilon > 0$, $\Pr[|g(\mathsf{t}) - g(\theta)| \geq \varepsilon] \to 0$. Choose for the given ε a δ as above, then $|g(\mathsf{t}) - g(\theta)| \geq \varepsilon$ implies $|\mathsf{t} - \theta| \geq \delta$, because all those values of t for with $|\mathsf{t} - \theta| < \delta$ lead to a $g(\mathsf{t})$ with $|g(\mathsf{t}) - g(\theta)| < \varepsilon$. This logical implication means that

$$\Pr[|g(\mathsf{t}) - g(\theta)| \geq \varepsilon] \leq \Pr[|\mathsf{t} - \theta| \geq \delta]. \tag{13.1.2}$$

Since the probability on the righthand side converges to zero, the one on the lefthand side converges too.

Different consistent estimators can have quite different speeds of convergence. Are there estimators which have optimal asymptotic properties among all consistent estimators? Yes, if one limits oneself to a fairly reasonable subclass of consistent estimators.

Here are the details: Most consistent estimators we will encounter are asymptotically normal, i.e., the "shape" of their distribution function converges towards the normal distribution, as we had it for the sample mean in the central limit theorem. In order to be able to use this asymptotic distribution for significance tests and confidence intervals, however, one needs more than asymptotic normality (and many textbooks are not aware of this): one needs the convergence to normality to be *uniform in compact intervals* [**Rao73**, p. 346–351]. Such estimators are called *consistent uniformly asymptotically normal* estimators (CUAN estimators)

If one limits oneself to CUAN estimators it can be shown that there are asymptotically "best" CUAN estimators. Since the distribution is asymptotically normal, there is no problem to define what it means to be asymptotically best: those estimators are asymptotically best whose asymptotic MSE = asymptotic variance is smallest. CUAN estimators whose MSE is asymptotically no larger than that of any other CUAN estimator, are called *asymptotically efficient*. Rao has shown that for CUAN estimators the lower bound for this asymptotic variance is the asymptotic limit of the Cramer Rao lower bound (CRLB). (More about the CRLB below). Maximum likelihood estimators are therefore usually efficient CUAN estimators. In this sense one can think of maximum likelihood estimators to be something like asymptotically best consistent estimators, compare a statement to this effect in [**Ame94**, p. 144]. And one can think of asymptotically efficient CUAN estimators as estimators who are in large samples as good as maximum likelihood estimators.

All these are large sample properties. Among the asymptotically efficient estimators there are still wide differences regarding the small sample properties. Asymptotic efficiency should therefore again be considered a minimum requirement: there must be very good reasons *not* to be working with an asymptotically efficient estimator.

Problem 195. *Can you think of situations in which an estimator is acceptable which is not asymptotically efficient?*

Answer. If robustness matters then the median may be preferable to the mean, although it is less efficient. □

13.2. Small Sample Properties

In order to judge how good an estimator is for small samples, one has two dilemmas: (1) there are many different criteria for an estimator to be "good"; (2) even if one has decided on one criterion, a given estimator may be good for some values of the unknown parameters and not so good for others.

If x and y are two estimators of the parameter θ, then each of the following conditions can be interpreted to mean that x is better than y:

$$\Pr[|\mathsf{x}-\theta| \leq |\mathsf{y}-\theta|] = 1 \tag{13.2.1}$$

$$\mathrm{E}[g(\mathsf{x}-\theta)] \leq \mathrm{E}[g(\mathsf{y}-\theta)] \tag{13.2.2}$$

for every continuous function g which is and nonincreasing for $x < 0$ and nondecreasing for $x > 0$

$$\mathrm{E}[g(|\mathsf{x}-\theta|)] \leq \mathrm{E}[g(|\mathsf{y}-\theta|)] \tag{13.2.3}$$

for every continuous and nondecreasing function g

$$\Pr[\{|\mathsf{x}-\theta| > \varepsilon\}] \leq \Pr[\{|\mathsf{y}-\theta| > \varepsilon\}] \quad \text{for every } \varepsilon \tag{13.2.4}$$

$$\mathrm{E}[(\mathsf{x}-\theta)^2] \leq \mathrm{E}[(\mathsf{y}-\theta)^2] \tag{13.2.5}$$

$$\Pr[|\mathsf{x}-\theta| < |\mathsf{y}-\theta|] \geq \Pr[|\mathsf{x}-\theta| > |\mathsf{y}-\theta|] \tag{13.2.6}$$

This list is from [**Ame94**, pp. 118–122]. But we will simply use the MSE.

Therefore we are left with dilemma (2). There is no single estimator that has uniformly the smallest MSE in the sense that its MSE is better than the MSE of *any* other estimator whatever the value of the parameter value. To see this, simply think of the following estimator t of θ: $\mathsf{t} = 10$; i.e., whatever the outcome of the experiments, t always takes the value 10. This estimator has zero MSE when θ happens to be 10, but is a bad estimator when θ is far away from 10. If an estimator existed which had uniformly best MSE, then it had to be better than all the constant estimators, i.e., have zero MSE whatever the value of the parameter, and this is only possible if the parameter itself is observed.

Although the MSE criterion cannot be used to pick one best estimator, it can be used to rule out estimators which are unnecessarily bad in the sense that other estimators exist which are never worse but sometimes better in terms of MSE whatever the true parameter values. Estimators which are dominated in this sense are called inadmissible.

But how can one choose between two admissible estimators? [**Ame94**, p. 124] gives two reasonable strategies. One is to integrate the MSE out over a distribution

of the likely values of the parameter. This is in the spirit of the Bayesians, although Bayesians would still do it differently. The other strategy is to choose a minimax strategy. Amemiya seems to consider this an alright strategy, but it is really too defensive. Here is a third strategy, which is often used but less well founded theoretically: Since there are no estimators which have minimum MSE among all estimators, one often looks for estimators which have minimum MSE among all estimators with a certain property. And the "certain property" which is most often used is unbiasedness. The MSE of an unbiased estimator is its variance; and an estimator which has minimum variance in the class of all unbiased estimators is called "efficient."

The class of unbiased estimators has a high-sounding name, and the results related with Cramer-Rao and Least Squares seem to confirm that it is an important class of estimators. However I will argue in these class notes that unbiasedness itself is not a desirable property.

13.3. Comparison Unbiasedness Consistency

Let us compare consistency with unbiasedness. If the estimator is unbiased, then its expected value for any sample size, whether large or small, is equal to the true parameter value. By the law of large numbers this can be translated into a statement about large samples: The mean of many independent replications of the estimate, *even if each replication only uses a small number of observations*, gives the true parameter value. Unbiasedness says therefore something about the small sample properties of the estimator, while consistency does not.

The following thought experiment may clarify the difference between unbiasedness and consistency. Imagine you are conducting an experiment which gives you every ten seconds an independent measurement, i.e., a measurement whose value is not influenced by the outcome of previous measurements. Imagine further that the experimental setup is connected to a computer which estimates certain parameters of that experiment, re-calculating its estimate every time twenty new observation have become available, and which displays the current values of the estimate on a screen. And assume that the estimation procedure used by the computer is consistent, but biased for any finite number of observations.

Consistency means: after a sufficiently long time, the digits of the parameter estimate displayed by the computer will be correct. That the estimator is biased, means: if the computer were to use every batch of 20 observations to form a *new* estimate of the parameter, without utilizing prior observations, and then would use the *average* of all these independent estimates as its updated estimate, it would end up displaying a wrong parameter value on the screen.

A biased extimator gives, even in the limit, an incorrect result as long as one's updating procedure is the simple taking the averages of all previous estimates. If an estimator is biased but consistent, then a better updating method is available,

which will end up in the correct parameter value. A biased estimator therefore is not necessarily one which gives incorrect information about the parameter value; but it is one which one cannot update by simply taking averages. But there is no reason to limit oneself to such a crude method of updating. Obviously the question whether the estimate is biased is of little relevance, as long as it is consistent. The moral of the story is: If one looks for desirable estimators, by no means should one restrict one's search to unbiased estimators! The high-sounding name "unbiased" for the technical property $\mathrm{E}[\mathsf{t}] = \theta$ has created a lot of confusion.

Besides having no advantages, the category of unbiasedness even has some inconvenient properties: In some cases, in which consistent estimators exist, there are no unbiased estimators. And if an estimator t is an unbiased estimate for the parameter θ, then the estimator $g(\mathsf{t})$ is usually no longer an unbiased estimator for $g(\theta)$. It depends on the way a certain quantity is measured whether the estimator is unbiased or not. However consistency carries over.

Unbiasedness is not the only possible criterion which ensures that the values of the estimator are centered over the value it estimates. Here is another plausible definition:

Definition 13.3.1. An estimator $\hat{\theta}$ of the scalar θ is called *median unbiased* for all $\theta \in \Theta$ iff

$$\Pr[\hat{\theta} < \theta] = \Pr[\hat{\theta} > \theta] = \frac{1}{2} \tag{13.3.1}$$

This concept is always applicable, even for estimators whose expected value does not exist.

Problem 196. *6 points (Not eligible for in-class exams) The purpose of the following problem is to show how restrictive the requirement of unbiasedness is. Sometimes no unbiased estimators exist, and sometimes, as in the example here, unbiasedness leads to absurd estimators. Assume the random variable* x *has the geometric distribution with parameter* p*, where* $0 \leq p \leq 1$*. In other words, it can only assume the integer values* $1, 2, 3, \ldots$*, with probabilities*

$$\Pr[\mathsf{x} = r] = (1-p)^{r-1}p. \tag{13.3.2}$$

Show that the unique *unbiased estimator of* p *on the basis of one observation of* x *is the random variable* $f(\mathsf{x})$ *defined by* $f(x) = 1$ *if* $x = 1$ *and 0 otherwise. Hint: Use the mathematical fact that a function* $\phi(q)$ *that can be expressed as a power series* $\phi(q) = \sum_{j=0}^{\infty} a_j q^j$*, and which takes the values* $\phi(q) = 1$ *for all* q *in some interval of nonzero length, is the power series with* $a_0 = 1$ *and* $a_j = 0$ *for* $j \neq 0$*. (You will need the hint at the end of your answer, don't try to start with the hint!)*

Answer. Unbiasedness means that $\mathrm{E}[f(\mathsf{x})] = \sum_{r=1}^{\infty} f(r)(1-p)^{r-1}p = p$ for all p in the unit interval, therefore $\sum_{r=1}^{\infty} f(r)(1-p)^{r-1} = 1$. This is a power series in $q = 1-p$, which must be

identically equal to 1 for all values of q between 0 and 1. An application of the hint shows that the constant term in this power series, corresponding to the value $r - 1 = 0$, must be $= 1$, and all other $f(r) = 0$. Here older formulation: An application of the hint with $q = 1-p$, $j = r-1$, and $a_j = f(j+1)$ gives $f(1) = 1$ and all other $f(r) = 0$. This estimator is absurd since it lies on the boundary of the range of possible values for q. □

Problem 197. *As in Question 61, you make two independent trials of a Bernoulli experiment with success probability* θ*, and you observe* t*, the number of successes.*

- a. *Give an unbiased estimator of* θ *based on* t *(i.e., which is a function of* t*).*
- b. *Give an unbiased estimator of* θ^2*.*
- c. *Show that there is no unbiased estimator of* θ^3*.*

Hint: Since t *can only take the three values 0, 1, and 2, any estimator* u *which is a function of* t *is determined by the values it takes when* t *is 0, 1, or 2, call them* u_0*,* u_1*, and* u_2*. Express* $\mathrm{E}[\mathsf{u}]$ *as a function of* u_0*,* u_1*, and* u_2*.*

Answer. $\mathrm{E}[\mathsf{u}] = u_0(1-\theta)^2 + 2u_1\theta(1-\theta) + u_2\theta^2 = u_0 + (2u_1 - 2u_0)\theta + (u_0 - 2u_1 + u_2)\theta^2$. This is always a second degree polynomial in θ, therefore whatever is not a second degree polynomial in θ cannot be the expected value of any function of t. For $\mathrm{E}[\mathsf{u}] = \theta$ we need $u_0 = 0$, $2u_1 - 2u_0 = 2u_1 = 1$, therefore $u_1 = 0.5$, and $u_0 - 2u_1 + u_2 = -1 + u_2 = 0$, i.e. $u_2 = 1$. This is, in other words, $\mathsf{u} = \mathsf{t}/2$. For $\mathrm{E}[\mathsf{u}] = \theta^2$ we need $u_0 = 0$, $2u_1 - 2u_0 = 2u_1 = 0$, therefore $u_1 = 0$, and $u_0 - 2u_1 + u_2 = u_2 = 1$, This is, in other words, $\mathsf{u} = \mathsf{t}(\mathsf{t}-1)/2$. From this equation one also sees that θ^3 and higher powers, or things like $1/\theta$, cannot be the expected values of any estimators. □

- d. *Compute the moment generating function of* t*.*

Answer.

$$\mathrm{E}[e^{\lambda\mathsf{t}}] = e^0 \cdot (1-\theta)^2 + e^{\lambda} \cdot 2\theta(1-\theta) + e^{2\lambda} \cdot \theta^2 = \left(1 - \theta + \theta e^{\lambda}\right)^2 \tag{13.3.3}$$

□

Problem 198. *This is* [**KS79**, Question 17.11 on p. 34]*, originally* [**Fis**, p. 700].

- a. *1 point Assume* t *and* u *are two unbiased estimators of the same unknown scalar nonrandom parameter* θ*.* t *and* u *have finite variances and satisfy* $\mathrm{var}[\mathsf{u} - \mathsf{t}] \neq 0$*. Show that a linear combination of* t *and* u*, i.e., an estimator of* θ *which can be written in the form* $\alpha\mathsf{t} + \beta\mathsf{u}$*, is unbiased if and only if* $\alpha = 1 - \beta$*. In other words, any unbiased estimator which is a linear combination of* t *and* u *can be written in the form*

$$\mathsf{t} + \beta(\mathsf{u} - \mathsf{t}). \tag{13.3.4}$$

- b. *2 points By solving the first order condition show that the unbiased linear combination of* t *and* u *which has lowest* MSE *is*

$$\hat{\theta} = \mathsf{t} - \frac{\mathrm{cov}[\mathsf{t}, \mathsf{u} - \mathsf{t}]}{\mathrm{var}[\mathsf{u} - \mathsf{t}]}(\mathsf{u} - \mathsf{t}) \tag{13.3.5}$$

Hint: your arithmetic will be simplest if you start with (13.3.4).

• c. *1 point If ρ^2 is the squared correlation coefficient between* t *and* $\mathsf{u}-\mathsf{t}$, *i.e.,*

$$\rho^2 = \frac{(\operatorname{cov}[\mathsf{t}, \mathsf{u}-\mathsf{t}])^2}{\operatorname{var}[\mathsf{t}]\operatorname{var}[\mathsf{u}-\mathsf{t}]} \tag{13.3.6}$$

show that $\operatorname{var}[\hat{\theta}] = \operatorname{var}[\mathsf{t}](1-\rho^2)$.

• d. *1 point Show that* $\operatorname{cov}[\mathsf{t}, \mathsf{u}-\mathsf{t}] \neq 0$ *implies* $\operatorname{var}[\mathsf{u}-\mathsf{t}] \neq 0$.

• e. *2 points Use* (13.3.5) *to show that if* t *is the minimum* MSE *unbiased estimator of* θ, *and* u *another unbiased estimator of* θ, *then*

$$\operatorname{cov}[\mathsf{t}, \mathsf{u}-\mathsf{t}] = 0. \tag{13.3.7}$$

• f. *1 point Use* (13.3.5) *to show also the opposite: if* t *is an unbiased estimator of* θ *with the property that* $\operatorname{cov}[\mathsf{t}, \mathsf{u}-\mathsf{t}] = 0$ *for every other unbiased estimator* u *of* θ, *then* t *has minimum* MSE *among all unbiased estimators of* θ.

There are estimators which are consistent but their bias does not converge to zero:

$$\hat{\theta}_n = \begin{cases} \theta & \text{with probability } 1-\frac{1}{n} \\ n & \text{with probability } \frac{1}{n} \end{cases} \tag{13.3.8}$$

Then $\Pr(\left|\hat{\theta}_n - \theta\right| \geq \varepsilon) \leq \frac{1}{n}$, i.e., the estimator is consistent, but $\operatorname{E}[\hat{\theta}] = \theta\frac{n-1}{n} + 1 \to \theta + 1 \neq 0$.

Problem 199. *4 points Is it possible to have a consistent estimator whose bias becomes unbounded as the sample size increases? Either prove that it is not possible or give an example.*

Answer. Yes, this can be achieved by making the rare outliers even wilder than in (13.3.8), say

$$\hat{\theta}_n = \begin{cases} \theta & \text{with probability } 1-\frac{1}{n} \\ n^2 & \text{with probability } \frac{1}{n} \end{cases} \tag{13.3.9}$$

Here $\Pr(\left|\hat{\theta}_n - \theta\right| \geq \varepsilon) \leq \frac{1}{n}$, i.e., the estimator is consistent, but $\operatorname{E}[\hat{\theta}] = \theta\frac{n-1}{n} + n \to \theta + n$. □

And of course there are estimators which are unbiased but not consistent: simply take the first observation x_1 as an estimator if $\operatorname{E}[\mathsf{x}]$ and ignore all the other observations.

13.4. The Cramer-Rao Lower Bound

Take a scalar random variable y with density function f_{y}. The entropy of y, if it exists, is $\operatorname{H}[\mathsf{y}] = -\operatorname{E}[\log(f_{\mathsf{y}}(\mathsf{y}))]$. This is the continuous equivalent of (3.11.2). The entropy is the measure of the amount of randomness in this variable. If there is little information and much noise in this variable, the entropy is high.

Now let $y \mapsto g(y)$ be the density function of a different random variable x. In other words, g is some function which satisfies $g(y) \geq 0$ for all y, and $\int_{-\infty}^{+\infty} g(y)\,dy = 1$. Equation (3.11.10) with $v = g(y)$ and $w = f_{\mathsf{y}}(y)$ gives

$$f_{\mathsf{y}}(y) - f_{\mathsf{y}}(y)\log f_{\mathsf{y}}(y) \leq g(y) - f_{\mathsf{y}}(y)\log g(y). \tag{13.4.1}$$

This holds for every value y, and integrating over y gives $1 - \operatorname{E}[\log f_{\mathsf{y}}(\mathsf{y})] \leq 1 - \operatorname{E}[\log g(\mathsf{y})]$ or

$$\operatorname{E}[\log f_{\mathsf{y}}(\mathsf{y})] \geq \operatorname{E}[\log g(\mathsf{y})]. \tag{13.4.2}$$

This is an important extremal value property which distinguishes the density function $f_{\mathsf{y}}(y)$ of y from all other density functions: That density function g which maximizes $\operatorname{E}[\log g(\mathsf{y})]$ is $g = f_{\mathsf{y}}$, the true density function of y.

This optimality property lies at the basis of the Cramer-Rao inequality, and it is also the reason why maximum likelihood estimation is so good. The difference between the left and right hand side in (13.4.2) is called the Kullback-Leibler discrepancy between the random variables y and x (where x is a random variable whose density is g).

The Cramer Rao inequality gives a lower bound for the MSE of an unbiased estimator of the parameter of a probability distribution (which has to satisfy certain regularity conditions). This allows one to determine whether a given unbiased estimator has a MSE as low as any other unbiased estimator (i.e., whether it is "efficient.")

Problem 200. *Assume the density function of* y *depends on a parameter* θ, *write it* $f_{\mathsf{y}}(y;\theta)$, *and* θ° *is the true value of* θ. *In this problem we will compare the expected value of* y *and of functions of* y *with what would be their expected value if the true parameter value were not* θ° *but would take some other value* θ. *If the random variable* t *is a function of* y, *we write* $\operatorname{E}_\theta[\mathsf{t}]$ *for what would be the expected value of* t *if the true value of the parameter were* θ *instead of* θ°. *Occasionally, we will use the subscript* $\circ$ *as in* $\operatorname{E}_\circ$ *to indicate that we are dealing here with the usual case in which the expected value is taken with respect to the true parameter value* θ°. *Instead of* $\operatorname{E}_\circ$ *one usually simply writes* E, *since it is usually self-understood that one has to plug the right parameter values into the density function if one takes expected values. The subscript* $\circ$ *is necessary here only because in the present problem, we*

sometimes take expected values with respect to the "wrong" parameter values. The same notational convention also applies to variances, covariances, and the MSE.

Throughout this problem we assume that the following regularity conditions hold: (a) the range of y *is independent of* θ*, and (b) the derivative of the density function with respect to* θ *is a continuous differentiable function of* θ*. These regularity conditions ensure that one can differentiate under the integral sign, i.e., for all function* $t(y)$ *follows*

$$\int_{-\infty}^{\infty} \frac{\partial}{\partial\theta} f_{\mathsf{y}}(y;\theta)t(y)\,dy = \frac{\partial}{\partial\theta}\int_{-\infty}^{\infty} f_{\mathsf{y}}(y;\theta)t(y)\,dy = \frac{\partial}{\partial\theta}\,\mathrm{E}_\theta[t(\mathsf{y})] \tag{13.4.3}$$

$$\int_{-\infty}^{\infty} \frac{\partial^2}{(\partial\theta)^2} f_{\mathsf{y}}(y;\theta)t(y)\,dy = \frac{\partial^2}{(\partial\theta)^2}\int_{-\infty}^{\infty} f_{\mathsf{y}}(y;\theta)t(y)\,dy = \frac{\partial^2}{(\partial\theta)^2}\,\mathrm{E}_\theta[t(\mathsf{y})]. \tag{13.4.4}$$

• a. *1 point The* score *is defined as the random variable*

$$q(\mathsf{y};\theta) = \frac{\partial}{\partial\theta}\log f_{\mathsf{y}}(\mathsf{y};\theta). \tag{13.4.5}$$

In other words, we do three things to the density function: take its logarithm, then take the derivative of this logarithm with respect to the parameter, and then plug the random variable into it. This gives us a random variable which also depends on the nonrandom parameter θ*. Show that the score can also be written as*

$$q(\mathsf{y};\theta) = \frac{1}{f_{\mathsf{y}}(\mathsf{y};\theta)}\frac{\partial f_{\mathsf{y}}(\mathsf{y};\theta)}{\partial\theta} \tag{13.4.6}$$

Answer. This is the chain rule for differentiation: for any differentiable function $g(\theta)$, $\frac{\partial}{\partial\theta}\log g(\theta) = \frac{1}{g(\theta)}\frac{\partial g(\theta)}{\partial\theta}$. □

• b. *1 point If the density function is member of an exponential dispersion family* (6.2.9)*, show that the score function has the form*

$$q(\mathsf{y};\theta) = \frac{\mathsf{y} - \frac{\partial b(\theta)}{\partial\theta}}{a(\psi)} \tag{13.4.7}$$

Answer. This is a simple substitution: if

$$f_{\mathsf{y}}(y;\theta,\psi) = \exp\Bigl(\frac{y\theta - b(\theta)}{a(\psi)} + c(y,\psi)\Bigr), \tag{13.4.8}$$

then

$$\frac{\partial\log f_{\mathsf{y}}(\mathsf{y};\theta,\psi)}{\partial\theta} = \frac{\mathsf{y} - \frac{\partial b(\theta)}{\partial\theta}}{a(\psi)} \tag{13.4.9}$$

□

• c. *3 points If* $f_{\mathsf{y}}(y;\theta^\circ)$ *is the true density function of* y*, then we know from* (13.4.2) *that* $\mathrm{E}_\circ[\log f_{\mathsf{y}}(\mathsf{y};\theta^\circ)] \geq \mathrm{E}_\circ[\log f(\mathsf{y};\theta)]$ *for all* θ*. This explains why the score is so important: it is the derivative of that function whose expected value is maximized if the true parameter is plugged into the density function. The first-order conditions in this situation read: the expected value of this derivative must be zero for the true parameter value. This is the next thing you are asked to show: If* θ° *is the true parameter value, show that* $\mathrm{E}_\circ[q(\mathsf{y};\theta^\circ)] = 0$*.*

Answer. First write for general θ

$$\mathrm{E}_\circ[q(\mathsf{y};\theta)] = \int_{-\infty}^{\infty} q(y;\theta)f_{\mathsf{y}}(y;\theta^\circ)\,dy = \int_{-\infty}^{\infty}\frac{1}{f_{\mathsf{y}}(y;\theta)}\frac{\partial f_{\mathsf{y}}(y;\theta)}{\partial\theta}f_{\mathsf{y}}(y;\theta^\circ)\,dy. \tag{13.4.10}$$

For $\theta = \theta^\circ$ this simplifies:

$$\mathrm{E}_\circ[q(\mathsf{y};\theta^\circ)] = \int_{-\infty}^{\infty}\frac{\partial f_{\mathsf{y}}(y;\theta)}{\partial\theta}\Bigg|_{\theta=\theta^\circ} dy = \frac{\partial}{\partial\theta}\int_{-\infty}^{\infty} f_{\mathsf{y}}(y;\theta)\,dy\Bigg|_{\theta=\theta^\circ} = \frac{\partial}{\partial\theta}1 = 0. \tag{13.4.11}$$

Here I am writing $\frac{\partial f_{\mathsf{y}}(y;\theta)}{\partial\theta}\Big|_{\theta=\theta^\circ}$ instead of the simpler notation $\frac{\partial f_{\mathsf{y}}(y;\theta^\circ)}{\partial\theta}$, in order to emphasize that one *first* has to take a derivative with respect to θ and *then* one plugs θ° into that derivative. □

• d. *Show that, in the case of the exponential dispersion family,*

$$\mathrm{E}_\circ[\mathsf{y}] = \frac{\partial b(\theta)}{\partial\theta}\Big|_{\theta=\theta^\circ} \tag{13.4.12}$$

Answer. Follows from the fact that the score function of the exponential family (13.4.7) has zero expected value. □

• e. *5 points If we differentiate the score, we obtain the* Hessian

$$\mathsf{h}(\theta) = \frac{\partial^2}{(\partial\theta)^2}\log f_{\mathsf{y}}(\mathsf{y};\theta). \tag{13.4.13}$$

From now on we will write the score function as $\mathsf{q}(\theta)$ *instead of* $q(\mathsf{y};\theta)$*; i.e., we will no longer make it explicit that* q *is a function of* y *but write it as a random variable which depends on the parameter* θ*. We also suppress the dependence of* h *on* y*; our notation* $\mathsf{h}(\theta)$ *is short for* $h(\mathsf{y};\theta)$*. Since there is only one parameter in the density function, score and Hessian are scalars; but in the general case, the score is a vector and the Hessian a matrix. Show that, for the true parameter value* θ°*, the negative of the expected value of the Hessian equals the variance of the score, i.e., the expected value of the square of the score:*

$$\mathrm{E}_\circ[\mathsf{h}(\theta^\circ)] = -\,\mathrm{E}_\circ[\mathsf{q}^2(\theta^\circ)]. \tag{13.4.14}$$

Answer. Start with the definition of the score

$$q(\mathsf{y};\theta) = \frac{\partial}{\partial\theta}\log f_{\mathsf{y}}(\mathsf{y};\theta) = \frac{1}{f_{\mathsf{y}}(\mathsf{y};\theta)}\frac{\partial}{\partial\theta}f_{\mathsf{y}}(\mathsf{y};\theta), \tag{13.4.15}$$

and differentiate the rightmost expression one more time:

$$h(\mathsf{y};\theta) = \frac{\partial}{(\partial\theta)}q(\mathsf{y};\theta) = -\frac{1}{f_{\mathsf{y}}^2(\mathsf{y};\theta)}\left(\frac{\partial}{\partial\theta}f_{\mathsf{y}}(\mathsf{y};\theta)\right)^2 + \frac{1}{f_{\mathsf{y}}(\mathsf{y};\theta)}\frac{\partial^2}{\partial\theta^2}f_{\mathsf{y}}(\mathsf{y};\theta) \tag{13.4.16}$$

$$= -q^2(\mathsf{y};\theta) + \frac{1}{f_{\mathsf{y}}(\mathsf{y};\theta)}\frac{\partial^2}{\partial\theta^2}f_{\mathsf{y}}(\mathsf{y};\theta) \tag{13.4.17}$$

Taking expectations we get

$$\mathrm{E}_\circ[h(\mathsf{y};\theta)] = -\mathrm{E}_\circ[q^2(\mathsf{y};\theta)] + \int_{-\infty}^{+\infty}\frac{1}{f_{\mathsf{y}}(y;\theta)}\left(\frac{\partial^2}{\partial\theta^2}f_{\mathsf{y}}(y;\theta)\right)f_{\mathsf{y}}(y;\theta^\circ)\,dy \tag{13.4.18}$$

Again, for $\theta = \theta^\circ$, we can simplify the integrand and differentiate under the integral sign:

$$\int_{-\infty}^{+\infty}\frac{\partial^2}{\partial\theta^2}f_{\mathsf{y}}(y;\theta)\,dy = \frac{\partial^2}{\partial\theta^2}\int_{-\infty}^{+\infty}f_{\mathsf{y}}(y;\theta)\,dy = \frac{\partial^2}{\partial\theta^2}1 = 0. \tag{13.4.19}$$

□

• f. *Derive from* (13.4.14) *that, for the exponential dispersion family* (6.2.9),

$$\mathrm{var}_\circ[\mathsf{y}] = \frac{\partial^2 b(\theta)}{\partial\theta^2}a(\phi)\Big|_{\theta=\theta^\circ} \tag{13.4.20}$$

Answer. Differentiation of (13.4.7) gives $\mathsf{h}(\theta) = -\frac{\partial^2 b(\theta)}{\partial\theta^2}\frac{1}{a(\phi)}$. This is constant and therefore equal to its own expected value. (13.4.14) says therefore

$$\frac{\partial^2 b(\theta)}{\partial\theta^2}\Big|_{\theta=\theta^\circ}\frac{1}{a(\phi)} = \mathrm{E}_\circ[\mathsf{q}^2(\theta^\circ)] = \frac{1}{\big(a(\phi)\big)^2}\,\mathrm{var}_\circ[\mathsf{y}] \tag{13.4.21}$$

from which (13.4.20) follows. □

Problem 201.

• a. *Use the results from question 200 to derive the following strange and interesting result: for any random variable* t *which is a function of* y, *i.e.,* $\mathsf{t} = t(\mathsf{y})$, *follows* $\mathrm{cov}_\circ[\mathsf{q}(\theta^\circ),\mathsf{t}] = \frac{\partial}{\partial\theta}\mathrm{E}_\theta[\mathsf{t}]\big|_{\theta=\theta^\circ}$.

Answer. The following equation holds for all θ:

$$\mathrm{E}_\circ[\mathsf{q}(\theta)\mathsf{t}] = \int_{-\infty}^{\infty}\frac{1}{f_{\mathsf{y}}(y;\theta)}\frac{\partial f_{\mathsf{y}}(y;\theta)}{\partial\theta}t(y)f_{\mathsf{y}}(y;\theta^\circ)\,dy \tag{13.4.22}$$

If the θ in $\mathsf{q}(\theta)$ is the right parameter value θ° one can simplify:

$$\mathrm{E}_\circ[\mathsf{q}(\theta^\circ)\mathsf{t}] = \int_{-\infty}^{\infty}\frac{\partial f_{\mathsf{y}}(y;\theta)}{\partial\theta}\Big|_{\theta=\theta^\circ}t(y)\,dy \tag{13.4.23}$$

$$= \frac{\partial}{\partial\theta}\int_{-\infty}^{\infty}f_{\mathsf{y}}(y;\theta)t(y)\,dy\Big|_{\theta=\theta^\circ} \tag{13.4.24}$$

$$= \frac{\partial}{\partial\theta}\mathrm{E}_\theta[\mathsf{t}]\Big|_{\theta=\theta^\circ} \tag{13.4.25}$$

This is at the same time the covariance: $\mathrm{cov}_\circ[\mathsf{q}(\theta^\circ),\mathsf{t}] = \mathrm{E}_\circ[\mathsf{q}(\theta^\circ)\mathsf{t}] - \mathrm{E}_\circ[\mathsf{q}(\theta^\circ)]\,\mathrm{E}_\circ[\mathsf{t}] = \mathrm{E}_\circ[\mathsf{q}(\theta^\circ)\mathsf{t}]$, since $\mathrm{E}_\circ[\mathsf{q}(\theta^\circ)] = 0$. □

Explanation, nothing to prove here: Now if t *is an unbiased estimator of* θ, *whatever the value of* θ, *then it follows* $\mathrm{cov}_\circ[\mathsf{q}(\theta^\circ),\mathsf{t}] = \frac{\partial}{\partial\theta}\theta = 1$. *From this follows by Cauchy-Schwartz* $\mathrm{var}_\circ[\mathsf{t}]\,\mathrm{var}_\circ[\mathsf{q}(\theta^\circ)] \geq 1$, *or* $\mathrm{var}_\circ[\mathsf{t}] \geq 1/\,\mathrm{var}_\circ[\mathsf{q}(\theta^\circ)]$. *Since* $\mathrm{E}_\circ[\mathsf{q}(\theta^\circ)] = 0$, *we know* $\mathrm{var}_\circ[\mathsf{q}(\theta^\circ)] = \mathrm{E}_\circ[\mathsf{q}^2(\theta^\circ)]$, *and since* t *is unbiased, we know* $\mathrm{var}_\circ[\mathsf{t}] = \mathrm{MSE}_\circ[\mathsf{t};\theta^\circ]$. *Therefore the Cauchy-Schwartz inequality reads*

$$\mathrm{MSE}_\circ[\mathsf{t};\theta^\circ] \geq 1/\,\mathrm{E}_\circ[\mathsf{q}^2(\theta^\circ)]. \tag{13.4.26}$$

This is the Cramer-Rao inequality. *The inverse of the variance of* $\mathsf{q}(\theta^\circ)$, $1/\,\mathrm{var}_\circ[\mathsf{q}(\theta^\circ)] = 1/\,\mathrm{E}_\circ[\mathsf{q}^2(\theta^\circ)]$, *is called the* Fisher information, *written* $I(\theta^\circ)$. *It is a lower bound for the* MSE *of any unbiased estimator of* θ. *Because of* (13.4.14), *the Cramer Rao inequality can also be written in the form*

$$\mathrm{MSE}[\mathsf{t};\theta^\circ] \geq -1/\,\mathrm{E}_\circ[\mathsf{h}(\theta^\circ)]. \tag{13.4.27}$$

(13.4.26) and (13.4.27) are usually written in the following form: Assume y has density function $f_{\mathsf{y}}(y;\theta)$ which depends on the unknown parameter θ, and and let $\mathsf{t}(\mathsf{y})$ be any unbiased estimator of θ. Then

$$\mathrm{var}[\mathsf{t}] \geq \frac{1}{\mathrm{E}[\big(\frac{\partial}{\partial\theta}\log f_{\mathsf{y}}(\mathsf{y};\theta)\big)^2]} = \frac{-1}{\mathrm{E}[\frac{\partial^2}{\partial\theta^2}\log f_{\mathsf{y}}(\mathsf{y};\theta)]}. \tag{13.4.28}$$

(Sometimes the first and sometimes the second expression is easier to evaluate.)

If one has a whole *vector* of observations then the Cramer-Rao inequality involves the *joint* density function:

$$\mathrm{var}[\mathsf{t}] \geq \frac{1}{\mathrm{E}[\big(\frac{\partial}{\partial\theta}\log f_{\mathbf{y}}(\mathbf{y};\theta)\big)^2]} = \frac{-1}{\mathrm{E}[\frac{\partial^2}{\partial\theta^2}\log f_{\mathbf{y}}(\mathbf{y};\theta)]}. \tag{13.4.29}$$

This inequality also holds if y is discrete and one uses its probability mass function instead of the density function. In small samples, this lower bound is not always attainable; in some cases there is no unbiased estimator with a variance as low as the Cramer Rao lower bound.

PROBLEM 202. *4 points Assume n independent observations of a variable $\mathsf{y} \sim \mathsf{N}(\mu, \sigma^2)$ are available, where σ^2 is known. Show that the sample mean $\bar{\mathsf{y}}$ attains the Cramer-Rao lower bound for μ.*

ANSWER. The density function of each y_i is

$$f_{\mathsf{y}_i}(y) = (2\pi\sigma^2)^{-1/2} \exp\Big(-\frac{(y-\mu)^2}{2\sigma^2}\Big) \tag{13.4.30}$$

therefore the log likelihood function of the whole vector is

$$\ell(\mathsf{y};\mu) = \sum_{i=1}^n \log f_{\mathsf{y}_i}(\mathsf{y}_i) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^n(\mathsf{y}_i-\mu)^2 \tag{13.4.31}$$

$$\frac{\partial}{\partial\mu}\ell(\mathsf{y};\mu) = \frac{1}{\sigma^2}\sum_{i=1}^n(\mathsf{y}_i-\mu) \tag{13.4.32}$$

In order to apply (13.4.29) you can either square this and take the expected value

$$\mathrm{E}[\Big(\frac{\partial}{\partial\mu}\ell(\mathsf{y};\mu)\Big)^2] = \frac{1}{\sigma^4}\sum \mathrm{E}[(\mathsf{y}_i-\mu)^2] = n/\sigma^2 \tag{13.4.33}$$

alternatively one may take one more derivative from (13.4.32) to get

$$\frac{\partial^2}{\partial\mu^2}\ell(\mathsf{y};\mu) = -\frac{n}{\sigma^2} \tag{13.4.34}$$

This is constant, therefore equal to its expected value. Therefore the Cramer-Rao Lower Bound says that $\mathrm{var}[\bar{\mathsf{y}}] \geq \sigma^2/n$. This holds with equality. □

PROBLEM 203. *Assume $\mathsf{y}_i \sim \mathrm{NID}(0,\sigma^2)$ (i.e., normally independently distributed) with unknown σ^2. The obvious estimate of σ^2 is $\mathsf{s}^2 = \frac{1}{n}\sum \mathsf{y}_i^2$.*

• a. *2 points Show that s^2 is an unbiased estimator of σ^2, is distributed $\sim \frac{\sigma^2}{n}\chi^2_n$, and has variance $2\sigma^4/n$. You are allowed to use the fact that a χ^2_n has variance $2n$, which is equation (5.9.5).*

ANSWER.

$$\mathrm{E}[y_i^2] = \mathrm{var}[y_i] + (\mathrm{E}[y_i])^2 = \sigma^2 + 0 = \sigma^2 \tag{13.4.35}$$

$$z_i = \frac{y_i}{\sigma} \sim \mathrm{NID}(0,1) \tag{13.4.36}$$

$$y_i = \sigma z_i \tag{13.4.37}$$

$$y_i^2 = \sigma^2 z_i^2 \tag{13.4.38}$$

$$\sum_{i=1}^n y_i^2 = \sigma^2\sum_{i=1}^n z_i^2 \sim \sigma^2\chi^2_n \tag{13.4.39}$$

$$\frac{1}{n}\sum_{i=1}^n y_i^2 = \frac{\sigma^2}{n}\sum_{i=1}^n z_i^2 \sim \frac{\sigma^2}{n}\chi^2_n \tag{13.4.40}$$

$$\mathrm{var}\Big[\frac{1}{n}\sum_{i=1}^n y_i^2\Big] = \frac{\sigma^4}{n^2}\mathrm{var}[\chi^2_n] = \frac{\sigma^4}{n^2}2n = \frac{2\sigma^4}{n} \tag{13.4.41}$$

□

• b. *4 points Show that this variance is at the same time the Cramer Rao lower bound.*

ANSWER.

$$\ell(y,\sigma^2) = \log f_{\mathsf{y}}(y;\sigma^2) = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma^2 - \frac{y^2}{2\sigma^2} \tag{13.4.42}$$

$$\frac{\partial\log f_{\mathsf{y}}}{\partial\sigma^2}(y;\sigma^2) = -\frac{1}{2\sigma^2} + \frac{y^2}{2\sigma^4} = \frac{y^2-\sigma^2}{2\sigma^4} \tag{13.4.43}$$

Since $\dfrac{\mathsf{y}^2-\sigma^2}{2\sigma^4}$ has zero mean, it follows

$$\mathrm{E}[\Big(\frac{\partial\log f_{\mathsf{y}}}{\partial\sigma^2}(\mathsf{y};\sigma^2)\Big)^2] = \frac{\mathrm{var}[\mathsf{y}^2]}{4\sigma^8} = \frac{1}{2\sigma^4}. \tag{13.4.44}$$

Alternatively, one can differentiate one more time:

$$\frac{\partial^2\log f_{\mathsf{y}}}{(\partial\sigma^2)^2}(y;\sigma^2) = -\frac{y^2}{\sigma^6} + \frac{1}{2\sigma^4} \tag{13.4.45}$$

$$\mathrm{E}[\frac{\partial^2\log f_{\mathsf{y}}}{(\partial\sigma^2)^2}(\mathsf{y};\sigma^2)] = -\frac{\sigma^2}{\sigma^6} + \frac{1}{2\sigma^4} = \frac{1}{2\sigma^4} \tag{13.4.46}$$

(13.4.47)

This makes the Cramer Rao lower bound $2\sigma^4/n$. □

PROBLEM 204. *4 points Assume $\mathsf{x}_1,\ldots,\mathsf{x}_n$ is a random sample of independent observations of a Poisson distribution with parameter λ, i.e., each of the x_i has probability mass function*

$$p_{\mathsf{x}_i}(x) = \Pr[\mathsf{x}_i = x] = \frac{\lambda^x}{x!}e^{-\lambda} \quad x = 0,1,2,\ldots. \tag{13.4.48}$$

A Poisson variable with parameter λ has expected value λ and variance λ. (You are not required to prove this here.) Is there an unbiased estimator of λ with lower variance than the sample mean $\bar{\mathsf{x}}$?

Here is a formulation of the Cramer Rao Inequality for probability mass functions, as you need it for Question 204. Assume $\mathsf{y}_1, \ldots, \mathsf{y}_n$ are n independent observations of a random variable y whose probability mass function depends on the unknown parameter θ and satisfies certain regularity conditions. Write the univariate probability mass function of each of the y_i as $p_{\mathsf{y}}(y;\theta)$ and let t be any unbiased estimator of θ. Then

$$\text{var}[\mathsf{t}] \geq \frac{1}{n\,\text{E}[\left(\frac{\partial}{\partial\theta}\ln p_{\mathsf{y}}(\mathsf{y};\theta)\right)^2]} = \frac{-1}{n\,\text{E}[\frac{\partial^2}{\partial\theta^2}\ln p_{\mathsf{y}}(\mathsf{y};\theta)]}. \tag{13.4.49}$$

Answer. The Cramer Rao lower bound says no.

$$\log p_{\mathsf{x}}(\mathsf{x};\lambda) = \mathsf{x}\log\lambda - \log\mathsf{x}! - \lambda \tag{13.4.50}$$

$$\frac{\partial \log p_{\mathsf{x}}}{\partial\lambda}(\mathsf{x};\lambda) = \frac{\mathsf{x}}{\lambda} - 1 = \frac{\mathsf{x}-\lambda}{\lambda} \tag{13.4.51}$$

$$\text{E}[\left(\frac{\partial \log p_{\mathsf{x}}}{\partial\lambda}(\mathsf{x};\lambda)\right)^2] = \text{E}[\frac{(\mathsf{x}-\lambda)^2}{\lambda^2}] = \frac{\text{var}[\mathsf{x}]}{\lambda^2} = \frac{1}{\lambda}. \tag{13.4.52}$$

Or alternatively, after (13.4.51) do

$$\frac{\partial^2 \log p_{\mathsf{x}}}{\partial\lambda^2}(\mathsf{x};\lambda) = -\frac{\mathsf{x}}{\lambda^2} \tag{13.4.53}$$

$$-\text{E}[\left(\frac{\partial^2 \log p_{\mathsf{x}}}{\partial\lambda^2}(\mathsf{x};\lambda)\right)] = \frac{\text{E}[\mathsf{x}]}{\lambda^2} = \frac{1}{\lambda}. \tag{13.4.54}$$

Therefore the Cramer Rao lower bound is $\frac{\lambda}{n}$, which is the variance of the sample mean. □

If the density function depends on more than one unknown parameter, i.e., if it has the form $f_{\mathsf{y}}(y;\theta_1,\ldots,\theta_k)$, the Cramer Rao Inequality involves the following steps: (1) define $\ell(\mathsf{y};\theta_1,\cdots,\theta_k) = \log f_{\mathsf{y}}(\mathsf{y};\theta_1,\ldots,\theta_k)$, (2) form the following matrix which is called the *information matrix*:

(13.4.55)

$$\boldsymbol{I} = \begin{bmatrix} -n\,\text{E}[\frac{\partial^2\ell}{\partial\theta_1^2}] & \cdots & -n\,\text{E}[\frac{\partial^2\ell}{\partial\theta_1\partial\theta_k}] \\ \vdots & \ddots & \vdots \\ -n\,\text{E}[\frac{\partial^2\ell}{\partial\theta_k\partial\theta_1}] & \cdots & -n\,\text{E}[\frac{\partial^2\ell}{\partial\theta_k^2}] \end{bmatrix} = \begin{bmatrix} n\,\text{E}[\left(\frac{\partial\ell}{\partial\theta_1}\right)^2] & \cdots & n\,\text{E}[\frac{\partial\ell}{\partial\theta_1}\frac{\partial\ell}{\partial\theta_k}] \\ \vdots & \ddots & \vdots \\ n\,\text{E}[\frac{\partial\ell}{\partial\theta_k}\frac{\partial\ell}{\partial\theta_1}] & \cdots & n\,\text{E}[\left(\frac{\partial\ell}{\partial\theta_k}\right)^2] \end{bmatrix},$$

and (3) form the matrix inverse $\boldsymbol{I}^{-1}$. If the vector random variable $\mathbf{t} = \begin{bmatrix} \mathsf{t}_1 \\ \vdots \\ \mathsf{t}_n \end{bmatrix}$ is an unbiased estimator of the parameter vector $\boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_n \end{bmatrix}$, then the inverse of the information matrix $\boldsymbol{I}^{-1}$ is a lower bound for the covariance matrix $\mathcal{V}[\mathbf{t}]$ in the following sense: the difference matrix $\mathcal{V}[\mathbf{t}] - \boldsymbol{I}^{-1}$ is always nonnegative definite.

From this follows in particular: if i^{ii} is the ith diagonal element of $\boldsymbol{I}^{-1}$, then $\text{var}[\mathsf{t}_i] \geq i^{ii}$.

13.5. Best Linear Unbiased Without Distribution Assumptions

If the x_i are Normal with unknown expected value and variance, their sample mean has lowest MSE among all unbiased estimators of μ. If one does not assume Normality, then the sample mean has lowest MSE in the class of all *linear* unbiased estimators of μ. This is true not only for the sample mean but also for all least squares estimates. This result needs remarkably weak assumptions: nothing is assumed about the distribution of the x_i other than the existence of mean and variance. Problem 205 shows that in some situations one can even dispense with the independence of the observations.

Problem 205. *5 points* [**Lar82**, example 5.4.1 on p 266] *Let y_1 and y_2 be two random variables with same mean μ and variance σ^2, but we do not assume that they are uncorrelated; their correlation coefficient is ρ, which can take any value $|\rho| \leq 1$. Show that $\bar{\mathsf{y}} = (\mathsf{y}_1 + \mathsf{y}_2)/2$ has lowest mean squared error among all linear unbiased estimators of μ, and compute its* MSE. *(An estimator $\tilde{\mu}$ of μ is linear iff it can be written in the form $\tilde{\mu} = \alpha_1\mathsf{y}_1 + \alpha_2\mathsf{y}_2$ with some constant numbers α_1 and α_2.)*

Answer.

$$\tilde{\mathsf{y}} = \alpha_1\mathsf{y}_1 + \alpha_2\mathsf{y}_2 \tag{13.5.1}$$

$$\text{var}\,\tilde{\mathsf{y}} = \alpha_1^2\,\text{var}[\mathsf{y}_1] + \alpha_2^2\,\text{var}[\mathsf{y}_2] + 2\alpha_1\alpha_2\,\text{cov}[\mathsf{y}_1,\mathsf{y}_2] \tag{13.5.2}$$

$$= \sigma^2(\alpha_1^2 + \alpha_2^2 + 2\alpha_1\alpha_2\rho). \tag{13.5.3}$$

Here we used (8.1.14). Unbiasedness means $\alpha_2 = 1 - \alpha_1$, therefore we call $\alpha_1 = \alpha$ and $\alpha_2 = 1 - \alpha$:

$$\text{var}[\tilde{\mathsf{y}}]/\sigma^2 = \alpha^2 + (1-\alpha)^2 + 2\alpha(1-\alpha)\rho \tag{13.5.4}$$

Now sort by the powers of α:

$$= 2\alpha^2(1-\rho) - 2\alpha(1-\rho) + 1 \tag{13.5.5}$$

$$= 2(\alpha^2 - \alpha)(1-\rho) + 1. \tag{13.5.6}$$

This takes its minimum value where the derivative $\frac{\partial}{\partial\alpha}(\alpha^2-\alpha)=2\alpha-1=0$. For the MSE plug $\alpha_1=\alpha_2-1/2$ into (13.5.3) to get $\frac{\sigma^2}{2}\big(1+\rho\big)$. □

PROBLEM 206. *You have two unbiased measurements with errors of the same quantity μ (which may or may not be random). The first measurement y_1 has mean squared error $\mathrm{E}[(\mathsf{y}_1-\mu)^2]=\sigma^2$, the other measurement y_2 has $\mathrm{E}[(\mathsf{y}_1-\mu)^2]=\tau^2$. The measurement errors $\mathsf{y}_1-\mu$ and $\mathsf{y}_2-\mu$ have zero expected values (i.e., the measurements are unbiased) and are independent of each other.*

• a. *2 points Show that the linear unbiased estimators of μ based on these two measurements are simply the weighted averages of these measurements, i.e., they can be written in the form $\tilde{\mu}=\alpha\mathsf{y}_1+(1-\alpha)\mathsf{y}_2$, and that the* MSE *of such an estimator is $\alpha^2\sigma^2+(1-\alpha)^2\tau^2$. Note: we are using the word "estimator" here even if μ is random. An estimator or predictor $\tilde{\mu}$ is unbiased if $\mathrm{E}[\tilde{\mu}-\mu]=0$. Since we allow μ to be random, the proof in the class notes has to be modified.*

ANSWER. The estimator $\tilde{\mu}$ is linear (more precisely: affine) if it can written in the form

$$\tilde{\mu}=\alpha_1\mathsf{y}_1+\alpha_2\mathsf{y}_2+\gamma \tag{13.5.7}$$

The measurements themselves are unbiased, i.e., $\mathrm{E}[\mathsf{y}_i-\mu]=0$, therefore

$$\mathrm{E}[\tilde{\mu}-\mu]=(\alpha_1+\alpha_2-1)\,\mathrm{E}[\mu]+\gamma=0 \tag{13.5.8}$$

for all possible values of $\mathrm{E}[\mu]$; therefore $\gamma=0$ and $\alpha_2=1-\alpha_1$. To simplify notation, we will call from now on $\alpha_1=\alpha$, $\alpha_2=1-\alpha$. Due to unbiasedness, the MSE is the variance of the estimation error

$$\mathrm{var}[\tilde{\mu}-\mu]=\alpha^2\sigma^2+(1-\alpha)^2\tau^2 \tag{13.5.9}$$

□

• b. *4 points Define ω^2 by*

$$\frac{1}{\omega^2}=\frac{1}{\sigma^2}+\frac{1}{\tau^2}\qquad\textit{which can be solved to give}\qquad \omega^2=\frac{\sigma^2\tau^2}{\sigma^2+\tau^2}. \tag{13.5.10}$$

Show that the Best (i.e., minimum MSE*) linear unbiased estimator (BLUE) of μ based on these two measurements is*

$$\hat{\mathsf{y}}=\frac{\omega^2}{\sigma^2}\mathsf{y}_1+\frac{\omega^2}{\tau^2}\mathsf{y}_2 \tag{13.5.11}$$

i.e., it is the weighted average of y_1 and y_2 where the weights are proportional to the inverses of the variances.

ANSWER. The variance (13.5.9) takes its minimum value where its derivative with respect of α is zero, i.e., where

$$\frac{\partial}{\partial\alpha}\Big(\alpha^2\sigma^2+(1-\alpha)^2\tau^2\Big)=2\alpha\sigma^2-2(1-\alpha)\tau^2=0 \tag{13.5.12}$$

$$\alpha\sigma^2=\tau^2-\alpha\tau^2 \tag{13.5.13}$$

$$\alpha=\frac{\tau^2}{\sigma^2+\tau^2} \tag{13.5.14}$$

In terms of ω one can write

$$\alpha=\frac{\tau^2}{\sigma^2+\tau^2}=\frac{\omega^2}{\sigma^2}\qquad\text{and}\qquad 1-\alpha=\frac{\sigma^2}{\sigma^2+\tau^2}=\frac{\omega^2}{\tau^2}. \tag{13.5.15}$$

□

• c. *2 points Show: the* MSE *of the BLUE ω^2 satisfies the following equation:*

$$\frac{1}{\omega^2}=\frac{1}{\sigma^2}+\frac{1}{\tau^2} \tag{13.5.16}$$

ANSWER. We already have introduced the notation ω^2 for the quantity defined by (13.5.16); therefore all we have to show is that the MSE or, equivalently, the variance of the estimation error is equal to this ω^2:

$$\mathrm{var}[\tilde{\mu}-\mu]=\Big(\frac{\omega^2}{\sigma^2}\Big)^2\sigma^2+\Big(\frac{\omega^2}{\tau^2}\Big)^2\tau^2=\omega^4\Big(\frac{1}{\sigma^2}+\frac{1}{\tau^2}\Big)=\omega^4\frac{1}{\omega^2}=\omega^2 \tag{13.5.17}$$

□

Examples of other classes of estimators for which a best estimator exists are: if one requires the estimator to be translation invariant, then the least squares estimators are best in the class of all translation invariant estimators. But there is no best *linear* estimator in the linear model. (Theil)

13.6. Maximum Likelihood Estimation

This is an excellent and very widely applicable estimation principle. Its main drawback is its computational complexity, but with modern computing power it becomes more and more manageable. Another drawback is that it requires a full specification of the distribution.

PROBLEM 207. *2 points What are the two greatest disadvantages of Maximum Likelihood Estimation?*

ANSWER. Its high information requirements (the functional form of the density function must be known), and computational complexity. □

In our discussion of entropy in Section 3.11 we derived an extremal value property which distinguishes the actual density function $f_{\mathsf{y}}(y)$ of a given random variable y from all other possible density functions of y, i.e., from all other functions $g\geq 0$ with $\int_{-\infty}^{+\infty}g(y)\,dy=1$. The true density function of y is the one which maximizes

$\mathrm{E}[\log g(\mathsf{y})]$. We showed that this principle can be used to design a payoff scheme by which it is in the best interest of a forecaster to tell the truth. Now we will see that this principle can also be used to design a good estimator. Say you have n independent observations of y. You know the density of y belongs to a given family $\mathcal{F}$ of density functions, but you don't know which member of $\mathcal{F}$ it is. Then form the arithmetic mean of $\log f(y_i)$ for all $f \in \mathcal{F}$. It converges towards $\mathrm{E}[\log f(\mathsf{y})]$. For the true density function, this expected value is higher than for all the other density functions. If one does not know which the true density function is, then it is a good strategy to select that density function f for which the sample mean of the $\log f(y_i)$ is largest. This is the maximum likelihood estimator.

Let us interject here a short note about the definitional difference between density function and likelihood function. If we know $\mu = \mu_0$, we can write down the density function as

$$f_{\mathsf{y}}(y;\mu_0) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-\mu_0)^2}{2}}. \tag{13.6.1}$$

It is a function of y, the possible values assumed by y, and the letter μ_0 symbolizes a constant, the true parameter value. The same function considered as a function of the *variable* μ, representing all possible values assumable by the true mean, with y being *fixed* at the actually observed value, becomes the likelihood function.

In the same way one can also turn probability mass functions $p_{\mathsf{x}}(x)$ into likelihood functions.

Now let us compute some examples of the MLE. You make n independent observations $\mathsf{y}_1, \ldots, \mathsf{y}_n$ from a $\mathsf{N}(\mu, \sigma^2)$ distribution. Write the likelihood function as

$$L(\mu, \sigma^2; \mathsf{y}_1, \ldots, \mathsf{y}_n) = \prod_{i=1}^{n} f_{\mathsf{y}}(\mathsf{y}_i) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2\sigma^2}\sum(\mathsf{y}_i-\mu)^2}. \tag{13.6.2}$$

Its logarithm is more convenient to maximize:

$$\ell = \ln L(\mu, \sigma^2; \mathsf{y}_1, \ldots, \mathsf{y}_n) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum(\mathsf{y}_i - \mu)^2. \tag{13.6.3}$$

To compute the maximum we need the partial derivatives:

$$\frac{\partial \ell}{\partial \mu} = \frac{1}{\sigma^2}\sum(\mathsf{y}_i - \mu) \tag{13.6.4}$$

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum(\mathsf{y}_i - \mu)^2. \tag{13.6.5}$$

The maximum likelihood estimators are those values $\hat{\mu}$ and $\hat{\sigma}^2$ which set these two partials zero. I.e., at the same time at which we set the partials zero we must put the hats on μ and σ^2. As long as $\hat{\sigma}^2 \neq 0$ (which is the case with probability one), the first equation determines $\hat{\mu}$: $\sum \mathsf{y}_i - n\hat{\mu} = 0$, i.e., $\hat{\mu} = \frac{1}{n}\sum \mathsf{y}_i = \bar{\mathsf{y}}$. (This would

be the MLE of μ even if σ^2 were known). Now plug this $\hat{\mu}$ into the second equation to get $\frac{n}{2} = \frac{1}{2\hat{\sigma}^2}\sum(\mathsf{y}_i - \bar{\mathsf{y}})^2$, or $\hat{\sigma}^2 = \frac{1}{n}\sum(\mathsf{y}_i - \bar{\mathsf{y}})^2$.

Here is another example: $\mathsf{t}_1, \ldots, \mathsf{t}_n$ are independent and follow an exponential distribution, i.e.,

$$f_{\mathsf{t}}(t;\lambda) = \lambda e^{-\lambda t} \quad (t > 0) \tag{13.6.6}$$

$$L(\mathsf{t}_1, \ldots, \mathsf{t}_n; \lambda) = \lambda^n e^{-\lambda(\mathsf{t}_1 + \cdots + \mathsf{t}_n)} \tag{13.6.7}$$

$$\ell(\mathsf{t}_1, \ldots, \mathsf{t}_n \lambda) = n\ln\lambda - \lambda(\mathsf{t}_1 + \cdots + \mathsf{t}_n) \tag{13.6.8}$$

$$\frac{\partial \ell}{\partial \lambda} = \frac{n}{\lambda} - (\mathsf{t}_1 + \cdots + \mathsf{t}_n). \tag{13.6.9}$$

Set this zero, and write $\hat{\lambda}$ instead of λ to get $\hat{\lambda} = \frac{n}{\mathsf{t}_1 + \cdots + \mathsf{t}_n} = 1/\bar{\mathsf{t}}$.

Usually the MLE is asymptotically unbiased and asymptotically normal. Therefore it is important to have an estimate of its asymptotic variance. Here we can use the fact that asymptotically the Cramer Rao Lower Bound is not merely a lower bound for this variance but is equal to its variance. (From this follows that the maximum likelihood estimator is asymptotically efficient.) The Cramer Rao lower bound itself depends on unknown parameters. In order to get a consistent estimate of the Cramer Rao lower bound, do the following: (1) Replace the unknown parameters in the second derivative of the log likelihood function by their maximum likelihood estimates. (2) Instead of taking expected values over the observed values x_i you may simply insert the sample values of the x_i into these maximum likelihood estimates, and (3) then invert this estimate of the information matrix.

MLE obeys an important functional invariance principle: if $\hat{\theta}$ is the MLE of θ, then $g(\hat{\theta})$ is the MLE of $g(\theta)$. E.g., $\mu = \frac{1}{\lambda}$ is the expected value of the exponential variable, and its MLE is $\bar{\mathsf{x}}$.

Problem 208. *$\mathsf{x}_1, \ldots, \mathsf{x}_m$ is a sample from a $\mathsf{N}(\mu_{\mathsf{x}}, \sigma^2)$, and $\mathsf{y}_1, \ldots, \mathsf{y}_n$ from a $\mathsf{N}(\mu_{\mathsf{y}}, \sigma^2)$ with different mean but same σ^2. All observations are independent of each other.*

• a. *2 points Show that the MLE of μ_{x}, based on the combined sample, is $\bar{\mathsf{x}}$. (By symmetry it follows that the MLE of μ_{y} is $\bar{\mathsf{y}}$.)*

Answer.

$$\begin{aligned}\ell(\mu_{\mathsf{x}}, \mu_{\mathsf{y}}, \sigma^2) = &-\frac{m}{2}\ln 2\pi - \frac{m}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{m}(\mathsf{x}_i - \mu_{\mathsf{x}})^2 \\ &- \frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{j=1}^{n}(\mathsf{y}_j - \mu_{\mathsf{y}})^2\end{aligned} \tag{13.6.10}$$

$$\frac{\partial \ell}{\partial \mu_{\mathsf{x}}} = -\frac{1}{2\sigma^2}\sum -2(\mathsf{x}_i - \mu_{\mathsf{x}}) \qquad = 0 \quad \text{for} \quad \mu_{\mathsf{x}} = \bar{\mathsf{x}} \tag{13.6.11}$$

□

• b. *2 points Derive the MLE of σ^2, based on the combined samples.*

Answer.

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{m+n}{2\sigma^2} + \frac{1}{2\sigma^4}\left(\sum_{i=1}^{m}(\mathsf{x}_i - \mu_{\mathsf{x}})^2 + \sum_{j=1}^{n}(\mathsf{y}_j - \mu_{\mathsf{y}})^2\right) \tag{13.6.12}$$

$$\hat{\sigma}^2 = \frac{1}{m+n}\left(\sum_{i=1}^{m}(\mathsf{x}_i - \bar{\mathsf{x}})^2 + \sum_{j=1}^{n}(\mathsf{y}_i - \bar{\mathsf{y}})^2\right). \tag{13.6.13}$$

□

13.7. Method of Moments Estimators

Method of moments estimators use the sample moments as estimates of the population moments. I.e., the estimate of μ is $\bar{\mathsf{x}}$, the estimate of the variance σ^2 is $\frac{1}{n}\sum(\mathsf{x}_i - \bar{\mathsf{x}})^2$, etc. If the parameters are a given function of the population moments, use the same function of the sample moments (using the lowest moments which do the job).

The advantage of method of moments estimators is their computational simplicity. Many of the estimators discussed above are method of moments estimators. However if the moments do not exist, then method of moments estimators are inconsistent, and in general method of moments estimators are not as good as maximum likelihood estimators.

13.8. M-Estimators

The class of M-estimators maximizes something other than a likelihood function: it includes nonlinear least squares, generalized method of moments, minimum distance and minimum chi-squared estimators. The purpose is to get a "robust" estimator which is good for a wide variety of likelihood functions. Many of these are asymptotically efficient; but their small-sample properties may vary greatly.

13.9. Sufficient Statistics and Estimation

Weak Sufficiency Principle: If $\boldsymbol{\mathsf{x}}$ has a p.d.f. $f_{\boldsymbol{\mathsf{x}}}(\boldsymbol{x};\theta)$ and if a sufficient statistic $\mathbf{s}(\boldsymbol{\mathsf{x}})$ exists for θ, then identical conclusions should be drawn from data $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ which have same value $\mathbf{s}(\boldsymbol{x}_1) = \mathbf{s}(\boldsymbol{x}_2)$.

Why? Sufficiency means: after knowing $\mathbf{s}(\boldsymbol{\mathsf{x}})$, the rest of the data $\boldsymbol{\mathsf{x}}$ can be regarded generated by a random mechanism not dependent on θ, and are therefore uninformative about θ.

This principle can be used to improve on given estimators. Without proof we will state here

Rao Blackwell Theorem: Let $\mathsf{t}(\boldsymbol{\mathsf{x}})$ be an estimator of θ and $\mathbf{s}(\boldsymbol{\mathsf{x}})$ a sufficient statistic for θ. Then one can get an estimator $\mathsf{t}^*(\boldsymbol{\mathsf{x}})$ of θ which has no worse a MSE than $\mathsf{t}(\boldsymbol{\mathsf{x}})$ by taking expectations conditionally on the sufficient statistic, i.e., $\mathsf{t}^*(\boldsymbol{\mathsf{x}}) = \mathrm{E}[\mathsf{t}(\boldsymbol{\mathsf{x}})|\mathbf{s}(\boldsymbol{\mathsf{x}})]$.

To recapitulate: $\mathsf{t}^*(\mathsf{x})$ is obtained by the following two steps: (1) Compute the conditional expectation $\mathsf{t}^{**}(\boldsymbol{s}) = \mathrm{E}[\mathsf{t}(\boldsymbol{\mathsf{x}})|\mathbf{s}(\boldsymbol{\mathsf{x}}) = \boldsymbol{s}]$, and (2) plug $\mathbf{s}(\boldsymbol{\mathsf{x}})$ into t^{**}, i.e., $\mathsf{t}^*(\boldsymbol{\mathsf{x}}) = \mathsf{t}^{**}(\mathbf{s}(\boldsymbol{\mathsf{x}}))$.

A statistic $\mathbf{s}$ is said to be *complete*, if the only real-valued function g defined on the range of $\mathbf{s}$, which satisfies $\mathrm{E}[g(\mathbf{s})] = 0$ whatever the value of θ, is the function which is identically zero. If a statistic $\mathbf{s}$ is complete and sufficient, then every function $g(\mathbf{s})$ is the minimum MSE unbiased estimator of its expected value $\mathrm{E}[g(\mathbf{s})]$.

If a complete and sufficient statistic exists, this gives a systematic approach to minimum MSE unbiased estimators (*Lehmann Scheffé Theorem*): if t is an *unbiased* estimator of θ and $\mathbf{s}$ is *complete* and sufficient, then $\mathsf{t}^*(\mathsf{x}) = \mathrm{E}[\mathsf{t}(\mathsf{x})|\mathbf{s}(\mathsf{x})]$ has lowest MSE in the class of all unbiased estimators of θ. Problem 209 steps you through the proof.

Problem 209. [**BD77**, Problem 4.2.6 on p. 144] *If a statistic* $\mathbf{s}$ *is* complete *and* sufficient*, then every function* $g(\mathbf{s})$ *is the minimum* MSE *unbiased estimator of* $\mathrm{E}[g(\mathbf{s})]$ *(*Lehmann-Scheffé theorem*). This gives a systematic approach to finding minimum* MSE *unbiased estimators. Here are the definitions:* $\mathbf{s}$ *is* sufficient *for* θ *if for any event* E *and any value* $\boldsymbol{s}$*, the conditional probability* $\Pr[E|\mathbf{s} \leq \boldsymbol{s}]$ *does not involve* θ*.* $\mathbf{s}$ *is* complete *for* θ *if the only function* $g(\mathbf{s})$ *of* $\mathbf{s}$*, which has zero expected value whatever the value of* θ*, is the function which is identically zero, i.e.,* $g(\boldsymbol{s}) = 0$ *for all* $\boldsymbol{s}$*.*

• a. *3 points Given an unknown parameter* θ*, and a complete sufficient statistic* $\mathbf{s}$*, how can one find that function of* $\mathbf{s}$ *whose expected value is* θ*? There is an easy trick: start with any statistic* p *with* $\mathrm{E}[\mathsf{p}] = \theta$*, and use the conditional expectation* $\mathrm{E}[\mathsf{p}|\mathbf{s}]$*. Argue why this conditional expectation does not depend on the unknown parameter* θ*, is an unbiased estimator of* θ*, and why this leads to the same estimate regardless which* p *one starts with.*

Answer. You need sufficiency for the first part of the problem, the law of iterated expectations for the second, and completeness for the third.

Set $E = \{\mathsf{p} \leq p\}$ in the definition of sufficiency given at the beginning of the Problem to see that the cdf of p conditionally on $\mathbf{s}$ being in any interval does not involve θ, therefore also $\mathrm{E}[\mathsf{p}|\mathbf{s}]$ does not involve θ.

Unbiasedness follows from the theorem of iterated expectations $\mathrm{E}\big[\mathrm{E}[\mathsf{p}|\mathbf{s}]\big] = \mathrm{E}[\mathsf{p}] = \theta$.

The independence on the choice of p can be shown as follows: Since the conditional expectation conditionally on $\mathbf{s}$ is a function of $\mathbf{s}$, we can use the notation $\mathrm{E}[\mathsf{p}|\mathbf{s}] = g_1(\mathsf{s})$ and $\mathrm{E}[\mathsf{q}|\mathbf{s}] = g_2(\mathsf{s})$. From $\mathrm{E}[\mathsf{p}] = \mathrm{E}[\mathsf{q}]$ follows by the law of iterated expectations $\mathrm{E}[g_1(\mathsf{s}) - g_2(\mathsf{s})] = 0$, therefore by completeness $g_1(\mathsf{s}) - g_2(\mathsf{s}) \equiv 0$. □

• b. *2 points Assume* $\mathsf{y}_i \sim \mathsf{NID}(\mu, 1)$ *(*$i = 1, \ldots, n$*), i.e., they are independent and normally distributed with mean* μ *and variance 1. Without proof you are allowed to use the fact that in this case, the sample mean* $\bar{\mathsf{y}}$ *is a complete sufficient statistic for* μ*. What is the minimum* MSE *unbiased estimate of* μ*, and what is that of* μ^2*?*

Answer. We have to find functions of $\bar{\mathsf{y}}$ with the desired parameters as expected values. Clearly, $\bar{\mathsf{y}}$ is that of μ, and $\bar{\mathsf{y}}^2 - 1/n$ is that of μ^2. □

• c. *1 point For a given* j*, let* π *be the probability that the* j^{th} *observation is nonnegative, i.e.,* $\pi = \Pr[\mathsf{y}_j \geq 0]$*. Show that* $\pi = \Phi(\mu)$ *where* Φ *is the cumulative distribution function of the standard normal. The purpose of the remainder of this Problem is to find a minimum* MSE *unbiased estimator of* π*.*

Answer.

$$\pi = \Pr[\mathsf{y}_i \geq 0] = \Pr[\mathsf{y}_i - \mu \geq -\mu] = \Pr[\mathsf{y}_i - \mu \leq \mu] = \Phi(\mu) \tag{13.9.1}$$

because $\mathsf{y}_i - \mu \sim \mathsf{N}(0, 1)$. We needed symmetry of the distribution to flip the sign. □

• d. *1 point As a first step we have to find an unbiased estimator of* π*. It does not have to be a good one, any ubiased estimator will do. And such an estimator is indeed implicit in the definition of* π*. Let* q *be the "indicator function" for nonnegative values, satisfying* $q(y) = 1$ *if* $y \geq 0$ *and 0 otherwise. We will be working with the random variable which one obtains by inserting the* j^{th} *observation* y_j *into* q*, i.e., with* $\mathsf{q} = q(\mathsf{y}_j)$*. Show that* q *is an unbiased estimator of* π*.*

Answer. $q(\mathsf{y}_j)$ has a discrete distribution and $\Pr[q(\mathsf{y}_j) = 1] = \Pr[\mathsf{y}_j \geq 0] = \pi$ by (13.9.1) and therefore $\Pr[q(\mathsf{y}_j) = 0] = 1 - \pi$

The expected value is $\mathrm{E}[q(\mathsf{y}_j)] = (1 - \pi) \cdot 0 + \pi \cdot 1 = \pi$. □

• e. *2 points Given* q *we can apply the Lehmann-Scheffé theorem:* $\mathrm{E}[q(\mathsf{y}_j)|\bar{\mathsf{y}}]$ *is the best unbiased estimator of* π*. We will compute* $\mathrm{E}[q(\mathsf{y}_j)|\bar{\mathsf{y}}]$ *in four steps which build on each other. First step: since for every indicator function follows* $\mathrm{E}[q(\mathsf{y}_j)|\bar{\mathsf{y}}] = \Pr[\mathsf{y}_j \geq 0|\bar{\mathsf{y}}]$*, we need for every given value* $\bar{y}$*, the conditional distribution of* y_j *conditionally on* $\bar{\mathsf{y}} = \bar{y}$*. (Not just the conditional mean but the whole conditional distribution.) In order to construct this, we first have to specify exactly the joint distribution of* y_j *and* $\bar{\mathsf{y}}$*:*

Answer. They are jointly normal:

$$\begin{bmatrix} \mathsf{y}_j \\ \bar{\mathsf{y}} \end{bmatrix} \sim N\left(\begin{bmatrix} \mu \\ \mu \end{bmatrix}, \begin{bmatrix} 1 & 1/n \\ 1/n & 1/n \end{bmatrix}\right) \tag{13.9.2}$$

□

• f. *2 points Second step: From this joint distribution derive the conditional distribution of* y_j *conditionally on* $\bar{\mathsf{y}} = \bar{y}$*. (Not just the conditional mean but the whole conditional distribution.) For this you will need formula* (10.3.18) *and* (10.3.20)*.*

Answer. Here are these two formulas: if u and v are jointly normal, then the conditional distribution of v conditionally on $\mathsf{u} = u$ is Normal with mean

$$\mathrm{E}[\mathsf{v}|\mathsf{u} = u] = \mathrm{E}[\mathsf{v}] + \frac{\mathrm{cov}[\mathsf{u}, \mathsf{v}]}{\mathrm{var}[\mathsf{u}]}(u - \mathrm{E}[\mathsf{u}]) \tag{13.9.3}$$

and variance

$$\mathrm{var}[\mathsf{v}|\mathsf{u} = u] = \mathrm{var}[\mathsf{v}] - \frac{(\mathrm{cov}[\mathsf{u}, \mathsf{v}])^2}{\mathrm{var}[\mathsf{u}]}. \tag{13.9.4}$$

Plugging $\mathsf{u} = \bar{\mathsf{y}}$ and $\mathsf{v} = \mathsf{y}_j$ into (10.3.18) and (10.3.20) gives: the conditional distribution of y_j conditionally on $\bar{\mathsf{y}} = \bar{y}$ has mean

$$\mathrm{E}[\mathsf{y}_j|\bar{\mathsf{y}} = \bar{y}] = \mathrm{E}[\mathsf{y}_j] + \frac{\mathrm{cov}[\bar{\mathsf{y}}, \mathsf{y}_j]}{\mathrm{var}[\bar{\mathsf{y}}]}(\bar{y} - \mathrm{E}[\bar{\mathsf{y}}]) \tag{13.9.5}$$

$$= \mu + \frac{1/n}{1/n}(\bar{y} - \mu) = \bar{y} \tag{13.9.6}$$

and variance

$$\mathrm{var}[\mathsf{y}_j|\bar{\mathsf{y}} = \bar{y}] = \mathrm{var}[\mathsf{y}_j] - \frac{(\mathrm{cov}[\bar{\mathsf{y}}, \mathsf{y}_j])^2}{\mathrm{var}[\bar{\mathsf{y}}]} \tag{13.9.7}$$

$$= 1 - \frac{(1/n)^2}{1/n} = 1 - \frac{1}{n}. \tag{13.9.8}$$

Therefore the conditional distribution of y_j conditional on $\bar{\mathsf{y}}$ is $\mathsf{N}(\bar{y}, (n-1)/n)$. How can this be motivated? if we know the actual arithmetic mean of the variables, then our best estimate is that each variable is equal to this arithmetic mean. And this additional knowledge cuts down the variance by $1/n$. □

• g. *2 points The variance decomposition* (8.6.6) *gives a decomposition of* $\mathrm{var}[\mathsf{y}_j]$*: give it here:*

Answer.

$$\mathrm{var}[\mathsf{y}_j] = \mathrm{var}\Big[\mathrm{E}[\mathsf{y}_j|\bar{\mathsf{y}}]\Big] + \mathrm{E}\Big[\mathrm{var}[\mathsf{y}_j|\bar{\mathsf{y}}]\Big] \tag{13.9.9}$$

$$= \mathrm{var}[\bar{\mathsf{y}}] + \mathrm{E}\left[\frac{n-1}{n}\right] = \frac{1}{n} + \frac{n-1}{n} \tag{13.9.10}$$

□

• h. *Compare the conditional with the unconditional distribution.*

Answer. Conditional distribution does not depend on unknown parameters, and it has smaller variance! □

• i. *2 points Third step: Compute the probability, conditionally on* $\bar{\mathsf{y}} = \bar{y}$*, that* $\mathsf{y}_j \geq 0$*.*

Answer. If $\mathsf{x} \sim \mathsf{N}(\bar{y}, (n-1)/n)$ (I call it x here instead of y_j since we use it not with its familiar unconditional distribution $N(\mu, 1)$ but with a conditional distribution), then $\Pr[\mathsf{x} \geq 0] = \Pr[\mathsf{x} - \bar{y} \geq -\bar{y}] = \Pr[\mathsf{x} - \bar{y} \leq \bar{y}] = \Pr\Big[(\mathsf{x} - \bar{y})\sqrt{n/(n-1)} \leq \bar{y}\sqrt{n/(n-1)}\Big] = \Phi(\bar{y}\sqrt{n/(n-1)})$ because $(\mathsf{x} - \bar{y})\sqrt{n/(n-1)} \sim \mathsf{N}(0, 1)$ conditionally on $\bar{y}$. Again we needed symmetry of the distribution to flip the sign. □

• j. *1 point Finally, put all the pieces together and write down* $\mathrm{E}[q(\mathsf{y}_j)|\bar{\mathsf{y}}]$*, the conditional expectation of* $q(\mathsf{y}_j)$ *conditionally on* $\bar{\mathsf{y}}$*, which by the Lehmann-Scheffé theorem is the minimum* MSE *unbiased estimator of* π*. The formula you should come up with is*

$$\hat{\pi} = \Phi(\bar{\mathsf{y}}\sqrt{n/(n-1)}), \tag{13.9.11}$$

where Φ *is the standard normal cumulative distribution function.*

Answer. The conditional expectation of $q(\mathsf{y}_j)$ conditionally on $\bar{\mathsf{y}} = \bar{y}$ is, by part d, simply the probability that $\mathsf{y}_j \geq 0$ under this conditional distribution. In part i this was computed as $\Phi(\bar{y}\sqrt{n/(n-1)})$. Therefore all we have to do is replace $\bar{y}$ by $\bar{\mathsf{y}}$ to get the minimum MSE unbiased estimator of π as $\Phi(\bar{\mathsf{y}}\sqrt{n/(n-1)})$. □

Remark: this particular example did not give any brand new estimators, but it can rather be considered a proof that certain obvious estimators are unbiased and efficient. But often this same procedure gives new estimators which one would not have been able to guess. Already when the variance is unknown, the above example becomes quite a bit more complicated, see [**Rao73**, p. 322, example 2]*. When the variables have an exponential distribution then this example (probability of early failure) is discussed in* [**BD77**, example 4.2.4 on pp. 124/5].

13.10. The Likelihood Principle

Consider two experiments whose likelihood functions depend on the same parameter vector $\boldsymbol{\theta}$. Suppose that for particular realizations of the data $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$, the respective likelihood functions are proportional to each other, i.e., $\ell_1(\boldsymbol{\theta}; \boldsymbol{y}_1) = \alpha\ell_2(\boldsymbol{\theta}; \boldsymbol{y}_2)$ where α does not depend on $\boldsymbol{\theta}$ although it may depend on $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$. Then the *likelihood principle* states that identical conclusions should be drawn from these two experiments about $\boldsymbol{\theta}$.

The likelihood principle is equivalent to the combination of two simpler principles: the weak sufficiency principle, and the following principle, which seems very plausible:

Weak Conditonality Principle: Given two possible experiments A and B. A mixed experiment is one in which one throws a coin and performs A if the coin shows head and B if it shows tails. The weak conditionality principle states: suppose it is known that the coin shows tails. Then the evidence of the mixed experiment is equivalent to the evidence gained had one not thrown the coin but performed B without the possible alternative of A. This principle says therefore that an experiment which one did not do but which one could have performed does not alter the information gained from the experiment actually performed.

As an application of the likelihood principle look at the following situation:

Problem 210. *3 points You have a Bernoulli experiment with unknown parameter* θ*,* $0 \leq \theta \leq 1$*. Person A was originally planning to perform this experiment 12 times, which she does. She obtains 9 successes and 3 failures. Person B was originally planning to perform the experiment until he has reached 9 successes, and it took him 12 trials to do this. Should both experimenters draw identical conclusions from these two experiments or not?*

Answer. The probability mass function in the first is by (3.7.1) $\binom{12}{9}\theta^9(1-\theta)^3$, and in the second it is by (5.1.13) $\binom{11}{8}\theta^9(1-\theta)^3$. They are proportional, the stopping rule therefore does not matter! □

13.11. Bayesian Inference

Real-life estimation usually implies the choice between competing estimation methods all of which have their advantages and disadvantages. Bayesian inference removes some of this arbitrariness.

Bayesians claim that "any inferential or decision process that does not follow from some likelihood function and some set of priors has objectively verifiable deficiencies" [**Cor69**, p. 617]. The "prior information" used by Bayesians is a formalization of the notion that the information about the parameter values never comes from the experiment alone. The Bayesian approach to estimation forces the researcher to cast his or her prior knowledge (and also the loss function for estimation errors) in a mathematical form, because in this way, unambiguous mathematical prescriptions can be derived as to how the information of an experiment should be evaluated.

To the objection that these are large information requirements which are often not satisfied, one might answer that it is less important whether these assumptions are actually the right ones. The formulation of prior density merely ensures that the researcher proceeds from a coherent set of beliefs.

The mathematics which the Bayesians do is based on a "final" instead of an "initial" criterion of precision. In other words, not an estimation procedure is evaluated which will be good in hypothetical repetitions of the experiment in the average, but one which is good for the given set of data and the given set of priors. Data which could have been observed but were not observed are not taken into consideration.

Both Bayesians and non-Bayesians define the probabilistic properties of an experiment by the density function (likelihood function) of the observations, which may depend on one or several unknown parameters. The non-Bayesian considers these parameters fixed but unknown, while the Bayesian considers the parameters random, i.e., he symbolizes his prior information about the parameters by a prior probability distribution.

An excellent example in which this prior probability distribution is discrete is given in [**Ame94**, pp. 168–172]. In the more usual case that the prior distribution

has a density function, a Bayesian is working with the joint density function of the parameter values and the data. Like all joint density function, it can be written as the product of a marginal and conditional density. The marginal density of the parameter value represents the beliefs the experimenter holds about the parameters before the experiment (prior density), and the likelihood function of the experiment is the conditional density of the data given the parameters. After the experiment has been conducted, the experimenter's belief about the parameter values is represented by their *conditional density given the data*, called the posterior density.

Let $\mathbf{y}$ denote the observations, $\boldsymbol{\theta}$ the unknown parameters, and $f(\boldsymbol{y}, \boldsymbol{\theta})$ their joint density. Then

$$f(\boldsymbol{y}, \boldsymbol{\theta}) = f(\boldsymbol{\theta}) f(\boldsymbol{y}|\boldsymbol{\theta}) \tag{13.11.1}$$

$$= f(\boldsymbol{y}) f(\boldsymbol{\theta}|\boldsymbol{y}). \tag{13.11.2}$$

Therefore

$$f(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{f(\boldsymbol{\theta}) f(\boldsymbol{y}|\boldsymbol{\theta})}{f(\boldsymbol{y})}. \tag{13.11.3}$$

In this formula, the value of $f(\boldsymbol{y})$ is irrelevant. It only depends on $\mathbf{y}$ but not on $\boldsymbol{\theta}$, but $\mathbf{y}$ is fixed, i.e., it is a constant. If one knows the posterior density function of $\boldsymbol{\theta}$ up to a constant, one knows it altogether, since the constant is determined by the requirement that the area under the density function is 1. Therefore (13.11.3) is usually written as ($\propto$ means "proportional to")

$$f(\boldsymbol{\theta}|\boldsymbol{y}) \propto f(\boldsymbol{\theta}) f(\boldsymbol{y}|\boldsymbol{\theta}); \tag{13.11.4}$$

here the lefthand side contains the posterior density function of the parameter, the righthand side the prior density function and the likelihood function representing the probability distribution of the experimental data.

The Bayesian procedure does not yield a point estimate or an interval estimate, but a whole probability distribution for the unknown parameters (which represents our *information about* these parameters) containing the "prior" information "updated" by the information yielded by the sample outcome.

Of course, such probability distributions can be summarized by various measures of location (mean, median), which can then be considered Bayesian point estimates. Such summary measures for a whole probability distribution are rather arbitrary. But if a loss function is given, then this process of distilling point estimates from the posterior distribution can once more be systematized. For a concrete decision it tells us that parameter value which minimizes the expected loss function under the posterior density function, the so-called "Bayes risk." This can be considered the Bayesian analog of a point estimate.

For instance, if the loss function is quadratic, then the *posterior mean* is the parameter value which minimizes expected loss.

There is a difference between Bayes risk and the notion of risk we applied previously. The frequentist minimizes expected loss in a large number of repetitions of the trial. This risk is dependent on the unknown parameters, and therefore usually no estimators exist which give minimum risk in all situations. The Bayesian conditions on the data (final criterion!) and minimizes the expected loss where the expectation is taken over the posterior density of the *parameter vector*.

The irreducibility of absence to presences: the absence of knowledge (or also the absence of regularity itself) cannot be represented by a probability distribution. Proof: if I give a certain random variable a neutral prior, then functions of this random variable have non-neutral priors. This argument is made in [**Roy97**, p. 174].

Many good Bayesians drift away from the subjective point of view and talk about a stratified world: their center of attention is no longer the world out there versus our knowledge of it, but the empirical world versus the underlying systematic forces that shape it.

Bayesians say that frequentists use subjective elements too; their outcomes depend on what the experimenter planned to do, even if he never did it. This again comes from [**Roy97**, p. ??]. Nature does not know about the experimenter's plans, and any evidence should be evaluated in a way independent of this.

CHAPTER 14

Interval Estimation

Look at our simplest example of an estimator, the sample mean of an independent sample from a normally distributed variable. Since the population mean of a normal variable is at the same time its median, the sample mean will in 50 percent of the cases be larger than the population mean, and in 50 percent of the cases it will be smaller. This is a statement about the procedure how the sample mean was obtained, not about any given observed value of the sample mean. Say in one particular sample the observed sample mean was 3.5. This number 3.5 is either larger or smaller than the true mean, there is no probability involved. But if one were to compute sample means of many different independent samples, then these means would in 50% of the cases lie above and in 50% of the cases below the population mean. This is why one can, from knowing how this one given number was obtained, derive the "confidence" of 50% that the actual mean lies above 3.5, and the same with below. The sample mean can therefore be considered a one-sided confidence bound, although one usually wants higher confidence levels than 50%. (I am 95% confident that ϕ is greater or equal than a certain value computed from the sample.) The concept of "confidence" is nothing but the usual concept of probability if one uses an initial criterion of precision.

The following thought experiment illustrates what is involved. Assume you bought a widget and want to know whether it is defective or not. The obvious way (which would correspond to a "final" criterion of precision) would be to open it up and look if it is defective or not. Now assume we cannot do it: there is no way telling by just looking at it whether it will work. Then another strategy would be to go by an "initial" criterion of precision: we visit the widget factory and look how they make them, how much quality control there is and such. And if we find out that 95% of all widgets coming out of the same factory have no defects, then we have the "confidence" of 95% that our particular widget is not defective either.

The matter becomes only slightly more mystified if one talks about intervals. Again, one should not forget that *confidence intervals are random intervals.* Besides confidence intervals and one-sided confidence bounds one can, if one regards several

parameters simultaneously, also construct confidence rectangles, ellipsoids and more complicated shapes. Therefore we will define in all generality:

Let $\mathbf{y}$ be a random vector whose distribution depends on some vector of unknown parameters $\boldsymbol{\phi} \in \Omega$. A *confidence region* is a prescription which assigns to every possible value $\boldsymbol{y}$ of $\mathbf{y}$ a subset $R(\boldsymbol{y}) \subset \Omega$ of parameter space, so that the probability that this subset covers the true value of $\boldsymbol{\phi}$ is at least a given confidence level $1-\alpha$, i.e.,

$$\Pr\big[R(\mathbf{y}) \ni \boldsymbol{\phi}_0 | \boldsymbol{\phi} = \boldsymbol{\phi}_0\big] \geq 1-\alpha \quad \text{for all } \boldsymbol{\phi}_0 \in \Omega. \tag{14.0.5}$$

The important thing to remember about this definition is that these regions $R(\mathbf{y})$ are random regions; every time one performs the experiment one obtains a different region.

Now let us go to the specific case of constructing an interval estimate for the parameter μ when we have n independent observations from a normally distributed population $\sim \mathsf{N}(\mu, \sigma^2)$ in which neither μ nor σ^2 are known. The vector of observations is therefore distributed as $\mathbf{y} \sim \mathsf{N}(\boldsymbol{\iota}\mu, \sigma^2\boldsymbol{I})$, where $\boldsymbol{\iota}\mu$ is the vector every component of which is μ.

I will give you now what I consider to be the cleanest argument deriving the so-called t-interval. It generalizes directly to the F-test in linear regression. It is not the same derivation which you will usually find, and I will bring the usual derivation below for comparison. Recall the observation made earlier, based on (12.1.1), that the sample mean $\bar{y}$ is that number $\bar{y} = a$ which minimizes the sum of squared deviations $\sum(y_i - a)^2$. (In other words, $\bar{y}$ is the "least squares estimate" in this situation.) This least squares principle also naturally leads to interval estimates for μ: we will say that a lies in the interval for μ if and only if

$$\frac{\sum(y_i - a)^2}{\sum(y_i - \bar{y})^2} \leq c \tag{14.0.6}$$

for some number $c \geq 1$. Of course, the value of c depends on the confidence level, but the beauty of this criterion here is that the value of c can be determined by the confidence level *alone* without knowledge of the true values of μ or σ^2.

To show this, note first that (14.0.6) is equivalent to

$$\frac{\sum(y_i - a)^2 - \sum(y_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} \leq c - 1 \tag{14.0.7}$$

and then apply the identity $\sum(y_i - a)^2 = \sum(y_i - \bar{y})^2 + n(\bar{y} - a)^2$ to the numerator to get the following equivalent formulation of (14.0.6):

$$\frac{n(\bar{y} - a)^2}{\sum(y_i - \bar{y})^2} \leq c - 1 \tag{14.0.8}$$

The confidence level of this interval is the probability that the true μ lies in an interval randomly generated using this principle. In other words, it is

$$\Pr\Big[\frac{n(\bar{\mathsf{y}}-\mu)^2}{\sum(\mathsf{y}_i-\bar{\mathsf{y}})^2} \leq c-1\Big] \tag{14.0.9}$$

Although for every *known* a, the probability that a lies in the confidence interval depends on the unknown μ and σ^2, we will show now that the probability that the *unknown* μ lies in the confidence interval does not depend on any unknown parameters. First look at the distribution of the numerator: Since $\bar{\mathsf{y}} \sim \mathsf{N}(\mu, \sigma^2/n)$, it follows $(\bar{\mathsf{y}}-\mu)^2 \sim (\sigma^2/n)\chi^2_1$. We also know the distribution of the denominator. Earlier we have shown that the variable $\sum(\mathsf{y}_i-\bar{\mathsf{y}})^2$ is a $\sigma^2\chi^2_{n-1}$. It is not enough to know the distribution of numerator and denominator separately, we also need their joint distribution. For this go back to our earlier discussion of variance estimation again; there we also showed that $\bar{\mathsf{y}}$ is independent of the vector $\begin{bmatrix}\mathsf{y}_1-\bar{\mathsf{y}} & \cdots & \mathsf{y}_n-\bar{\mathsf{y}}\end{bmatrix}^\top$; therefore any function of $\bar{\mathsf{y}}$ is also independent of any function of this vector, from which follows that numerator and denominator in our fraction are *independent*. Therefore this fraction is distributed as an $\sigma^2\chi^2_1$ over an independent $\sigma^2\chi^2_{n-1}$, and since the σ^2's cancel out, this is the same as a χ^2_1 over an independent χ^2_{n-1}. In other words, this distribution does not depend on any unknown parameters!

The definition of a F-distribution with k and m degrees of freedom is the distribution of a ratio of a χ^2_k/k divided by a χ^2_m/m; therefore if we divide the sum of squares in the numerator by $n-1$ we get a F distribution with 1 and $n-1$ d.f.:

$$\frac{(\bar{\mathsf{y}}-\mu)^2}{\frac{1}{n}\frac{1}{n-1}\sum(\mathsf{y}_i-\bar{\mathsf{y}})^2} \sim \mathsf{F}_{1,n-1} \tag{14.0.10}$$

If one does not take the square in the numerator, i.e., works with $\bar{\mathsf{y}}-\mu$ instead of $(\bar{\mathsf{y}}-\mu)^2$, and takes square root in the denominator, one obtains a t-distribution:

$$\frac{\bar{\mathsf{y}}-\mu}{\sqrt{\frac{1}{n}}\sqrt{\frac{1}{n-1}\sum(\mathsf{y}_i-\bar{\mathsf{y}})^2}} \sim \mathsf{t}_{n-1} \tag{14.0.11}$$

The left hand side of this last formula has a suggestive form. It can be written as $(\bar{\mathsf{y}}-\mu)/\mathsf{s}_{\bar{\mathsf{y}}}$, where $\mathsf{s}_{\bar{\mathsf{y}}}$ is an estimate of the standard deviation of $\bar{\mathsf{y}}$ (it is the square root of the unbiased estimate of the variance of $\bar{\mathsf{y}}$). In other words, this t-statistic can be considered an estimate of the number of standard deviations the observed value of $\bar{\mathsf{y}}$ is away from μ.

Now we will give, as promised, the usual derivation of the t-confidence intervals, which is based on this interpretation. This usual derivation involves the following two steps:

(1) First assume that σ^2 is known. Then it is obvious what to do; for every observation $\boldsymbol{y}$ of $\boldsymbol{\mathsf{y}}$ construct the following interval:

$$R(\boldsymbol{y}) = \{u \in \mathbb{R}\colon\ |u-\bar{y}| \leq N_{(\alpha/2)}\sigma_{\bar{\mathsf{y}}}\}. \tag{14.0.12}$$

What do these symbols mean? The interval R (as in region) has $\boldsymbol{y}$ as an argument, i.e.. it is denoted $R(\boldsymbol{y})$, because it depends on the observed value $\boldsymbol{y}$. $\mathbb{R}$ is the set of real numbers. $N_{(\alpha/2)}$ is the upper $\alpha/2$-quantile of the Normal distribution, i.e., it is that number c for which a standard Normal random variable z satisfies $\Pr[\mathsf{z} \geq c] = \alpha/2$. Since by the symmetry of the Normal distribution, $\Pr[\mathsf{z} \leq -c] = \alpha/2$ as well, one obtains for a two-sided test:

$$\Pr[|\mathsf{z}| \geq N_{(\alpha/2)}] = \alpha. \tag{14.0.13}$$

From this follows the coverage probability:

$$\Pr[\mathsf{R}(\boldsymbol{\mathsf{y}}) \ni \mu] = \Pr[|\mu-\bar{\mathsf{y}}| \leq N_{(\alpha/2)}\sigma_{\bar{\mathsf{y}}}] \tag{14.0.14}$$

$$= \Pr[|(\mu-\bar{\mathsf{y}})/\sigma_{\bar{\mathsf{y}}}| \leq N_{(\alpha/2)}] = \Pr[|-\mathsf{z}| \leq N_{(\alpha/2)}] = 1-\alpha \tag{14.0.15}$$

since $\mathsf{z} = (\bar{\mathsf{y}}-\mu)/\sigma_{\bar{\mathsf{y}}}$ is a standard Normal. I.e., $R(\boldsymbol{y})$ is a confidence interval for μ with confidence level $1-\alpha$.

(2) Second part: what if σ^2 is not known? Here a seemingly ad-hoc way out would be to replace σ^2 by its unbiased estimate s^2. Of course, then the Normal distribution no longer applies. However if one replaces the normal critical values by those of the t_{n-1} distribution, one still gets, by miraculous coincidence, a confidence level which is independent of any unknown parameters.

Problem 211. *If $\mathsf{y}_i \sim \mathsf{NID}(\mu, \sigma^2)$ (normally independently distributed) with μ and σ^2 unknown, then the confidence interval for μ has the form*

$$R(\boldsymbol{y}) = \{u \in \mathbb{R}\colon\ |u-\bar{y}| \leq t_{(n-1;\alpha/2)}s_{\bar{\mathsf{y}}}\}. \tag{14.0.16}$$

Here $t_{(n-q;\alpha/2)}$ is the upper $\alpha/2$-quantile of the t distribution with $n-1$ degrees of freedom, i.e., it is that number c for which a random variable t which has a t distribution with $n-1$ degrees of freedom satisfies $\Pr[\mathsf{t} \geq c] = \alpha/2$. And $s_{\bar{\mathsf{y}}}$ is obtained as follows: write down the standard deviation of $\bar{\mathsf{y}}$ and replace σ by s. One can also say $s_{\bar{\mathsf{y}}} = \sigma_{\bar{\mathsf{y}}}\frac{s}{\sigma}$ where $\sigma_{\bar{\mathsf{y}}}$ is an abbreviated notation for $\operatorname{std.dev}[\boldsymbol{\mathsf{y}}] = \sqrt{\operatorname{var}[\boldsymbol{\mathsf{y}}]}$.

- a. *1 point Write down the formula for $s_{\bar{\mathsf{y}}}$.*

Answer. Start with $\sigma^2_{\bar{\mathsf{y}}} = \operatorname{var}[\bar{\mathsf{y}}] = \frac{\sigma^2}{n}$, therefore $\sigma_{\bar{\mathsf{y}}} = \sigma/\sqrt{n}$, and

$$s_{\bar{\mathsf{y}}} = s/\sqrt{n} = \sqrt{\sum\frac{(\mathsf{y}_i-\bar{\mathsf{y}})^2}{n(n-1)}} \tag{14.0.17}$$

□

- b. *2 points Compute the coverage probability of the interval* (14.0.16).

TABLE 1. Percentiles of Student's t Distribution. Table entry x satisfies $\Pr[\mathrm{t}_n \leq x] = p$.

	$p =$					
n	.750	.900	.950	.975	.990	.995
1	1.000	3.078	6.314	12.706	31.821	63.657
2	0.817	1.886	2.920	4.303	6.965	9.925
3	0.765	1.638	2.354	3.182	4.541	5.841
4	0.741	1.533	2.132	2.776	3.747	4.604
5	0.727	1.476	2.015	2.571	3.365	4.032

ANSWER. The coverage probability is

$$\Pr[\mathsf{R}(\mathbf{y}) \ni \mu] = \Pr[\left|\mu - \bar{\mathrm{y}}\right| \leq t_{(n-1;\alpha/2)}\, \mathrm{s}_{\bar{\mathrm{y}}}] \tag{14.0.18}$$

$$= \Pr[\left|\frac{\mu - \bar{\mathrm{y}}}{\mathrm{s}_{\bar{\mathrm{y}}}}\right| \leq t_{(n-1;\alpha/2)}] \tag{14.0.19}$$

$$= \Pr[\left|\frac{(\mu - \bar{\mathrm{y}})/\sigma_{\bar{\mathrm{y}}}}{\mathrm{s}_{\bar{\mathrm{y}}}/\sigma_{\bar{\mathrm{y}}}}\right| \leq t_{(n-1;\alpha/2)}] \tag{14.0.20}$$

$$= \Pr[\left|\frac{(\bar{\mathrm{y}} - \mu)/\sigma_{\bar{\mathrm{y}}}}{\mathrm{s}/\sigma}\right| \leq t_{(n-1;\alpha/2)}] \tag{14.0.21}$$

$$= 1 - \alpha, \tag{14.0.22}$$

because the expression in the numerator is a standard normal, and the expression in the denominator is the square root of an independent χ^2_{n-1} divided by $n-1$. The random variable between the absolute signs has therefore a t-distribution, and (14.0.22) follows from (41.4.8).

□

• c. *2 points Four independent observations are available of a normal random variable with unknown mean μ and variance σ^2: the values are -2, $-\sqrt{2}$, $+\sqrt{2}$, and $+2$. (These are not the kind of numbers you are usually reading off a measurement instrument, but they make the calculation easy). Give a 95% confidence interval for μ. Table 1 gives the percentiles of the* t*-distribution.*

ANSWER. In our situation

$$\frac{\bar{\mathrm{x}} - \mu}{\mathrm{s}/\sqrt{n}} \sim \mathrm{t}_3 \tag{14.0.23}$$

According to table 1, for $b = 3.182$ follows

$$\Pr[\mathrm{t}_3 \leq b] = 0.975 \tag{14.0.24}$$

therefore

$$\Pr[\mathrm{t}_3 > b] = 0.025 \tag{14.0.25}$$

and by symmetry of the t-distribution

$$\Pr[\mathrm{t}_3 < -b] = 0.025 \tag{14.0.26}$$

Now subtract (14.0.26) from (14.0.24) to get

$$\Pr[-b \leq \mathrm{t}_3 \leq b] = 0.95 \tag{14.0.27}$$

or

$$\Pr[|\mathrm{t}_3| \leq b] = 0.95 \tag{14.0.28}$$

or, plugging in the formula for t_3,

$$\Pr\Big[\Big|\frac{\bar{\mathrm{x}} - \mu}{\mathrm{s}/\sqrt{n}}\Big| \leq b\Big] = .95 \tag{14.0.29}$$

$$\Pr[|\bar{\mathrm{x}} - \mu| \leq b\mathrm{s}/\sqrt{n}] = .95 \tag{14.0.30}$$

$$\Pr[-b\mathrm{s}/\sqrt{n} \leq \mu - \bar{\mathrm{x}} \leq b\mathrm{s}/\sqrt{n}] = .95 \tag{14.0.31}$$

$$\Pr[\bar{\mathrm{x}} - b\mathrm{s}/\sqrt{n} \leq \mu \leq \bar{\mathrm{x}} + b\mathrm{s}/\sqrt{n}] = .95 \tag{14.0.32}$$

the confidence interval is therefore $\left[\bar{\mathrm{x}} - b\mathrm{s}/\sqrt{n}, \bar{\mathrm{x}} + b\mathrm{s}/\sqrt{n}\right]$. In our sample, $\bar{\mathrm{x}} = 0$, $\mathrm{s}^2 = \frac{12}{3} = 4$, $n = 4$, therefore $\mathrm{s}^2/n = 1$, therefore also $\mathrm{s}/\sqrt{n} = 1$. So the sample value of the confidence interval is $[-3.182, +3.182]$.

□

PROBLEM 212. *Using* R*, construct 20 samples of 12 observation each from a* $\mathsf{N}(0,1)$ *distribution, construct the 95% confidence t-intervals for the mean based on these 20 samples, plot these intervals, and count how many intervals contain the true mean.*

Here are the commands: `stdnorms<-matrix(rnorm(240),nrow=12,ncol=20` *gives a* 12×20 *matrix containing 240 independent random normals. You get the vector containing the midpoints of the confidence intervals by the assignment* `midpts <- apply(stdnorms,2,mean)`*. About* `apply` *see* [**BCW96**, p. 130]*. The vector containing the half width of each confidence interval can be obtained by another use of* `apply`*:* `halfwidth <- (qt(0.975,11)/sqrt(12)) * sqrt(apply(stdnorms,2,var));` *To print the values on the screen you may simply issue the command* `cbind(midpts-halfwidth,m` *But it is much better to plot them. Since such a plot does not have one of the usual formats, we have to put it together with some low-level commands. See* [**BCW96**, page 325]*. At the very minimum we need the following:* `frame()` *starts a new plot.* `par(usr = c(1,20, range(c(midpts-halfwidth,midpts+halfwidth))` *sets a coordinate system which accommodates all intervals. The 20 confidence intervals are constructed by* `segments(1:20, midpts-halfwidth, 1:20, midpts+halfwidth)`*. Finally,* `abline(0,0)` *adds a horizontal line, so that you can see how many intervals contain the true mean.*

The `ecmet` *package has a function* `confint.segments` *which draws such plots automatically. Choose how many observations in each experiment (the argument*

n), and how many confidence intervals (the argument `rep`*), and the confidence level* `level` *(the default is here 95%), and then issue, e.g. the command* `confint.segments(n=`

Here is the transcript of the function:

```
confint.segments <- function(n, rep, level = 95/100)
{
  stdnormals <- matrix(rnorm(n * rep), nrow = n, ncol = rep)
  midpts <- apply(stdnormals, 2, mean)
  halfwidth <- qt(p=(1 + level)/2, df= n - 1) * sqrt(1/n)* sqrt(apply(std
  frame()
  x <- c(1:rep, 1:rep)
  y <- c(midpts + halfwidth, midpts - halfwidth)
  par(usr = c(1, rep, range(y)))
  segments(1:rep, midpts - halfwidth, 1:rep, midpts + halfwidth)
  abline(0, 0)
  invisible(cbind(x,y))
}
```

This function draws the plot as a "side effect," but it also returns a matrix with the coordinates of the endpoints of the plots (without printing them on the screen). This matrix can be used as input for the `identify` *function. If you do for instance* `iddata<-confint.segments(12,20)` *and then* `identify(iddata,labels=iddata[,2]`*, then the following happens: if you move the mouse cursor on the graph near one of the endpoints of one of the intervals, and click the left button, then it will print on the graph the coordinate of the bounday of this interval. Clicking any other button of the mouse gets you out of the* `identify` *function.*

CHAPTER 15

Hypothesis Testing

Imagine you are a business person considering a major investment in order to launch a new product. The sales prospects of this product are not known with certainty. You have to rely on the outcome of n marketing surveys that measure the demand for the product once it is offered. If μ is the actual (unknown) rate of return on the investment, each of these surveys here will be modeled as a random variable, which has a Normal distribution with this mean μ and known variance 1. Let $y_1, y_2, \ldots, y_n$ be the observed survey results. How would you decide whether to build the plant?

The intuitively reasonable thing to do is to go ahead with the investment if the sample mean of the observations is greater than a given value c, and not to do it otherwise. This is indeed an optimal decision rule, and we will discuss in what respect it is, and how c should be picked.

Your decision can be the wrong decision in two different ways: either you decide to go ahead with the investment although there will be no demand for the product, or you fail to invest although there would have been demand. There is no decision rule which eliminates both errors at once; the first error would be minimized by the rule never to produce, and the second by the rule always to produce. In order to determine the right tradeoff between these errors, it is important to be aware of their *asymmetry*. The error to go ahead with production although there is no demand has potentially disastrous consequences (loss of a lot of money), while the other error may cause you to miss a profit opportunity, but there is no actual loss involved, and presumably you can find other opportunities to invest your money.

To express this asymmetry, the error with the potentially disastrous consequences is called "error of type one," and the other "error of type two." The distinction between type one and type two errors can also be made in other cases. Locking up an innocent person is an error of type one, while letting a criminal go unpunished is an error of type two; publishing a paper with false results is an error of type one, while foregoing an opportunity to publish is an error of type two (at least this is what it ought to be).

Such an asymmetric situation calls for an asymmetric decision rule. One needs strict safeguards against committing an error of type one, and if there are several decision rules which are equally safe with respect to errors of type one, then one will select among those that decision rule which minimizes the error of type two.

Let us look here at decision rules of the form: make the investment if $\bar{\mathsf{y}} > c$. An error of type one occurs if the decision rule advises you to make the investment while there is no demand for the product. This will be the case if $\bar{\mathsf{y}} > c$ but $\mu \leq 0$. The probability of this error depends on the unknown parameter μ, but it is at most $\alpha = \Pr[\bar{\mathsf{y}} > c \,|\, \mu = 0]$. This maximum value of the type one error probability is called the significance level, and you, as the director of the firm, will have to decide on α depending on how tolerable it is to lose money on this venture, which presumably depends on the chances to lose money on alternative investments. It is a serious shortcoming of the classical theory of hypothesis testing that it does not provide good guidelines how α should be chosen, and how it should change with sample size. Instead, there is the tradition to choose α to be either 5% or 1% or 0.1%. Given α, a table of the cumulative standard normal distribution function allows you to find that c for which $\Pr[\bar{\mathsf{y}} > c \,|\, \mu = 0] = \alpha$.

Problem 213. *2 points Assume each* $\mathsf{y}_i \sim N(\mu, 1)$, $n = 400$ *and* $\alpha = 0.05$, *and different* y_i *are independent. Compute the value* c *which satisfies* $\Pr[\bar{\mathsf{y}} > c \,|\, \mu = 0] = \alpha$. *You shoule either look it up in a table and include a xerox copy of the table with the entry circled and the complete bibliographic reference written on the xerox copy, or do it on a computer, writing exactly which commands you used. In* R, *the function* `qnorm` *does what you need, find out about it by typing* `help(qnorm)`.

Answer. In the case $n = 400$, $\bar{\mathsf{y}}$ has variance $1/400$ and therefore standard deviation $1/20 = 0.05$. Therefore $20\bar{\mathsf{y}}$ is a standard normal: from $\Pr[\bar{\mathsf{y}} > c \,|\, \mu = 0] = 0.05$ follows $\Pr[20\bar{\mathsf{y}} > 20c \,|\, \mu = 0] = 0.05$. Therefore $20c = 1.645$ can be looked up in a table, perhaps use [**JHG+88**, p. 986], the row for ∞ d.f.

Let us do this in R. The p-"quantile" of the distribution of the random variable y is defined as that value q for which $\Pr[\mathsf{y} \leq q] = p$. If y is normally distributed, this quantile is computed by the R-function `qnorm(p, mean=0, sd=1, lower.tail=TRUE)`. In the present case we need either `qnorm(p=1-0.05, mean=0, sd=0.05)` or `qnorm(p=0.05, mean=0, sd=0.05, lower.tail=FALSE)` which gives the value 0.08224268.

□

Choosing a decision which makes a loss unlikely is not enough; your decision must also give you a chance of success. E.g., the decision rule to build the plant if $-0.06 \leq \bar{y} \leq -0.05$ and not to build it otherwise is completely perverse, although the significance level of this decision rule is approximately 4% (if $n = 100$). In other words, the significance level is not enough information for evaluating the performance of the test. You also need the "power function," which gives you the probability with which the test advises you to make the "critical" decision, as a function of

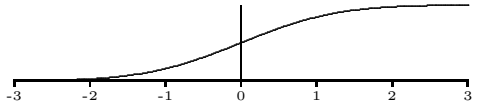


FIGURE 1. Eventually this Figure will show the Power function of a one-sided normal test, i.e., the probability of error of type one as a function of μ; right now this is simply the cdf of a Standard Normal

the true parameter values. (Here the "critical" decision is that decision which might potentially lead to an error of type one.) By the definition of the significance level, the power function does not exceed the significance level for those parameter values for which going ahead would lead to a type 1 error. But only those tests are "powerful" whose power function is high for those parameter values for which it would be correct to go ahead. In our case, the power function must be below 0.05 when $\mu \leq 0$, and we want it as high as possible when $\mu > 0$. Figure 1 shows the power function for the decision rule to go ahead whenever $\bar{\mathrm{y}} \geq c$, where c is chosen in such a way that the significance level is 5%, for $n = 100$.

The hypothesis whose *rejection*, although it is true, constitutes an error of type one, is called the *null hypothesis*, and its alternative the *alternative hypothesis*. (In the examples the null hypotheses were: the return on the investment is zero or negative, the defendant is innocent, or the results about which one wants to publish a research paper are wrong.) The null hypothesis is therefore the hypothesis that nothing is the case. The test tests whether this hypothesis should be rejected, will safeguard against the hypothesis one *wants to reject* but one is afraid to reject *erroneously*. If you reject the null hypothesis, you don't want to regret it.

Mathematically, every test can be identified with its null hypothesis, which is a region in parameter space (often consisting of one point only), and its "critical region," which is the *event* that the test comes out in favor of the "critical decision," i.e., rejects the null hypothesis. The critical region is usually an event of the form that the value of a certain random variable, the "test statistic," is within a given range, usually that it is too high. The power function of the test is the probability of the critical region as a function of the unknown parameters, and the significance level is the maximum (or, if this maximum depends on unknown parameters, any upper bound) of the power function over the null hypothesis.

PROBLEM 214. *Mr. Jones is on trial for counterfeiting Picasso paintings, and you are an expert witness who has developed fool-proof statistical significance tests for identifying the painter of a given painting.*

• a. *2 points There are two ways you can set up your test.*

a: *You can either say: The null hypothesis is that the painting was done by Picasso, and the alternative hypothesis that it was done by Mr. Jones.*

b: *Alternatively, you might say: The null hypothesis is that the painting was done by Mr. Jones, and the alternative hypothesis that it was done by Picasso.*

Does it matter which way you do the test, and if so, which way is the correct one. Give a reason to your answer, i.e., say what would be the consequences of testing in the incorrect way.

ANSWER. The determination of what the null and what the alternative hypothesis is depends on what is considered to be the catastrophic error which is to be guarded against. On a trial, Mr. Jones is considered innocent until proven guilty. Mr. Jones should not be convicted unless he can be proven guilty beyond "reasonable doubt." Therefore the test must be set up in such a way that the hypothesis that the painting is by Picasso will only be rejected if the chance that it is actually by Picasso is very small. The error of type one is that the painting is considered counterfeited although it is really by Picasso. Since the error of type one is always the error to reject the null hypothesis although it is true, solution a. is the correct one. You are not proving, you are testing. □

• b. *2 points After the trial a customer calls you who is in the process of acquiring a very expensive alleged Picasso painting, and who wants to be sure that this painting is not one of Jones's falsifications. Would you now set up your test in the same way as in the trial or in the opposite way?*

ANSWER. It is worse to spend money on a counterfeit painting than to forego purchasing a true Picasso. Therefore the null hypothesis would be that the painting was done by Mr. Jones, i.e., it is the opposite way. □

PROBLEM 215. *7 points Someone makes an extended experiment throwing a coin 10,000 times. The relative frequency of heads in these 10,000 throws is a random variable. Given that the probability of getting a head is p, what are the mean and standard deviation of the relative frequency? Design a test, at 1% significance level, of the null hypothesis that the coin is fair, against the alternative hypothesis that $p < 0.5$. For this you should use the central limit theorem. If the head showed 4,900 times, would you reject the null hypothesis?*

ANSWER. Let x_i be the random variable that equals one when the i-th throw is a head, and zero otherwise. The expected value of x is p, the probability of throwing a head. Since $\mathrm{x}^2 = \mathrm{x}$, $\mathrm{var}[\mathrm{x}] = \mathrm{E}[\mathrm{x}] - (\mathrm{E}[\mathrm{x}])^2 = p(1-p)$. The relative frequency of heads is simply the average of all x_i, call it $\bar{\mathrm{x}}$. It has mean p and variance $\sigma^2_{\bar{\mathrm{x}}} = \frac{p(1-p)}{10{,}000}$. Given that it is a fair coin, its mean is 0.5 and its standard deviation is 0.005. Reject if the actual frequency $< 0.5 - 2.326\sigma_{\bar{\mathrm{x}}} = .48857$. Another approach:

$$\Pr(\bar{\mathrm{x}} \leq 0.49) = \Pr\left(\frac{\bar{\mathrm{x}} - 0.5}{0.005} \leq -2\right) = 0.0227 \tag{15.0.33}$$

since the fraction is, by the central limit theorem, approximately a standard normal random variable. Therefore do not reject. □

15.1. Duality between Significance Tests and Confidence Regions

There is a duality between confidence regions with confidence level $1-\alpha$ and certain significance tests. Let us look at a family of significance tests, which all have a significance level $\leq \alpha$, and which define for every possible value of the parameter $\boldsymbol{\phi}_0 \in \Omega$ a critical region $C(\boldsymbol{\phi}_0)$ for rejecting the simple null hypothesis that the true parameter is equal to $\boldsymbol{\phi}_0$. The condition that all significance levels are $\leq \alpha$ means mathematically

$$\Pr\big[C(\boldsymbol{\phi}_0)|\boldsymbol{\phi}=\boldsymbol{\phi}_0\big] \leq \alpha \quad \text{for all } \boldsymbol{\phi}_0 \in \Omega. \tag{15.1.1}$$

Mathematically, confidence regions and such families of tests are one and the same thing: if one has a confidence region $R(\mathsf{y})$, one can define a test of the null hypothesis $\boldsymbol{\phi}=\boldsymbol{\phi}_0$ as follows: for an observed outcome $\boldsymbol{y}$ reject the null hypothesis if and only if $\boldsymbol{\phi}_0$ is not contained in $R(\boldsymbol{y})$. On the other hand, given a family of tests, one can build a confidence region by the prescription: $R(\boldsymbol{y})$ is the set of all those parameter values which would not be rejected by a test based on observation $\boldsymbol{y}$.

Problem 216. *Show that with these definitions, equations* (14.0.5) *and* (15.1.1) *are equivalent.*

Answer. Since $\boldsymbol{\phi}_0 \in R(\boldsymbol{y})$ iff $\boldsymbol{y} \in C'(\boldsymbol{\phi}_0)$ (the complement of the critical region rejecting that the parameter value is $\boldsymbol{\phi}_0$), it follows $\Pr[R(\mathsf{y}) \in \boldsymbol{\phi}_0|\boldsymbol{\phi}=\boldsymbol{\phi}_0] = 1-\Pr[C(\boldsymbol{\phi}_0)|\boldsymbol{\phi}=\boldsymbol{\phi}_0] \geq 1-\alpha$. □

This duality is discussed in [**BD77**, pp. 177–182].

15.2. The Neyman Pearson Lemma and Likelihood Ratio Tests

Look one more time at the example with the fertilizer. Why are we considering only regions of the form $\bar{\mathsf{y}} \geq \mu_0$, why not one of the form $\mu_1 \leq \bar{\mathsf{y}} \leq \mu_2$, or maybe not use the mean but decide to build if $\mathsf{y}_1 \geq \mu_3$? Here the μ_1, μ_2, and μ_3 can be chosen such that the probability of committing an error of type one is still α.

It seems intuitively clear that these alternative decision rules are not reasonable. The Neyman Pearson lemma proves this intuition right. It says that the critical regions of the form $\bar{\mathsf{y}} \geq \mu_0$ are uniformly most powerful, in the sense that every other critical region with same probability of type one error has equal or higher probability of committing error of type two, regardless of the true value of μ.

Here are formulation and proof of the Neyman Pearson lemma, first for the case that both null hypothesis and alternative hypothesis are simple: $H_0: \theta=\theta_0$, $H_A: \theta=\theta_1$. In other words, we want to determine on the basis of the observations of the random variables $\mathsf{y}_1, \ldots, \mathsf{y}_n$ whether the true θ was θ_0 or θ_1, and a determination $\theta=\theta_1$ when in fact $\theta=\theta_0$ is an error of type one. The critical region C is the set of all outcomes that lead us to conclude that the parameter has value θ_1.

The Neyman Pearson lemma says that a uniformly most powerful test exists in this situation. It is a so-called likelihood-ratio test, which has the following critical region:

$$\mathsf{C} = \{y_1, \ldots, y_n \colon L(y_1, \ldots, y_n; \theta_1) \geq kL(y_1, \ldots, y_n; \theta_0)\}. \tag{15.2.1}$$

C consists of those outcomes for which θ_1 is at least k times as likely as θ_0 (where k is chosen such that $\Pr[\mathsf{C}|\theta_0]=\alpha$).

To prove that this decision rule is uniformly most powerful, assume D is the critical region of a different test with same significance level α, i.e., if the null hypothesis is correct, then C and D reject (and therefore commit an error of type one) with equally low probabilities α. In formulas, $\Pr[\mathsf{C}|\theta_0]=\Pr[\mathsf{D}|\theta_0]=\alpha$. Look at figure 2 with $\mathsf{C}=\mathsf{U}\cup\mathsf{V}$ and $\mathsf{D}=\mathsf{V}\cup\mathsf{W}$. Since C and D have the same significance level, it follows

$$\Pr[\mathsf{U}|\theta_0] = \Pr[\mathsf{W}|\theta_0]. \tag{15.2.2}$$

Also

$$\Pr[\mathsf{U}|\theta_1] \geq k\Pr[\mathsf{U}|\theta_0], \tag{15.2.3}$$

since $\mathsf{U} \subset \mathsf{C}$ and C were chosen such that the likelihood (density) function of the alternative hypothesis is high relatively to that of the null hypothesis. Since W lies outside C, the same argument gives

$$\Pr[\mathsf{W}|\theta_1] \leq k\Pr[\mathsf{W}|\theta_0]. \tag{15.2.4}$$

Linking those two inequalities and the equality gives

$$\Pr[\mathsf{W}|\theta_1] \leq k\Pr[\mathsf{W}|\theta_0] = k\Pr[\mathsf{U}|\theta_0] \leq \Pr[\mathsf{U}|\theta_1], \tag{15.2.5}$$

hence $\Pr[\mathsf{D}|\theta_1] \leq \Pr[\mathsf{C}|\theta_1]$. In other words, if θ_1 is the correct parameter value, then C will discover this and reject at least as often as D. Therefore C is at least as powerful as D, or the type two error probability of C is at least as small as that of D.

Back to our fertilizer example. To make both null and alternative hypotheses simple, assume that either $\mu=0$ (fertilizer is ineffective) or $\mu=t$ for some fixed

FIGURE 2. Venn Diagram for Proof of Neyman Pearson Lemma ec660.1005

$t > 0$. Then the likelihood ratio critical region has the form

(15.2.6)
$$C = \{y_1, \dots, y_n\colon \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}((y_1-t)^2+\cdots+(y_n-t)^2)} \geq k\left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}(y_1^2+\cdots+y_n^2)}\}$$

(15.2.7)
$$= \{y_1, \dots, y_n\colon -\frac{1}{2}((y_1-t)^2+\cdots+(y_n-t)^2) \geq \ln k - \frac{1}{2}(y_1^2+\cdots+y_n^2)\}$$

(15.2.8)
$$= \{y_1, \dots, y_n\colon t(y_1+\cdots+y_n) - \frac{t^2 n}{2} \geq \ln k\}$$

(15.2.9)
$$= \{y_1, \dots, y_n\colon \bar{y} \geq \frac{\ln k}{nt} + \frac{t}{2}\}$$

i.e., C has the form $\bar{y} \geq$ some constant. The dependence of this constant on k is not relevant, since this constant is usually chosen such that the maximum probability of error of type one is equal to the given significance level.

PROBLEM 217. *8 points You have four independent observations $y_1, \dots, y_4$ from an $\mathsf{N}(\mu, 1)$, and you are testing the null hypothesis $\mu = 0$ against the alternative hypothesis $\mu = 1$. For your test you are using the likelihood ratio test with critical region*

(15.2.10) $$C = \{y_1, \dots, y_4\colon L(y_1, \dots, y_4; \mu = 1) \geq 3.633 \cdot L(y_1, \dots, y_4; \mu = 0)\}.$$

Compute the significance level of this test. (According to the Neyman-Pearson lemma, this is the uniformly most powerful test for this significance level.) Hints: In order to show this you need to know that $\ln 3.633 = 1.29$*, everything else can be done without a calculator. Along the way you may want to show that C can also be written in the form $C = \{y_1, \dots, y_4 :\ y_1 + \cdots + y_4 \geq 3.290\}$.*

ANSWER. Here is the equation which determines when $y_1, \dots, y_4$ lie in C:

(15.2.11) $$(2\pi)^{-2}\exp -\frac{1}{2}\Big((y_1-1)^2+\cdots+(y_4-1)^2\Big) \geq 3.633 \cdot (2\pi)^{-2}\exp -\frac{1}{2}\Big(y_1^2+\cdots+y_4^2\Big)$$

(15.2.12) $$-\frac{1}{2}\Big((y_1-1)^2+\cdots+(y_4-1)^2\Big) \geq \ln(3.633) - \frac{1}{2}\Big(y_1^2+\cdots+y_4^2\Big)$$

(15.2.13) $$y_1+\cdots+y_4-2 \geq 1.290$$

Since $\Pr[\mathsf{y}_1+\cdots+\mathsf{y}_4 \geq 3.290] = \Pr[\mathsf{z} = (\mathsf{y}_1+\cdots+\mathsf{y}_4)/2 \geq 1.645]$ and z is a standard normal, one obtains the significance level of 5% from the standard normal table or the t-table. □

Note that due to the properties of the Normal distribution, this critical region, for a given significance level, does not depend at all on the value of t. Therefore this test is uniformly most powerful against the composite hypothesis $\mu > 0$.

One can als write the null hypothesis as the composite hypothesis $\mu \leq 0$, because the highest probability of type one error will still be attained when $\mu = 0$. This completes the proof that the test given in the original fertilizer example is uniformly most powerful.

Most other distributions discussed here are equally well behaved, therefore uniformly most powerful one-sided tests exist not only for the mean of a normal with known variance, but also the variance of a normal with known mean, or the parameters of a Bernoulli and Poisson distribution.

However the given one-sided hypothesis is the only situation in which a uniformly most powerful test exists. In other situations, the *generalized likelihood ratio test* has good properties even though it is no longer uniformly most powerful. Many known tests (e.g., the F test) are generalized likelihood ratio tests.

Assume you want to test the composite null hypothesis $H_0 : \theta \in \omega$, where ω is a subset of the parameter space, against the alternative $H_A : \theta \in \Omega$, where $\Omega \supset \omega$ is a more comprehensive subset of the parameter space. ω and Ω are defined by functions with continuous first-order derivatives. The generalized likelihood ratio critical region has the form

(15.2.14) $$C = \{x_1, \dots, x_n\colon \frac{\sup_{\theta\in\Omega} L(x_1, \dots, x_n; \theta)}{\sup_{\theta\in\omega} L(x_1, \dots, x_n; \theta)} \geq k\}$$

where k is chosen such that the probability of the critical region when the null hypothesis is true has as its maximum the desired significance level. It can be shown that twice the log of this quotient is asymptotically distributed as a χ^2_{q-s}, where q is the dimension of Ω and s the dimension of ω. (Sometimes the likelihood ratio is defined as the inverse of this ratio, but whenever possible we will define our test statistics so that the null hypothjesis is rejected if the value of the test statistic is too large.)

In order to perform a likelihood ratio test, the following steps are necessary: First construct the MLE's for $\theta \in \Omega$ and $\theta \in \omega$, then take twice the difference of the attained levels of the log likelihoodfunctions, and compare with the χ^2 tables.

15.3. The Runs Test

[**Spr98**, pp. 171–175] is a good introductory treatment, similar to the one given here. More detail in [**GC92**, Chapter 3] (not in University of Utah Main Library) and even more in [**Bra68**, Chapters 11 and 23] (which is in the Library).

Each of your three research assistants has to repeat a certain experiment 9 times, and record whether each experiment was a success (1) or a failure (0). In all cases, the experiments happen to have been successful 4 times. Assistant A has the following sequence of successes and failures: $0, 1, 0, 0, 1, 0, 1, 1, 0$, B has $0, 1, 0, 1, 0, 1, 0, 1, 0$, and C has $1, 1, 1, 1, 0, 0, 0, 0, 0$.

On the basis of these results, you suspect that the experimental setup used by B and C is faulty: for C, it seems that something changed over time so that the first experiments were successful and the latter experiments were not. Or perhaps the fact that a given experiment was a success (failure) made it more likely that also the next experiment would be a success (failure). For B, the opposite effect seems to have taken place.

From the pattern of successes and failures you made inferences about whether the outcomes were independent or followed some regularity. A mathematical formalization of this inference counts "runs" in each sequence of outcomes. A run is a sucession of several ones or zeros. The first outcome had 7 runs, the second 9, and the third only 2. Given that the number of successes is 4 and the number of failures is 5, 9 runs seem too many and 2 runs too few.

The "runs test" (sometimes also called "run test") exploits this in the following way: it counts the number of runs, and then asks if this is a reasonable number of runs to expect given the total number of successes and failures. It rejects whenever the number of runs is either too large or too low.

The choice of the number of runs as test statistic cannot be derived from a likelihood ratio principle, since we did not specify the joint distribution of the outcome of the experiment. But the above argument says that it will probably detect at least some of the cases we are interested in.

In order to compute the error of type one, we will first derive the probability distribution of the number of runs conditionally on the outcome that the number of successes is 4. This conditional distribution can be computed, even if we do not know the probability of success of each experiment, as long as their joint distribution has the following property (which holds under the null hypothesis of statistical independence): the probability of a given sequence of failures and successes only depends on

the number of failures and successes, not on the order in which they occur. Then the conditional distribution of the number of runs can be obtained by simple counting.

How many arrangements of 5 zeros and 4 ones are there? The answer is $\binom{9}{4} = \frac{(9)(8)(7)(6)}{(1)(2)(3)(4)} = 126$. How many of these arrangements have 9 runs? Only one, i.e., the probability of having 9 runs (conditionally on observing 4 successes) is 1/126. The probability of having 2 runs is 2/126, since one can either have the zeros first, or the ones first.

In order to compute the probability of 7 runs, lets first ask: what is the probability of having 4 runs of ones and 3 runs of zeros? Since there are only 4 ones, each run of ones must have exactly one element. So the distribution of ones and zeros must be:

$$1 - \text{one or more zeros} - 1 - \text{one or more zeros} - 1 - \text{one or more zeros} - 1.$$

In order to specify the distribution of ones and zeros completely, we must therefore count how many ways there are to split the sequence of 5 zeros into 3 nonempty batches. Here are the possibilities:

$$\begin{array}{ccccccccc} 0 & & 0 & & 0 & | & 0 & | & 0 \\ 0 & & 0 & | & 0 & & 0 & | & 0 \\ 0 & & 0 & | & 0 & | & 0 & & 0 \\ 0 & | & 0 & & 0 & & 0 & | & 0 \\ 0 & | & 0 & & 0 & | & 0 & & 0 \\ 0 & | & 0 & | & 0 & & 0 & & 0 \end{array} \tag{15.3.1}$$

Generally, the number of possibilities is $\binom{4}{2}$ because there are 4 spaces between those 5 zeros, and we have to put in two dividers.

We have therfore 6 possibilities to make 4 runs of zeros and 3 runs of ones. Now how many possiblities are there to make 3 runs of zeros and 4 runs of ones? There are 4 ways to split the 5 zeros into 4 batches, and there are 3 ways to split the 4 ones into 3 batches, represented by the schemes

$$\begin{array}{ccccccccc} 0 & & 0 & | & 0 & | & 0 & | & 0 \\ 0 & | & 0 & & 0 & | & 0 & | & 0 \\ 0 & | & 0 & | & 0 & & 0 & | & 0 \\ 0 & | & 0 & | & 0 & | & 0 & & 0 \end{array} \quad \text{and} \quad \begin{array}{ccccccc} 1 & & 1 & | & 1 & | & 1 \\ 1 & | & 1 & & 1 & | & 1 \\ 1 & | & 1 & | & 1 & & 1 \end{array} \tag{15.3.2}$$

One can combine any of the first with any of the second, i.e., one obtains 12 possibilities. Together the probability of seven runs is therefore 18/126.

One can do the same thing for all other possibilities of runs and will get a distribution of runs similar to that depicted in the diagram (which is for 7 instead of 9 trials). Mathematically one gets two different formulas according to whether the number of runs is odd or even: we have a total of m zeros and n ones (it could also

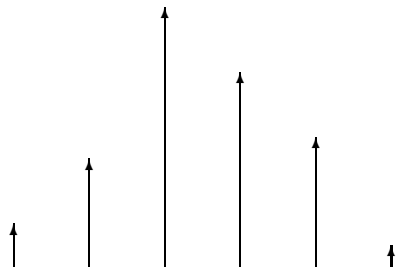


FIGURE 3. Distribution of runs in 7 trials, if there are 4 successes and 3 failures

be the other way round), and $\mathtt{r}$ is the number of runs:

$$\Pr[\mathtt{r} = 2s+1] = \frac{\binom{m-1}{s-1}\binom{n-1}{s} + \binom{m-1}{s}\binom{n-1}{s-1}}{\binom{m+n}{m}} \tag{15.3.3}$$

$$\Pr[\mathtt{r} = 2s] = 2\frac{\binom{m-1}{s-1}\binom{n-1}{s-1}}{\binom{m+n}{m}} \tag{15.3.4}$$

Some computer programs (`StatXact`, `www.cytel.com`) compute these probabilities exactly or by monte carlo simulation; but there is also an asymptotic test based on the facts that

$$\mathrm{E}[\mathtt{r}] = 1 + \frac{2mn}{m+n} \qquad \mathrm{var}[\mathtt{r}] = \frac{2mn(2mn-m-n)}{(m+n)^2(m+n-1)} \tag{15.3.5}$$

and that the standardized number of runs is asymptotically a Normal distribution. (see [**GC92**, section 3.2])

We would therefore reject when the observed number of runs is in the tails of this distribution. Since the exact test statistic is discrete, we cannot make tests for every arbitrary significance level. In the given example, if the critical region is $\{r = 9\}$, then the significance level is 1/126. If the critical region is $\{r = 2 \text{ or } 9\}$, the significance level is 3/126.

We said before that we could not make precise statements about the power of the test, i.e., the error of type two. But we will show that it is possible to make precise statements about the error of type one.

Right now we only have the *conditional* probability of errors of type one, given that there are exactly 4 successes in our 9 trials. And we have no information about

the probability of having indeed four successes, it might be 1 in a million. However in certain situations, the conditional significance level is exactly what is needed. And even if the unconditional significance level is needed, there is one way out. If we were to specify a decision rule for every number of successes in such a way that the conditional probability of rejecting is the same in all of them, then this conditional probability is also equal to the unconditional probability. The only problem here is that, due to discreteness, we can make the probability of type one errors only approximately equal; but with increasing sample size this problem disappears.

PROBLEM 218. *Write approximately 200 x'es and o's on a piece of paper trying to do it in a random manner. Then make a run test whether these x's and o's were indeed random. Would you want to run a two-sided or one-sided test?*

The law of rare events literature can be considered a generalization of the run test. For epidemiology compare [**Cha96**], [**DH94**], [**Gri79**], and [**JL97**].

15.4. Pearson's Goodness of Fit Test.

Given an experiment with r outcomes, which have probabilities $p_1, \ldots, p_r$, where $\sum p_i = 1$. You make n independent trials and the ith outcome occurred $\mathtt{x}_i$ times. The $\mathtt{x}_1, \ldots, \mathtt{x}_r$ have the multinomial distribution with parameters n and $p_1, \ldots, p_r$. Their mean and covariance matrix are given in equation (8.4.2) above. How do you test $H_0 : p_1 = p_1^0, \ldots, p_r = p_r^0$?

Pearson's Goodness of Fit test uses as test statistic a weighted sum of the squared deviations of the observed values from their expected values:

$$\sum_{i=1}^{r} \frac{(\mathtt{x}_i - np_i^0)^2}{np_i^0}. \tag{15.4.1}$$

This test statistic is often called the Chi-Square statistic. It is asymptotically distributed as a χ^2_{r-1}; reject the null hypothesis when the observed value of this statistic is too big, the critical region can be read off a table of the χ^2.

Why does one get a χ^2 distribution in the limiting case? Because the $\mathtt{x}_i$ themselves are asymptotically normal, and certain quadratic forms of normal distributions are χ^2. The matter is made a little complicated by the fact that the $\mathtt{x}_i$ are linearly dependent, since $\sum \mathtt{x}_j = n$, and therefore their covariance matrix is singular. There are two ways to deal with such a situation. One is to drop one observation; one will not lose any information by this, and the remaining $r-1$ observations are well behaved. (This explains, by the way, why one has a χ^2_{r-1} instead of a χ^2_r.)

We will take an alternative route, namely, use theorems which are valid even if the covariance matrix is singular. This is preferable because it leads to more unified theories. In equation (10.4.9), we characterized all the quadratic forms of multivariate normal variables that are χ^2's. Here it is again: Assume $\mathbf{y}$ is a jointly

normal vector random variable with mean vector $\boldsymbol{\mu}$ and covariance matrix $\sigma^2\boldsymbol{\Psi}$, and $\boldsymbol{\Omega}$ is a symmetric nonnegative definite matrix. Then $(\mathbf{y}-\boldsymbol{\mu})^\top\boldsymbol{\Omega}(\mathbf{y}-\boldsymbol{\mu}) \sim \sigma^2\chi^2_k$ iff $\boldsymbol{\Psi}\boldsymbol{\Omega}\boldsymbol{\Psi}\boldsymbol{\Omega}\boldsymbol{\Psi} = \boldsymbol{\Psi}\boldsymbol{\Omega}\boldsymbol{\Psi}$ and k is the rank of $\boldsymbol{\Omega}$. If $\boldsymbol{\Psi}$ is singular, i.e., does not have an inverse, and $\boldsymbol{\Omega}$ is a g-inverse of $\boldsymbol{\Psi}$, then condition (10.4.9) holds. A matrix $\boldsymbol{\Omega}$ is a g-inverse of $\boldsymbol{\Psi}$ iff $\boldsymbol{\Psi}\boldsymbol{\Omega}\boldsymbol{\Psi} = \boldsymbol{\Psi}$. Every matrix has at least one g-inverse, but may have more than one.

Now back to our multinomial distribution. By the central limit theorem, the x_i are asymptotically jointly normal; their mean and covariance matrix are given by equation (8.4.2). This covariance matrix is singular (has rank $r-1$), and a g-inverse is given by (15.4.2), which has in its diagonal exactly the weighting factors used in the statistic for the goodness of fit test.

Problem 219. *2 points A matrix $\boldsymbol{\Omega}$ is a g-inverse of $\boldsymbol{\Psi}$ iff $\boldsymbol{\Psi}\boldsymbol{\Omega}\boldsymbol{\Psi} = \boldsymbol{\Psi}$. Show that the following matrix*

$$\frac{1}{n}\begin{bmatrix} \frac{1}{p_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{p_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{p_r} \end{bmatrix} \tag{15.4.2}$$

is a g-inverse of the covariance matrix of the multinomial distribution given in (8.4.2).

Answer. Postmultiplied by the g-inverse given in (15.4.2), the covariance matrix from (8.4.2) becomes

(15.4.3)

$$\begin{bmatrix} p_1-p_1^2 & -p_1p_2 & \cdots & -p_1p_r \\ -p_2p_1 & p_2-p_2^2 & \cdots & -p_2p_r \\ \vdots & \vdots & \ddots & \vdots \\ -p_rp_1 & -p_rp_2 & \cdots & p_r-p_r^2 \end{bmatrix} \frac{1}{n}\begin{bmatrix} \frac{1}{p_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{p_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{p_r} \end{bmatrix} = \begin{bmatrix} 1-p_1 & -p_1 & \cdots & -p_1 \\ -p_2 & 1-p_2 & \cdots & -p_2 \\ \vdots & \vdots & \ddots & \vdots \\ -p_r & -p_r & \cdots & 1-p_r \end{bmatrix},$$

and if one postmultiplies this again by the covariance matrix, one gets the covariance matrix back:

$$\begin{bmatrix} 1-p_1 & -p_1 & \cdots & -p_1 \\ -p_2 & 1-p_2 & \cdots & -p_2 \\ \vdots & \vdots & \ddots & \vdots \\ -p_r & -p_r & \cdots & 1-p_r \end{bmatrix} \begin{bmatrix} p_1-p_1^2 & -p_1p_2 & \cdots & -p_1p_r \\ -p_2p_1 & p_2-p_2^2 & \cdots & -p_2p_r \\ \vdots & \vdots & \ddots & \vdots \\ -p_rp_1 & -p_rp_2 & \cdots & p_r-p_r^2 \end{bmatrix} = \begin{bmatrix} p_1-p_1^2 & -p_1p_2 & \cdots & -p \\ -p_2p_1 & p_2-p_2^2 & \cdots & -p \\ \vdots & \vdots & \ddots & \\ -p_rp_1 & -p_rp_2 & \cdots & p_r \end{bmatrix}$$

In the left upper corner, matrix multiplication yields the element $p_1 - 2p_1^2 + p_1^3 + p_1^2(p_2 + \cdots + p_r) = p_1 - 2p_1^2 + p_1^2(p_1 + p_2 + \cdots + p_n) = p_1 - p_1^2$, and the first element in the second row is $-2p_1p_2 + p_1p_2(p_1 + p_2 + \cdots + p_r) = -p_1p_2$. Since the product matrix is symmetric, this gives all the typical elements. □

Problem 220. *Show that the covariance matrix of the multinomial distribution given in* (8.4.2) *has rank $n-1$.*

Answer. Use the fact that the rank of $\boldsymbol{\Psi}$ is $\operatorname{tr}(\boldsymbol{\Psi}\boldsymbol{\Psi}^-)$, and one sees that the trace of the matrix on the rhs of (15.4.3) is $n-1$. □

From this follows that asymptotically, $\sum_{i+1}^r \frac{(\mathsf{x}_i - np_i)^2}{np_i}$ has a χ^2 distribution with $r-1$ degrees of freedom. This is only *asymptotically* a χ^2; the usual rule is that n must be large enough so that np_i is at least 5 for all i. Others refined this rule: if $r \geq 5$, it is possible to let one of the np_i be as small as 1 (requiring the others to be 5 or more) and still get a "pretty good" approximation.

Problem 221. [**HC70**, pp. 310/11] *I throw a die 60 times and get the following frequencies for the six faces of the die:* $13, 19, 11, 8, 5, 4$. *Test at the 5% significance level whether the probability of each face is* $\frac{1}{6}$.

Answer. The hypothesis must be rejected: observed value is 15.6, critical value 11.1. □

Until now we assumed that the probabilities of the r outcomes $p_1, \ldots, p_r$ are known. If these probabilities depend on certain unknown parameters, then this estimation procedure can be extended in an amazingly simple way. In this case we are allowed to use the observed values of the random sample to compute maximum likelihood estimates of these parameters, and plug these point estimates (instead of the known parameter values themselves) into the quadratic form. If we do that, we have to deduct the number of estimates from the number of degrees of freedom of the χ^2. An example of this is the contingency table:

Assume your sample is categorized according to two criteria:

$$\begin{array}{rcc} & \text{smoker} & \text{nonsmoker} \\ \text{lung cancer} & \mathsf{y}_{11} & \mathsf{y}_{12} \\ \text{no lung cancer} & \mathsf{y}_{21} & \mathsf{y}_{22} \end{array} \tag{15.4.4}$$

The procedure described would allow us to test whether the data is compabile with the four cell probabilities being any given set of values $\begin{matrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{matrix}$. But the question which is most often asked is not whether the data are compatible with a specific set of cell probabilities, but whether the criteria tabled are independent. I.e., whether there are two numbers p and q so that the cell probabilities have the form $\begin{matrix} pq & p(1-q) \\ (1-p)q & (1-p)(1-q) \end{matrix}$. If this is so, then p is the probability of being a smoker, and q the probability of having lung cancer. Their MLE's are $\tilde{p} = (\mathsf{y}_{11} + \mathsf{y}_{12})/n$, which is usually written as $\mathsf{y}_{1.}/n$, and $\tilde{q} = (\mathsf{y}_{11} + \mathsf{y}_{21})/n = \mathsf{y}_{.1}/n$. Therefore the MLE's of all four cell probabilities are

$$\begin{matrix} \mathsf{y}_{1.}\mathsf{y}_{.1}/n^2 & \mathsf{y}_{1.}\mathsf{y}_{.2}/n^2 \\ \mathsf{y}_{2.}\mathsf{y}_{.1}/n^2 & \mathsf{y}_{2.}\mathsf{y}_{.2}/n^2 \end{matrix}. \tag{15.4.5}$$

Plugging these MLE's into the formula for the goodness of fit test statistic, we get

$$\sum_{i,j} \frac{(\mathsf{x}_{ij} - \mathsf{x}_{.i}\mathsf{x}_{.j}/n)^2}{\mathsf{x}_{.i}\mathsf{x}_{.j}/n} \sim \chi^2_1. \tag{15.4.6}$$

Since there are four cells, i.e., three independent counts, and we estimated two parameters, the number of degrees of freedom is $3-2=1$. Again, this is only *asymptotically* a χ^2.

15.5. Permutation Tests

Of five subjects, each first pulls a ball out of an urn which contains three black and two red balls, and does not put this ball back. The result of this ball pulling is represented by a dummy 5-vector $\mathbf{d}$, where $\mathsf{d}_i = 1$ if the ith subject pulled a black ball, and $\mathsf{d}_i = 0$ if a red ball. Those who pull a black ball receive treatment A, and those pulling a red ball treatment B. The responses to this treatment are represented by the random 5-vector $\mathbf{y}$, and the following t-statistic is computed, which we will encounter again in (42.2.22):

$$\mathsf{t} = \frac{\bar{\mathsf{a}} - \bar{\mathsf{b}}}{\sqrt{\mathsf{s}^2/n_{\mathsf{a}} + \mathsf{s}^2/n_{\mathsf{b}}}} \tag{15.5.1}$$

where

$$n_{\mathsf{a}} = \sum_{i=1}^{n} \mathsf{d}_i = 3 \tag{15.5.2}$$

$$n_{\mathsf{b}} = \sum_{i=1}^{n} (1 - \mathsf{d}_i) = 2 \tag{15.5.3}$$

$$\bar{\mathsf{a}} = \frac{1}{n_{\mathsf{a}}} \sum_{i:\ \mathsf{d}_i=1} \mathsf{y}_i \tag{15.5.4}$$

$$\bar{\mathsf{b}} = \frac{1}{n_{\mathsf{b}}} \sum_{i:\ \mathsf{d}_i=0} \mathsf{y}_i \tag{15.5.5}$$

$$\mathsf{s}^2 = \frac{\sum_{i:\ \mathsf{d}_i=1}(\mathsf{y}_i - \bar{\mathsf{a}})^2 + \sum_{i:\ \mathsf{d}_i=0}(\mathsf{y}_i - \bar{\mathsf{b}})^2}{n_{\mathsf{a}} + n_{\mathsf{b}} - 2} \tag{15.5.6}$$

Assume for example that the observed values of the two random variables $\mathbf{d}$ and $\mathbf{y}$ are $\boldsymbol{d} = \begin{bmatrix} 0 & 0 & 1 & 1 & 1 \end{bmatrix}$ and $\boldsymbol{y} = \begin{bmatrix} 6 & 12 & 18 & 30 & 54 \end{bmatrix}$. I.e., the three subjects receiving treatment A had the results 18, 30, and 54, and the two subjects receiving treatment B the results 6 and 12. This gives a value of $t = 1.81$. Does this indicate a significant difference between A and B?

The usual approach to answer this question is discussed in chapter/section 42, p. 493. It makes the following assumptions: under the null hypothesis, $\mathbf{y} \sim \mathsf{N}(\boldsymbol{\iota}\mu, \sigma^2\boldsymbol{I})$ and under the alternative hypothesis, the means are different under the two treatments but the σ^2 is the same. Since the means are asymptotically Normal, t has asymptotically a t-distribution with 3 degrees of freedom. The probability that the value of this t is greater than 1.81 is 0.08399529.

The *permutation test* approach (Fisher's exact test) uses the same t-statistic as is familiar from the comparison of two means in a Normal population, but it computes the significance levels in a different way.

Under the null hypothesis, the treatments have no effect on the outcomes, i.e., every other assignment of the subjects to the treatments would have led to the same outcomes; but the values of the t-statistic would have been different. Going through all the possibilities, we see that the following other assignments of subjects to treatments would have been possible:

d_1	d_2	d_3	d_4	d_5	$\bar{\mathsf{a}}$	b	t
1	1	1	0	0	12	42	–3.00
1	1	0	1	0	16	36	–1.22
1	1	0	0	1	24	24	0.00
1	0	1	1	0	18	33	–0.83
1	0	1	0	1	26	21	0.25
1	0	0	1	1	30	15	0.83
0	1	1	1	0	20	30	–0.52
0	1	1	0	1	28	18	0.52
0	1	0	1	1	32	12	1.22
0	0	1	1	1	34	9	1.81

Since there are 10 possible outcomes, and the outcome observed is the most extreme of them, the significance level is 0.1. i.e., in this case, the permutation test, gives less evidence for a treatment effect than the ordinary t-test.

PROBLEM 222. *This is an example adapted from* [**GG95**], *which is also discussed in* [**Spr98**, pp. 2/3 and 375–379]. *Table 1 contains artificial data about two firms hiring in the same labor market. For the sake of the argument it is assumed that both firms receive the exact same number of applications (100), and both firms hire 11 new employees. Table 1 shows how many of the applicants and how many of the new hires were Minorities.*

• a. *3 points Let p_1 be the proportion of Minorities hired, and p_2 the proportion Majorities hired. Compute the difference $p_1 - p_2$ and the odds ratio $\big(p_1/(1-p_1)\big)/\big(p_2/(1-p_2)\big)$ for each firm. Which of the two firms seems to discriminate more? Is the difference of probabilities or the odds ratio the more relevant statistic here?*

Hirings by Two Different Firms with 100 Applications Each				
	Firm A		*Firm B*	
	Minority	*Majority*	*Minority*	*Majority*
Hired	*1*	*10*	*2*	*9*
Not Hired	*31*	*58*	*46*	*43*

TABLE 1. Which Firm's Hiring Policies are More Equitable?

ANSWER. In firm A, 3.125% of the minority applicants and 14.7% of the Majority applicants were hired. The difference of the probabilities is 11.581% and the odds ratio is $\frac{29}{155} = 0.1871$. In firm B, 4.167% of the minority applicants and 17.308% of the majority applicants were hired. The difference is 13.141% and the odds ratio $\frac{43}{207} = 0.2077$. On both accounts, firm A seems to discriminate more.

In order to decide which statistic is more relevant we need to know the purpose of the comparison. The *difference* is more relevant if one wants to assess the macroeconomic implications of discrimination. The *odds ratio* is more relevant if one wants to know the impact of discrimination on one individual. □

• b. *1 point Government agencies enforcing discrimination laws traditionally have been using the* selection ratio p_1/p_2. *Compute the selection ratio for both firms.*

ANSWER. In firm A, the selection ratio is $\frac{1}{32}\frac{68}{10} = \frac{17}{80} = 0.2125$. In firm B, it is $\frac{13}{54} = 0.2407$. □

• c. *3 points Statisticians argue that the selection ratio is a flawed measure of discrimination, see* [**Gas88**, pp. 207–11 of vol. 1]. *Demonstrate this by comparing firm A with firm C which hires 5 out of 32 black and 40 out of 68 white applicants.*

Hirings by Two Different Firms with 100 Applications Each				
	Firm A		Firm C	
	Minority	Majority	Minority	Majority
Hired	1	10	5	40
Not Hired	31	58	27	28

TABLE 2. Selection Ratio gives Conflicting Verdicts

ANSWER. In Firm C the selection ratio is $\frac{5}{32}\frac{68}{40} = \frac{17}{64} = 0.265625$. In firm A, the chances for blacks to be hired is 24% that of whites, and in firm C it is 26%. Firm C seems better. But if we compare the chances *not* to get hired we get a conflicting verdict: In firm A the ratio is $\frac{31}{32}\frac{68}{58} = 1.1357$. In firm C it is $\frac{27}{32}\frac{68}{28} = 2.0491$. In firm C, the chances not to get hired is twice as high for Minorities as it is for Whites, in firm A the chances not to get hired are more equal. Here A seems better.

This illustrates an important drawback of the selection ratio: if we compare the chances of *not* being hired instead of those of being hired, we get $(1-p_1)/(1-p_2)$ instead of p_1/p_2. There is no simple relationship between these two numbers, indeed $(1-p_1)/(1-p_2)$ is not a function of p_1/p_2, although both ratios should express the same concept. This is why one can get conflicting

information if one looks at the selection ratio for a certain event or the selection ratio for its complement.

The odds ratio and the differences in probabilities do not give rise to such discrepancies: the odds ratio for not being hired is just the inverse of the odds ratio for being hired, and the difference in the probabilities of not being hired is the negative of the difference in the probabilities of being hired.

As long as p_1 and p_2 are both close to zero, the odds ratio is approximately equal to the selection ratio, therefore in this case the selection ratio is acceptable despite the above criticism. □

• d. *3 points Argue whether Fisher's exact test, which is a conditional test, is appropriate in this example.*

ANSWER. The firms do not have control over the number of job applications, and they also do not have control over how many job openings they have. Here is a situation in which Fisher's exact test, which is conditional on the row sums and column sums of the table, is entirely appropriate. Note that this criterion has nothing to do with sample size. □

• e. *4 points Compute the significance levels for rejecting the null hypothesis of equal treatment with the one-sided alternative of discrimination for each firm using Fisher's exact test. You will get a counterintuitive result. How can you explain this result?*

ANSWER. The R-commands can be run as `ecmet.script(hiring)`. Although firm A hired a lower percentage of applicants than firm B, the significance level for discrimination on Fisher's exact test is 0.07652 for firm A and 0.03509 for firm B. I.e., in a court of law, firm B might be convicted of discrimination, but firm A, which hired a lower percentage of its minority applicants, could not.

[**Spr98**, p. 377] explains this as follows: "the smaller number of minority hirings reduces the power of Fisher's exact test applied to firm A relative to the power where there is a surplus of minority hirings (firm B). This extra power is enough to produce a significant result despite the higher percentage of promotions among minority hirings (or the higher odds ratio if one makes the comparison on that basis)." □

Promotions by Two Different Firms with 100 Employees Each				
	Firm A		*Firm B*	
	Minority	*Majority*	*Minority*	*Majority*
Promoted	*1*	*10*	*2*	*9*
Not Promoted	*31*	*58*	*46*	*43*

TABLE 3. Which Firm's Promotion Policies are More Equitable?

• f. *5 points Now let's change the example. Table 3 has the same numbers as Table 1, but now these numbers do not count hirings but promotions from the pool of existing employees, and instead of the number of applicants, the column totals are*

the total numbers of employees of each firm. Let us first look at the overall race composition of the employees in each firm. Let us assume that 40% of the population are minorities, and 32 of the 100 employees of firm A are minorities, and 48 of the 100 employees of firm B are minorities. Is there significant evidence that the firms discriminated in hiring*?*

Answer. Assuming that the population is infinite, the question is: if one makes 100 independent random drawings from a population that contains 40% minorities, what is the probability to end up with 32 or less minorities? The R-command is `pbinom(q=32,size=100,prob=0.4)` which is 0.06150391. The other firm has more than 40 black employees; here one might wonder if there is evidence of discrimination against whites. `pbinom(q=48,size=100,prob=0.4)` gives $0.9576986 = 1 - 0.0423$, i.e., it is significant at the 5% level. But here we should apply a two-sided test. A one-sided test about discrimination against Blacks can be justified by the assumption "if there is discrimination at all, it is against blacks." This assumption cannot be made in the case of discrimination against Whites. We have to allow for the possibility of discrimination against Minorities *and* against Whites; therefore the critical value is at probability 0.975, and the observed result is not significant.

□

• g. *2 points You want to use Table 3 to investigate whether the firms discriminated in* promotion*, and you are considering Fisher's exact test. Do the arguments made above with respect to Fisher's exact still apply?*

Answer. No. A conditional test is no longer appropriate here because the proportion of candidates for promotion is under control of the firms. Firm A not only promoted a smaller percentage of their minority employees, but it also hired fewer minority workers in the first place. These two acts should be considered together to gauge the discrimination policies. The above Sprent-quote [**Spr98**, p. 377] continues: "There is a timely warning here about the need for care when using conditional tests when the marginal totals used for conditioning may themselves be conditional upon a further factor, in this case hiring policy."

□

15.5.1. Cornfield's Lemma. In a court, businesses which hire fewer blacks than whites are asked to explain how they obtained this outcome by nondiscriminatory actions. It is illegal to deny blacks a job because they are black, but it is legal to deny blacks a job because they have less job experience or lack other qualifications. The possibility whether some factor x that is relevant for the hiring decision and that is distributed unevenly between minority and majority applicants could have explained the observed disparity is assessed by a variant of Cornfield's lemma proved in [**Gas88**, p. 296 in vol. 1].

We need the following definitions:

p_1	chance for a minority applicant without x to get hired
$p_1 r_{\mathsf{x}}$	chance for a minority applicant with x to get hired
p_2	chance for a majority applicant without x to get hired
$p_2 r_{\mathsf{x}}$	chance for a majority applicant with x to get hired
f_1	proportion of minority applicants who have x
f_2	proportion of majority applicants who have x

The probability that a majority group member is hired is $p_2 r_{\mathsf{x}} f_2 + p_2(1 - f_2)$. The probability that a minority group member is hired is $p_1 r_{\mathsf{x}} f_1 + p_1(1 - f_1)$. The ratio between those probabilities, i.e., the relative advantage of the majority over the minority group members, is

$$r_m = \frac{p_2 r_{\mathsf{x}} f_2 + p_2(1 - f_2)}{p_1 r_{\mathsf{x}} f_1 + p_1(1 - f_1)} = \frac{p_2}{p_1}\frac{(r_{\mathsf{x}} - 1)f_2 + 1}{(r_{\mathsf{x}} - 1)f_1 + 1} \tag{15.5.7}$$

There is no discrimination, i.e., r_m is fully explained by the fact that more whites have x than blacks, if $p_1 \geq p_2$ and therefore

$$r_m \leq \frac{(r_{\mathsf{x}} - 1)f_2 + 1}{(r_{\mathsf{x}} - 1)f_1 + 1} \tag{15.5.8}$$

i.e.,

$$r_m(r_{\mathsf{x}} - 1)f_1 + r_m \leq (r_{\mathsf{x}} - 1)f_2 + 1 \tag{15.5.9}$$

which is equivalent to

$$r_m f_1 + \frac{r_m - 1}{r_{\mathsf{x}} - 1} \leq f_2 \tag{15.5.10}$$

Problem 223. *Suppose that 60% of whites are hired, while only 40% of a minority group are hired. Suppose that a certain type of training or education was related to the job in question, and it is believed that at least 10% of the minority group had this training.*

• a. *3 points Assuming that persons with this training had twice the chance of getting the job, which percentage of whites would have had this qualification in order to explain the disparity in the hiring rates?*

Answer. Since 60% of whites are hired and 40% of the minority group, $r_m = 60/40 = 1.5$. Training is the factor x. Sind persons with training had twice the chance of getting the job, $r_{\mathsf{x}} = 2$. Since 10% of the minority group had this training, $f_1 = 0.1$. Therefore (15.5.10) implies that at least $1.5 \cdot 0.1 + \frac{0.5}{1} = 65\%$ of whites had to have this qualification in order to explain the observed disparity in hiring rates. □

• b. *1 point What would this percentage have to be if training tripled (instead of doubling) one's chances of getting the job?*

Answer. If training tripled one's chances of being hired, then the training would explain the disparity if $1.5 \cdot 0.1 + \frac{0.5}{2} = 40\%$ or more of whites had this training. □

Cornfield's lemma is an example of a retroductive argument, i.e., of inference from an observed outcome (disparity in hiring rates) to unobserved causes which might have generated this outcome (disparate distribution of the factor $\mathfrak{x}$ between majority and minority).

15.6. The Wald, Likelihood Ratio, and Lagrange Multiplier Tests

Let us start with the generalized Wald test. Assume $\tilde{\theta}$ is an asymptotically normal estimator of $\boldsymbol{\theta}$, whose asymptotic distribution is $\mathrm{N}(\boldsymbol{\theta}, \boldsymbol{\Psi})$. Assume furthermore that $\hat{\boldsymbol{\Psi}}$ is a consistent estimate of $\boldsymbol{\Psi}$. Then the following statistic is called the generalized Wald statistic. It can be used for an asymtotic test of the hypothesis $\boldsymbol{h}(\boldsymbol{\theta}) = \boldsymbol{o}$, where $\boldsymbol{h}$ is a q-vector-valued differentiable function:

$$\text{G.W.} = \boldsymbol{h}(\tilde{\theta})^\top \left\{ \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{\theta}^\top}\Big|_{\tilde{\theta}} \hat{\boldsymbol{\Psi}} \frac{\partial \boldsymbol{h}^\top}{\partial \boldsymbol{\theta}}\Big|_{\tilde{\theta}} \right\}^{-1} \boldsymbol{h}(\tilde{\theta}) \tag{15.6.1}$$

Under the null hypothesis, this test statistic is asymptotically distributed as a χ^2_q. To understand this, note that for all θ close to $\tilde{\boldsymbol{\theta}}$, $\boldsymbol{h}(\theta) \asymp \boldsymbol{h}(\tilde{\boldsymbol{\theta}}) + \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{\theta}^\top}\Big|_{\tilde{\boldsymbol{\theta}}} (\theta - \tilde{\boldsymbol{\theta}})$. Taking covariances

$$\frac{\partial \boldsymbol{h}}{\partial \boldsymbol{\theta}^\top}\Big|_{\tilde{\theta}} \hat{\boldsymbol{\Psi}} \frac{\partial \boldsymbol{h}^\top}{\partial \boldsymbol{\theta}}\Big|_{\tilde{\theta}} \tag{15.6.2}$$

is an estimate of the covariance matrix of $\boldsymbol{h}(\tilde{\theta})$. I.e., one takes $\boldsymbol{h}(\tilde{\theta})$ twice and "divides" it by its covariance matrix.

Now let us make more stringent assumptions. Assume the density $f_{\mathfrak{x}}(\boldsymbol{x}; \boldsymbol{\theta})$ of $\mathfrak{x}$ depends on the parameter vector $\boldsymbol{\theta}$. We are assuming that the conditions are satisfied which ensure asymptotic normality of the maximum likelihood estimator $\hat{\theta}$ and also of $\bar{\theta}$, the constrained maximum likelihood estimator subject to the constraint $\boldsymbol{h}(\theta) = \boldsymbol{o}$.

There are three famous tests to test this hypothesis, which asymptotically are all distributed like χ^2_q. The likehihood-ratio test is

$$LRT = -2 \log \frac{\max_{\boldsymbol{h}(\boldsymbol{\theta})=\boldsymbol{o}} f_{\mathfrak{y}}(\mathfrak{y}; \boldsymbol{\theta})}{\max_{\boldsymbol{\theta}} f_{\mathfrak{y}}(\mathfrak{y}; \boldsymbol{\theta})} = 2(\log f_{\mathfrak{y}}(\mathfrak{y}, \hat{\theta}) - \log f_{\mathfrak{y}}(\mathfrak{y}, \bar{\theta})) \tag{15.6.3}$$

It rejects if imposing the constraint reduces the attained level of the likelihood function too much.

The Wald test has the form

$$\text{Wald} = -\boldsymbol{h}(\hat{\theta})^\top \left\{ \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{\theta}^\top}\Big|_{\hat{\theta}} \left(\frac{\partial^2 \log f(\mathfrak{y}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}\Big|_{\hat{\theta}} \right)^{-1} \frac{\partial \boldsymbol{h}^\top}{\partial \boldsymbol{\theta}}\Big|_{\hat{\theta}} \right\}^{-1} \boldsymbol{h}(\hat{\theta}) \tag{15.6.4}$$

To understand this formula, note that $-\left(\mathcal{E}\left[\frac{\partial^2 \log f(\mathfrak{y}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}\right]\right)^{-1}$ is the Cramer Rao lower bound, and since all maximum likelihood estimators asymptotically attain the CRLB, it is the asymptotic covariance matrix of $\hat{\theta}$. If one does not take the expected value but plugs $\hat{\theta}$ into these partial derivatives of the log likelihood function, one gets a consistent estimate of the asymtotic covariance matrix. Therefore the Wald test is a special case of the generalized Wald test.

Finally the score test has the form

$$\text{Score} = -\frac{\partial \log f(\mathfrak{y}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}\Big|_{\bar{\theta}} \left(\frac{\partial^2 \log f(\mathfrak{y}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}\Big|_{\bar{\theta}} \right)^{-1} \frac{\partial \log f(\mathfrak{y}; \boldsymbol{\theta})^\top}{\partial \boldsymbol{\theta}}\Big|_{\bar{\theta}} \tag{15.6.5}$$

This test tests whether the score, i.e., the gradient of the unconstrained log likelihood function, evaluated at the constrained maximum likelihood estimator, is too far away from zero. To understand this formula, remember that we showed in the proof of the Cramer-Rao lower bound that the negative of the expected value of the Hessian $-\mathcal{E}\left[\frac{\partial^2 \log f(\mathfrak{y}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}\right]$ is the covariance matrix of the score, i.e., here we take the score twice and divide it by its estimated covariance matrix.

CHAPTER 16

General Principles of Econometric Modelling

[**Gre97**, 6.1 on p. 220] says: "An econometric study begins with a set of propositions about some aspect of the economy. The theory specifies a set of precise, deterministic relationships among variables. Familiar examples are demand equations, production functions, and macroeconomic models. The empirical investigation provides estimates of unknown parameters in the model, such as elasticities or the marginal propensity to consume, and usually attempts to measure the validity of the theory against the behavior of the observable data."

[**Hen95**, p. 6] distinguishes between two extremes: "'Theory-driven' approaches, in which the model is derived from a priori theory and calibrated from data evidence. They suffer from theory dependence in that their credibility depends on the credibility of the theory from which they arose—when that theory is discarded, so is the associated evidence." The other extreme is "'Data-driven' approaches, where models are developed to closely describe the data . . . These suffer from sample dependence in that accidental and transient data features are embodied as tightly in the model as permanent aspects, so that extension of the data set often reveal predictive failure."

Hendry proposes the following useful distinction of 4 levels of knowledge:

A Consider the situation where we know the complete structure of the process which gernerates economic data and the values of all its parameters. This is the equivalent of a probability theory course (example: rolling a perfect die), but involves economic theory and econometric concepts.

B consider a known economic structure with unknown values of the parameters. Equivalent to an estimation and inference course in statistics (example: independent rolls of an imperfect die and estimating the probabilities of the different faces) but focusing on econometrically relevant aspects.

C is "the empirically relevant situation where neither the form of the data-generating process nor its parameter values are known. (Here one does not know whether the rolls of the die are independent, or whether the probabilities of the different faces remain constant.) Model discovery, evaluation, data mining, model-search procedures, and associated methodological issues.

D Forecasting the future when the data outcomes are unknown. (Model of money demand under financial innovation).

The example of Keynes's consumption function in [**Gre97**, pp. 221/22] sounds at the beginning as if it was close to *B*, but in the further discussion Greene goes more and more over to *C*. It is remarkable here that economic theory usually does not yield functional forms. Greene then says: the most common functional form is the linear one $\mathsf{c} = \alpha + \beta\mathsf{x}$ with $\alpha > 0$ and $0 < \beta < 1$. He does not mention the aggregation problem hidden in this. Then he says: "But the linear function is only approximate; in fact, it is unlikely that consumption and income can be connected by any simple relationship. The **deterministic** relationship is clearly inadequate." Here Greene uses a random relationship to model a relationship which is quantitatively "fuzzy." This is an interesting and relevant application of randomness.

A sentence later Green backtracks from this insight and says: "We are not so ambitious as to attempt to capture every influence in the relationship, but only those that are substantial enough to model directly." The "fuzziness" is not due to a lack of ambition of the researcher, but the world is inherently quantiatively fuzzy. It is not that we don't know the law, but there is no law; not everything that happens in an economy is driven by economic laws. Greene's own example, in Figure 6.2, that during the war years consumption was below the trend line, shows this.

Greene's next example is the relationship between income and education. This illustrates multiple instead of simple regression: one must also include age, and then also the square of age, even if one is not interested in the effect which age has, but in order to "control" for this effect, so that the effects of education and age will not be confounded.

Problem 224. *Why should a regression of income on education include not only age but also the square of age?*

Answer. Because the effect of age becomes smaller with increases in age. □

Critical Realist approaches are [**Ron02**] and [**Mor02**].

CHAPTER 17

Causality and Inference

This chapter establishes the connection between critical realism and Holland and Rubin's modelling of causality in statistics as explained in [**Hol86**] and [**WM83**, pp. 3–25] (and the related paper [**LN81**] which comes from a Bayesian point of view). A different approach to causality and inference, [**Roy97**], is discussed in chapter/section 2.8. Regarding critical realism and econometrics, also [**Dow99**] should be mentioned: this is written by a Post Keynesian econometrician working in an explicitly realist framework.

Everyone knows that correlation does not mean causality. Nevertheless, experience shows that statisticians can on occasion make valid inferences about causality. It is therefore legitimate to ask: how and under which conditions can causal conclusions be drawn from a statistical experiment or a statistical investigation of nonexperimental data?

Holland starts his discussion with a description of the "logic of association" (= a flat empirical realism) as opposed to causality (= depth realism). His model for the "logic of association" is essentially the conventional mathematical model of probability by a set U of "all possible outcomes," which we described and criticized on p. 6 above.

After this, Rubin describes his own model (developed together with Holland). Rubin introduces "counterfactual" (or, as Bhaskar would say, "transfactual") elements since he is not only talking about the value a variable takes for a given individual, but also the value this variable would have taken for the same individual if the causing variables (which Rubin also calls "treatments") had been different. For simplicity, Holland assumes here that the treatment variable has only two levels: either the individual receives the treatment, or he/she does not (in which case he/she belongs to the "control" group). The correlational view would simply measure the average response of those individuals who receive the treatment, and of those who don't. Rubin recognizes in his model that the same individual may or may not be subject to the treatment, therefore the response variable has two values, one being the individual's response if he or she receives the treatment, the other the response if he or she does not.

A third variable indicates who receives the treatment. I.e, he has the "causal indicator" s which can take two values, t (treatment) and c (control), and two variables y_t and y_c, which, evaluated at individual ω, indicate the responses this individual would give in case he was subject to the treatment, and in case he was or not.

Rubin defines $\mathsf{y}_t - \mathsf{y}_c$ to be the causal effect of treatment t versus the control c. But this causal effect cannot be observed. We cannot observe how those indiviuals who received the treatement would have responded if they had not received the treatment, despite the fact that this non-actualized response is just as real as the response which they indeed gave. This is what Holland calls the *Fundamental Problem of Causal Inference*.

Problem 225. *Rubin excludes race as a cause because the individual cannot do anything about his or her race. Is this argument justified?*

Does this Fundamental Problem mean that causal inference is impossible? Here are several scenarios in which causal inference is possible after all:

- Temporal stability of the response, and transience of the causal effect.
- Unit homogeneity.
- Constant effect, i.e., $y_t(\omega) - y_c(\omega)$ is the same for all ω.
- Independence of the response with respect to the selection process regarding who gets the treatment.

For an example of this last case, say

Problem 226. *Our universal set* U *consists of patients who have a certain disease. We will explore the causal effect of a given treatment with the help of three events,* T, C, *and* S, *the first two of which are* counterfactual, *compare* [**Hol86**]. *These events are defined as follows:* T *consists of all patients who would recover if given treatment;* C *consists of all patients who would recover if not given treatment (i.e., if included in the control group). The event* S *consists of all patients actually receiving treatment. The average causal effect of the treatment is defined as* $\Pr[\mathsf{T}] - \Pr[\mathsf{C}]$.

- a. *2 points Show that*

$$\Pr[\mathsf{T}] = \Pr[\mathsf{T}|S]\Pr[\mathsf{S}] + \Pr[\mathsf{T}|S'](1 - \Pr[\mathsf{S}]) \tag{17.0.6}$$

and that

$$\Pr[\mathsf{C}] = \Pr[\mathsf{C}|S]\Pr[\mathsf{S}] + \Pr[\mathsf{C}|S'](1 - \Pr[\mathsf{S}]) \tag{17.0.7}$$

Which of these probabilities can be estimated as the frequencies of observable outcomes and which cannot?

Answer. This is a direct application of (2.7.9). The problem here is that for all $\omega \in \mathsf{C}$, i.e., for those patients who do not receive treatment, we do not know whether they would have recovered

if given treatment, and for all $\omega \in \mathsf{T}$, i.e., for those patients who do receive treatment, we do not know whether they would have recovered if not given treatment. In other words, neither $\Pr[\mathsf{T}|S]$ nor $\mathrm{E}[\mathsf{C}|S']$ can be estimated as the frequencies of observable outcomes. □

• b. *2 points Assume now that* S *is independent of* T *and* C*, because the subjects are assigned randomly to treatment or control. How can this be used to estimate those elements in the equations* (17.0.6) *and* (17.0.7) *which could not be estimated before?*

Answer. In this case, $\Pr[\mathsf{T}|S] = \Pr[\mathsf{T}|S']$ and $\Pr[\mathsf{C}|S'] = \Pr[\mathsf{C}|S]$. Therefore, the average causal effect can be simplified as follows:

$$\begin{aligned}\Pr[\mathsf{T}] - \Pr[\mathsf{C}] &= \Pr[\mathsf{T}|S]\Pr[S] + \Pr[\mathsf{T}|S'](1-\Pr[S]) - \Pr[\mathsf{C}|S]\Pr[S] + \Pr[\mathsf{C}|S'](1-\Pr[S]) \\ &= \Pr[\mathsf{T}|S]\Pr[S] + \Pr[\mathsf{T}|S](1-\Pr[S]) - \Pr[\mathsf{C}|S']\Pr[S] + \Pr[\mathsf{C}|S'](1-\Pr[S]) \\ &= \Pr[\mathsf{T}|S] - \Pr[\mathsf{C}|S'] \qquad (17.0.8)\end{aligned}$$

□

• c. *2 points Why were all these calculations necessary? Could one not have defined from the beginning that the causal effect of the treatment is* $\Pr[\mathsf{T}|S] - \Pr[\mathsf{C}|S']$*?*

Answer. $\Pr[\mathsf{T}|S] - \Pr[\mathsf{C}|S']$ is only the empirical difference in recovery frequencies between those who receive treatment and those who do not. It is always possible to measure these differences, but these differences are not necessarily due to the treatment but may be due to other reasons. □

The main message of the paper is therefore: before drawing causal conclusions one should acertain whether one of these conditions apply which make causal conclusions possible.

In the rest of the paper, Holland compares his approach with other approaches. Suppes's definitions of causality are interesting:

- If $r < s$ denote two time values, event C_r is a *prima facie cause* of E_s iff $\Pr[E_s|C_r] > \Pr[E_s]$.
- C_r is a *spurious cause* of E_s iff it is a prima facie cause of E_s and for some $q < r < s$ there is an event D_q so that $\Pr[E_s|C_r, D_q] = \Pr[E_s|D_q]$ and $\Pr[E_s|C_r, D_q] \geq \Pr[E_s|C_r]$.
- Event C_r is a *genuine cause* of E_s iff it is a prima facie but not a spurious cause.

This is quite different than Rubin's analysis. Suppes concentrates on the causes of a given effect, not the effects of a given cause. Suppes has a Popperian falsificationist view: a hypothesis is good if one cannot falsify it, while Holland has the depth-realist view which says that the empirical is only a small part of reality, and which looks at the underlying mechanisms.

Problem 227. *Construct an example of a probability field with a spurious cause.*

Granger causality (see chapter/section 67.2.1) is based on the idea: knowing a cause ought to improve our ability to predict. It is more appropriate to speak here of "noncausality" instead of causality: a variable does *not* cause another if knowing that variable does *not* improve our ability to predict the other variable. Granger formulates his theory in terms of a specific predictor, the BLUP, while Holland extends it to all predictors. Granger works on it in a time series framework, while Holland gives a more general formulation. Holland's formulation strips off the unnecessary detail in order to get at the essence of things. Holland defines: x is not a Granger cause of y relative to the information in $\boldsymbol{z}$ (which in the timeseries context contains the past values of y) if and only if x and y are conditionally independent given $\boldsymbol{z}$. Problem 40 explains why this can be tested by testing predictive power.

CHAPTER 18

Mean-Variance Analysis in the Linear Model

In the present chapter, the only distributional assumptions are that means and variances exist. (From this follows that also the covariances exist).

18.1. Three Versions of the Linear Model

As background reading please read [**CD97**, Chapter 1].

Following [**JHG+88**, Chapter 5], we will start with three different linear statistical models. Model 1 is the simplest estimation problem already familiar from chapter 12, with n independent observations from the same distribution, call them $\mathsf{y}_1, \ldots, \mathsf{y}_n$. The only thing known about the distribution is that mean and variance exist, call them μ and σ^2. In order to write this as a special case of the "linear model," define $\varepsilon_i = \mathsf{y}_i - \mu$, and define the vectors $\boldsymbol{\mathsf{y}} = \begin{bmatrix}\mathsf{y}_1 & \mathsf{y}_2 & \cdots & \mathsf{y}_n\end{bmatrix}^\top$, $\boldsymbol{\varepsilon} = \begin{bmatrix}\varepsilon_1 & \varepsilon_2 & \cdots & \varepsilon_n\end{bmatrix}^\top$, and $\boldsymbol{\iota} = \begin{bmatrix}1 & 1 & \cdots & 1\end{bmatrix}^\top$. Then one can write the model in the form

$$\boldsymbol{\mathsf{y}} = \boldsymbol{\iota}\mu + \boldsymbol{\varepsilon} \qquad \boldsymbol{\varepsilon} \sim (\boldsymbol{o}, \sigma^2\boldsymbol{I}) \tag{18.1.1}$$

The notation $\boldsymbol{\varepsilon} \sim (\boldsymbol{o}, \sigma^2\boldsymbol{I})$ is shorthand for $\mathcal{E}[\boldsymbol{\varepsilon}] = \boldsymbol{o}$ (the null vector) and $\mathcal{V}[\boldsymbol{\varepsilon}] = \sigma^2\boldsymbol{I}$ (σ^2 times the identity matrix, which has 1's in the diagonal and 0's elsewhere). μ is the deterministic part of all the y_i, and ε_i is the random part.

Model 2 is "simple regression" in which the deterministic part μ is not constant but is a function of the nonrandom variable x. The assumption here is that this function is differentiable and can, in the range of the variation of the data, be approximated by a linear function [**Tin51**, pp. 19–20]. I.e., each element of $\boldsymbol{\mathsf{y}}$ is a constant α plus a constant multiple of the corresponding element of the nonrandom vector $\boldsymbol{x}$ plus a random error term: $\mathsf{y}_t = \alpha + x_t\beta + \varepsilon_t$, $t = 1, \ldots, n$. This can be written as

$$\begin{bmatrix}\mathsf{y}_1\\ \vdots \\ \mathsf{y}_n\end{bmatrix} = \begin{bmatrix}1\\ \vdots \\ 1\end{bmatrix}\alpha + \begin{bmatrix}x_1\\ \vdots \\ x_n\end{bmatrix}\beta + \begin{bmatrix}\varepsilon_1\\ \vdots \\ \varepsilon_n\end{bmatrix} = \begin{bmatrix}1 & x_1\\ \vdots & \vdots \\ 1 & x_n\end{bmatrix}\begin{bmatrix}\alpha\\ \beta\end{bmatrix} + \begin{bmatrix}\varepsilon_1\\ \vdots \\ \varepsilon_n\end{bmatrix} \tag{18.1.2}$$

or

$$\boldsymbol{\mathsf{y}} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \boldsymbol{\varepsilon} \sim (\boldsymbol{o}, \sigma^2\boldsymbol{I}) \tag{18.1.3}$$

Problem 228. *1 point Compute the matrix product*

$$\begin{bmatrix}1 & 2 & 5\\ 0 & 3 & 1\end{bmatrix}\begin{bmatrix}4 & 0\\ 2 & 1\\ 3 & 8\end{bmatrix}$$

Answer.

$$\begin{bmatrix}1 & 2 & 5\\ 0 & 3 & 1\end{bmatrix}\begin{bmatrix}4 & 0\\ 2 & 1\\ 3 & 8\end{bmatrix} = \begin{bmatrix}1\cdot 4 + 2\cdot 2 + 5\cdot 3 & 1\cdot 0 + 2\cdot 1 + 5\cdot 8\\ 0\cdot 4 + 3\cdot 2 + 1\cdot 3 & 0\cdot 0 + 3\cdot 1 + 1\cdot 8\end{bmatrix} = \begin{bmatrix}23 & 42\\ 9 & 11\end{bmatrix}$$

□

If the systematic part of y depends on more than one variable, then one needs multiple regression, model 3. Mathematically, multiple regression has the same form (18.1.3), but this time $\boldsymbol{X}$ is *arbitrary* (except for the restriction that all its columns are linearly independent). Model 3 has Models 1 and 2 as special cases.

Multiple regression is also used to "correct for" disturbing influences. Let me explain. A functional relationship, which makes the systematic part of y dependent on some other variable x will usually only hold if other relevant influences are kept constant. If those other influences vary, then they may affect the form of this functional relation. For instance, the marginal propensity to consume may be affected by the interest rate, or the unemployment rate. This is why some econometricians (Hendry) advocate that one should start with an "encompassing" model with many explanatory variables and then narrow the specification down by hypothesis tests. Milton Friedman, by contrast, is very suspicious about multiple regressions, and argues in [**FS91**, pp. 48/9] against the encompassing approach.

Friedman does not give a theoretical argument but argues by an example from Chemistry. Perhaps one can say that the variations in the other influences may have more serious implications than just modifying the form of the functional relation: they may destroy this functional relation altogether, i.e., prevent any systematic or predictable behavior.

	observed	unobserved
random	$\boldsymbol{\mathsf{y}}$	$\boldsymbol{\varepsilon}$
nonrandom	$\boldsymbol{X}$	$\boldsymbol{\beta}$, σ^2

18.2. Ordinary Least Squares

In the model $\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim (\boldsymbol{o}, \sigma^2\boldsymbol{I})$, the OLS-estimate $\hat{\boldsymbol{\beta}}$ is defined to be that value $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ which minimizes

$$\mathsf{SSE} = (\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}) = \mathbf{y}^\top\mathbf{y} - 2\mathbf{y}^\top\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta}. \tag{18.2.1}$$

Problem 184 shows that in model 1, this principle yields the arithmetic mean.

Problem 229. *2 points Prove that, if one predicts a random variable* y *by a constant* a*, the constant which gives the best* MSE *is* $a = \mathrm{E}[\mathsf{y}]$*, and the best* MSE *one can get is* $\mathrm{var}[\mathsf{y}]$.

Answer. $\mathrm{E}[(\mathsf{y} - a)^2] = \mathrm{E}[\mathsf{y}^2] - 2a\,\mathrm{E}[\mathsf{y}] + a^2$. Differentiate with respect to a and set zero to get $a = \mathrm{E}[\mathsf{y}]$. One can also differentiate first and then take expected value: $\mathrm{E}[2(\mathsf{y} - a)] = 0$. □

We will solve this minimization problem using the first-order conditions in vector notation. As a preparation, you should read the beginning of Appendix C about matrix differentiation and the connection between matrix differentiation and the Jacobian matrix of a vector function. All you need at this point is the two equations (C.1.6) and (C.1.7). The chain rule (C.1.23) is enlightening but not strictly necessary for the present derivation.

The matrix differentiation rules (C.1.6) and (C.1.7) allow us to differentiate (18.2.1) to get

$$\partial\mathsf{SSE}/\partial\boldsymbol{\beta}^\top = -2\mathbf{y}^\top\boldsymbol{X} + 2\boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{X}. \tag{18.2.2}$$

Transpose it (because it is notationally simpler to have a relationship between column vectors), set it zero while at the same time replacing $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}$, and divide by 2, to get the "normal equation"

$$\boldsymbol{X}^\top\mathbf{y} = \boldsymbol{X}^\top\boldsymbol{X}\hat{\boldsymbol{\beta}}. \tag{18.2.3}$$

Due to our assumption that all columns of $\boldsymbol{X}$ are linearly independent, $\boldsymbol{X}^\top\boldsymbol{X}$ has an inverse and one can premultiply both sides of (18.2.3) by $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\mathbf{y}. \tag{18.2.4}$$

If the columns of $\boldsymbol{X}$ are not linearly independent, then (18.2.3) has more than one solution, and the normal equation is also in this case a necessary and sufficient condition for $\hat{\boldsymbol{\beta}}$ to minimize the SSE (proof in Problem 232).

Problem 230. *4 points Using the matrix differentiation rules*

$$\partial\boldsymbol{w}^\top\boldsymbol{x}/\partial\boldsymbol{x}^\top = \boldsymbol{w}^\top \tag{18.2.5}$$

$$\partial\boldsymbol{x}^\top\boldsymbol{M}\boldsymbol{x}/\partial\boldsymbol{x}^\top = 2\boldsymbol{x}^\top\boldsymbol{M} \tag{18.2.6}$$

for symmetric $\boldsymbol{M}$*, compute the least-squares estimate* $\hat{\boldsymbol{\beta}}$ *which minimizes*

$$\mathsf{SSE} = (\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}) \tag{18.2.7}$$

You are allowed to assume that $\boldsymbol{X}^\top\boldsymbol{X}$ *has an inverse.*

Answer. First you have to multiply out

$$(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}) = \mathbf{y}^\top\mathbf{y} - 2\mathbf{y}^\top\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta}. \tag{18.2.8}$$

The matrix differentiation rules (18.2.5) and (18.2.6) allow us to differentiate (18.2.8) to get

$$\partial\mathsf{SSE}/\partial\boldsymbol{\beta}^\top = -2\mathbf{y}^\top\boldsymbol{X} + 2\boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{X}. \tag{18.2.9}$$

Transpose it (because it is notationally simpler to have a relationship between column vectors), set it zero while at the same time replacing $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}$, and divide by 2, to get the "normal equation"

$$\boldsymbol{X}^\top\mathbf{y} = \boldsymbol{X}^\top\boldsymbol{X}\hat{\boldsymbol{\beta}}. \tag{18.2.10}$$

Since $\boldsymbol{X}^\top\boldsymbol{X}$ has an inverse, one can premultiply both sides of (18.2.10) by $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\mathbf{y}. \tag{18.2.11}$$

□

Problem 231. *2 points Show the following: if the columns of* $\boldsymbol{X}$ *are linearly independent, then* $\boldsymbol{X}^\top\boldsymbol{X}$ *has an inverse. (*$\boldsymbol{X}$ *itself is not necessarily square.) In your proof you may use the following criteria: the columns of* $\boldsymbol{X}$ *are linearly independent (this is also called:* $\boldsymbol{X}$ *has full column rank) if and only if* $\boldsymbol{X}\boldsymbol{a} = \boldsymbol{o}$ *implies* $\boldsymbol{a} = \boldsymbol{o}$*. And a square matrix has an inverse if and only if its columns are linearly independent.*

Answer. We have to show that any $\boldsymbol{a}$ which satisfies $\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{a} = \boldsymbol{o}$ is itself the null vector. From $\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{a} = \boldsymbol{o}$ follows $\boldsymbol{a}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{a} = 0$ which can also be written $\|\boldsymbol{X}\boldsymbol{a}\|^2 = 0$. Therefore $\boldsymbol{X}\boldsymbol{a} = \boldsymbol{o}$, and since the columns of $\boldsymbol{X}$ are linearly independent, this implies $\boldsymbol{a} = \boldsymbol{o}$. □

Problem 232. *3 points In this Problem we do not assume that* $\boldsymbol{X}$ *has full column rank, it may be arbitrary.*

• a. *The normal equation* (18.2.3) *has always at least one solution. Hint: you are allowed to use, without proof, equation* (A.3.3) *in the mathematical appendix.*

Answer. With this hint it is easy: $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^-\boldsymbol{X}^\top\mathbf{y}$ is a solution. □

• b. *If* $\hat{\boldsymbol{\beta}}$ *satisfies the normal equation and* $\boldsymbol{\beta}$ *is an arbitrary vector, then*

$$(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}) = (\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^\top\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}). \tag{18.2.12}$$

Answer. This is true even if $\boldsymbol{X}$ has deficient rank, and it will be shown here in this general case. To prove (18.2.12), write (18.2.1) as $\mathsf{SSE} = \big((\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) - \boldsymbol{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\big)^\top\big((\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) - \boldsymbol{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\big)$; since $\hat{\boldsymbol{\beta}}$ satisfies (18.2.3), the cross product terms disappear. □

• c. *Conclude from this that the normal equation is a necessary and sufficient condition characterizing the values* $\hat{\boldsymbol{\beta}}$ *minimizing the sum of squared errors* (18.2.12).

Answer. (18.2.12) shows that the normal equations are sufficient. For necessity of the normal equations let $\hat{\boldsymbol{\beta}}$ be an arbitrary solution of the normal equation, we have seen that there is always at least one. Given $\hat{\boldsymbol{\beta}}$, it follows from (18.2.12) that for any solution $\boldsymbol{\beta}^*$ of the minimization, $\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}) = \boldsymbol{o}$. Use (18.2.3) to replace $(\boldsymbol{X}^\top\boldsymbol{X})\hat{\boldsymbol{\beta}}$ by $\boldsymbol{X}^\top\mathbf{y}$ to get $\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta}^* = \boldsymbol{X}^\top\mathbf{y}$. □

It is customary to use the notation $\boldsymbol{X}\hat{\boldsymbol{\beta}} = \hat{\mathbf{y}}$ for the so-called *fitted values*, which are the estimates of the vector of means $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$. Geometrically, $\hat{\mathbf{y}}$ is the orthogonal projection of $\mathbf{y}$ on the space spanned by the columns of $\boldsymbol{X}$. See Theorem A.6.1 about projection matrices.

The vector of differences between the actual and the fitted values is called the vector of "residuals" $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \hat{\mathbf{y}}$. The residuals are "predictors" of the actual (but unobserved) values of the disturbance vector $\boldsymbol{\varepsilon}$. An estimator of a random magnitude is usually called a "predictor," but in the linear model estimation and prediction are treated on the same footing, therefore it is not necessary to distinguish between the two.

You should understand the difference between disturbances and residuals, and between the two decompositions

$$\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\varepsilon}} \tag{18.2.13}$$

Problem 233. *2 points Assume that $\boldsymbol{X}$ has full column rank. Show that $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{M}\mathbf{y}$ where $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$. Show that $\boldsymbol{M}$ is symmetric and idempotent.*

Answer. By definition, $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = \mathbf{y} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}\mathbf{y} = \big(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}\big)\mathbf{y}$. Idempotent, i.e. $\boldsymbol{M}\boldsymbol{M} = \boldsymbol{M}$:

$$\boldsymbol{M}\boldsymbol{M} = \big(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\big)\big(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\big) = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top + \tag{18.2.14}$$

□

Problem 234. *Assume $\boldsymbol{X}$ has full column rank. Define $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$.*

• a. *1 point Show that the space $\boldsymbol{M}$ projects on is the space orthogonal to all columns in $\boldsymbol{X}$, i.e., $\boldsymbol{M}\boldsymbol{q} = \boldsymbol{q}$ if and only if $\boldsymbol{X}^\top\boldsymbol{q} = \boldsymbol{o}$.*

Answer. $\boldsymbol{X}^\top\boldsymbol{q} = \boldsymbol{o}$ clearly implies $\boldsymbol{M}\boldsymbol{q} = \boldsymbol{q}$. Conversely, $\boldsymbol{M}\boldsymbol{q} = \boldsymbol{q}$ implies $\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{q} = \boldsymbol{o}$. Premultiply this by $\boldsymbol{X}^\top$ to get $\boldsymbol{X}^\top\boldsymbol{q} = \boldsymbol{o}$. □

• b. *1 point Show that a vector $\boldsymbol{q}$ lies in the range space of $\boldsymbol{X}$, i.e., the space spanned by the columns of $\boldsymbol{X}$, if and only if $\boldsymbol{M}\boldsymbol{q} = \boldsymbol{o}$. In other words, $\{\boldsymbol{q}\colon\ \boldsymbol{q} = \boldsymbol{X}\boldsymbol{a}$ for some $\boldsymbol{a}\} = \{\boldsymbol{q}\colon\ \boldsymbol{M}\boldsymbol{q} = \boldsymbol{o}\}$.*

Answer. First assume $\boldsymbol{M}\boldsymbol{q} = \boldsymbol{o}$. This means $\boldsymbol{q} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{q} = \boldsymbol{X}\boldsymbol{a}$ with $\boldsymbol{a} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{q}$. Conversely, if $\boldsymbol{q} = \boldsymbol{X}\boldsymbol{a}$ then $\boldsymbol{M}\boldsymbol{q} = \boldsymbol{M}\boldsymbol{X}\boldsymbol{a} = \boldsymbol{O}\boldsymbol{a} = \boldsymbol{o}$. □

Problem 235. *In 2-dimensional space, write down the projection matrix on the diagonal line $y = x$ (call it $\boldsymbol{E}$), and compute $\boldsymbol{E}\boldsymbol{z}$ for the three vectors $\boldsymbol{a} = \left[\begin{smallmatrix}2\\1\end{smallmatrix}\right]$, $\boldsymbol{b} = \left[\begin{smallmatrix}2\\2\end{smallmatrix}\right]$, and $\boldsymbol{c} = \left[\begin{smallmatrix}3\\2\end{smallmatrix}\right]$. Draw these vectors and their projections.*

Assume we have a dependent variable $\mathbf{y}$ and two regressors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, each with 15 observations. Then one can visualize the data either as 15 points in 3-dimensional space (a 3-dimensional scatter plot), or 3 points in 15-dimensional space. In the first case, each point corresponds to an observation, in the second case, each point corresponds to a variable. In this latter case the points are usually represented as vectors. You only have 3 vectors, but each of these vectors is a vector in 15-dimensional space. But you do not have to draw a 15-dimensional space to draw these vectors; these 3 vectors span a 3-dimensional subspace, and $\hat{\mathbf{y}}$ is the projection of the vector $\mathbf{y}$ on the space spanned by the two regressors not only in the original 15-dimensional space, but already in this 3-dimensional subspace. In other words, [**DM93**, Figure 1.3] is valid in all dimensions! In the 15-dimensional space, each dimension represents one observation. In the 3-dimensional subspace, this is no longer true.

Problem 236. *"Simple regression" is regression with an intercept and one explanatory variable only, i.e.,*

$$\mathbf{y}_t = \alpha + \beta x_t + \varepsilon_t \tag{18.2.15}$$

Here $\boldsymbol{X} = \begin{bmatrix}\boldsymbol{\iota} & \boldsymbol{x}\end{bmatrix}$ and $\boldsymbol{\beta} = \begin{bmatrix}\alpha & \beta\end{bmatrix}^\top$. Evaluate (18.2.4) *to get the following formulas for $\hat{\boldsymbol{\beta}} = \begin{bmatrix}\hat{\alpha} & \hat{\beta}\end{bmatrix}^\top$:*

$$\hat{\alpha} = \frac{\sum x_t^2 \sum \mathbf{y}_t - \sum x_t \sum x_t\mathbf{y}_t}{n\sum x_t^2 - (\sum x_t)^2} \tag{18.2.16}$$

$$\hat{\beta} = \frac{n\sum x_t\mathbf{y}_t - \sum x_t \sum \mathbf{y}_t}{n\sum x_t^2 - (\sum x_t)^2} \tag{18.2.17}$$

Answer.

$$\boldsymbol{X}^\top\boldsymbol{X} = \begin{bmatrix}\boldsymbol{\iota}^\top\\ \boldsymbol{x}^\top\end{bmatrix}\begin{bmatrix}\boldsymbol{\iota} & \boldsymbol{x}\end{bmatrix} = \begin{bmatrix}\boldsymbol{\iota}^\top\boldsymbol{\iota} & \boldsymbol{\iota}^\top\boldsymbol{x}\\ \boldsymbol{x}^\top\boldsymbol{\iota} & \boldsymbol{x}^\top\boldsymbol{x}\end{bmatrix} = \begin{bmatrix}n & \sum x_t\\ \sum x_t & \sum x_t^2\end{bmatrix} \tag{18.2.18}$$

$$\boldsymbol{X}^\top\boldsymbol{X}^{-1} = \frac{1}{n\sum x_t^2 - (\sum x_t)^2}\begin{bmatrix}\sum x_t^2 & -\sum x_t\\ -\sum x_t & n\end{bmatrix} \tag{18.2.19}$$

$$\boldsymbol{X}^\top\mathbf{y} = \begin{bmatrix}\boldsymbol{\iota}^\top\mathbf{y}\\ \boldsymbol{x}^\top\mathbf{y}\end{bmatrix} = \begin{bmatrix}\sum \mathbf{y}_t\\ \sum x_i\mathbf{y}_t\end{bmatrix} \tag{18.2.20}$$

Therefore $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\mathbf{y}$ gives equations (18.2.16) and (18.2.17). □

PROBLEM 237. *Show that*

$$\sum_{t=1}^{n}(x_t-\bar{x})(\mathsf{y}_t-\bar{\mathsf{y}})=\sum_{t=1}^{n}x_t\mathsf{y}_t-n\bar{x}\bar{\mathsf{y}} \tag{18.2.21}$$

(Note, as explained in [**DM93**, pp. 27/8] *or* [**Gre97**, Section 5.4.1]*, that the left hand side is computationally much more* stable *than the right.)*

ANSWER. Simply multiply out. □

PROBLEM 238. *Show that* (18.2.17) *and* (18.2.16) *can also be written as follows:*

$$\hat{\beta}=\frac{\sum(x_t-\bar{x})(\mathsf{y}_t-\bar{\mathsf{y}})}{\sum(x_t-\bar{x})^2} \tag{18.2.22}$$

$$\hat{\alpha}=\bar{\mathsf{y}}-\hat{\beta}\bar{x} \tag{18.2.23}$$

ANSWER. Using $\sum x_i=n\bar{x}$ and $\sum \mathsf{y}_i=n\bar{\mathsf{y}}$ in (18.2.17), it can be written as

$$\hat{\beta}=\frac{\sum x_t\mathsf{y}_t-n\bar{x}\bar{\mathsf{y}}}{\sum x_t^2-n\bar{x}^2} \tag{18.2.24}$$

Now apply Problem 237 to the numerator of (18.2.24), and Problem 237 with $\mathsf{y}=\boldsymbol{x}$ to the denominator, to get (18.2.22).

To prove equation (18.2.23) for $\hat{\alpha}$, let us work backwards and plug (18.2.24) into the righthand side of (18.2.23):

$$\bar{\mathsf{y}}-\bar{x}\hat{\beta}=\frac{\bar{\mathsf{y}}\sum x_t^2-\bar{\mathsf{y}}n\bar{x}^2-\bar{x}\sum x_t\mathsf{y}_t+n\bar{x}\bar{x}\bar{\mathsf{y}}}{\sum x_t^2-n\bar{x}^2} \tag{18.2.25}$$

The second and the fourth term in the numerator cancel out, and what remains can be shown to be equal to (18.2.16). □

PROBLEM 239. *3 points Show that in the simple regression model, the fitted regression line can be written in the form*

$$\hat{\mathsf{y}}_t=\bar{\mathsf{y}}+\hat{\beta}(x_t-\bar{x}). \tag{18.2.26}$$

From this follows in particular that the fitted regression line always goes through the point $\bar{x},\bar{\mathsf{y}}$.

ANSWER. Follows immediately if one plugs (18.2.23) into the defining equation $\hat{\mathsf{y}}_t=\hat{\alpha}+\hat{\beta}x_t$. □

Formulas (18.2.22) and (18.2.23) are interesting because they express the regression coefficients in terms of the sample means and covariances. Problem 240 derives the properties of the population equivalents of these formulas:

PROBLEM 240. *Given two random variables* x *and* y *with finite variances, and* $\operatorname{var}[\mathsf{x}]>0$. *You know the expected values, variances and covariance of* x *and* y*, and you observe* x*, but* y *is unobserved. This question explores the properties of the Best Linear Unbiased Predictor (BLUP) of* y *in this situation.*

• a. *4 points Give a direct proof of the following, which is a special case of theorem 27.1.1: If you want to predict* y *by an affine expression of the form* $a+b\mathsf{x}$*, you will get the lowest mean squared error* MSE *with* $b=\operatorname{cov}[\mathsf{x},\mathsf{y}]/\operatorname{var}[\mathsf{x}]$ *and* $a=\operatorname{E}[\mathsf{y}]-b\operatorname{E}[\mathsf{x}]$.

ANSWER. The MSE is variance plus squared bias (see e.g. problem 193), therefore

$$\operatorname{MSE}[a+b\mathsf{x};\mathsf{y}]=\operatorname{var}[a+b\mathsf{x}-\mathsf{y}]+(\operatorname{E}[a+b\mathsf{x}-\mathsf{y}])^2=\operatorname{var}[b\mathsf{x}-\mathsf{y}]+(a-\operatorname{E}[\mathsf{y}]+b\operatorname{E}[\mathsf{x}])^2. \tag{18.2.27}$$

Therefore we choose a so that the second term is zero, and then you only have to minimize the first term with respect to b. Since

$$\operatorname{var}[b\mathsf{x}-\mathsf{y}]=b^2\operatorname{var}[\mathsf{x}]-2b\operatorname{cov}[\mathsf{x},\mathsf{y}]+\operatorname{var}[\mathsf{y}] \tag{18.2.28}$$

the first order condition is

$$2b\operatorname{var}[\mathsf{x}]-2\operatorname{cov}[\mathsf{x},\mathsf{y}]=0 \tag{18.2.29}$$

□

• b. *2 points For the first-order conditions you needed the partial derivatives* $\frac{\partial}{\partial a}\operatorname{E}[(\mathsf{y}-a-b\mathsf{x})^2]$ *and* $\frac{\partial}{\partial b}\operatorname{E}[(\mathsf{y}-a-b\mathsf{x})^2]$. *It is also possible, and probably shorter, to interchange taking expected value and partial derivative, i.e., to compute* $\operatorname{E}\left[\frac{\partial}{\partial a}(\mathsf{y}-a-b\mathsf{x})^2\right]$ *and* $\operatorname{E}\left[\frac{\partial}{\partial b}(\mathsf{y}-a-b\mathsf{x})^2\right]$ *and set those zero. Do the above proof in this alternative fashion.*

ANSWER. $\operatorname{E}\left[\frac{\partial}{\partial a}(\mathsf{y}-a-b\mathsf{x})^2\right]=-2\operatorname{E}[\mathsf{y}-a-b\mathsf{x}]=-2(\operatorname{E}[\mathsf{y}]-a-b\operatorname{E}[\mathsf{x}])$. Setting this zero gives the formula for a. Now $\operatorname{E}\left[\frac{\partial}{\partial b}(\mathsf{y}-a-b\mathsf{x})^2\right]=-2\operatorname{E}[\mathsf{x}(\mathsf{y}-a-b\mathsf{x})]=-2(\operatorname{E}[\mathsf{x}\mathsf{y}]-a\operatorname{E}[\mathsf{x}]-b\operatorname{E}[\mathsf{x}^2])$. Setting this zero gives $\operatorname{E}[\mathsf{x}\mathsf{y}]-a\operatorname{E}[\mathsf{x}]-b\operatorname{E}[\mathsf{x}^2]=0$. Plug in formula for a and solve for b:

$$b=\frac{\operatorname{E}[\mathsf{x}\mathsf{y}]-\operatorname{E}[\mathsf{x}]\operatorname{E}[\mathsf{y}]}{\operatorname{E}[\mathsf{x}^2]-(\operatorname{E}[\mathsf{x}])^2}=\frac{\operatorname{cov}[\mathsf{x},\mathsf{y}]}{\operatorname{var}[\mathsf{x}]}. \tag{18.2.30}$$

□

• c. *2 points Compute the* MSE *of this predictor.*

ANSWER. If one plugs the optimal a into (18.2.27), this just annulls the last term of (18.2.27) so that the MSE is given by (18.2.28). If one plugs the optimal $b=\operatorname{cov}[\mathsf{x},\mathsf{y}]/\operatorname{var}[\mathsf{x}]$ into (18.2.28), one gets

$$\operatorname{MSE}=\left(\frac{\operatorname{cov}[\mathsf{x},\mathsf{y}]}{\operatorname{var}[\mathsf{x}]}\right)^2\operatorname{var}[\mathsf{x}]-2\frac{(\operatorname{cov}[\mathsf{x},\mathsf{y}])}{\operatorname{var}[\mathsf{x}]}\operatorname{cov}[\mathsf{x},\mathsf{y}]+\operatorname{var}[\mathsf{x}] \tag{18.2.31}$$

$$=\operatorname{var}[\mathsf{y}]-\frac{(\operatorname{cov}[\mathsf{x},\mathsf{y}])^2}{\operatorname{var}[\mathsf{x}]}. \tag{18.2.32}$$

□

• d. *2 points Show that the prediction error is uncorrelated with the observed* x.

Answer.

$$\operatorname{cov}[\mathsf{x}, \mathsf{y} - a - b\mathsf{x}] = \operatorname{cov}[\mathsf{x}, \mathsf{y}] - a \operatorname{cov}[\mathsf{x}, \mathsf{x}] = 0 \tag{18.2.33}$$

□

• e. *4 points If* $\operatorname{var}[\mathsf{x}] = 0$, *the quotient* $\operatorname{cov}[\mathsf{x}, \mathsf{y}]/\operatorname{var}[\mathsf{x}]$ *can no longer be formed, but if you replace the inverse by the g-inverse, so that the above formula becomes*

$$b = \operatorname{cov}[\mathsf{x}, \mathsf{y}](\operatorname{var}[\mathsf{x}])^{-} \tag{18.2.34}$$

then it always gives the minimum MSE *predictor, whether or not* $\operatorname{var}[\mathsf{x}] = 0$, *and regardless of which g-inverse you use (in case there are more than one). To prove this, you need to answer the following four questions: (a) what is the BLUP if* $\operatorname{var}[\mathsf{x}] = 0$*? (b) what is the g-inverse of a nonzero scalar? (c) what is the g-inverse of the scalar number 0? (d) if* $\operatorname{var}[\mathsf{x}] = 0$, *what do we know about* $\operatorname{cov}[\mathsf{x}, \mathsf{y}]$*?*

Answer. (a) If $\operatorname{var}[\mathsf{x}] = 0$ then $\mathsf{x} = \mu$ almost surely, therefore the observation of x does not give us any new information. The BLUP of y is ν in this case, i.e., the above formula holds with $b = 0$.

(b) The g-inverse of a nonzero scalar is simply its inverse.

(c) Every scalar is a g-inverse of the scalar 0.

(d) if $\operatorname{var}[\mathsf{x}] = 0$, then $\operatorname{cov}[\mathsf{x}, \mathsf{y}] = 0$.

Therefore pick a g-inverse 0, an arbitrary number will do, call it c. Then formula (18.2.34) says $b = 0 \cdot c = 0$. □

Problem 241. *3 points Carefully state the specifications of the random variables involved in the linear regression model. How does the model in Problem 240 differ from the linear regression model? What do they have in common?*

Answer. In the regression model, you have several observations, in the other model only one. In the regression model, the $\boldsymbol{x}_i$ are nonrandom, only the y_i are random, in the other model both x and y are random. In the regression model, the expected value of the y_i are not fully known, in the other model the expected values of both x and y are fully known. Both models have in common that the second moments are known only up to an unknown factor. Both models have in common that only first and second moments need to be known, and that they restrict themselves to linear estimators, and that the criterion function is the MSE (the regression model minimaxes it, but the other model minimizes it since there is no unknown parameter whose value one has to minimax over. But this I cannot say right now, for this we need the Gauss-Markov theorem. Also the Gauss-Markov is valid in both cases!)

□

Problem 242. *2 points We are in the multiple regression model* $\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ *with intercept, i.e.,* $\boldsymbol{X}$ *is such that there is a vector* $\boldsymbol{a}$ *with* $\boldsymbol{\iota} = \boldsymbol{X}\boldsymbol{a}$. *Define the row vector* $\bar{\boldsymbol{x}}^\top = \frac{1}{n}\boldsymbol{\iota}^\top\boldsymbol{X}$, *i.e., it has as its* j*th component the sample mean of the* j*th independent variable. Using the normal equations* $\boldsymbol{X}^\top\mathbf{y} = \boldsymbol{X}^\top\boldsymbol{X}\hat{\boldsymbol{\beta}}$, *show that*

$\bar{\mathsf{y}} = \bar{\boldsymbol{x}}^\top\hat{\boldsymbol{\beta}}$ *(i.e., the regression plane goes through the center of gravity of all data points).*

Answer. Premultiply the normal equation by $\boldsymbol{a}^\top$ to get $\boldsymbol{\iota}^\top\mathbf{y} - \boldsymbol{\iota}^\top\boldsymbol{X}\hat{\boldsymbol{\beta}} = 0$. Premultiply by $1/n$ to get the result. □

Problem 243. *The fitted values* $\hat{\mathbf{y}}$ *and the residuals* $\hat{\boldsymbol{\varepsilon}}$ *are "orthogonal" in two different ways.*

• a. *2 points Show that the inner product* $\hat{\mathbf{y}}^\top\hat{\boldsymbol{\varepsilon}} = 0$. *Why should you expect this from the geometric intuition of Least Squares?*

Answer. Use $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{M}\mathbf{y}$ and $\hat{\mathbf{y}} = (\boldsymbol{I}-\boldsymbol{M})\mathbf{y}$: $\hat{\mathbf{y}}^\top\hat{\boldsymbol{\varepsilon}} = \mathbf{y}^\top(\boldsymbol{I}-\boldsymbol{M})\boldsymbol{M}\mathbf{y} = 0$ because $\boldsymbol{M}(\boldsymbol{I}-\boldsymbol{M}) = \boldsymbol{O}$. This is a consequence of the more general result given in problem **??**.

□

• b. *2 points Sometimes two random variables are called "orthogonal" if their covariance is zero. Show that* $\hat{\mathbf{y}}$ *and* $\hat{\boldsymbol{\varepsilon}}$ *are orthogonal also in this sense, i.e., show that for every* i *and* j, $\operatorname{cov}[\hat{\mathsf{y}}_i, \hat{\varepsilon}_j] = 0$. *In matrix notation this can also be written* $\mathcal{C}[\hat{\mathbf{y}}, \hat{\boldsymbol{\varepsilon}}] = \boldsymbol{O}$.

Answer. $\mathcal{C}[\hat{\mathbf{y}}, \hat{\boldsymbol{\varepsilon}}] = \mathcal{C}[(\boldsymbol{I}-\boldsymbol{M})\mathbf{y}, \boldsymbol{M}\mathbf{y}] = (\boldsymbol{I}-\boldsymbol{M})\,\mathcal{V}[\mathbf{y}]\boldsymbol{M}^\top = (\boldsymbol{I}-\boldsymbol{M})(\sigma^2\boldsymbol{I})\boldsymbol{M} = \sigma^2(\boldsymbol{I}-\boldsymbol{M})\boldsymbol{M} = \boldsymbol{O}$. This is a consequence of the more general result given in question 300. □

18.3. The Coefficient of Determination

Among the criteria which are often used to judge whether the model is appropriate, we will look at the "coefficient of determination" R^2, the "adjusted" $\bar{R}^2$, and later also at Mallow's $\mathsf{C_p}$ statistic. Mallow's $\mathsf{C_p}$ comes later because it is not a final but an initial criterion, i.e., it does not measure the fit of the model to the given data, but it estimates its MSE. Let us first look at R^2.

A value of R^2 always is based (explicitly or implicitly) on a comparison of two models, usually nested in the sense that the model with fewer parameters can be viewed as a specialization of the model with more parameters. The value of R^2 is then 1 minus the ratio of the smaller to the larger sum of squared residuals.

Thus, there is no such thing as the R^2 from a single fitted model—one must always think about what model (perhaps an implicit "null" model) is held out as a standard of comparison. Once that is determined, the calculation is straightforward, based on the sums of squared residuals from the two models. This is particularly appropriate for nls(), which minimizes a sum of squares.

The treatment which follows here is a little more complete than most. Some textbooks, such as [**DM93**], never even give the leftmost term in formula (18.3.6) according to which R^2 is the sample correlation coefficient. Other textbooks, such that [**JHG+88**] and [**Gre97**], do give this formula, but it remains a surprise: there is no explanation why the same quantity R^2 can be expressed mathematically in

two quite different ways, each of which has a different interpretation. The present treatment explains this.

If the regression has a constant term, then the OLS estimate $\hat{\boldsymbol{\beta}}$ has a third optimality property (in addition to minimizing the SSE and being the BLUE): no other linear combination of the explanatory variables has a higher squared sample correlation with $\boldsymbol{y}$ than $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$.

In the proof of this optimality property we will use the symmetric and idempotent projection matrix $\boldsymbol{D} = \boldsymbol{I} - \frac{1}{n}\boldsymbol{\iota}\boldsymbol{\iota}^\top$. Applied to any vector $\boldsymbol{z}$, $\boldsymbol{D}$ gives $\boldsymbol{D}\boldsymbol{z} = \boldsymbol{z} - \boldsymbol{\iota}\bar{z}$, which is $\boldsymbol{z}$ with the mean taken out. Taking out the mean is therefore a projection, on the space orthogonal to $\boldsymbol{\iota}$. See Problem 189.

PROBLEM 244. *In the* `reggeom` *visualization, see Problem 350, in which* $\boldsymbol{x}_1$ *is the vector of ones, which are the vectors* $\boldsymbol{D}\boldsymbol{x}_2$ *and* $\boldsymbol{D}\mathbf{y}$*?*

ANSWER. $\boldsymbol{D}\boldsymbol{x}_2$ is *og*, the dark blue line starting at the origin, and $\boldsymbol{D}\mathbf{y}$ is *cy*, the red line starting on $\boldsymbol{x}_1$ and going up to the peak. □

As an additional mathematical tool we will need the Cauchy-Schwartz inequality for the vector product:

$$(\boldsymbol{u}^\top\boldsymbol{v})^2 \leq (\boldsymbol{u}^\top\boldsymbol{u})(\boldsymbol{v}^\top\boldsymbol{v}) \tag{18.3.1}$$

PROBLEM 245. *If* $\boldsymbol{Q}$ *is any nonnegative definite matrix, show that also*

$$(\boldsymbol{u}^\top\boldsymbol{Q}\boldsymbol{v})^2 \leq (\boldsymbol{u}^\top\boldsymbol{Q}\boldsymbol{u})(\boldsymbol{v}^\top\boldsymbol{Q}\boldsymbol{v}). \tag{18.3.2}$$

ANSWER. This follows from the fact that any nnd matrix $\boldsymbol{Q}$ can be written in the form $\boldsymbol{Q} = \boldsymbol{R}^\top\boldsymbol{R}$. □

In order to prove that $\hat{\boldsymbol{y}}$ has the highest squared sample correlation, take any vector $\boldsymbol{c}$ and look at $\tilde{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{c}$. We will show that the sample correlation of $\boldsymbol{y}$ with $\tilde{\boldsymbol{y}}$ cannot be higher than that of $\boldsymbol{y}$ with $\hat{\boldsymbol{y}}$. For this let us first compute the sample *covariance*. By (12.3.17), n times the sample covariance between $\tilde{\boldsymbol{y}}$ and $\boldsymbol{y}$ is

$$n \text{ times sample covariance}(\tilde{\boldsymbol{y}}, \boldsymbol{y}) = \tilde{\boldsymbol{y}}^\top\boldsymbol{D}\boldsymbol{y} = \boldsymbol{c}^\top\boldsymbol{X}^\top\boldsymbol{D}(\hat{\boldsymbol{y}} + \hat{\boldsymbol{\varepsilon}}). \tag{18.3.3}$$

By Problem 246, $\boldsymbol{D}\hat{\boldsymbol{\varepsilon}} = \hat{\boldsymbol{\varepsilon}}$, hence $\boldsymbol{X}^\top\boldsymbol{D}\hat{\boldsymbol{\varepsilon}} = \boldsymbol{X}^\top\hat{\boldsymbol{\varepsilon}} = \boldsymbol{o}$ (this last equality is equivalent to the Normal Equation (18.2.3)), therefore (18.3.3) becomes $\tilde{\boldsymbol{y}}^\top\boldsymbol{D}\boldsymbol{y} = \tilde{\boldsymbol{y}}^\top\boldsymbol{D}\hat{\boldsymbol{y}}$. Together with (18.3.2) this gives

$$\big(n \text{ times sample covariance}(\tilde{\boldsymbol{y}}, \boldsymbol{y})\big)^2 = (\tilde{\boldsymbol{y}}^\top\boldsymbol{D}\hat{\boldsymbol{y}})^2 \leq (\tilde{\boldsymbol{y}}^\top\boldsymbol{D}\tilde{\boldsymbol{y}})(\hat{\boldsymbol{y}}^\top\boldsymbol{D}\hat{\boldsymbol{y}}) \tag{18.3.4}$$

In order to get from n^2 times the squared sample covariance to the squared sample correlation coefficient we have to divide it by n^2 times the sample variances

of $\tilde{\boldsymbol{y}}$ and of $\boldsymbol{y}$:

$$\big(\text{sample correlation}(\tilde{\boldsymbol{y}}, \boldsymbol{y})\big)^2 = \frac{(\tilde{\boldsymbol{y}}^\top\boldsymbol{D}\boldsymbol{y})^2}{(\tilde{\boldsymbol{y}}^\top\boldsymbol{D}\tilde{\boldsymbol{y}})(\boldsymbol{y}^\top\boldsymbol{D}\boldsymbol{y})} \leq \frac{\hat{\boldsymbol{y}}^\top\boldsymbol{D}\hat{\boldsymbol{y}}}{\boldsymbol{y}^\top\boldsymbol{D}\boldsymbol{y}} = \frac{\sum(\hat{y}_j - \bar{\hat{y}})^2}{\sum(y_j - \bar{y})^2} = \frac{\sum(\hat{y}_j - \bar{y})^2}{\sum(y_j - \bar{y})^2}. \tag{18.3.5}$$

For the rightmost equal sign in (18.3.5) we need Problem 247.

If $\tilde{\boldsymbol{y}} = \hat{\boldsymbol{y}}$, inequality (18.3.4) becomes an equality, and therefore also (18.3.5) becomes an equality throughout. This completes the proof that $\hat{\boldsymbol{y}}$ has the highest possible squared sample correlation with $\boldsymbol{y}$, and gives at the same time two different formulas for the same entity

$$R^2 = \frac{\big(\sum(\hat{y}_j - \bar{\hat{y}})(y_j - \bar{y})\big)^2}{\sum(\hat{y}_j - \bar{\hat{y}})^2\sum(y_j - \bar{y})^2} = \frac{\sum(\hat{y}_j - \bar{y})^2}{\sum(y_j - \bar{y})^2}. \tag{18.3.6}$$

PROBLEM 246. *1 point Show that, if* $\boldsymbol{X}$ *contains a constant term, then* $\boldsymbol{D}\hat{\boldsymbol{\varepsilon}} = \hat{\boldsymbol{\varepsilon}}$*. You are allowed to use the fact that* $\boldsymbol{X}^\top\hat{\boldsymbol{\varepsilon}} = \boldsymbol{o}$*, which is equivalent to the normal equation* (18.2.3)*.*

ANSWER. Since $\boldsymbol{X}$ has a constant term, a vector $\boldsymbol{a}$ exists such that $\boldsymbol{X}\boldsymbol{a} = \boldsymbol{\iota}$, therefore $\boldsymbol{\iota}^\top\hat{\boldsymbol{\varepsilon}} = \boldsymbol{a}^\top\boldsymbol{X}^\top\hat{\boldsymbol{\varepsilon}} = \boldsymbol{a}^\top\boldsymbol{o} = 0$. From $\boldsymbol{\iota}^\top\hat{\boldsymbol{\varepsilon}} = 0$ follows $\boldsymbol{D}\hat{\boldsymbol{\varepsilon}} = \hat{\boldsymbol{\varepsilon}}$. □

PROBLEM 247. *1 point Show that, if* $\boldsymbol{X}$ *has a constant term, then* $\bar{\hat{y}} = \bar{y}$

ANSWER. Follows from $0 = \boldsymbol{\iota}^\top\hat{\boldsymbol{\varepsilon}} = \boldsymbol{\iota}^\top\boldsymbol{y} - \boldsymbol{\iota}^\top\hat{\boldsymbol{y}}$. In the visualization, this is equivalent with the fact that both *ocb* and *ocy* are right angles. □

PROBLEM 248. *Instead of* (18.3.6) *one often sees the formula*

$$\frac{\big(\sum(\hat{y}_j - \bar{y})(y_j - \bar{y})\big)^2}{\sum(\hat{y}_j - \bar{y})^2\sum(y_j - \bar{y})^2} = \frac{\sum(\hat{y}_j - \bar{y})^2}{\sum(y_j - \bar{y})^2}. \tag{18.3.7}$$

Prove that they are equivalent. Which equation is better?

The denominator in the righthand side expression of (18.3.6), $\sum(y_j - \bar{y})^2$, is usually called "SST," the total (corrected) sum of squares. The numerator $\sum(\hat{y}_j - \bar{y})^2$ is usually called "SSR," the sum of squares "explained" by the regression. In order to understand SSR better, we will show next the famous "Analysis of Variance" identity SST = SSR + SSE.

PROBLEM 249. *In the* `reggeom` *visualization, again with* $\boldsymbol{x}_1$ *representing the vector of ones, show that* SST = SSR + SSE*, and show that* $\mathsf{R}^2 = \cos^2\alpha$ *where* α *is the angle between two lines in this visualization. Which lines?*

ANSWER. $\hat{\varepsilon}$ is the *by*, the green line going up to the peak, and SSE is the squared length of it. SST is the squared length of $\mathbf{y} - \boldsymbol{\iota}\bar{\mathbf{y}}$. Sincer $\boldsymbol{\iota}\bar{\mathbf{y}}$ is the projection of $\mathbf{y}$ on $\boldsymbol{x}_1$, i.e., it is *oc*, the part of $\boldsymbol{x}_1$ that is red, one sees that SST is the squared length of *cy*. SSR is the squared length of *cb*. The

analysis of variance identity follows because cby is a right angle. $\mathsf{R}^2 = \cos^2\alpha$ where α is the angle between bcy in this same triangle.

□

Since the regression has a constant term, the decomposition

$$\boldsymbol{y} = (\boldsymbol{y} - \hat{\boldsymbol{y}}) + (\hat{\boldsymbol{y}} - \boldsymbol{\iota}\bar{y}) + \boldsymbol{\iota}\bar{y} \tag{18.3.8}$$

is an orthogonal decomposition (all three vectors on the righthand side are orthogonal to each other), therefore in particular

$$(\boldsymbol{y} - \hat{\boldsymbol{y}})^\top(\hat{\boldsymbol{y}} - \boldsymbol{\iota}\bar{y}) = 0. \tag{18.3.9}$$

Geometrically this follows from the fact that $\boldsymbol{y} - \hat{\boldsymbol{y}}$ is orthogonal to the column space of $\boldsymbol{X}$, while $\hat{\boldsymbol{y}} - \boldsymbol{\iota}\bar{y}$ lies in that column space.

Problem 250. *Show the decomposition 18.3.8 in the* `reggeom`*-visualization.*

Answer. From y take the green line down to b, then the light blue line to c, then the red line to the origin. □

This orthogonality can also be explained in terms of sequential projections: instead of projecting $\mathbf{y}$ on $\boldsymbol{x}_1$ directly I can first project it on the plane spanned by $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, and then project this projection on $\boldsymbol{x}_1$.

From (18.3.9) follows (now the same identity written in three different notations):

$$(\boldsymbol{y} - \boldsymbol{\iota}\bar{y})^\top(\boldsymbol{y} - \boldsymbol{\iota}\bar{y}) = (\boldsymbol{y} - \hat{\boldsymbol{y}})^\top(\boldsymbol{y} - \hat{\boldsymbol{y}}) + (\hat{\boldsymbol{y}} - \boldsymbol{\iota}\bar{y})^\top(\hat{\boldsymbol{y}} - \boldsymbol{\iota}\bar{y}) \tag{18.3.10}$$

$$\sum_t (y_t - \bar{y})^2 = \sum_t (y_t - \hat{y}_t)^2 + \sum_t (\hat{y}_t - \bar{y})^2 \tag{18.3.11}$$

$$\mathsf{SST} = \mathsf{SSE} + \mathsf{SSR} \tag{18.3.12}$$

Problem 251. *5 points Show that the "analysis of variance" identity* $\mathsf{SST} = \mathsf{SSE} + \mathsf{SSR}$ *holds in a regression with intercept, i.e., prove one of the two following equations:*

$$(\boldsymbol{y} - \boldsymbol{\iota}\bar{y})^\top(\boldsymbol{y} - \boldsymbol{\iota}\bar{y}) = (\boldsymbol{y} - \hat{\boldsymbol{y}})^\top(\boldsymbol{y} - \hat{\boldsymbol{y}}) + (\hat{\boldsymbol{y}} - \boldsymbol{\iota}\bar{y})^\top(\hat{\boldsymbol{y}} - \boldsymbol{\iota}\bar{y}) \tag{18.3.13}$$

$$\sum_t (y_t - \bar{y})^2 = \sum_t (y_t - \hat{y}_t)^2 + \sum_t (\hat{y}_t - \bar{y})^2 \tag{18.3.14}$$

Answer. Start with

$$\mathsf{SST} = \sum (y_t - \bar{y})^2 = \sum (y_t - \hat{y}_t + \hat{y}_t - \bar{y})^2 \tag{18.3.15}$$

and then show that the cross product term $\sum(y_t - \hat{y}_t)(\hat{y}_t - \bar{y}) = \sum \hat{\varepsilon}_t(\hat{y}_t - \bar{y}) = \hat{\boldsymbol{\varepsilon}}^\top(\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{\iota}\frac{1}{n}\boldsymbol{\iota}^\top\boldsymbol{y}) = 0$ since $\hat{\boldsymbol{\varepsilon}}^\top\boldsymbol{X} = \boldsymbol{o}^\top$ and in particular, since a constant term is included, $\hat{\boldsymbol{\varepsilon}}^\top\boldsymbol{\iota} = 0$.

□

From the so-called "analysis of variance" identity (18.3.12), together with (18.3.6), one obtains the following three alternative expressions for the maximum possible correlation, which is called R^2 and which is routinely used as a measure of the "fit" of the regression:

$$R^2 = \frac{\left(\sum(\hat{y}_j - \bar{\hat{y}})(y_j - \bar{y})\right)^2}{\sum(\hat{y}_j - \bar{\hat{y}})^2 \sum(y_j - \bar{y})^2} = \frac{\mathsf{SSR}}{\mathsf{SST}} = \frac{\mathsf{SST} - \mathsf{SSE}}{\mathsf{SST}} \tag{18.3.16}$$

The first of these three expressions is the squared sample correlation coefficient between $\hat{\boldsymbol{y}}$ and $\boldsymbol{y}$, hence the notation R^2. The usual interpretation of the middle expression is the following: SST can be decomposed into a part SSR which is "explained" by the regression, and a part SSE which remains "unexplained," and R^2 measures that fraction of SST which can be "explained" by the regression. [**Gre97**, pp. 250–253] and also [**JHG+88**, pp. 211/212] try to make this notion plausible. Instead of using the vague notions "explained" and "unexplained," I prefer the following reading, which is based on the third expression for R^2 in (18.3.16): $\boldsymbol{\iota}\bar{y}$ is the vector of fitted values if one regresses $\boldsymbol{y}$ on a constant term only, and SST is the SSE in this "restricted" regression. R^2 measures therefore the proportionate reduction in the SSE if one adds the nonconstant regressors to the regression. From this latter formula one can also see that $\mathsf{R}^2 = \cos^2\alpha$ where α is the angle between $\mathbf{y} - \boldsymbol{\iota}\bar{\mathbf{y}}$ and $\hat{\mathbf{y}} - \boldsymbol{\iota}\bar{\mathbf{y}}$.

Problem 252. *Given two data series $\boldsymbol{x}$ and $\boldsymbol{y}$. Show that the regression of $\boldsymbol{y}$ on $\boldsymbol{x}$ has the same R^2 as the regression of $\boldsymbol{x}$ on $\boldsymbol{y}$. (Both regressions are assumed to include a constant term.) Easy, but you have to think!*

Answer. The symmetry comes from the fact that, in this particular case, R^2 is the squared sample correlation coefficient between $\boldsymbol{x}$ and $\boldsymbol{y}$. Proof: $\hat{\boldsymbol{y}}$ is an affine transformation of $\boldsymbol{x}$, and correlation coefficients are invariant under affine transformations (compare Problem 254). □

Problem 253. *This Problem derives some relationships which are valid in simple regression $y_t = \alpha + \beta x_t + \varepsilon_t$ but their generalization to multiple regression is not obvious.*

- a. *2 points Show that*

$$R^2 = \hat{\beta}^2 \frac{\sum(x_t - \bar{x})^2}{\sum(y_t - \bar{y})^2} \tag{18.3.17}$$

Hint: show first that $\hat{y}_t - \bar{y} = \hat{\beta}(x_t - \bar{x})$.

Answer. From $\hat{y}_t = \hat{\alpha} + \hat{\beta}x_t$ and $\bar{y} = \hat{\alpha} + \hat{\beta}\bar{x}$ follows $\hat{y}_t - \bar{y} = \hat{\beta}(x_t - \bar{x})$. Therefore

$$R^2 = \frac{\sum(\hat{y}_t - \bar{y})^2}{\sum(y_t - \bar{y})^2} = \hat{\beta}^2 \frac{\sum(x_t - \bar{x})^2}{\sum(y_t - \bar{y})^2} \tag{18.3.18}$$

□

• b. *2 points Furthermore show that R^2 is the sample correlation coefficient between $\boldsymbol{y}$ and $\boldsymbol{x}$, i.e.,*

$$R^2 = \frac{\left(\sum(x_t-\bar{x})(y_t-\bar{y})\right)^2}{\sum(x_t-\bar{x})^2\sum(y_t-\bar{y})^2}. \tag{18.3.19}$$

Hint: you are allowed to use (18.2.22).

Answer.

$$R^2 = \hat{\beta}^2\frac{\sum(x_t-\bar{x})^2}{\sum(y_t-\bar{y})^2} = \frac{\left(\sum(x_t-\bar{x})(y_t-\bar{y})\right)^2\sum(x_t-\bar{x})^2}{\left(\sum(x_t-\bar{x})^2\right)^2\sum(y_t-\bar{y})^2} \tag{18.3.20}$$

which simplifies to (18.3.19). □

• c. *1 point Finally show that $R^2 = \hat{\beta}_{xy}\hat{\beta}_{yx}$, i.e., it is the product of the two slope coefficients one gets if one regresses $\boldsymbol{y}$ on $\boldsymbol{x}$ and $\boldsymbol{x}$ on $\boldsymbol{y}$.*

If the regression does not have a constant term, but a vector $\boldsymbol{a}$ exists with $\boldsymbol{\iota} = \boldsymbol{X}\boldsymbol{a}$, then the above mathematics remains valid. If $\boldsymbol{a}$ does not exist, then the identity $\mathsf{SST} = \mathsf{SSR} + \mathsf{SSE}$ no longer holds, and (18.3.16) is no longer valid. The fraction $\frac{\mathsf{SST}-\mathsf{SSE}}{\mathsf{SST}}$ can assume negative values. Also the sample correlation coefficient between $\hat{\boldsymbol{y}}$ and $\boldsymbol{y}$ loses its motivation, since there will usually be other linear combinations of the columns of $\boldsymbol{X}$ that have higher sample correlation with $\boldsymbol{y}$ than the fitted values $\hat{\boldsymbol{y}}$.

Equation (18.3.16) is still puzzling at this point: why do two quite different simple concepts, the sample correlation and the proportionate reduction of the SSE, give the same numerical result? To explain this, we will take a short digression about correlation coefficients, in which it will be shown that correlation coefficients *always* denote proportionate reductions in the MSE. Since the SSE is (up to a constant factor) the sample equivalent of the MSE of the prediction of $\boldsymbol{y}$ by $\hat{\boldsymbol{y}}$, this shows that (18.3.16) is simply the sample equivalent of a general fact about correlation coefficients.

But first let us take a brief look at the Adjusted R^2.

18.4. The Adjusted R-Square

The coefficient of determination R^2 is often used as a criterion for the selection of regressors. There are several drawbacks to this. [**KA69**, Chapter 8] shows that the distribution function of R^2 depends on both the unknown error variance and the values taken by the explanatory variables; therefore the R^2 belonging to different regressions cannot be compared.

A further drawback is that inclusion of more regressors always increases the R^2. The adjusted $\bar{R}^2$ is designed to remedy this. Starting from the formula $R^2 = 1 - \mathsf{SSE}/\mathsf{SST}$, the "adjustment" consists in dividing both SSE and SST by their degrees of freedom:

$$\bar{R}^2 = 1 - \frac{\mathsf{SSE}/(n-k)}{\mathsf{SST}/(n-1)} = 1-(1-R^2)\frac{n-1}{n-k}. \tag{18.4.1}$$

For given SST, i.e., when one looks at alternative regressions with the same dependent variable, $\bar{R}^2$ is therefore a declining function of s^2, the unbiased estimator of σ^2. Choosing the regression with the highest $\bar{R}^2$ amounts therefore to selecting that regression which yields the lowest value for s^2.

$\bar{R}^2$ has the following interesting property: (which we note here only for reference, because we have not yet discussed the F-test:) Assume one adds i more regressors: then $\bar{R}^2$ increases only if the F statistic for these additional regressors has a value greater than one. One can also say: s^2 decreases only if $\mathsf{F} > 1$. To see this, write this F statistic as

$$\mathsf{F} = \frac{(\mathsf{SSE}_k - \mathsf{SSE}_{k+i})/i}{\mathsf{SSE}_{k+i}/(n-k-i)} = \frac{n-k-i}{i}\Big(\frac{\mathsf{SSE}_k}{\mathsf{SSE}_{k+i}} - 1\Big) \tag{18.4.2}$$

$$= \frac{n-k-i}{i}\Big(\frac{(n-k)\mathsf{s}_k^2}{(n-k-i)\mathsf{s}_{k+i}^2} - 1\Big) \tag{18.4.3}$$

$$= \frac{(n-k)\mathsf{s}_k^2}{i\mathsf{s}_{k+i}^2} - \frac{n-k}{i} + 1 \tag{18.4.4}$$

$$= \frac{(n-k)}{i}\Big(\frac{\mathsf{s}_k^2}{\mathsf{s}_{k+i}^2} - 1\Big) + 1 \tag{18.4.5}$$

From this the statement follows.

Minimizing the adjusted $\bar{R}^2$ is equivalent to minimizing the unbiased variance estimator s^2; it still does not penalize the loss of degrees of freedom heavily enough, i.e., it still admits too many variables into the model.

Alternatives minimize Amemiya's prediction criterion or Akaike's information criterion, which minimize functions of the estimated variances and n and k. Akaike's information criterion minimizes an estimate of the Kullback-Leibler discrepancy, which was discussed on p. 188.

CHAPTER 19

Digression about Correlation Coefficients

19.1. A Unified Definition of Correlation Coefficients

Correlation coefficients measure linear association. The usual definition of the simple correlation coefficient between two variables ρ_{xy} (sometimes we also use the notation $\operatorname{corr}[\mathsf{x},\mathsf{y}]$) is their standardized covariance

$$\rho_{\mathsf{xy}} = \frac{\operatorname{cov}[\mathsf{x},\mathsf{y}]}{\sqrt{\operatorname{var}[\mathsf{x}]}\sqrt{\operatorname{var}[\mathsf{y}]}}. \tag{19.1.1}$$

Because of Cauchy-Schwartz, its value lies between -1 and 1.

Problem 254. *Given the constant scalars $a \neq 0$ and $c \neq 0$ and b and d arbitrary. Show that* $\operatorname{corr}[\mathsf{x},\mathsf{y}] = \pm\operatorname{corr}[a\mathsf{x}+b, c\mathsf{y}+d]$, *with the $+$ sign being valid if a and c have the same sign, and the $-$ sign otherwise.*

Answer. Start with $\operatorname{cov}[a\mathsf{x}+b, c\mathsf{y}+d] = ac\operatorname{cov}[\mathsf{x},\mathsf{y}]$ and go from there. □

Besides the *simple* correlation coefficient ρ_{xy} between two scalar variables y and x, one can also define the squared *multiple* correlation coefficient $\rho^2_{\mathsf{y}(\boldsymbol{\mathsf{x}})}$ between one scalar variable y and a whole vector of variables $\boldsymbol{\mathsf{x}}$, and the *partial* correlation coefficient $\rho_{12.\mathsf{x}}$ between two scalar variables y_1 and y_2, with a vector of other variables $\boldsymbol{\mathsf{x}}$ "partialled out." The multiple correlation coefficient measures the strength of a linear association between y and *all* components of $\boldsymbol{\mathsf{x}}$ together, and the partial correlation coefficient measures the strength of that part of the linear association between y_1 and y_2 which cannot be attributed to their joint association with $\boldsymbol{\mathsf{x}}$. One can also define partial multiple correlation coefficients. If one wants to measure the linear association between two *vectors*, then one number is no longer enough, but one needs several numbers, the "canonical correlations."

The multiple or partial correlation coefficients are usually defined as simple correlation coefficients involving the best linear predictor or its residual. But all these correlation coefficients share the property that they indicate a proportionate reduction in the MSE. See e.g. [**Rao73**, pp. 268–70]. Problem 255 makes this point for the *simple* correlation coefficient:

Problem 255. *4 points Show that the proportionate reduction in the* MSE *of the best predictor of y, if one goes from predictors of the form $\mathsf{y}^* = a$ to predictors of the form $\mathsf{y}^* = a + b\mathsf{x}$, is equal to the squared correlation coefficient between y and x. You are allowed to use the results of Problems 229 and 240. To set notation, call the minimum* MSE *in the first prediction (Problem 229)* MSE[*constant term*; y], *and the minimum* MSE *in the second prediction (Problem 240)* MSE[*constant term and* x; y]. *Show that*

$$\frac{\text{MSE}[\textit{constant term};\mathsf{y}] - \text{MSE}[\textit{constant term and } \mathsf{x};\mathsf{y}]}{\text{MSE}[\textit{constant term};\mathsf{y}]} = \frac{(\operatorname{cov}[\mathsf{y},\mathsf{x}])^2}{\operatorname{var}[\mathsf{y}]\operatorname{var}[\mathsf{x}]} = \rho^2_{\mathsf{yx}}. \tag{19.1.2}$$

Answer. The minimum MSE with only a constant is $\operatorname{var}[\mathsf{y}]$ and (18.2.32) says that MSE[constant term and x; y] $= \operatorname{var}[\mathsf{y}] - (\operatorname{cov}[\mathsf{x},\mathsf{y}])^2/\operatorname{var}[\mathsf{x}]$. Therefore the difference in MSE's is $(\operatorname{cov}[\mathsf{x},\mathsf{y}])^2/\operatorname{var}[\mathsf{x}]$, and if one divides by $\operatorname{var}[\mathsf{y}]$ to get the relative difference, one gets exactly the squared correlation coefficient. □

Multiple Correlation Coefficients. Now assume $\boldsymbol{\mathsf{x}}$ is a vector while y remains a scalar. Their joint mean vector and dispersion matrix are

$$\begin{bmatrix}\boldsymbol{\mathsf{x}}\\ \mathsf{y}\end{bmatrix} \sim \begin{bmatrix}\boldsymbol{\mu}\\ \nu\end{bmatrix}, \sigma^2\begin{bmatrix}\boldsymbol{\Omega}_{\mathsf{xx}} & \boldsymbol{\omega}_{\mathsf{xy}}\\ \boldsymbol{\omega}^\top_{\mathsf{xy}} & \omega_{\mathsf{yy}}\end{bmatrix}. \tag{19.1.3}$$

By theorem **??**, the best linear predictor of y based on $\boldsymbol{\mathsf{x}}$ has the formula

$$\mathsf{y}^* = \nu + \boldsymbol{\omega}^\top_{\mathsf{xy}}\boldsymbol{\Omega}^-_{\mathsf{xx}}(\boldsymbol{\mathsf{x}} - \boldsymbol{\mu}) \tag{19.1.4}$$

y^* has the following additional extremal value property: no linear combination $\boldsymbol{b}^\top\boldsymbol{\mathsf{x}}$ has a higher squared correlation with y than y^*. This maximal value of the squared correlation is called the squared multiple correlation coefficient

$$\rho^2_{\mathsf{y}(\boldsymbol{\mathsf{x}})} = \frac{\boldsymbol{\omega}^\top_{\mathsf{xy}}\boldsymbol{\Omega}^-_{\mathsf{xx}}\boldsymbol{\omega}_{\mathsf{xy}}}{\omega_{\mathsf{yy}}} \tag{19.1.5}$$

The multiple correlation coefficient itself is the positive square root, i.e., it is always nonnegative, while some other correlation coefficients may take on negative values.

The squared multiple correlation coefficient can also defined in terms of proportionate reduction in MSE. It is equal to the proportionate reduction in the MSE of the best predictor of y if one goes from predictors of the form $\mathsf{y}^* = a$ to predictors of the form $\mathsf{y}^* = a + \boldsymbol{b}^\top\boldsymbol{\mathsf{x}}$, i.e.,

$$\rho^2_{\mathsf{y}(\boldsymbol{\mathsf{x}})} = \frac{\text{MSE}[\text{constant term};\mathsf{y}] - \text{MSE}[\text{constant term and } \boldsymbol{\mathsf{x}};\mathsf{y}]}{\text{MSE}[\text{constant term};\mathsf{y}]} \tag{19.1.6}$$

There are therefore two natural definitions of the multiple correlation coefficient. These two definitions correspond to the two formulas for R^2 in (18.3.6).

Partial Correlation Coefficients. Now assume $\mathbf{y} = \begin{bmatrix} \mathsf{y}_1 & \mathsf{y}_2 \end{bmatrix}^\top$ is a vector with two elements and write

$$\begin{bmatrix} \mathbf{x} \\ \mathsf{y}_1 \\ \mathsf{y}_2 \end{bmatrix} \sim \begin{bmatrix} \boldsymbol{\mu} \\ \nu_1 \\ \nu_2 \end{bmatrix}, \sigma^2 \begin{bmatrix} \boldsymbol{\Omega}_{\mathbf{xx}} & \boldsymbol{\omega}_{\mathsf{y}1} & \boldsymbol{\omega}_{\mathsf{y}2} \\ \boldsymbol{\omega}_{\mathsf{y}1}^\top & \omega_{11} & \omega_{12} \\ \boldsymbol{\omega}_{\mathsf{y}2}^\top & \omega_{21} & \omega_{22} \end{bmatrix}. \tag{19.1.7}$$

Let $\mathbf{y}^*$ be the best linear predictor of $\mathbf{y}$ based on $\mathbf{x}$. The partial correlation coefficient $\rho_{12.\mathbf{x}}$ is defined to be the simple correlation between the residuals $\operatorname{corr}[(\mathsf{y}_1 - \mathsf{y}_1^*), (\mathsf{y}_2 - \mathsf{y}_2^*)]$. This measures the correlation between y_1 and y_2 which is "local," i.e., which does not follow from their association with $\mathbf{x}$. Assume for instance that both y_1 and y_2 are highly correlated with $\mathbf{x}$. Then they will also have a high correlation with each other. Subtracting y_i^* from y_i eliminates this dependency on $\mathbf{x}$, therefore any remaining correlation is "local." Compare [**Krz88**, p. 475].

The partial correlation coefficient can be defined as the relative reduction in the MSE if one adds y_2 to $\mathbf{x}$ as a predictor of y_1:

$$\rho^2_{12.\mathbf{x}} = \frac{\text{MSE[constant term and } \mathbf{x}; \mathsf{y}_2] - \text{MSE[constant term, } \mathbf{x}\text{, and } \mathsf{y}_1; \mathsf{y}_2]}{\text{MSE[constant term and } \mathbf{x}; \mathsf{y}_2]}. \tag{19.1.8}$$

Problem 256. *Using the definitions in terms of* MSE*'s, show that the following relationship holds between the squares of multiple and partial correlation coefficients:*

$$1 - \rho^2_{2(\mathbf{x},1)} = (1 - \rho^2_{21.\mathbf{x}})(1 - \rho^2_{2(\mathbf{x})}) \tag{19.1.9}$$

Answer. In terms of the MSE, (19.1.9) reads

$$\frac{\text{MSE[constant term, } \mathbf{x}\text{, and } \mathsf{y}_1; \mathsf{y}_2]}{\text{MSE[constant term; } \mathsf{y}_2]} = \frac{\text{MSE[constant term, } \mathbf{x}\text{, and } \mathsf{y}_1; \mathsf{y}_2]}{\text{MSE[constant term and } \mathbf{x}; \mathsf{y}_2]} \frac{\text{MSE[constant term and } \mathbf{x}; \mathsf{y}_2]}{\text{MSE[constant term; } \mathsf{y}_2]} \tag{19.1.10}$$

□

From (19.1.9) follows the following weighted average formula:

$$\rho^2_{2(\mathbf{x},1)} = \rho^2_{2(\mathbf{x})} + (1 - \rho^2_{2(\mathbf{x})})\rho^2_{21.\mathbf{x}} \tag{19.1.11}$$

An alternative proof of (19.1.11) is given in [**Gra76**, pp. 116/17].

Mixed cases: One can also form multiple correlations coefficients with some of the variables partialled out. The dot notation used here is due to Yule, [**Yul07**]. The notation, definition, and formula for the squared correlation coefficient is

$$\rho^2_{\mathsf{y}(\mathbf{x}).\mathbf{z}} = \frac{\text{MSE[constant term and } \mathbf{z}; \mathsf{y}] - \text{MSE[constant term, } \mathbf{z}\text{, and } \mathbf{x}; \mathsf{y}]}{\text{MSE[constant term and } \mathbf{z}; \mathsf{y}]} \tag{19.1.12}$$

$$= \frac{\boldsymbol{\omega}^\top_{\mathbf{x}\mathsf{y}.\mathbf{z}} \boldsymbol{\Omega}^-_{\mathbf{xx}.\mathbf{z}} \boldsymbol{\omega}_{\mathbf{x}\mathsf{y}.\mathbf{z}}}{\omega_{\mathsf{y}\mathsf{y}.\mathbf{z}}} \tag{19.1.13}$$

19.2. Correlation Coefficients and the Associated Least Squares Problem

One can define the correlation coefficients also as proportionate reductions in the objective functions of the associated GLS problems. However one must reverse predictor and predictand, i.e., one must look at predictions of a vector $\mathbf{x}$ by linear functions of a scalar y.

Here it is done for multiple correlation coefficients: The value of the GLS objective function if one predicts $\mathbf{x}$ by the best linear predictor $\mathbf{x}^*$, which is the minimum attainable when the scalar observation y is given and the vector $\boldsymbol{x}$ can be chosen freely, as long as it satisfies the constraint $\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{\Omega}_{\mathbf{xx}}\boldsymbol{q}$ for some $\boldsymbol{q}$, is

$$\mathsf{SSE}[y; \text{best } \boldsymbol{x}] = \min_{\boldsymbol{x} s.t. \ldots} \begin{bmatrix} (\boldsymbol{x} - \boldsymbol{\mu})^\top & (y - \nu)^\top \end{bmatrix} \begin{bmatrix} \boldsymbol{\Omega}_{\mathbf{xx}} & \boldsymbol{\omega}_{\mathbf{x}\mathsf{y}} \\ \boldsymbol{\omega}^\top_{\mathbf{x}\mathsf{y}} & \omega_{\mathsf{y}\mathsf{y}} \end{bmatrix}^- \begin{bmatrix} \boldsymbol{x} - \boldsymbol{\mu} \\ y - \nu \end{bmatrix} = (y - \nu)^\top \omega^-_{\mathsf{y}\mathsf{y}}(y - \nu). \tag{19.2.1}$$

On the other hand, the value of the GLS objective function when one predicts $\mathbf{x}$ by the best constant $\boldsymbol{x} = \boldsymbol{\mu}$ is

$$\mathsf{SSE}[y; \boldsymbol{x} = \boldsymbol{\mu}] = \begin{bmatrix} \boldsymbol{o}^\top & (y - \nu)^\top \end{bmatrix} \begin{bmatrix} \boldsymbol{\Omega}^-_{\mathbf{xx}} + \boldsymbol{\Omega}^-_{\mathbf{xx}}\boldsymbol{\omega}_{\mathbf{x}\mathsf{y}}\omega^-_{\mathsf{y}\mathsf{y}.\mathbf{x}}\boldsymbol{\omega}^\top_{\mathbf{x}\mathsf{y}}\boldsymbol{\Omega}^-_{\mathbf{xx}} & -\boldsymbol{\Omega}^-_{\mathbf{xx}}\boldsymbol{\omega}_{\mathbf{x}\mathsf{y}}\omega^-_{\mathsf{y}\mathsf{y}.\mathbf{x}} \\ -\omega^-_{\mathsf{y}\mathsf{y}.\mathbf{x}}\boldsymbol{\omega}^\top_{\mathbf{x}\mathsf{y}}\boldsymbol{\Omega}^-_{\mathbf{xx}} & \omega^-_{\mathsf{y}\mathsf{y}.\mathbf{x}} \end{bmatrix} \begin{bmatrix} \boldsymbol{o} \\ y - \nu \end{bmatrix} = \tag{19.2.2}$$

$$= (y - \nu)^\top \omega^-_{\mathsf{y}\mathsf{y}.\mathbf{x}}(y - \nu). \tag{19.2.3}$$

The proportionate reduction in the objective function is

$$\frac{\mathsf{SSE}[y; \boldsymbol{x} = \boldsymbol{\mu}] - \mathsf{SSE}[y; \text{best } \boldsymbol{x}]}{\mathsf{SSE}[y; \boldsymbol{x} = \boldsymbol{\mu}]} = \frac{(y - \nu)^2/\omega_{\mathsf{y}\mathsf{y}.\mathbf{x}} - (y - \nu)^2/\omega_{\mathsf{y}\mathsf{y}}}{(y - \nu)^2/\omega_{\mathsf{y}\mathsf{y}.\mathbf{x}}} = \tag{19.2.4}$$

$$= \frac{\omega_{\mathsf{y}\mathsf{y}} - \omega_{\mathsf{y}\mathsf{y}.\mathbf{x}}}{\omega_{\mathsf{y}\mathsf{y}}} = \rho^2_{\mathsf{y}(\mathbf{x})} = 1 - \frac{\omega_{\mathsf{y}\mathsf{y}.\mathbf{x}}}{\omega_{\mathsf{y}\mathsf{y}}} = 1 - \frac{1}{\omega^{\mathsf{y}\mathsf{y}}\omega_{\mathsf{y}\mathsf{y}}} = \rho^2_{\mathsf{y}(\mathbf{x})} \tag{19.2.5}$$

19.3. Canonical Correlations

Now what happens with the correlation coefficients if both predictor and predictand are vectors? In this case one has more than one correlation coefficient. One first finds those two linear combinations of the two vectors which have highest correlation, then those which are uncorrelated with the first and have second highest correlation, and so on. Here is the mathematical construction needed:

Let $\mathbf{x}$ and $\mathbf{y}$ be two column vectors consisting of p and q scalar random variables, respectively, and let

$$\mathcal{V}[\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}] = \sigma^2 \begin{bmatrix} \boldsymbol{\Omega}_{\mathbf{xx}} & \boldsymbol{\Omega}_{\mathbf{x}\mathbf{y}} \\ \boldsymbol{\Omega}_{\mathbf{y}\mathbf{x}} & \boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}} \end{bmatrix}, \tag{19.3.1}$$

where $\boldsymbol{\Omega}_{\mathbf{xx}}$ and $\boldsymbol{\Omega}_{\mathbf{yy}}$ are nonsingular, and let r be the rank of $\boldsymbol{\Omega}_{\mathbf{xy}}$. Then there exist two separate transformations

$$\mathbf{u} = \boldsymbol{L}\mathbf{x}, \quad \boldsymbol{v} = \boldsymbol{M}\mathbf{y} \tag{19.3.2}$$

such that

$$\mathcal{V}[\begin{bmatrix}\mathbf{u}\\ \boldsymbol{v}\end{bmatrix}] = \sigma^2 \begin{bmatrix}\boldsymbol{I}_p & \boldsymbol{\Lambda}\\ \boldsymbol{\Lambda}^\top & \boldsymbol{I}_q\end{bmatrix} \tag{19.3.3}$$

where $\boldsymbol{\Lambda}$ is a (usually rectangular) diagonal matrix with only r diagonal elements positive, and the others zero, and where these diagonal elements are sorted in descending order.

Proof: One obtains the matrix $\boldsymbol{\Lambda}$ by a singular value decomposition of $\boldsymbol{\Omega}_{\mathbf{xx}}^{-1/2}\boldsymbol{\Omega}_{\mathbf{xy}}\boldsymbol{\Omega}_{\mathbf{yy}}^{-1/2}$ $\boldsymbol{A}$, say. Let $\boldsymbol{A} = \boldsymbol{P}^\top\boldsymbol{\Lambda}\boldsymbol{Q}$ be its singular value decomposition with fully orthogonal matrices, as in equation (A.9.8). Define $\boldsymbol{L} = \boldsymbol{P}\boldsymbol{\Omega}_{\mathbf{xx}}^{-1/2}$ and $\boldsymbol{M} = \boldsymbol{Q}\boldsymbol{\Omega}_{\mathbf{yy}}^{-1/2}$. Therefore $\boldsymbol{L}\boldsymbol{\Omega}_{\mathbf{xx}}\boldsymbol{L}^\top = \boldsymbol{I}$, $\boldsymbol{M}\boldsymbol{\Omega}_{\mathbf{yy}}\boldsymbol{M}^\top = \boldsymbol{I}$, and $\boldsymbol{L}\boldsymbol{\Omega}_{\mathbf{xy}}\boldsymbol{M}^\top = \boldsymbol{P}\boldsymbol{\Omega}_{\mathbf{xx}}^{-1/2}\boldsymbol{\Omega}_{\mathbf{xy}}\boldsymbol{\Omega}_{\mathbf{yy}}^{-1/2}\boldsymbol{Q}^\top = \boldsymbol{P}\boldsymbol{A}\boldsymbol{Q}^\top = \boldsymbol{\Lambda}$.

The next problems show how one gets from this the maximization property of the canonical correlation coefficients:

Problem 257. *Show that for every p-vector $\boldsymbol{l}$ and q-vector $\boldsymbol{m}$,*

$$\left|\operatorname{corr}(\boldsymbol{l}^\top\mathbf{x}, \boldsymbol{m}^\top\mathbf{y})\right| \leq \lambda_1 \tag{19.3.4}$$

where λ_1 is the first (and therefore biggest) diagonal element of $\boldsymbol{\Lambda}$. Equality in (19.3.4) *holds if $\boldsymbol{l} = \boldsymbol{l}_1$, the first row in $\boldsymbol{L}$, and $\boldsymbol{m} = \boldsymbol{m}_1$, the first row in $\boldsymbol{M}$.*

Answer: If $\boldsymbol{l}$ or $\boldsymbol{m}$ is the null vector, then there is nothing to prove. If neither of them is a null vector, then one can, without loss of generality, multiply them with appropriate scalars so that $\boldsymbol{p} = (\boldsymbol{L}^{-1})^\top\boldsymbol{l}$ and $\boldsymbol{q} = (\boldsymbol{M}^{-1})^\top\boldsymbol{m}$ satisfy $\boldsymbol{p}^\top\boldsymbol{p} = 1$ and $\boldsymbol{q}^\top\boldsymbol{q} = 1$. Then

$$\mathcal{V}[\begin{bmatrix}\boldsymbol{l}^\top\mathbf{x}\\ \boldsymbol{m}^\top\mathbf{y}\end{bmatrix}] = \mathcal{V}[\begin{bmatrix}\boldsymbol{p}^\top\boldsymbol{L}\mathbf{x}\\ \boldsymbol{q}^\top\boldsymbol{M}\mathbf{y}\end{bmatrix}] = \mathcal{V}[\begin{bmatrix}\boldsymbol{p}^\top & \boldsymbol{o}^\top\\ \boldsymbol{o}^\top & \boldsymbol{q}^\top\end{bmatrix}\begin{bmatrix}\mathbf{u}\\ \boldsymbol{v}\end{bmatrix}] = \sigma^2\begin{bmatrix}\boldsymbol{p}^\top & \boldsymbol{o}^\top\\ \boldsymbol{o}^\top & \boldsymbol{q}^\top\end{bmatrix}\begin{bmatrix}\boldsymbol{I}_p & \boldsymbol{\Lambda}\\ \boldsymbol{\Lambda}^\top & \boldsymbol{I}_q\end{bmatrix}\begin{bmatrix}\boldsymbol{p} & \boldsymbol{o}\\ \boldsymbol{o} & \boldsymbol{q}\end{bmatrix} = \sigma^2 \tag{19.3.5}$$

Since the matrix at the righthand side has ones in the diagonal, it is the correlation matrix, i.e., $\boldsymbol{p}^\top\boldsymbol{\Lambda}\boldsymbol{q} = \operatorname{corr}(\boldsymbol{l}^\top\mathbf{x}, \boldsymbol{m}^\top\mathbf{y})$. Therefore (19.3.4) follows from Problem 258.

Problem 258. *If $\sum p_i^2 = \sum q_i^2 = 1$, and $\lambda_i \geq 0$, show that $|\sum p_i\lambda_i q_i| \leq \max\lambda_i$. Hint: first get an upper bound for $|\sum p_i\lambda_i q_i|$ through a Cauchy-Schwartz-type argument.*

Answer. $(\sum p_i\lambda_i q_i)^2 \leq \sum p_i^2\lambda_i \sum q_i^2\lambda_i \leq (\max\lambda_i)^2$. □

Problem 259. *Show that for every p-vector $\boldsymbol{l}$ and q-vector $\boldsymbol{m}$ such that $\boldsymbol{l}^\top\mathbf{x}$ is uncorrelated with $\boldsymbol{l}_1^\top\mathbf{x}$, and $\boldsymbol{m}^\top\mathbf{y}$ is uncorrelated with $\boldsymbol{m}_1^\top\mathbf{y}$,*

$$\left|\operatorname{corr}(\boldsymbol{l}^\top\mathbf{x}, \boldsymbol{m}^\top\mathbf{y})\right| \leq \lambda_2 \tag{19.3.6}$$

where λ_2 is the second diagonal element of $\boldsymbol{\Lambda}$. Equality in (19.3.6) *holds if $\boldsymbol{l} = \boldsymbol{l}_2$, the second row in $\boldsymbol{L}$, and $\boldsymbol{m} = \boldsymbol{m}_2$, the second row in $\boldsymbol{M}$.*

Answer. If $\boldsymbol{l}$ or $\boldsymbol{m}$ is the null vector, then there is nothing to prove. If neither of them is a null vector, then one can, without loss of generality, multiply them with appropriate scalars so that $\boldsymbol{p} = (\boldsymbol{L}^{-1})^\top\boldsymbol{l}$ and $\boldsymbol{q} = (\boldsymbol{M}^{-1})^\top\boldsymbol{m}$ satisfy $\boldsymbol{p}^\top\boldsymbol{p} = 1$ and $\boldsymbol{q}^\top\boldsymbol{q} = 1$. Now write $\boldsymbol{e}_1$ for the first unit vector, which has a 1 as first component and zeros everywhere else:

$$\operatorname{cov}[\boldsymbol{l}^\top\mathbf{x}, \boldsymbol{l}_1^\top\mathbf{x}] = \operatorname{cov}[\boldsymbol{p}^\top\boldsymbol{L}\mathbf{x}, \boldsymbol{e}_1^\top\boldsymbol{L}\mathbf{x}] = \boldsymbol{p}^\top\boldsymbol{\Lambda}\boldsymbol{e}_1 = \boldsymbol{p}^\top\boldsymbol{e}_1\lambda_1. \tag{19.3.7}$$

This covariance is zero iff $p_1 = 0$. Furthermore one also needs the following, directly from the proof of Problem 257:

$$\mathcal{V}[\begin{bmatrix}\boldsymbol{l}^\top\mathbf{x}\\ \boldsymbol{m}^\top\mathbf{y}\end{bmatrix}] = \mathcal{V}[\begin{bmatrix}\boldsymbol{p}^\top\boldsymbol{L}\mathbf{x}\\ \boldsymbol{q}^\top\boldsymbol{M}\mathbf{y}\end{bmatrix}] = \mathcal{V}[\begin{bmatrix}\boldsymbol{p}^\top & \boldsymbol{o}^\top\\ \boldsymbol{o}^\top & \boldsymbol{q}^\top\end{bmatrix}\begin{bmatrix}\mathbf{u}\\ \boldsymbol{v}\end{bmatrix}] = \sigma^2\begin{bmatrix}\boldsymbol{p}^\top & \boldsymbol{o}^\top\\ \boldsymbol{o}^\top & \boldsymbol{q}^\top\end{bmatrix}\begin{bmatrix}\boldsymbol{I}_p & \boldsymbol{\Lambda}\\ \boldsymbol{\Lambda}^\top & \boldsymbol{I}_q\end{bmatrix}\begin{bmatrix}\boldsymbol{p} & \boldsymbol{o}\\ \boldsymbol{o} & \boldsymbol{q}\end{bmatrix} = \sigma^2\begin{bmatrix}\boldsymbol{p}^\top\boldsymbol{p} & \boldsymbol{p}^\top\boldsymbol{\Lambda}\boldsymbol{q}\\ \boldsymbol{q}^\top\boldsymbol{\Lambda}\boldsymbol{p} & \boldsymbol{q}^\top\boldsymbol{q}\end{bmatrix} \tag{19.3.8}$$

Since the matrix at the righthand side has ones in the diagonal, it is the correlation matrix, i.e., $\boldsymbol{p}^\top\boldsymbol{\Lambda}\boldsymbol{q} = \operatorname{corr}(\boldsymbol{l}^\top\mathbf{x}, \boldsymbol{m}^\top\mathbf{y})$. Equation (19.3.6) follows from Problem 258 if one lets the subscript i start at 2 instead of 1. □

Problem 260. *(Not eligible for in-class exams) Extra credit question for good mathematicians: Reformulate the above treatment of canonical correlations without the assumption that $\boldsymbol{\Omega}_{\mathbf{xx}}$ and $\boldsymbol{\Omega}_{\mathbf{yy}}$ are nonsingular.*

19.4. Some Remarks about the Sample Partial Correlation Coefficients

The definition of the partial sample correlation coefficients is analogous to that of the partial population correlation coefficients: Given two data vectors $\mathbf{y}$ and $\mathbf{z}$, and the matrix $\boldsymbol{X}$ (which includes a constant term), and let $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$ be the "residual maker" with respect to $\boldsymbol{X}$. Then the squared partial sample correlation is the squared simple correlation between the least squares residuals:

$$r^2_{\mathbf{zy}.\boldsymbol{X}} = \frac{(\mathbf{z}^\top\boldsymbol{M}\mathbf{y})^2}{(\mathbf{z}^\top\boldsymbol{M}\mathbf{z})(\mathbf{y}^\top\boldsymbol{M}\mathbf{y})} \tag{19.4.1}$$

Alternatively, one can define it as the proportionate reduction in the SSE. Although $\boldsymbol{X}$ is assumed to incorporate a constant term, I am giving it here separately, in order to show the analogy with (19.1.8):

$$r^2_{\mathbf{zy}.\boldsymbol{X}} = \frac{\mathsf{SSE}[\text{constant term and } \boldsymbol{X}; \mathbf{y}] - \mathsf{SSE}[\text{constant term, } \boldsymbol{X}, \text{ and } \mathbf{z}; \mathbf{y}]}{\mathsf{SSE}[\text{constant term and } \boldsymbol{X}; \mathbf{y}]}. \tag{19.4.2}$$

[**Gre97**, p. 248] considers it unintuitive that this can be computed using t-statistics. Our approach explains why this is so. First of all, note that the square of the t-statistic is the F-statistic. Secondly, the formula for the F-statistic for the inclusion of $\boldsymbol{z}$ into the regression is

(19.4.3)

$$t^2 = F = \frac{\mathsf{SSE}[\text{constant term and } \boldsymbol{X}; \mathbf{y}] - \mathsf{SSE}[\text{constant term, } \boldsymbol{X}\text{, and } \boldsymbol{z}; \mathbf{y}]}{\mathsf{SSE}[\text{constant term, } \boldsymbol{X}\text{, and } \boldsymbol{z}; \mathbf{y}]/(n-k-1)}$$

This is very similar to the formula for the squared partial correlation coefficient. From (19.4.3) follows

$$F + n - k - 1 = \frac{\mathsf{SSE}[\text{constant term and } \boldsymbol{X}; \mathbf{y}](n-k-1)}{\mathsf{SSE}[\text{constant term, } \boldsymbol{X}\text{, and } \boldsymbol{z}; \mathbf{y}]} \tag{19.4.4}$$

and therefore

$$r^2_{\boldsymbol{z}\mathbf{y}.\boldsymbol{X}} = \frac{F}{F + n - k - 1} \tag{19.4.5}$$

which is [**Gre97**, (6-29) on p. 248].

It should also be noted here that [**Gre97**, (6-36) on p. 254] is the sample equivalent of (19.1.11).

CHAPTER 20

Numerical Methods for computing OLS Estimates

20.1. QR Decomposition

One precise and fairly efficient method to compute the Least Squares estimates is the $\boldsymbol{QR}$ decomposition. It amounts to going over to an orthonormal basis in $\mathrm{R}[\boldsymbol{X}]$. It uses the following mathematical fact:

Every matrix $\boldsymbol{X}$, which has full column rank, can be decomposed in the product of two matrices $\boldsymbol{QR}$, where $\boldsymbol{Q}$ has the same number of rows and columns as $\boldsymbol{X}$, and is "suborthogonal" or "incomplete orthogonal," i.e., it satisfies $\boldsymbol{Q}^\top\boldsymbol{Q} = \boldsymbol{I}$. The other factor $\boldsymbol{R}$ is upper triangular and nonsingular.

To construct the least squares estimates, make a $\boldsymbol{QR}$ decomposition of the matrix of explanatory variables $\boldsymbol{X}$ (which is assumed to have full column rank). With $\boldsymbol{X} = \boldsymbol{QR}$, the normal equations read

$$\boldsymbol{X}^\top\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^\top\mathbf{y} \tag{20.1.1}$$

$$\boldsymbol{R}^\top\boldsymbol{Q}^\top\boldsymbol{QR}\hat{\boldsymbol{\beta}} = \boldsymbol{R}^\top\boldsymbol{Q}^\top\mathbf{y} \tag{20.1.2}$$

$$\boldsymbol{R}^\top\boldsymbol{R}\hat{\boldsymbol{\beta}} = \boldsymbol{R}^\top\boldsymbol{Q}^\top\mathbf{y} \tag{20.1.3}$$

$$\boldsymbol{R}\hat{\boldsymbol{\beta}} = \boldsymbol{Q}^\top\mathbf{y} \tag{20.1.4}$$

This last step can be made because $\boldsymbol{R}$ is nonsingular. (20.1.4) is a triangular system of equations, which can be solved easily. Note that it is not necessary for this procedure to compute the matrix $\boldsymbol{X}^\top\boldsymbol{X}$, which is a big advantage, since this computation is numerically quite unstable.

Problem 261. *2 points You have a $\boldsymbol{QR}$-decomposition $\boldsymbol{X} = \boldsymbol{QR}$, where $\boldsymbol{Q}^\top\boldsymbol{Q} = \boldsymbol{I}$, and $\boldsymbol{R}$ is upper triangular and nonsingular. For an estimate of $\mathcal{V}[\hat{\boldsymbol{\beta}}]$ you need $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$. How can this be computed without computing $\boldsymbol{X}^\top\boldsymbol{X}$? And why would you want to avoid computing $\boldsymbol{X}^\top\boldsymbol{X}$?*

Answer. $\boldsymbol{X}^\top\boldsymbol{X} = \boldsymbol{R}^\top\boldsymbol{Q}^\top\boldsymbol{QR} = \boldsymbol{R}^\top\boldsymbol{R}$, its inverse is therefore $\boldsymbol{R}^{-1}\boldsymbol{R}^{-1\top}$. □

Problem 262. *Compute the $\boldsymbol{QR}$ decomposition of*

$$\boldsymbol{X} = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 5 & -2 \\ 1 & 1 & 0 \\ 1 & 5 & -4 \end{bmatrix} \tag{20.1.5}$$

Answer.

$$\boldsymbol{Q} = \frac{1}{2}\begin{bmatrix} 1 & -1 & 1 \\ 1 & 1 & 1 \\ 1 & -1 & -1 \\ 1 & 1 & -1 \end{bmatrix} \qquad \boldsymbol{R} = 2\begin{bmatrix} 1 & 3 & -1 \\ 0 & 2 & -2 \\ 0 & 0 & 1 \end{bmatrix} \tag{20.1.6}$$

How to get it? We need a decomposition

$$\begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \boldsymbol{x}_3 \end{bmatrix} = \begin{bmatrix} \boldsymbol{q}_1 & \boldsymbol{q}_2 & \boldsymbol{q}_3 \end{bmatrix}\begin{bmatrix} r_{11} & r_{12} & r_{13} \\ 0 & r_{22} & r_{23} \\ 0 & 0 & r_{33} \end{bmatrix} \tag{20.1.7}$$

where $\boldsymbol{q}_1^\top\boldsymbol{q}_1 = \boldsymbol{q}_2^\top\boldsymbol{q}_2 = \boldsymbol{q}_3^\top\boldsymbol{q}_3 = 1$ and $\boldsymbol{q}_1^\top\boldsymbol{q}_2 = \boldsymbol{q}_1^\top\boldsymbol{q}_3 = \boldsymbol{q}_2^\top\boldsymbol{q}_3 = 0$. First column: $\boldsymbol{x}_1 = \boldsymbol{q}_1 r_{11}$ and $\boldsymbol{q}_1$ must have unit length. This gives $\boldsymbol{q}_1^\top = \begin{bmatrix} 1/2 & 1/2 & 1/2 & 1/2 \end{bmatrix}$ and $r_{11} = 2$. Second column:

$$\boldsymbol{x}_2 = \boldsymbol{q}_1 r_{12} + \boldsymbol{q}_2 r_{22} \tag{20.1.8}$$

and $\boldsymbol{q}_1^\top\boldsymbol{q}_2 = 0$ and $\boldsymbol{q}_2^\top\boldsymbol{q}_2 = 1$. Premultiply (20.1.8) by $\boldsymbol{q}_1^\top$ to get $\boldsymbol{q}_1^\top\boldsymbol{x}_2 = r_{12}$, i.e., $r_{12} = 6$. Thus we know $\boldsymbol{q}_2 r_{22} = \boldsymbol{x}_2 - \boldsymbol{q}_1 \cdot 6 = \begin{bmatrix} -2 & 2 & -2 & 2 \end{bmatrix}^\top$. Now we have to normalize it, to get $\boldsymbol{q}_2 = \begin{bmatrix} -1/2 & 1/2 & -1/2 & 1/2 \end{bmatrix}$ and $r_{22} = 4$. The rest remains a homework problem. But I am not sure if my numbers are right. □

Problem 263. *2 points Compute trace and determinant of* $\begin{bmatrix} 1 & 3 & -1 \\ 0 & 2 & -2 \\ 0 & 0 & 1 \end{bmatrix}$. *Is this matrix symmetric and, if so, is it nonnegative definite? Are its column vectors linearly dependent? Compute the matrix product*

$$\begin{bmatrix} 1 & -1 & 1 \\ 1 & 1 & 1 \\ 1 & -1 & -1 \\ 1 & 1 & -1 \end{bmatrix}\begin{bmatrix} 1 & 3 & -1 \\ 0 & 2 & -2 \\ 0 & 0 & 1 \end{bmatrix} \tag{20.1.9}$$

20.2. The LINPACK Implementation of the QR Decomposition

This is all we need, but numerically it is possible to construct, without much additional computing time, all the information which adds the missing orthogonal columns to $\boldsymbol{Q}$. In this way $\boldsymbol{Q}$ is square and $\boldsymbol{R}$ is conformable with $\boldsymbol{X}$. This is

sometimes called the "complete" $\boldsymbol{QR}$-decomposition. In terms of the decomposition above, we have now

$$\boldsymbol{X} = \begin{bmatrix}\boldsymbol{Q} & \boldsymbol{S}\end{bmatrix}\begin{bmatrix}\boldsymbol{R}\\ \boldsymbol{O}\end{bmatrix} \tag{20.2.1}$$

For every matrix $\boldsymbol{X}$ one can find an orthogonal matrix $\boldsymbol{Q}$ such that $\boldsymbol{Q}^\top\boldsymbol{X}$ has zeros below the diagonal, call that matrix $\boldsymbol{R}$. Alternatively one may say: every matrix $\boldsymbol{X}$ can be written as the product of two matrices $\boldsymbol{QR}$, where $\boldsymbol{R}$ is conformable with $\boldsymbol{X}$ and has zeros below the diagonal, and $\boldsymbol{Q}$ is orthogonal.

To prove this, and also for the numerical procedure, we will build $\boldsymbol{Q}^\top$ as the product of several orthogonal matrices, each converting one column of $\boldsymbol{X}$ into one with zeros below the diagonal.

First note that for every vector $\boldsymbol{v}$, the matrix $\boldsymbol{I} - \frac{2}{\boldsymbol{v}^\top\boldsymbol{v}}\boldsymbol{v}\boldsymbol{v}^\top$ is orthogonal. Given $\boldsymbol{X}$, let $\boldsymbol{x}$ be the first column of $\boldsymbol{X}$. If $\boldsymbol{x} = \boldsymbol{o}$, then go on to the next column. Otherwise choose $\boldsymbol{v} = \begin{bmatrix} x_{11} + \sigma\sqrt{\boldsymbol{x}^\top\boldsymbol{x}} \\ x_{21} \\ \vdots \\ x_{n1}\end{bmatrix}$, where $\sigma = 1$ if $x_{11} \geq 0$ and $\sigma = -1$ otherwise. (Mathematically, either $\sigma - +1$ or $\sigma = -1$ would do; but if one gives σ the same sign as x_{11}, then the first element of $\boldsymbol{v}$ gets largest possible absolute value, which improves numerical accuracy.) Then

$$\boldsymbol{v}^\top\boldsymbol{v} = (x_{11}^2 + 2\sigma x_{11}\sqrt{\boldsymbol{x}^\top\boldsymbol{x}} + \boldsymbol{x}^\top\boldsymbol{x}) + x_{21}^2 + \cdots + x_{n1}^2 \tag{20.2.2}$$

$$= 2(\boldsymbol{x}^\top\boldsymbol{x} + \sigma x_{11}\sqrt{\boldsymbol{x}^\top\boldsymbol{x}}) \tag{20.2.3}$$

$$\boldsymbol{v}^\top\boldsymbol{x} = \boldsymbol{x}^\top\boldsymbol{x} + \sigma x_{11}\sqrt{\boldsymbol{x}^\top\boldsymbol{x}} \tag{20.2.4}$$

therefore $2\boldsymbol{v}^\top\boldsymbol{x}/\boldsymbol{v}^\top\boldsymbol{v} = 1$, and

$$(\boldsymbol{I} - \frac{2}{\boldsymbol{v}^\top\boldsymbol{v}}\boldsymbol{v}\boldsymbol{v}^\top)\boldsymbol{x} = \boldsymbol{x} - \boldsymbol{v} = \begin{bmatrix} -\sigma\sqrt{\boldsymbol{x}^\top\boldsymbol{x}} \\ 0 \\ \vdots \\ 0\end{bmatrix}. \tag{20.2.5}$$

Premultiplication of $\boldsymbol{X}$ by $\boldsymbol{I} - \frac{2}{\boldsymbol{v}^\top\boldsymbol{v}}\boldsymbol{v}\boldsymbol{v}^\top$ gets therefore the first column into the desired shape. By the same principle one can construct a second vector $\boldsymbol{w}$, which has a zero in the first place, and which annihilates all elements below the second element in the second column of $\boldsymbol{X}$, etc. These successive orthogonal transformations will convert $\boldsymbol{X}$ into a matrix which has zeros below the diagonal; their product is therefore $\boldsymbol{Q}^\top$.

The LINPACK implementation of this starts with $\boldsymbol{X}$ and modifies its elements in place. For each column it generates the corresponding $\boldsymbol{v}$ vector and premultipies the

matrix by $\boldsymbol{I} - \frac{2}{\boldsymbol{v}^\top\boldsymbol{v}}\boldsymbol{v}\boldsymbol{v}^\top$. This generates zeros below the diagonal. Instead of writing the zeros into that matrix, it uses the "free" space to store the vector $\boldsymbol{v}$. There is almost enough room; the first nonzero element of $\boldsymbol{v}$ must be stored elsewhere. This is why the $\boldsymbol{QR}$ decomposition in `Splus` has two main components: `qr` is a matrix like `a`, and `qraux` is a vector of length `ncols(a)`.

LINPACK does not use or store exactly the same $\boldsymbol{v}$ as given here, but uses $\boldsymbol{u} = \boldsymbol{v}/(\sigma\sqrt{\boldsymbol{x}^\top\boldsymbol{x}})$ instead. The normalization does not affect the resulting orthogonal transformation; its advantage is that the leading element of each vector, that which is stored in `qraux`, is at the same time equal $\boldsymbol{u}^\top\boldsymbol{u}/2$. In other words, `qraux` doubles up as the divisor in the construction of the orthogonal matrices.

In `Splus` type `help(qr)`. At the end of the help file a program is given which shows how the $\boldsymbol{Q}$ might be constructed from the fragments `qr` and `qraux`.

CHAPTER 21

About Computers

21.1. General Strategy

With the fast-paced development of computer hardware and software, anyone who uses computers profesionally needs a strategy about how to allocate their time and money for hardware and software.

21.1.1. Operating System. In my view, there are two alternatives today: either do everything in Microsoft Windows and other commercial software, or use GNU/Linux, the free unix operating system together with the free software built on top of it, see www.linux.org, in addition to Microsoft Windows. I will argue here for the second route. It is true, GNU/Linux has a steeper learning curve than Windows, but this also means that you have a more powerful tool, and serious efforts are under way to make GNU/Linux more and more user friendly. Windows, on the other hand, has the following disadvantages:

- Microsoft Windows and the other commercial software are expensive.
- The philosophy of Microsoft Windows is to keep the user in the dark about how the computer is working, i.e., turn the computer user into a passive consumer. This severely limits the range of things you can do with your computer. The source code of the programs you are using is usually unavailable, therefore you never know exactly what you are doing and you cannot modify the program for your own uses. The unavailability of source code also makes the programs more vulnerable to virus attacks and breakins. In Linux, the user is the master of the computer and can exploit its full potential.
- You spend too much time pointing and clicking. In GNU/Linux and other unix systems, it is possible to set up menus too,m but everything that can be done through a menu can also be done on the command line or through a script.
- Windows and the commercial software based on it are very resource-hungry; they require powerful computers. Computers which are no longer fast and big enough to run the latest version of Windows are still very capable to run Linux.
- It is becoming more and more apparent that free software is more stable and of higher quality than commercial software. Free software is developed by programmers throughout the world who want good tools for themselves.
- Most Linux distributions have excellent systems which allows the user to automatically download always the latest versions of the software; this automates the tedious task of software maintenance, i.e., updating and fitting together the updates.

Some important software is not available on Linux or is much better on Windows. Certain tasks, like scanning, voice recognition, and www access, which have mass markets, are better done on Microsoft Windows than on Linux. Therefore you will probably not be able to eliminate Microsoft Windows completely; however it is possible to configure your PC so that you can run MS-Windows and Linux on it, or to have a Linux machine be the network server for a network which has Windows machines on it (this is more stable, faster, and cheaper than Windows NT).

There are several versions of Linux available, and the one which is most independent of commercial interests, and which is one of the most quality-conscious distributions, in my view, is Debian GNU/Linux, `http://www.debian.org`. The Linux route is more difficult at the beginning but will pay off in the long run, and I recommend it especially if you are going to work outside the USA. The Salt Lake Linux Users Group `http://www.sllug.org/index.html` meets on the third Wednesday of every month, usually on the University of Utah campus.

In order to demonstrate the usefulness of Linux I loaded Debian GNU/Linux on an old computer with one of the early Pentium processors, which became available at the Econ Department because it was too slow for Windows 98. It is by the window in the Econ Computer Lab. When you log onto this computer you are in the X-windows system. In Linux and other unix systems, the mouse usually has 3 buttons: left, right, and middle. The mouse which comes with the computer in the computer lab has 2 bottons: left and right, but if you press both buttons simultaneously you get the same effect as pressing the middle button on a unix mouse.

If the cursor is in front of the background, then you will get 3 different menus by pressing the different mouse buttons. The left mouse button gives you the different programs, if you press both buttons at the same time you can perform operations on the windows, and the right button gives you a list of all open windows.

Another little tidbit you need to know about unix systems is this: There are no drives as in Microsoft Dos or Windows, but all files are in one hierarchical directory tree. Instead of a backslash \ you have a forward slash `/`. In order to use the floppy disk, you have to insert the disk in the disk drive and then give the command `mount /floppy`. Then the disk is accessible to you as the contents of the directory `/floppy`.

Before taking the disk out you should give the command `umount /floppy`. You can do this only if `/floppy` is not the current directory.

In order to remotely access X-windows from Microsoft-Windows, you have to go through the following steps.

- click on the exceed icon which is in the network-neighborhood folder.
- then open a telnet session to the unix station you want to access.
- at the unix station give the `who -l` command so that you know the id of the machine from which you are telnetting from; assume it is `econlab9.econ.utah.`
- then give the command (if you are in a bash shell as you probably will be if it is linux)

  ```
  DISPLAY=econlab9.econ.utah.edu:0; export DISPLAY
  ```

  or, if it is the C-shell:

  ```
  setenv DISPLAY econlab9.econ.utah.edu:0
  DISPLAY=buc-17.econ.utah.edu:0; export DISPLAY
  ```

Something else: if I use the usual telnet program which comes with windows, in order to telnet into a unix machine, and then I try to edit a file using emacs, it does not work, it seems that some of the key sequences used by emacs make telnet hang. Therefore I use a different telnet program, `Teraterm Pro`, with downloading instructions at `http://www.egr.unlv.ecu/stock_answers/remote_access/install_ttssh.ht`

21.1.2. Application Software. I prefer learning a few pieces of software well instead of learning lots of software superficially. Therefore the choice of software is an especially important question.

I am using the editor `emacs` for reading mail, for writing papers which are then printed in TEX, for many office tasks, such as appointment calendar, address book, etc., for browsing the www, and as a frontend for running SAS or R/Splus and also the shell and C. `Emacs` shows that free software can have unsurpassed quality. The webpage for GNU is www.gnu.org.

With personal computers becoming more and more powerful, `emacs` and much of the Gnu-software is available not only on unix systems but also on Windows. As a preparation to a migration to Linux, you may want to install these programs on Microsoft Windows first. On the other hand, netscape and wordperfect are now both available for free on Linux.

Besides `emacs` I am using the typesetting system TEX, or, to be precise, the TEX-macro-package $\mathcal{A}\mathcal{M}\mathcal{S}$-LATEX. This is the tool which mathematicians use to write their articles and books, and many econometrics and statistics textbooks was written using TEX. Besides its math capabilities, another advantage of TEX is that it supports many different alphabets and languages.

For statistical software I recommend the combination of SAS and `Splus`, and it is easy to have a copy of the GNU-version of `Splus`, called R, on your computer. R is

not as powerful as `Splus`, but it is very similar, in the simple tasks almost identical. There is also a GNU version of `SPSS` in preparation.

21.1.3. Other points. With modern technology it is easy to keep everything you ever write, all your class notes, papers, book excerpts, etc. It will just take one or perhaps a handful of CD-roms to have it available, and it allows you greater continuity in your work.

In my view, windowing systems are overrated: they are necessary for web browsing or graphics applications, but I am still using character-based terminals most of the time. I consider them less straining on the eye, and in this way I also have worldwide access to my unix account through telnet. Instead of having several windows next to each other I do my work in several emacs buffers which I can display at will (i.e., the windows are on top of each other, but if necessary I can also display them side by side on the screen).

In an earlier version of these notes, in 1995, I had written the following:

> I do not consider it desirable to have a computer at home in which I buy and install all the software for myself. The installation of the regular updates, and then all the adjustments that are necesary so that the new software works together again like the old software did, is a lot of work, which should be centralized. I keep all my work on a unix account at the university. In this way it is accessible to me wherever I go, and it is backed up regularly.

In the meanwhile, I changed my mind about that. After switching to Debian GNU/Linux, with its excellent automatic updating of the software, I realized how outdated the unix workstations at the Econ Department have become. My Linux workstations have more modern software than the Sun stations. In my own situation as a University Professor, there is an additional benefit if I do my work on my own Linux workstation at home: as long as I am using University computers, the University will claim copyright for the software which I develop, even if I do it on my own time. If I have my own Linux workstation at home, it is more difficult for the University to appropriate work which they do not pay for.

21.2. The Emacs Editor

You can use emacs either on a character-based terminal or in X-windows. On a character-based terminal you simply type `emacs`. In a windows setting, it is probably available in one of the menus, but you can also get into it by just typing `emacs &` in one of the x-terminal windows. The ampersand means that you are running emacs in the "background." This is sufficient since emacs opens its own window. If you issue the command without the ampersand, then the X-terminal window from which

you invoked `local` will not accept any other commands, i.e., will be useless, until you leave emacs again.

The emacs commands which you have to learn first are the help commands. They all start with a `C-h`, i.e., control-`h`: type `h` while holding the control button down. The first thing you may want to do at a quiet moment is go through the `emacs` tutorial: get into `emacs` and then type `C-h t` and then follow instructions. Another very powerful resource at your fingertip is emacs-info. To get into it type `C-h i`. It has information pages for you to browse through, not only about `emacs` itself, but also a variety of other subjects. The parts most important for you is the `Emacs` menu item, which gives the whole Emacs-manual, and the `ESS` menu item, which explains how to run `Splus` and SAS from inside `emacs`.

Another important `emacs` key is the "quit" command `C-g`. If you want to abort a command, this will usually get you out. Also important command is the changing of the buffer, `C-x b`. Usually you will have many buffers in `emacs`, and switch between them if needed. The command `C-x C-c` terminates emacs.

Another thing I recommend you to learn is how to send and receive electronic mail from inside emacs. To send mail, give the command `C-x m`. Then fill out address and message field, and send it by typing `C-c C-c`. In order to receive mail, type `M-x rmail`. There are a few one-letter commands which allow you to move around in your messages: `n` is next message, `p` is previous message, `d` is delete the message, `r` means: reply to this message.

21.3. How to Enter and Exit SAS

From one of the computers on the Econ network, go into the Windows menu and double-click on the SAS icon. It will give you two windows, the command window on the bottom and a window for output on the top. Type your commands into the command window, and click on the button with the runner on it in order to submit the commands.

If you log on to the workstation `marx` or `keynes`, the first command you have to give is `openwin` in order to start the X-window-system. Then go to the `local` window and give the command `sas &`. The ampersand means that `sas` is run in the background; if you forget it you won't be able to use the `local` window until you exist `sas` again. As SAS starts up, it creates 3 windows, and you have to move those windows where you want them and then click the left mouse button.

From any computer with telnet access, get into the DOS prompt and then type `telnet marx.econ.utah.edu`. Then sign on with your user-id and your password, and then issue the command `sas`. Over telnet, those SAS commands which use function keys etc. will probably not work, and you have to do more typing. SAS over telnet is more feasible if you use SAS from inside emacs for instance.

The book [**Ell95**] is a simple introduction into SAS written by an instructor of the University of Utah and used by Math 317/318.

21.4. How to Transfer SAS Data Sets Between Computers

The following instructions work even if the computers have different operating systems. In order to transfer all SAS data files in the `/home/econ/ehrbar/sas` directory on `smith` to your own computer, you have to first enter SAS on `smith` and give the following commands:

```
libname ec7800 '/home/econ/ehrbar/ec7800/sasdata';
proc cport L=ec7800;
run;
```

This creates a file in the directory you were in when you started SAS (usually your home directory) by the name `sascat.dat`. Then you must transport the file `sascat.dat` to your own computer. If you want to put it onto your account on the novell network, you must log to your novell account and ftp from there to `smith` and get the file this way. For this you have to login to your account and then `cd ehrbar/ec7800/sasdata`. and then first give the command `binary` because it is a binary file, and then `get sascat.dat`. Or you can download it from the www by `http://www.cc.utah.edu/ ehrbar/sascat.dat`. but depending on your web browser it may not arrive in the right format. And the following SAS commands deposit the data sets into your directory `sasdata` on your machine:

```
libname myec7800  'mysasdata';
proc cimport L=myec7800;
run;
```

21.5. Instructions for Statistics 5969, Hans Ehrbar's Section

21.5.1. How to Download and Install the free Statistical Package R. The main archive for R is at `http://cran.r-project.org`, and the mirror for the USA is at `http://cran.us.r-project.org`. Here are instructions, current as of May 30, 2001, how to install R on a Microsoft Windows machine: click on "Download R for Windows"; this leads you into a directory; go to the subdirectory "base" and from there download the two file `SetupR.exe`. I.e., from Microsoft Internet Explorer right-click on the above link and choose the menu option: "save target as." It will ask you where to save it; the default will probably be a file of the same name in the "My Documents" folder, which is quite alright.

The next step is to run SetupR.exe. For this it close Internet Explorer and any other applications that may be running on your computer. Then go into the Start Menu, click on "Run", and then click on "Browse" and find the file SetupR.exe in the "My Documents" folder, and press OK to run it.

It may be interesting for you to read the license, which is the famous and influential GNU Public License.

Then you get to a screen "Select Destination Directory". It is ok to choose the default `C:\Program Files\R\rw1023`, click on Next.

Then it asks you to select the components to install, again the default is fine, but you may choose more or fewer components.

Under "Select Start Menu Folder" again select the default.

You may also want to install `wget` for windows from `http://www.stats.ox.ac.uk/pub/Rtools/wget.zip`. Interesting is also the FAQ at `http://www.stats.ox.ac.uk/pub/R/rw-FAQ.html`.

21.5.2. The text used in Stat 5969. This text is the R-manual called "An Introduction to R" version 1.2.3 which you will have on your computer as a pdf file after installing R. If you used all the defaults above, the path is `C:\Program Files\R\rw1023\doc\manual\R-intro.pdf`. This manual is also on the www at `http://cran.us.r-project.org/doc/manuals/R-intro.pdf`.

21.5.3. Syllabus for Stat 5969. Wednesday June 13: Your reading assignment for June 13 is some background reading about the GNU-Project and the concept of Free Software. Please read `http://www.fsf.org/gnu/thegnuproject.html`. There will be a mini quiz on Wednesday testing whether you have read it. In class we will go through the Sample Session pp. 80–84 in the Manual, and then discuss the basics of the R language, chapters 1–6 of the Manual. The following homework problems apply these basic language features:

PROBLEM 264. *3 points In the dataset* `LifeCycleSavings`*, which* R*-command returns a vector with the names of all countries for which the savings rate is smaller than 10 percent.*

ANSWER. `row.names(LifeCycleSavings)[LifeCycleSavings$sr < 10]`. □

PROBLEM 265. *6 points* `x <- 1:26; names(x) <- letters; vowels <- c("a", "e", "i", "o", "u'')` *Which* R*-expression returns the subvector of* `x` *corresponding to all consonants?*

ANSWER. `x[-x[vowels]]` □

PROBLEM 266. *4 points* `x` *is a numerical vector. Construct the vector of first differences of* `x`*, whose ith element is* $x_i - x_{i-1}$ *(*$i > 2$*), and whose first element is* `NA`*. Do not use the function* `diff(x)` *but the tools described in Section 2.7 of* R*-intro.*

ANSWER. `x-c(NA, x[-1])` or `c(NA, x[-1]-x[-length(x)])` □

PROBLEM 267. *2 points* `x` *is a vector with missing values. which* R*-expression replaces all missing values by 0?*

ANSWER. `x[is.na(x)] <- 0` or `ifelse(is.na(x), 0, x.` □

PROBLEM 268. *2 points Use* `paste` *to get the character vector* `"1999:1" "1999:2" "1999:3" "1999:4"`

ANSWER. `paste(1999, 1:4, sep=":")` □

PROBLEM 269. *5 points Do the exercise described on the middle of p. 17, i.e., compute the 95 percent confidence limits for the state mean incomes. You should be getting the following intervals:*

act	*nsw*	*nt*	*qld*	*sa*	*tas*	*vic*	*wa*
63.56	*68.41*	*112.68*	*65.00*	*63.72*	*66.85*	*70.56*	*60.71*
25.44	*46.25*	*-1.68*	*42.20*	*46.28*	*54.15*	*41.44*	*43.79*

ANSWER. `state <- c("tas", "sa", "qld", "nsw", "nsw", "nt", "wa", "wa", "qld", "vic", "nsw", "vic", "qld", "qld", "sa", "tas", "sa", "nt", "wa", "vic", "qld", "nsw", "nsw", "wa", "sa", "act", "nsw", "vic", "vic", "act"); statef <- factor(state); incomes <- c(60, 49, 40, 61, 64, 60, 59, 54, 62, 69, 70, 42, 56, 61, 61, 61, 58, 51, 48, 65, 49, 49, 41, 48, 52, 46, 59, 46, 58, 43); incmeans <- tapply(incomes, statef, mean); stderr <- function(x) sqrt(var(x)/length(x)); incster <- tapply(incomes, statef, stderr); sampsize <- tapply(incomes, statef, length);` Use 2-tail 5 percent, each tail has 2.5 percent: `critval <- qt(0.975,sampsize-1); conflow <- incmeans - critval * incster; confhigh <- incmeans + critval * incster;` To print the confidence intervals use `rbind(confhigh, conflow)` which gives the following output:

	act	nsw	nt	qld	sa	tas	vic	wa
confhigh	63.55931	68.41304	112.677921	65.00034	63.7155	66.8531	70.5598	60.70747
conflow	25.44069	46.25363	-1.677921	42.19966	46.2845	54.1469	41.4402	43.79253

□

PROBLEM 270. *4 points Use the* `cut` *function to generate a factor from the variable* `ddpi` *in the data frame* `LifeCycleSavings`*. This factor should have the three levels* `low` *for values* `ddpi` ≤ 3*, medium for values* $3 <$ `ddpi` ≤ 6*, and* `high` *for the other values.*

ANSWER. `cut(LifeCycleSavings$ddpi, c(0,3,6,20), c("low", "medium", "high"))` □

Monday June 18: graphical procedures, chapter 12. Please read this chapter before coming to class, there will be a mini quiz again. For the following homework it is helpful to do `demo(graphics)` and to watch closely which commands were used there.

PROBLEM 271. *5 points The data frame* `LifeCycleSavings` *has some egregious outliers. Which plots allow you to identify those? Use those plots to determine which of the data you consider outliers.*

ANSWER. Do `pairs(LifeCycleSavings)` and look for panels which have isolated points. In order to see which observation this is, do `attach(LifeCycleSavings)`, then `plot(sr,ddpi)`, then `identify(sr,ddpi)`. You see that 49 is clearly an outlier, and perhaps 47 and 23. Looking at some other panels in the scatter plot matrix you will find that 49 always stands out, with also 47 and 44. □

PROBLEM 272. *5 points* `x <- 1:40 + rnorm(40) + c(1,3,0,-4)` *Assume* `x` *is quarterly data. Make a plot of* `x` *in which each of the seasons is marked by a hollow dot filled in with a different color.*

ANSWER. `plot(x, type="n"); lines(x, lty="dotted"); points(x, bg=c("tan", "springgreen' "tomato", "orange"), pch= 21)` □

Wednesday June 20: More language features, chapters 6–10, and the beginning of statistical models, chapter 11. A Mini Quiz will check that you read chapters 6–10 before coming to class. Homework is an estimation problem.

Monday June 25: Mini Quiz about chapter 11. We will finish chapter 11. After this session you will have a take-home final exam for this part of the class, using the features of R. It will be due on Monday, July 2nd, at the beginning of class.

If you have installed `wget` in a location R can find it in (I think no longer necessary).

In unix, it is possible to start R or `Splus` just by typing R or `Splus`, whether you are in the X-windows system or on a character-based terminal.

But for serious work I prefer to run it from inside the editor `emacs`. Emacs provides a very convenient front end for `Splus` and `SAS` (and other languages will be added in the future). After entering Emacs, all you have to do is type `M-x S` (for `Splus` version 5 which we have on our workstations) or `M-x SAS` (for `SAS`). Here `M-x` means meta-`x`. On the workstations, the meta-key is the key to the left of the space bar. It works like the control key. Hold down this key and then type `x`. If you telnet in from your own computer, you need a two-key sequence for all meta-characters: first type the escape-key, then release it and then type `x`. If you do `M-x S` or `M-x SAS`, emacs will ask you: "from which directory?" This is the directory to which you would have cd'd before starting up `Splus` or `SAS`. Just type a return as a response, in this way your home directory will be the default directory. Then you can type and submit the `Splus`-commands given below from inside `emacs`.

Here are some common procedures for `Splus`: To dump a function into an edit buffer do `C-c C-d`, to compile it do `C-c C-l`, for parsing errors `C-x ‘`, for help about R/Splus `C-c C-v`, and for help on ess `C-h i`, and then `m ESS`.

The interface with `SAS` is at this point less well developed than that with `Splus`. You have to write a file with your `SAS`-commands in it, typically it is called `myfile.sas`. The file name extension should conventionally be `sas`, and if it is, emacs will help you writing the `SAS` code with the proper indentation. Say you have such a sas file in your current buffer and you want to submit it to `SAS`. First do `M-x SAS` to start `SAS`. This creates some other windows but your cursor should stay in the original window with the sas-file. Then to `C-c C-b` to submit the whole buffer to `SAS`.

There are some shortcuts to switch between the buffers: `C-c C-t` switches you into `*SAS.lst*` which lists the results of your computation.

For further work you may have to create a `region` in your buffer; go to the beginning of the region and type `C-@` (emacs will respond with the message in the minibuffer: "mark set"), and then go to the end of the region. Before using the region for editing, it is always good to do the command `C-x C-x` (which puts the cursor where the mark was and the marker where the cursor was) to make sure the region is what you want it to be. There is apparently a bug in many emacs versions where the point jumps by a word when you do it the first time, but when you correct it then it will stay. Emacs may also be configured in such a way that the region becomes inactive if other editing is done before it is used; the command `C-x C-x` re-activates the region. Then type `C-c C-r` to submit the region to the `SAS` process.

In order to make high resolution gs-plots, you have to put the following two lines into your batch files. For interactive use on X-terminals you must comment them out again (by putting `/*` in front and `*/` after them).

```
filename grafout 'temp.ps';
goptions device=ps gsfname=grafout gsfmode=append gaccess=sasgastd;
```

The emacs interface for `Splus` is much more sophisticated. Here are some commands to get you started. Whenever you type a command on the last line starting with `>` and hit return, this command will be submitted to `Splus`. The key combination `M-p` puts the previous command on the last line with the prompt; you may then edit it and resubmit it simply by typing the return key (the cursor does not have to be at the end of the line to do this). Earlier commands can be obtained by repeated `M-p`, and `M-n` will scroll the commands in the other direction. `C-c C-v` will display the help files for any object of your choice in a split screen. This is easy to remember, the two keys are right next to each other, and you will probably use this key sequence a lot. You can use the usual emacs commands to switch between buffers. Inside `S-mode` there is name completion for all objects, by just typing the tab key. There are very nice commands which allow you to write and debug your own `Splus`-functions. The command `C-c C-d` "dumps" a `Splus`-object into a separate buffer, so that you can change it with the editor. Then when you are done, type `C-c C-l` to "load" the new code. This will generate a new `Splus`-object, and if this is successful, you no longer need the special edit buffer. These are well designed powerful tools, but you have to study them, by accessing the documentation about `S-mode` in Emacs-info. They cannot be learned by trial and error, and they cannot be learned in one or two sessions.

If you are sitting at the console, then you must give the command `openwin()` to tell `Splus` to display high resolution graphs in a separate window. You will get a postscript printout simply by clicking the mouse on the `print` button in this window.

If you are logged in over telnet and access `Splus` through `emacs`, then it is possible to get some crude graphs on your screen after giving the command `printer(width=79)`. Your plotting commands will not generate a plot until you give the command `show()` in order to tell `Splus` that now is the time to send a character-based plot to the screen.

Splus has a very convenient routine to translate `SAS`-datasets into `Splus`-datasets. Assume there is a `SAS` dataset `cobbdoug` in the unix directory `/home/econ/ehrbar/ec780` i.e., this dataset is located in a unix file by the name `/home/econ/ehrbar/ec7800/sasdat` Then the `Splus`-command `mycobbdoug <- sas.get("/home/econ/ehrbar/ec7800/sa` `"cobbdoug")` will create a `Splus`-dataframe with the same data in it.

In order to transfer Splus-files from one computer to another, use the `data.dump` and `data.restore` commands.

To get out of `Splus` again, issue the command `C-c C-q`. It will ask you if you want all temporary files and buffers deleted, and you should answer `yes`. This will *not* delete the buffer with your `Splus`-commands in it. If you want a record of your `Splus`-session, you should save this buffer in a file, by giving the command `C-x C-s` (it will prompt you for a filename).

By the way, it is a good idea to do your unix commands through an emacs buffer too. In this way you have a record of your session and you have easier facilities to recall commands, which are usually the same as the commands you use in your `*S*`-buffer. To do this you have to give the command `M-x shell`.

Books on `Splus` include the "Blue book" [**BCW96**] which unfortunately does not discuss some of the features recently introduced into S, and the "White book" [**CH93**] which covers what is new in the 1991 release of S. The files `book.errata` and `model.errata` in the directory `/usr/local/splus-3.1/doc/` specify known errors in the Blue and White book.

Textbooks for using `Splus` include [**VR99**] which has an url www.stats.oz.ac.uk/pub/MAS [**Spe94**], [**Bur98**] (downloadable for free from the internet), and [**Eve94**].

`R` has now a very convenient facility to automatically download and update packages from CRAN. Look at the help page for `update.packages`.

21.6. The Data Step in SAS

We will mainly discuss here how to create new SAS data sets from already existing data sets. For this you need the `set` and `merge` statements.

Assume you have a dataset `mydata` which includes the variable `year`, and you want to run a regression procedure only for the years 1950–59. This you can do by including the following data step before running the regression:

```
data fifties;
   set mydata;
   if 1950 <= year <= 1959;
```

This works because the data step executes every command once for every observation. When it executes the `set` statement, It starts with the first observation and includes every variable from the data set `mydata` into the new data set `fifties`; but if the expression `1950 <= year <= 1959` is not true, then it throws this observation out again.

Another example is: you want to transform some of the variables in your data set. For instance you want to get aggregate capital stock, investment, and output for all industries. Then you might issue the commands:

```
data aggregate;
   set ec781.invconst;
   kcon00=sum(of kcon20-kcon39);
   icon00=sum(of icon20-icon39);
   ocon00=sum(of ocon20-ocon39);
   keep kcon00, icon00, ocon00, year;
```

The `keep` statement tells SAS to drop all the other variables, otherwise all variables in `ec781.invconst` would also be in `aggregate`.

Assume you need some variables from `ec781.invconst` and some from `ec781.invmis` Let us assume both have the same variable `year`. Then you can use the `merge` statement:

```
data mydata;
    merge ec781.invcost ec781.invmisc;
    by year;
    keep kcon20, icon20, ocon20, year, prate20, primeint;
```

For this step it is sometimes necessary to rename variables before merging. This can be done by the `rename` option.

The `by` statement makes sure that the years in the different datasets do not get mixed up. This allows you to use the `merge` statement also to get variables from the Citybase, even if the starting end ending years are not the same as in our datasets.

An alternative, but not so good method would be to use two set statements:

```
data mydata;
    set ec781.invcost;
    set ec781.invmisc;
    keep kcon20, icon20, ocon20, year, prate20, primeint;
```

If the year variable is in both datasets, SAS will first take the year from `invconst`, and overwrite it with the year data from `invmisc`, but it will not check whether the years match. Since both datasets start and stop with the same year, the result will still be correct.

If you use only one `set` statement with two datasets as arguments, the result will not be what you want. The following is therefore wrong:

```
data mydata;
    set ec781.invcost ec781.invmisc;
    keep kcon20, icon20, ocon20, year, prate20, primeint;
```

Here SAS first reads all observations from the first dataset and then all observations from the second dataset. Those variables in the first dataset which are not present in the second dataset get missing values for the second dataset, and vice versa. So you would end up with the variable `year` going twice from 1947 to 1985, and the variables `kcon20` having 39 missing values at the end, and `prate` having 39 missing values at the beginning.

People who want to use some Citibase data should include the following options on the `proc citibase` line: `beginyr=47 endyr=85`. If their data starts later, they will add missing values at the beginning, but the data will still be lined up with your data.

The `retain` statement tells SAS to retain the value of the variable from one loop through the data step to the next (instead of re-initializing it as a missing value.) The variable `monthtot` initially contains a missing value; if the data set does not

start with a January, then the total value for the first year will be a missing value, since adding something to a missing value gives a missing value again. If the dataset does not end with a December, then the (partial) sum of the months of the last year will not be read into the new data set.

The variable `date` which comes with the citibase data is a special data type. Internally it is the number of days since Jan 1, 1960, but it prints in several formats directed by a `format` statement which is automatically given by the citibase procedure. In order to get years, quarters, or months, use `year(date)`, `qtr(date)`, or `month(date)`. Therefore the conversion of monthly to yearly data would be now:

```
data annual;
    set monthly;
    retain monthtot;
    if month(date)=1 then monthtot=0;
    monthtot=monthtot+timeser;
    if month(date)=12 then output;
    yr=year(date);
    keep yr monthtot;
```

CHAPTER 22

Specific Datasets

22.1. Cobb Douglas Aggregate Production Function

PROBLEM 273. *2 points The Cobb-Douglas production function postulates the following relationship between annual output* q_t *and the inputs of labor* ℓ_t *and capital* k_t*:*

$$\mathsf{q}_t = \mu \ell_t^{\beta} k_t^{\gamma} \exp(\varepsilon_t). \tag{22.1.1}$$

q_t*,* ℓ_t*, and* k_t *are observed, and* μ*,* β*,* γ*, and the* ε_t *are to be estimated. By the variable transformation* $\mathsf{x}_t = \log \mathsf{q}_t$*,* $y_t = \log \ell_t$*,* $z_t = \log k_t$*, and* $\alpha = \log \mu$*, one obtains the linear regression*

$$\mathsf{x}_t = \alpha + \beta y_t + \gamma z_t + \varepsilon_t \tag{22.1.2}$$

Sometimes the following alternative variable transformation is made: $\mathsf{u}_t = \log(\mathsf{q}_t/\ell_t)$*,* $v_t = \log(k_t/\ell_t)$*, and the regression*

$$\mathsf{u}_t = \alpha + \gamma v_t + \varepsilon_t \tag{22.1.3}$$

is estimated. How are the regressions (22.1.2) *and* (22.1.3) *related to each other?*

ANSWER. Write (22.1.3) as

$$\mathsf{x}_t - y_t = \alpha + \gamma(z_t - y_t) + \varepsilon_t \tag{22.1.4}$$

and collect terms to get

$$\mathsf{x}_t = \alpha + (1-\gamma) y_t + \gamma z_t + \varepsilon_t \tag{22.1.5}$$

From this follows that running the regression (22.1.3) is equivalent to running the regression (22.1.2) with the constraint $\beta + \gamma = 1$ imposed. □

The assumption here is that output is the only random variable. The regression model is based on the assumption that the dependent variables have more noise in them than the independent variables. One can justify this by the argument that any noise in the independent variables will be transferred to the dependent variable, and also that variables which affect other variables have more steadiness in them than variables which depend on others. This justification often has merit, but in the specific case, there is much more measurement error in the labor and capital inputs

than in the outputs. Therefore the assumption that only the output has an error term is clearly wrong, and problem 275 below will look for possible alternatives.

PROBLEM 274. *Table 1 shows the data used by Cobb and Douglas in their original article* [**CD28**] *introducing the production function which would bear their name.* `output` *is "Day's index of the physical volume of production (1899 = 100)" described in* [**DP20**]*,* `capital` *is the capital stock in manufacturing in millions of 1880 dollars* [**CD28**, p. 145]*,* `labor` *is the "probable average number of wage earners employed in manufacturing"* [**CD28**, p. 148]*, and* `wage` *is an index of the real wage (1899–1908 = 100).*

year	1899	1900	1901	1902	1903	1904	1905	1906	1907	1908	1909	1910
output	100	101	112	122	124	122	143	152	151	126	155	159
capital	4449	4746	5061	5444	5806	6132	6626	7234	7832	8229	8820	9240
labor	4713	4968	5184	5554	5784	5468	5906	6251	6483	5714	6615	6807
wage	99	98	101	102	100	99	103	101	99	94	102	104

year	1911	1912	1913	1914	1915	1916	1917	1918	1919	1920	1921	1922
output	153	177	184	169	189	225	227	223	218	231	179	240
capital	9624	10067	10520	10873	11840	13242	14915	16265	17234	18118	18542	19192
labor	6855	7167	7277	7026	7269	8601	9218	9446	9096	9110	6947	7602
wage	97	99	100	99	99	104	103	107	111	114	115	119

TABLE 1. Cobb Douglas Original Data

• a. *A text file with the data is available on the web at www.econ.utah.edu/ehrbar/data/cobbdoug.txt, and a SDML file (XML for statistical data which can be read by* R*, Matlab, and perhaps also SPSS) is available at www.econ.utah.edu/ehrbar/data/cobbdoug.sdml. Load these data into your favorite statistics package.*

ANSWER. In R, you can simply issue the command `cobbdoug <- read.table("http://www.econ.utah.edu/ehrbar/data/cobbdoug.txt", header=TRUE)`. If you run R on unix, you can also do the following: download `cobbdoug.sdml` from the www, and then first issue the command `library(StatDataML)` and then `readSDML("cobbdoug.sdml")`. When I tried this last, the `XML` package necessary for `StatDataML` was not available on windows, but chances are it will be when you read this.

In SAS, you must issue the commands

```
data cobbdoug;
    infile 'cobbdoug.txt';
    input year output capital labor;
run;
```

But for this to work you must delete the first line in the file `cobbdoug.txt` which contains the variable names. (Is it possible to tell SAS to skip the first line?) And you may have to tell SAS

the full pathname of the text file with the data. If you want a permanent instead of a temporary dataset, give it a two-part name, such as `ecmet.cobbdoug`.

Here are the instructions for SPSS: 1) Begin SPSS with a blank spreadsheet. 2) Open up a file with the following commands and run:

```
SET
BLANKS=SYSMIS
UNDEFINED=WARN.
DATA LIST
FILE='A:\Cbbunst.dat' FIXED RECORDS=1 TABLE /1 year 1-4 output 5-9 capital
10-16 labor 17-22 wage 23-27 .
EXECUTE.
```

This files assume the data file to be on the same directory, and again the first line in the data file with the variable names must be deleted. Once the data are entered into SPSS the procedures (regression, etc.) are best run from the point and click environment.

□

• b. *The next step is to look at the data. On* [**CD28**, p. 150]*, Cobb and Douglas plot* ***capital****,* ***labor****, and* ***output*** *on a logarithmic scale against time, all 3 series normalized such that they start in 1899 at the same level =100. Reproduce this graph using a modern statistics package.*

• c. *Run both regressions* (22.1.2) *and* (22.1.3) *on Cobb and Douglas's original dataset. Compute 95% confidence intervals for the coefficients of capital and labor in the unconstrained and the cconstrained models.*

Answer. SAS does not allow you to transform the data on the fly, it insists that you first go through a data step creating the transformed data, before you can run a regression on them. Therefore the next set of commands creates a temporary dataset `cdtmp`. The data step `data cdtmp` includes all the data from `cobbdoug` into `cdtemp` and then creates some transformed data as well. Then one can run the regressions. Here are the commands; they are in the file `cbbrgrss.sas` in your data disk:

```
data cdtmp;
    set cobbdoug;
    logcap = log(capital);
    loglab = log(labor);
    logout = log(output);
    logcl = logcap-loglab;
    logol = logout-loglab;
run;
proc reg data = cdtmp;
    model logout = logcap loglab;
run;
proc reg data = cdtmp;
    model logol = logcl;
run;
```


Careful! In R, the command `lm(log(output)-log(labor) ~ log(capital)-log(labor), data=cobbdoug`
does not give the right results. It does not complain but the result is wrong nevertheless. The right way to write this command is `lm(I(log(output)-log(labor)) ~ I(log(capital)-log(labor)), data=cobbdo`

□

• d. *The regression results are graphically represented in Figure 1. The big ellipse is a joint 95% confidence region for β and γ. This ellipse is a level line of the* SSE*. The vertical and horizontal bands represent univariate 95% confidence regions for β and γ separately. The diagonal line is the set of all β and γ with $\beta + \gamma = 1$, representing the constraint of constant returns to scale. The small ellipse is that level line of the* SSE *which is tangent to the constraint. The point of tangency represents the constrained estimator. Reproduce this graph (or as much of this graph as you can) using your statistics package.*

Remark: In order to make the hand computations easier, Cobb and Douglass reduced the data for ***capital*** *and* ***labor*** *to index numbers (1899=100) which were rounded to integers, before running the regressions, and Figure 1 was constructed using these rounded data. Since you are using the nonstandardized data, you may get slightly different results.*

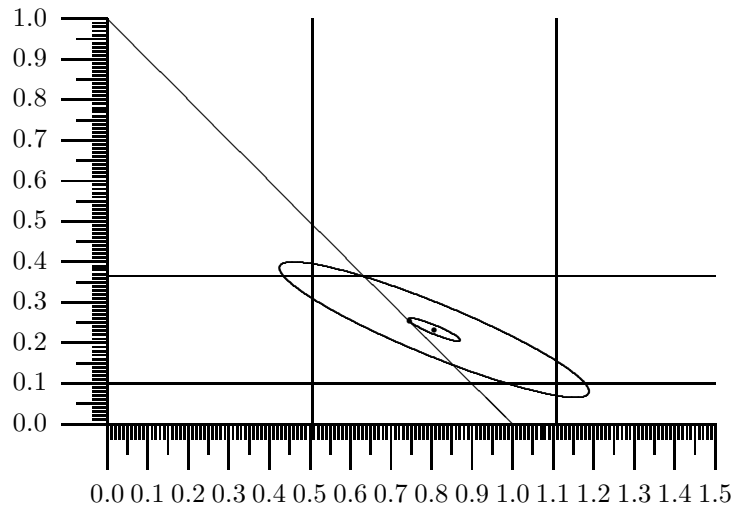


Figure 1. Coefficients of `capital` (vertical) and `labor` (horizontal), dependent variable `output`, unconstrained and constrained, 1899–1922

Answer. `lines(ellipse.lm(cbbfit, which=c(2, 3)))` □

PROBLEM 275. *In this problem we will treat the Cobb-Douglas data as a dataset with errors in all three variables. See chapter 53.4 and problem 476 about that.*

• a. *Run the three elementary regressions for the whole period, then choose at least two subperiods and run it for those. Plot all regression coefficients as points in a plane, using different colors for the different subperiods (you have to normalize them in a special way that they all fit on the same plot).*

ANSWER. Here are the results in R:

```
> outputlm<-lm(log(output)~log(capital)+log(labor),data=cobbdoug)
> capitallm<-lm(log(capital)~log(labor)+log(output),data=cobbdoug)
> laborlm<-lm(log(labor)~log(output)+log(capital),data=cobbdoug)
> coefficients(outputlm)
 (Intercept) log(capital)   log(labor)
  -0.1773097    0.2330535    0.8072782
> coefficients(capitallm)
 (Intercept)  log(labor) log(output)
 -2.72052726 -0.08695944  1.67579357
> coefficients(laborlm)
 (Intercept)  log(output) log(capital)
  1.27424214   0.73812541  -0.01105754

#Here is the information for the confidence ellipse:
> summary(outputlm,correlation=T)

Call:
lm(formula = log(output) ~ log(capital) + log(labor), data = cobbdoug)

Residuals:
      Min        1Q    Median        3Q       Max
-0.075282 -0.035234 -0.006439  0.038782  0.142114

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.17731  0.43429   -0.408  0.68721
log(capital)  0.23305  0.06353    3.668  0.00143  **
log(labor)    0.80728  0.14508    5.565  1.6e-05  ***
---
Signif. codes:  0  '***'  0.001  '**'  0.01  '*'  0.05  '.'  0.1  ' '  1

Residual standard error: 0.05814 on 21 degrees of freedom
Multiple R-Squared: 0.9574,Adjusted R-squared: 0.9534
F-statistic: 236.1 on 2 and 21 degrees of freedom,p-value: 3.997e-15

Correlation of Coefficients:
             (Intercept) log(capital)
log(capital)      0.7243
```



```
log(labor)       -0.9451      -0.9096


#Quantile of the F-distribution:
> qf(p=0.95, df1=2, df2=21)

[1] 3.4668
```

□

• b. *The elementary regressions will give you three fitted equations of the form*

$$\boldsymbol{output} = \hat{\alpha}_1 + \hat{\beta}_{12}\,\boldsymbol{labor} + \hat{\beta}_{13}\,\boldsymbol{capital} + \mathit{residual}_1 \tag{22.1.6}$$

$$\boldsymbol{labor} = \hat{\alpha}_2 + \hat{\beta}_{21}\,\boldsymbol{output} + \hat{\beta}_{23}\,\boldsymbol{capital} + \mathit{residual}_2 \tag{22.1.7}$$

$$\boldsymbol{capital} = \hat{\alpha}_3 + \hat{\beta}_{31}\,\boldsymbol{output} + \hat{\beta}_{32}\,\boldsymbol{labor} + \mathit{residual}_3. \tag{22.1.8}$$

In order to compare the slope parameters in these regressions, first rearrange them in the form

$$-\boldsymbol{output} + \hat{\beta}_{12}\,\boldsymbol{labor} + \hat{\beta}_{13}\,\boldsymbol{capital} + \hat{\alpha}_1 + \mathit{residual}_1 = 0 \tag{22.1.9}$$

$$\hat{\beta}_{21}\,\boldsymbol{output} - \boldsymbol{labor} + \hat{\beta}_{23}\,\boldsymbol{capital} + \hat{\alpha}_2 + \mathit{residual}_2 = 0 \tag{22.1.10}$$

$$\hat{\beta}_{31}\,\boldsymbol{output} + \hat{\beta}_{32}\,\boldsymbol{labor} - \boldsymbol{capital} + \hat{\alpha}_3 + \mathit{residual}_3 = 0 \tag{22.1.11}$$

This gives the following table of coefficients:

output	*labor*	*capital*	*intercept*
-1	0.8072782	0.2330535	-0.1773097
0.73812541	-1	-0.01105754	1.27424214
1.67579357	-0.08695944	-1	-2.72052726

Now divide the second and third rows by the negative of their first coefficient, so that the coefficient of output becomes -1:

out	*labor*	*capital*	*intercept*
-1	0.8072782	0.2330535	-0.1773097
-1	1/0.73812541	0.01105754/0.73812541	$-1.27424214/0.73812541$
-1	0.08695944/1.67579357	1/1.67579357	2.72052726/1.67579357

After performing the divisions the following numbers are obtained:

output	*labor*	*capital*	*intercept*
-1	0.8072782	0.2330535	-0.1773097
-1	1.3547833	0.014980570	-1.726322
-1	0.05189149	0.59673221	1.6234262

These results can also be re-written in the form given by Table 2.

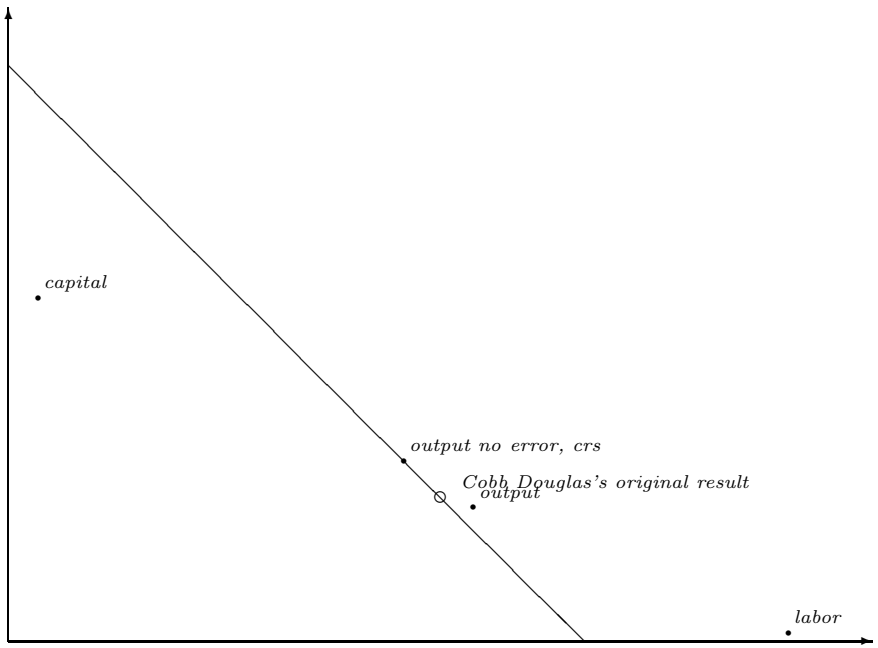


FIGURE 2. Coefficients of `capital` (vertical) and `labor` (horizontal), dependent variable `output`, 1899–1922

	Intercept	*Slope of* `output` *wrt* `labor`	*Slope of* `output` *wrt* `capital`
Regression of `output` *on* `labor` *and* `capital`			
Regression of `labor` *on* `output` *and* `capital`			
Regression of `capital` *on* `output` *and* `labor`			

TABLE 2. Comparison of coefficients in elementary regressions

Fill in the values for the whole period and also for several sample subperiods. Make a scatter plot of the contents of this table, i.e., represent each regression result as a point in a plane, using different colors for different sample periods.

PROBLEM 276. *Given a univariate problem with three variables all of which have zero mean, and a linear constraint that the coefficients of all variables sum to 0. (This*

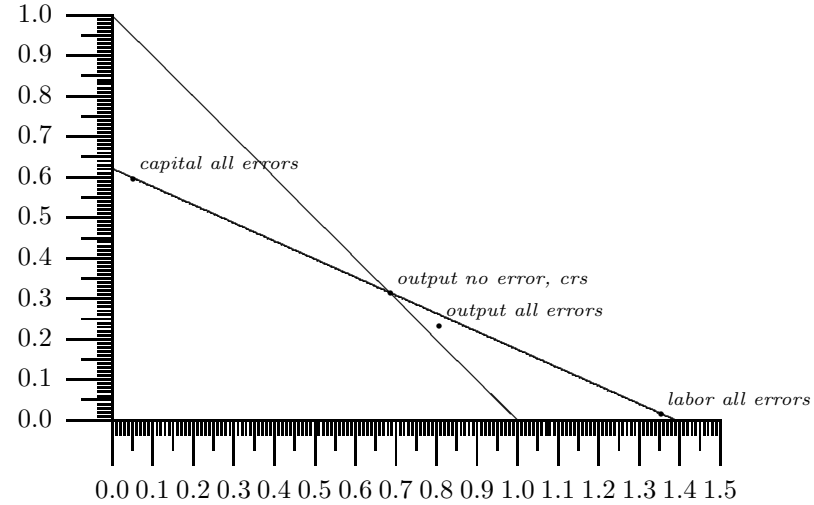


FIGURE 3. Coefficient of `capital` (vertical) and `labor` (horizontal) in the elementary regressions, dependent variable `output`, 1899–1922

is the model apparently appropriate to the Cobb-Douglas data, with the assumption of constant returns to scale, after taking out the means.) Call the observed variables $\mathtt{x}$, $\mathtt{y}$, *and* $\mathtt{z}$, *with underlying systematic variables* $\mathtt{x}^*$, $\mathtt{y}^*$, *and* $\mathtt{z}^*$, *and errors* $\mathtt{u}$, $\mathtt{v}$, *and* $\mathtt{w}$.

• a. *Write this model in the form* (53.3.1).

ANSWER.

$$\begin{bmatrix}\mathtt{x}^* & \mathtt{y}^* & \mathtt{z}^*\end{bmatrix}\begin{bmatrix}-1\\ \beta\\ 1-\beta\end{bmatrix}=0 \qquad \begin{bmatrix}\mathtt{x} & \mathtt{y} & \mathtt{z}\end{bmatrix}=\begin{bmatrix}\mathtt{x}^* & \mathtt{y}^* & \mathtt{z}^*\end{bmatrix}+\begin{bmatrix}\mathtt{u} & \mathtt{v} & \mathtt{w}\end{bmatrix} \quad \text{or} \quad \begin{aligned}\mathtt{x}^*&=\beta\mathtt{y}^*+(1-\beta)\mathtt{z}^*\\ \mathtt{x}&=\mathtt{x}^*+\mathtt{u}\\ \mathtt{y}&=\mathtt{y}^*+\mathtt{v}\\ \mathtt{z}&=\mathtt{z}^*+\mathtt{w}.\end{aligned} \tag{22.1.12}$$

□

• b. *The moment matrix of the systematic variables can be written fully in terms of* $\sigma^2_{\mathtt{y}^*}$, $\sigma^2_{\mathtt{z}^*}$, $\sigma_{\mathtt{y}^*\mathtt{z}^*}$, *and the unknown parameter* β. *Write out the moment matrix and therefore the Frisch decomposition.*

ANSWER. The moment matrix is the middle matrix in the following Frisch decomposition:

$$\begin{bmatrix}\sigma^2_{\mathtt{x}} & \sigma_{\mathtt{xy}} & \sigma_{\mathtt{xz}}\\ \sigma_{\mathtt{xy}} & \sigma^2_{\mathtt{y}} & \sigma_{\mathtt{yz}}\\ \sigma_{\mathtt{xz}} & \sigma_{\mathtt{yz}} & \sigma^2_{\mathtt{z}}\end{bmatrix}= \tag{22.1.13}$$

(22.1.14)

$$= \begin{bmatrix} \beta^2\sigma^2_{y^*} + 2\beta(1-\beta)\sigma_{y^*z^*} + (1-\beta)^2\sigma^2_{z^*} & \beta\sigma^2_{y^*} + (1-\beta)\sigma_{y^*z^*} & \beta\sigma_{y^*z^*} + (1-\beta)\sigma^2_{z^*} \\ \beta\sigma^2_{y^*} + (1-\beta)\sigma_{y^*z^*} & \sigma^2_{y^*} & \sigma_{y^*z^*} \\ \beta\sigma_{y^*z^*} + (1-\beta)\sigma^2_{z^*} & \sigma^2_{y^*} & \sigma^2_{z^*} \end{bmatrix} + \begin{bmatrix} \sigma^2_u & 0 \\ 0 & \sigma^2_v \\ 0 & 0 \end{bmatrix}$$

□

• c. *Show that the unknown parameters are not yet identified. However, if one makes the additional assumption that one of the three error variances σ^2_u, σ^2_v, or σ^2_w is zero, then the equations are identified. Since the quantity of output presumably has less error than the other two variables, assume $\sigma^2_u = 0$. Under this assumption, show that*

$$\beta = \frac{\sigma^2_x - \sigma_{xz}}{\sigma_{xy} - \sigma_{xz}} \tag{22.1.15}$$

and this can be estimated by replacing the variances and covariances by their sample counterparts. In a similar way, derive estimates of all other parameters of the model.

Answer. Solving (22.1.14) one gets from the yz element of the covariance matrix

$$\sigma_{y^*z^*} = \sigma_{yz} \tag{22.1.16}$$

and from the xz element

$$\sigma^2_{z^*} = \frac{\sigma_{xz} - \beta\sigma_{yz}}{1-\beta} \tag{22.1.17}$$

Similarly, one gets from the xy element:

$$\sigma^2_{y^*} = \frac{\sigma_{xy} - (1-\beta)\sigma_{yz}}{\beta} \tag{22.1.18}$$

Now plug (22.1.16), (22.1.17), and (22.1.18) into the equation for the xx element:

$$\sigma^2_x = \beta(\sigma_{xy} - (1-\beta)\sigma_{yz}) + 2\beta(1-\beta)\sigma_{yz} + (1-\beta)(\sigma_{xz} - \beta\sigma_{yz}) + \sigma^2_u \tag{22.1.19}$$

$$= \beta\sigma_{xy} + (1-\beta)\sigma_{xz} + \sigma^2_u \tag{22.1.20}$$

Since we are assuming $\sigma^2_u = 0$ this last equation can be solved for β:

$$\beta = \frac{\sigma^2_x - \sigma_{xz}}{\sigma_{xy} - \sigma_{xz}} \tag{22.1.21}$$

If we replace the variances and covariances by the sample variances and covariances, this gives an estimate of β.

□

• d. *Evaluate these formulas numerically. In order to get the sample means and the sample covariance matrix of the data, you may issue the SAS commands*

```
proc corr cov nocorr data=cdtmp;
    var logout loglab logcap;
    run;
```


These commands are in the file `cbbcovma.sas` *on the disk.*

Answer. Mean vector and covariance matrix are

$$\begin{bmatrix} \texttt{LOGOUT} \\ \texttt{LOGLAB} \\ \texttt{LOGCAP} \end{bmatrix} \sim \left(\begin{bmatrix} 5.07734 \\ 4.96272 \\ 5.35648 \end{bmatrix}, \begin{bmatrix} 0.0724870714 & 0.0522115563 & 0.1169330807 \\ 0.0522115563 & 0.0404318579 & 0.0839798588 \\ 0.1169330807 & 0.0839798588 & 0.2108441826 \end{bmatrix} \right) \tag{22.1.22}$$

Therefore equation (22.1.15) gives

$$\hat\beta = \frac{0.0724870714 - 0.1169330807}{0.0522115563 - 0.1169330807} = 0.686726861149148 \tag{22.1.23}$$

In Figure 3, the point $(\hat\beta, 1-\hat\beta)$ is exactly the intersection of the long dotted line with the constraint.

□

• e. *The fact that all 3 points lie almost on the same line indicates that there may be 2 linear relations: log labor is a certain coefficient times log output, and log capital is a different coefficient times log output. I.e., $y^* = \delta_1 + \gamma_1 x^*$ and $z^* = \delta_2 + \gamma_2 x^*$. In other words, there is no substitution. What would be the two coefficients γ_1 and γ_2 if this were the case?*

Answer. Now the Frisch decomposition is

$$\begin{bmatrix} \sigma^2_x & \sigma_{xy} & \sigma_{xz} \\ \sigma_{xy} & \sigma^2_y & \sigma_{yz} \\ \sigma_{xz} & \sigma_{yz} & \sigma^2_z \end{bmatrix} = \sigma^2_{x^*} \begin{bmatrix} 1 & \gamma_1 & \gamma_2 \\ \gamma_1 & \gamma_1^2 & \gamma_1\gamma_2 \\ \gamma_2 & \gamma_1\gamma_2 & \gamma_2^2 \end{bmatrix} + \begin{bmatrix} \sigma^2_u & 0 & 0 \\ 0 & \sigma^2_v & 0 \\ 0 & 0 & \sigma^2_w \end{bmatrix}. \tag{22.1.24}$$

Solving this gives (obtain γ_1 by dividing the 32-element by the 31-element, γ_2 by dividing the 32-element by the 12-element, $\sigma^2_{x^*}$ by dividing the 21-element by γ_1, etc.

(22.1.25)

$$\gamma_1 = \frac{\sigma_{yz}}{\sigma_{xy}} = \frac{0.0839798588}{0.1169330807} = 0.7181873452513939 \qquad \sigma^2_u = \sigma^2_x - \frac{\sigma_{yx}\sigma_{xz}}{\sigma_{yz}} = 0.0724870714 - 0.07$$

$$\gamma_2 = \frac{\sigma_{yz}}{\sigma_{xz}} = \frac{0.0839798588}{0.0522115563} = 1.608453467992104 \qquad \sigma^2_v = \sigma^2_y - \frac{\sigma_{xy}\sigma_{yz}}{\sigma_{xz}}$$

$$\sigma^2_{x^*} = \frac{\sigma_{yx}\sigma_{xz}}{\sigma_{yz}} = \frac{0.0522115563 \cdot 0.1169330807}{0.0839798588} = 0.0726990758 \qquad \sigma^2_w = \sigma^2_z - \frac{\sigma_{xz}\sigma_{zy}}{\sigma_{xy}}$$

This model is just barely rejected by the data since it leads to a slightly negative variance for $\mathbf{u}$. □

• f. *The assumption that there are two linear relations is represented as the light-blue line in Figure 3. What is the equation of this line?*

Answer. If $y = \gamma_1 x$ and $z = \gamma_2 x$ then the equation $x = \beta_1 y + \beta_2 z$ holds whenever $\beta_1\gamma_1 + \beta_2\gamma_2 = 1$. This is a straight line in the β_1, β_2-plane, going through the points and $(0, 1/\gamma_2) = (0, \frac{0.0522115563}{0.0839798588} = 0.6217152189353289)$ and $(1/\gamma_1, 0) = (\frac{0.1169330807}{0.0839798588} = 1.3923943475361023, 0)$. This line is in the figure, and it is just a tiny bit on the wrong side of the dotted line connecting the two estimates. □

22.2. Houthakker's Data

For this example we will use Berndt's textbook [**Ber91**], which discusses some of the classic studies in the econometric literature.

One example described there is the estimation of a demand function for electricity [**Hou51**], which is the first multiple regression with several variables run on a computer. In this exercise you are asked to do all steps in exercise 1 and 3 in chapter 7 of Berndt, and use the additional facilities of R to perform other steps of data analysis which Berndt did not ask for, such as, for instance, explore the best subset of regressors using `leaps` and the best nonlinear transformation using `avas`, do some diagnostics, search for outliers or influential observations, and check the normality of residuals by a probability plot.

Problem 277. *4 points The electricity demand date from* [**Hou51**] *are available on the web at* `www.econ.utah.edu/ehrbar/data/ukelec.txt`. *Import these data into your favorite statistics package. For* R *you need the command* `ukelec <- read.table("http://www.econ.utah.edu/ehrbar/data/ukelec.txt")`. *Make a scatterplot matrix of these data using e.g.* `pairs(ukelec)` *and describe what you see.*

Answer. inc and cap are negatively correlated. cap is capacity of rented equipment and not equipment owned. Apparently customers with higher income buy their equipment instead of renting it.

gas6 and gas8 are very highly correlated. mc4, mc6, and mc8 are less hightly correlated, the corrlation between mc6 and mc8 is higher than that between mc4 and mc6. It seem electicity prices have been coming down.

kwh, inc, and exp are strongly positively correlated.

the stripes in all the plots which have mc4, mc6, or mc8 in them come from the fact that the marginal cost of electricity is a round number.

electricity prices and kwh are negatively correlated.

There is no obvious positive correlation between kwh and cap or expen and cap.

Prices of electricity and gas are somewhat positively correlated, but not much.

When looking at the correlations of inc with the other variables, there are several outliers which could have a strong "leverage" effect.

in 1934, those with high income had lower electricity prices than those with low income. This effect dissipated by 1938.

No strong negative correlations anywhere.

cust negatively correlated with inc, because rich people live in smaller cities?

□

If you simply type `ukelec` in R, it will print the data on the screen. The variables have the following meanings:

`cust` Average number of consumers with two-part tariffs for electricity in 1937–38, in thousands. Two-part tariff means: they pay a fixed monthly sum plus a certain "running charge" times the number of kilowatt hours they use.

`inc` Average income of two-part consumers, in pounds per year. (Note that one pound had 240 pence at that time.)

`mc4` The running charge (marginal cost) on domestic two-part tariffs in 1933–34, in pence per KWH. (The marginal costs are the costs that depend on the number of kilowatt hours only, it is the cost of one additional kilowatt hour.

`mc6` The running charge (marginal cost) on domestic two-part tariffs in 1935–36, in pence per KWH

`mc8` The running charge (marginal cost) on domestic two-part tariffs in 1937–38, in pence per KWH

`gas6` The marginal price of gas in 1935–36, in pence per therm

`gas8` The marginal price of gas in 1937–38, in pence per therm

`kwh` Consumption on domestic two-part tariffs per consumer in 1937–38, in kilowatt hours

`cap` The average holdings (capacity) of heavy electric equipment bought on hire purchase (leased) by domestic two-part consumers in 1937–38, in kilowatts

`expen` The average total expenditure on electricity by two-part consumers in 1937–38, in pounds

The function `summary(ukelec)` displays summary statistics about every variable.

Since every data frame in R is a list, it is possible to access the variables in `ukelec` by typing `ukelec$mc4` etc. Try this; if you type this and then a return, you will get a listing of `mc4`. In order to have all variables available as separate objects and save typing `ukelec$` all the time, one has to "mount" the data frame by the command `attach(ukelec)`. After this, the individual data series can simply be printed on the screen by typing the name of the variable, for instance `mc4`, and then the return key.

Problem 278. *2 points Make boxplots of* `mc4`, `mc6`, *and* `mc6` *in the same graph next to each other, and the same with* `gas6` *and* `gas8`.

Problem 279. *2 points How would you answer the question whether marginal prices of gas vary more or less than those of electricity (say in the year 1936)?*

Answer. Marginal gas prices vary a little more than electricity prices, although electricity was the newer technology, and although gas prices are much more stable over time than the electricity prices. Compare `sqrt(var(mc6))/mean(mc6)` with `sqrt(var(gas6))/mean(gas6)`. You get 0.176 versus 0.203. Another way would be to compute `max(mc6)/min(mc6)` and compare with `max(gas6)/min(gas6)`: you get 2.27 versus 2.62. In any case this is a lot of variation. □

Problem 280. *2 points Make a plot of the (empirical) density function of* `mc6` *and* `gas6` *and interpret the results.*

Problem 281. *2 points Is electricity a big share of total income? Which command is better:* `mean(expen/inc)` *or* `mean(expen)/mean(inc)`? *What other options*

are there? Actually, there is a command which is clearly better than at least one of the above, can you figure out what it is?

ANSWER. The proportion is small, less than 1 percent. The two above commands give 0.89% and 0.84%. The command `sum(cust*expen) / sum(cust*inc)` is better than `mean(expen) / mean(inc)`, because each component in `expen` and `inc` is the mean over many households, the number of households given by `cust`. `mean(expen)` is therefore an average over averages over different population sizes, not a good idea. `sum(cust*expen)` is total expenditure in all households involved, and `sum(cust*inc)` is total income in all households involved. `sum(cust*expen) / sum(cust*inc)` gives the value 0.92%. Another option is `median(expen/inc)` which gives 0.91%. A good way to answer this question is to plot it: `plot(expen,inc)`. You get the line where expenditure is 1 percent of income by `abline(0,0.01)`. For higher incomes expenditure for electricity levels off and becomes a lower share of income. □

PROBLEM 282. *Have your computer compute the sample correlation matrix of the data. The* R*-command is* `cor(ukelec)`

• a. *4 points Are there surprises if one looks at the correlation matrix?*

ANSWER. Electricity consumption `kwh` is slightly negatively correlated with gas prices and with the capacity. If one takes the correlation matrix of the logarithmic data, one gets the expected positive signs.

marginal prices of gas and electricity are positively correlated in the order of 0.3 to 0.45.

higher correlation between mc6 and mc8 than between mc4 and mc6.

Correlation between `expen` and `cap` is negative and low in both matrices, while one should expect positive correlation. But in the logarithmic matrix, `mc6` has negative correlation with `expen`, i.e., elasticity of electricity demand is less than 1.

In the logarithmic data, `cust` has higher correlations than in the non-logarithmic data, and it is also more nearly normally distributed.

`inc` has negative correlation with `mc4` but positive correlation with `mc6` and `mc8`. (If one looks at the scatterplot matrix this seems just random variations in an essentially zero correlation).

mc6 and expen are positively correlated, and so are mc8 and expen. This is due to the one outlier with high expen and high income and also high electricity prices.

The marginal prices of electricity are not strongly correlated with `expen`, and in 1934, they are negatively correlated with `income`.

From the scatter plot of `kwh` versus `cap` it seems there are two datapoints whose removal might turn the sign around. To find out which they are do `plot(kwh,cap)` and then use the identify function: `identify(kwh,cap,labels=row.names(ukelec))`. The two outlying datapoints are Halifax and Wallase. Wallase has the highest income of all towns, namely, 1422, while Halifax's income of 352 is close to the minimum, which is 279. High income customers do not lease their equipment but buy it. □

• b. *3 points The correlation matrix says that* `kwh` *is negatively related with* `cap`*, but the correlation of the logarithm gives the expected positive sign. Can you explain this behavior?*

ANSWER. If one plots the date using `plot(cap,kwh)` one sees that the negative correlation comes from the two outliers. In a logarithmic scale, these two are no longer so strong outliers. □

PROBLEM 283. *Berndt on p. 338 defines the intramarginal expenditure* `f <- expen-mc8*kwh/240`*. What is this, and what do you find out looking at it?*

After this preliminary look at the data, let us run the regressions.

PROBLEM 284. *6 points Write up the main results from the regressions which in* R *are run by the commands*

```
houth.olsfit <- lm(formula = kwh ~ inc+I(1/mc6)+gas6+cap)
houth.glsfit <- lm(kwh ~ inc+I(1/mc6)+gas6+cap, weight=cust)
houth.olsloglogfit <- lm(log(kwh) ~
log(inc)+log(mc6)+log(gas6)+log(cap))
```

Instead of `1/mc6` *you had to type* `I(1/mc6)` *because the slash has a special meaning in formulas, creating a nested design, therefore it had to be "protected" by applying the function* `I()` *to it.*

If you then type `houth.olsfit`*, a short summary of the regression results will be displayed on the screen. There is also the command* `summary(houth.olsfit)`*, which gives you a more detailed summary. If you type* `plot(houth.olsfit)` *you will get a series of graphics relevant for this regression.*

ANSWER. All the expected signs.

Gas prices do not play a great role in determining electricity consumption, despite the "cookers" Berndt talks about on p. 337. Especially the logarithmic regression makes gas prices highly insignificant!

The weighted estimation has a higher R^2. □

PROBLEM 285. *2 points The output of the OLS regression gives as standard error of* `inc` *the value of 0.18, while in the GLS regression it is 0.20. For the other variables, the standard error as given in the GLS regression is lower than that in the OLS regression. Does this mean that one should use for* `inc` *the OLS estimate and for the other variables the GLS estimates?*

PROBLEM 286. *5 points Show, using the* `leaps` *procedure om* R *or some other selection of regressors, that the variables Houthakker used in his GLS-regression are the "best" among the following:* `inc, mc4, mc6, mc8, gas6, gas8, cap` *using either the* C_p *statistic or the adjusted* R^2*. (At this stage, do not transform the variables but just enter them into the regression untransformed, but do use the weights, which are theoretically well justified).*

To download the `leaps` *package, use* `install.packages("leaps", lib="C:/Documents and Settings/420lab.420LAB/My Documents")` *and to call it up, use* `library(leaps, lib.loc="C:/Documents and Settings/420lab.420LAB/My Documents")`*. If the library* `ecmet` *is available, the command* `ecmet.script(houthsel)` *runs the following script:*

```
library(leaps)
data(ukelec)
attach(ukelec)
houth.glsleaps<-leaps(x=cbind(inc,mc4,mc6,mc8,gas6,gas8,cap),
                      y=kwh, wt=cust, method="Cp",
                      nbest=5, strictly.compatible=F)
ecmet.prompt("Plot Mallow's Cp against number of regressors:")
plot(houth.glsleaps$size, houth.glsleaps$Cp)
ecmet.prompt("Throw out all regressions with a Cp > 50 (big gap)")
plot(houth.glsleaps$size[houth.glsleaps$Cp<50],
     houth.glsleaps$Cp[houth.glsleaps$Cp<50])
ecmet.prompt("Cp should be roughly equal the number of regressors")
abline(0,1)
cat("Does this mean the best regression is overfitted?")
ecmet.prompt("Click at the points to identify them, left click to quit")
## First construct the labels
lngth <- dim(houth.glsleaps$which)[1]
included <- as.list(1:lngth)
for (ii in 1:lngth)
  included[[ii]] <- paste(
   colnames(houth.glsleaps$which)[houth.glsleaps$which[ii,]],
   collapse=",")
identify(x=houth.glsleaps$size, y=houth.glsleaps$Cp, labels=included)
ecmet.prompt("Now use regsubsets instead of leaps")
houth.glsrss<- regsubsets.default(x=cbind(inc,mc4,mc6,mc8,gas6,gas8,cap),
                    y=kwh, weights=cust, method="exhaustive")
print(summary.regsubsets(houth.glsrss))
plot.regsubsets(houth.glsrss, scale="Cp")
ecmet.prompt("Now order the variables")
houth.glsrsord<- regsubsets.default(x=cbind(inc,mc6,cap,gas6,gas8,mc8,mc4
                      y=kwh, weights=cust, method="exhaustive")
print(summary.regsubsets(houth.glsrsord))
plot.regsubsets(houth.glsrsord, scale="Cp")
```

PROBLEM 287. *Use* `avas` *to determine the "best" nonlinear transformations of the explanatory and the response variable. Since the weights are theoretically well justified, one should do it for the weighted regression. Which functions do you think one should use for the different regressors?*

PROBLEM 288. *3 points Then, as a check whether the transformation interferred with data selection, redo* `leaps`*, but now with the transformed variables. Show that*

the GLS-regression Houthakker actually ran is the "best" regression among the following variables: `inc, 1/mc4, 1/mc6, 1/mc8, gas6, gas8, cap` *using either the* C_p *statistic or the adjusted* R^2.

PROBLEM 289. *Diagnostics, the identification of outliers or influential observations is something which we can do easily with* R*, although Berndt did not ask for it. The command* `houth.glsinf<-lm.influence(houth.glsfit)` *gives you the building blocks for many of the regression disgnostics statistics. Its output is a list if three objects: A matrix whose rows are all the the least squares estimates* $\hat{\boldsymbol{\beta}}(i)$ *when the* i*th observation is dropped, a vector with all the* $\mathsf{s}(i)$*, and a vector with all the* h_{ii}*. A more extensive function is* `influence.measures(houth.glsfit)`*, it has Cook's distance and others.*

In order to look at the residuals, use the command `plot(resid(houth.glsfit), type="h")` or `plot(rstandard(houth.glsfit), type="h")` or `plot(rstudent(houth.glsf type="h")`. To add the axis do `abline(0,0)`. If you wanted to check the residuals for normality, you would use `qqnorm(rstandard(houth.glsfit))`.

PROBLEM 290. *Which commands do you need to plot the predictive residuals?*

PROBLEM 291. *4 points Although there is good theoretical justification for using* `cust` *as weights, one might wonder if the data bear this out. How can you check this?*

ANSWER. Do `plot(cust, rstandard(houth.olsfit))` and `plot(cust, rstandard(houth.glsfit))`. In the first plot, smaller numbers of customers have larger residuals, in the second plot this is mitigated. Also the OLS plot has two terrible outliers, which are brought more into range with GLS. □

PROBLEM 292. *The variable* `cap` *does not measure the capacity of* all *electrical equipment owned by the households, but only those appliances which were leased from the Electric Utility company. A plot shows that people with higher income do not lease as much but presumably purchase their appliances outright. Does this mean the variable should not be in the regression?*

22.3. Long Term Data about US Economy

The dataset `uslt` is described in [**DL91**]. Home page of the authors is `www.cepremap.cnrs` `uslt` has the variables `kn`, `kg` (net and gross capital stock in current \$), `kn2`, `kg2` (the same in 1982\$), `hours` (hours worked), `wage` (hourly wage in current dollars), `gnp`, `gnp2`, `nnp`, `inv2` (investment in 1982 dollars), `r` (profit rate (`nnp`$-$`wage`$\times$`hours`)/`kn`), `u` (capacity utilization), `kne`, `kge`, `kne2`, `kge2`, `inve2` (capital stock and investment data for equipment), `kns`, `kgs`, `kns2`, `kgs2`, `invs2` (the same for structures).

Capital stock data were estimated separately for structures and equipment and then added up, i.e., `kn2` = `kne2` + `kns2` etc. Capital stock since 1925 has been constructed from annual investment data, and prior to 1925 the authors of the series

apparently went the other direction: they took someone's published capital stock estimates and constructed investment from it. In the 1800s, only a few observations were available, which were then interpolated. The capacity utilization ratio is equal to the ratio of `gnp2` to its trend, i.e., it may be negative.

Here are some possible commands for your R-session: `data(uslt)` makes the data available; `uslt.clean<-na.omit(uslt)` removes missing values; this dataset starts in 1869 (instead of 1805). `attach(uslt.clean)` makes the variables in this dataset available. Now you can plot various series, for instance `plot((nnp-hours*wage)/nnp, type="l")` plots the profit share, or `plot(gnp/gnp2, kg/kg2, type="l")` gives you a scatter plot of the price level for capital goods versus that for gnp. The command `plot(r, kn2/hours, type="b")` gives both points and dots; `type = "o"` will have the dots overlaid the line. After the plot you may issue the command `identify(r, kn2/hours, label=1869:1989)` and then click with the left mouse button on the plot those data points for which you want to have the years printed.

If you want more than one timeseries on the same plot, you may do `matplot(1869:198 cbind(kn2,kns2), type="l")`. If you want the y-axis logarithmic, say `matplot(1869:19 cbind(gnp/gnp2,kns/kns2,kne/kne2), type="l", log="y")`.

Problem 293. *Computer assignment: Make a number of such plots on the screen, and import the most interesting ones into your wordprocessor. Each class participant should write a short paper which shows the three most insteresting plots, together with a written explanation why these plots seem interesting.*

To use `pairs` or `xgobi`, you should carefully select the variables you want to include, and then you need the following preparations: `usltsplom <- cbind(gnp2=gnp2, kn2=kn2, inv2=inv2, hours=hours, year=1869:1989)` `dimnames(usltsplom)[[1]] <- paste(1869:1989)` The `dimnames` function adds the row labels to the matrix, so that you can see which year it is. `pairs(usltsplom)` or `library(xgobi)` and then `xgobi(usltsplom)`

You can also run regressions with commands of the following sort: `lm.fit <- lm(formula = gnp2 ~ hours + kne2 + kns2)`. You can also fit a "generalized additive model" with the formula `gam.fit <- gam(formula = gnp2 ~ s(hours) + s(kne2) + s(kns2))`. This is related to the `avas` command we talked about in class. It is discussed in [**CH93**].

22.4. Dougherty Data

We have a new dataset, in both `SAS` and `Splus`, namely the data described in [**Dou92**].

There are more data than in the tables at the end of the book; prelcosm for instance is the relative price of cosmetics, it is `100*pcosm/ptpe`, but apparently truncated at 5 digits.

22.5. Wage Data

The two datasets used in [**Ber91**, pp. 191–209] are available in R as the data frames `cps78` and `cps85`. In R on unix, the data can be downloaded by `cps78 <- readSDML("http://www.econ.utah.edu/ehrbar/data/cps78.sdml")`, and the corresponding for `cps85`. The original data provided by Berndt contain many dummy variables. The data frames in R have the same data coded as "factor" variables instead of dummies. These "factor" variables automatically generate dummies when included in the `model` statement.

`cps78` consists of 550 randomly selected employed workers from the May 1978 current population survey, and `cps85` consists of 534 randomly selected employed workers from the May 1985 current population survey. These are surveys of 50,000 households conducted monthly by the U.S. Department of Commerce. They serve as the basis for the national employment and unemployment statistics. Data are collected on a number of individual characteristics as well as employment status. The present extracts were performed by Leslie Sundt of the University of Arizona.

`ed` = years of education

`ex` = years of labor market experience (= `age` − `ed` − 6, or 0 if this is a negative number).

`lnwage` = natural logarithm of average hourly earnings

`age` = age in years

`ndep` = number of dependent children under 18 in household (only in cps78).

`region` has levels North, South

`race` has levels Other, Nonwhite, Hispanic. Nonwhite is mainly the Blacks, and Other is mainly the Non-Hispanic Whites.

`gender` has levels Male, Female

`marr` has levels Single, Married

`union` has levels Nonunion, Union

`industry` has levels Other, Manuf, and Constr

`occupation` has levels Other, Manag, Sales, Cler, Serv, and Prof

Here is a log of my commands for exercises 1 and 2 in [**Ber91**, pp. 194–197].

```
> cps78 <- readSDML("http://www.econ.utah.edu/ehrbar/data/cps78.sdml")
> attach(cps78)
> ###Exercise 1a (2 points) in chapter V of Berndt, p. 194
> #Here is the arithmetic mean of hourly wages:
> mean(exp(lnwage))
[1] 6.062766
> #Here is the geometric mean of hourly wages:
> #(Berndt's instructions are apparently mis-formulated):
> exp(mean(lnwage))
```

```
[1] 5.370935
> #Geometric mean is lower than arithmetic, due to Jensen's inequality
> #if the year has 2000 hours, this gives an annual wage of
> 2000*exp(mean(lnwage))
[1] 10741.87
> #What are arithmetic mean and standard deviation of years of schooling
> #and years of potential experience?
> mean(ed)
[1] 12.53636
> sqrt(var(ed))
[1] 2.772087
> mean(ex)
[1] 18.71818
> sqrt(var(ex))
[1] 13.34653
> #experience has much higher standard deviation than education, not surp
> ##Exercise 1b (1 point) can be answered with the two commands
> table(race)
 Hisp Nonwh Other
   36    57   457
> table(race, gender)
       gender
race    Female Male
  Hisp      12   24
  Nonwh     28   29
  Other    167  290
> #Berndt also asked for the sample means of certain dummy variables;
> #This has no interest in its own right but was an intermediate
> #step in order to compute the numbers of cases as above.
> ##Exercise 1c (2 points) can be answered using tapply
> tapply(ed,gender,mean)
  Female     Male
12.76329 12.39942
> #now the standard deviation:
> sqrt(tapply(ed,gender,var))
  Female     Male
2.220165 3.052312
> #Women do not have less education than men; it is about equal,
> #but their standard deviation is smaller
> #Now the geometric mean of the wage rate:
```



```
> exp(tapply(lnwage,gender,mean))
  Female     Male
4.316358 6.128320
> #Now do the same with race
> ##Exercise 1d (4 points)
> detach()
> ##This used to be my old command:
> cps85 <- read.table("~/dpkg/ecmet/usr/share/ecmet/usr/lib/R/library/ecmet/d
> #But this should work for everyone (perhaps only on linux):
> cps85 <- readSDML("http://www.econ.utah.edu/ehrbar/data/cps85.sdml")
> attach(cps85)
> mean(exp(lnwage))
[1] 9.023947
> sqrt(var(lnwage))
[1] 0.5277335
> exp(mean(lnwage))
[1] 7.83955
> 2000*exp(mean(lnwage))
[1] 15679.1
> 2000*exp(mean(lnwage))/1.649
[1] 9508.248
> #real wage has fallen
> tapply(exp(lnwage), gender, mean)
  Female     Male
7.878743 9.994794
> tapply(exp(lnwage), gender, mean)/1.649
  Female     Male
4.777891 6.061125
> #Compare that with 4.791237 6.830132 in 1979:
> #Male real wages dropped much more than female wages
> ##Exercise 1e (3 points)
> #using cps85
> w <- mean(lnwage); w
[1] 2.059181
> s <- sqrt(var(lnwage)); s
[1] 0.5277335
> lnwagef <- factor(cut(lnwage, breaks = w+s*c(-4,-2,-1,0,1,2,4)))
> table(lnwagef)
lnwagef
(-0.0518,1]    (1,1.53] (1.53,2.06] (2.06,2.59] (2.59,3.11] (3.11,4.17]
```

```
          3          93         174         180          72          12
> ks.test(lnwage, "pnorm")

 One-sample Kolmogorov-Smirnov test

data:  lnwage
D = 0.8754, p-value = < 2.2e-16
alternative hypothesis: two.sided

> ks.test(lnwage, "pnorm", mean=w, sd =s)

 One-sample Kolmogorov-Smirnov test

data:  lnwage
D = 0.0426, p-value = 0.2879
alternative hypothesis: two.sided

> #Normal distribution not rejected
>
> #If we do the same thing with
> wage <- exp(lnwage)
> ks.test(wage, "pnorm", mean=mean(wage), sd =sqrt(var(wage)))

 One-sample Kolmogorov-Smirnov test

data:  wage
D = 0.1235, p-value = 1.668e-07
alternative hypothesis: two.sided

> #This is highly significant, therefore normality rejected
>

> #An alternative, simpler way to answer question 1e is by using qqnorm
> qqnorm(lnwage)
> qqnorm(exp(lnwage))
> #Note that the SAS proc univariate rejects that wage is normally distril
> #but does not reject that lnwage is normally distributed.
> ###Exercise 2a (3 points), p. 196
> summary(lm(lnwage ~ ed, data = cps78))
```



```
Call:
lm(formula = lnwage ~ ed, data = cps78)

Residuals:
      Min        1Q    Median        3Q       Max
-2.123168 -0.331368 -0.007296  0.319713  1.594445

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.030445   0.092704  11.115  < 2e-16 ***
ed          0.051894   0.007221   7.187 2.18e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.469 on 548 degrees of freedom
Multiple R-Squared: 0.08613,Adjusted R-squared: 0.08447
F-statistic: 51.65 on 1 and 548 degrees of freedom,p-value: 2.181e-12

> #One year of education increases wages by 5 percent, but low R^2.
> #Mincer (5.18) had 7 percent for 1959
> #Now we need a 95 percent confidence interval for this coefficient
> 0.051894 + 0.007221*qt(0.975, 548)
[1] 0.06607823
> 0.051894 - 0.007221*qt(0.975, 548)
[1] 0.03770977
> ##Exercise 2b (3 points):  Include union participation
> summary(lm(lnwage ~ union + ed, data=cps78))

Call:
lm(formula = lnwage ~ union + ed, data = cps78)

Residuals:
      Min        1Q    Median        3Q       Max
-2.331754 -0.294114  0.001475  0.263843  1.678532

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.859166   0.091630   9.376  < 2e-16 ***
unionUnion  0.305129   0.041800   7.300 1.02e-12 ***
ed          0.058122   0.006952   8.361 4.44e-16 ***
```

```
---
Signif. codes:  0  ‘***’  0.001  ‘**’  0.01  ‘*’  0.05  ‘.’  0.1  ‘ ’  1

Residual standard error: 0.4481 on 547 degrees of freedom
Multiple R-Squared: 0.1673,Adjusted R-squared: 0.1642
F-statistic: 54.93 on 2 and 547 degrees of freedom,p-value:     0

> exp(0.058)
[1] 1.059715
> exp(0.305129)
[1] 1.3568
> # Union members have 36 percent higher wages
> # The test whether union and nonunion members have the same intercept
> # is the same as the test whether the union dummy is 0.
> # t-value = 7.300 which is highly significant,
> # i.e., they are different.

> #The union variable is labeled unionUnion, because
> #it is labeled 1 for Union and 0 for Nonun. Check with the command
> contrasts(cps78$union)
      Union
Nonun     0
Union     1
> #One sees it also if one runs
> model.matrix(lnwage ~ union + ed, data=cps78)
   (Intercept) union ed
 1           1     0 12
 2           1     1 12
 3           1     1  6
 4           1     1 12
 5           1     0 12
> #etc, rest of output flushed
> #and compares this with
> cps78$union[1:5]
[1] Nonun Union Union Union Nonun
Levels:  Nonun Union
> #Consequently, the intercept for nonunion is 0.8592
> #and the intercept for union is 0.8592+0.3051=1.1643.
> #Can I have a different set of dummies constructed from this factor?
> #We will first do
```



```
> ##Exercise 2e (2 points)
> contrasts(union)<-matrix(c(1,0),nrow=2,ncol=1)
> #This generates a new contrast matrix
> #which covers up that in cps78
> #Note that I do not say "data=cps78" in the next command:
> summary(lm(lnwage ~ union + ed))

Call:
lm(formula = lnwage ~ union + ed)

Residuals:
      Min        1Q    Median        3Q       Max
-2.331754 -0.294114  0.001475  0.263843  1.678532

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.164295   0.090453  12.872  < 2e-16 ***
union1      -0.305129   0.041800  -7.300 1.02e-12 ***
ed           0.058122   0.006952   8.361 4.44e-16 ***
---
Signif. codes:  0  ‘***’  0.001  ‘**’  0.01  ‘*’  0.05  ‘.’  0.1  ‘ ’  1

Residual standard error: 0.4481 on 547 degrees of freedom
Multiple R-Squared: 0.1673,Adjusted R-squared: 0.1642
F-statistic: 54.93 on 2 and 547 degrees of freedom,p-value:     0

> #Here the coefficients are different,
> #but it is consistent with the above result.
> ##Ecercise 2c (2 points):  If I want to have two contrasts from this one du
> contrasts(union,2)<-matrix(c(1,0,0,1),nrow=2,ncol=2)
> #The additional argument 2
> #specifies different number of contrasts than it expects
> #Now I have to supress the intercept in the regression
> summary(lm(lnwage ~ union + ed - 1))

Call:
lm(formula = lnwage ~ union + ed - 1)

Residuals:
      Min        1Q    Median        3Q       Max
```

```
-2.331754 -0.294114  0.001475  0.263843  1.678532

Coefficients:
       Estimate Std. Error t value Pr(>|t|)
union1 0.859166   0.091630   9.376  < 2e-16 ***
union2 1.164295   0.090453  12.872  < 2e-16 ***
ed     0.058122   0.006952   8.361 4.44e-16 ***
---
Signif. codes:  0  ‘***’  0.001  ‘**’  0.01  ‘*’  0.05  ‘.’  0.1  ‘ ’  1

Residual standard error: 0.4481 on 547 degrees of freedom
Multiple R-Squared: 0.9349,Adjusted R-squared: 0.9345
F-statistic:  2617 on 3 and 547 degrees of freedom,p-value:     0

> #actually it was unnecessary to construct the contrast matrix.
> #If we regress with a categorical variable without
> #an intercept, R will automatically use dummies for all levels:
> lm(lnwage ~ union + ed - 1, data=cps85)

Call:
lm(formula = lnwage ~ union + ed - 1, data = cps85)

Coefficients:
unionNonunion     unionUnion             ed
       0.9926         1.2909         0.0778


> ##Exercise 2d (1 point) Why is it not possible to include two dummies p
> # an intercept?  Because the two dummies add to 1,
> # you have perfect collinearity


> ###Exercise 3a (2 points):
> summary(lm(lnwage ~ ed + ex + I(ex^2), data=cps78))
> #All coefficients are highly significant, but the R^2 is only 0.2402
> #Returns to experience are positive and decline with increase in experi
> ##Exercise 3b (2 points):
> summary(lm(lnwage ~ gender + ed + ex + I(ex^2), data=cps78))
> contrasts(cps78$gender)
> #We see here that gender is coded 0 for female and 1 for male;
```



```
> #by default, the levels in a factor variable occur in alphabetical order.
> #Intercept in our regression = 0.1909203 (this is for female),
> #genderMale has coefficient = 0.3351771,
> #i.e., the intercept for women is 0.5260974
> #Gender is highly significant
> ##Exercise 3c (2 points):
> summary(lm(lnwage ~ gender + marr + ed + ex + I(ex^2), data=cps78))
> #Coefficient of marr in this is insignificant
> ##Exercise 3d (1 point) asks to construct a variable which we do
> #not need when we use factor variables
> ##Exercise 3e (3 points): For interaction term do
> summary(lm(lnwage ~ gender * marr + ed + ex + I(ex^2), data=cps78))

Call:
lm(formula = lnwage ~ gender * marr + ed + ex + I(ex^2), data = cps78)

Residuals:
     Min       1Q   Median       3Q      Max
-2.45524 -0.24566  0.01969  0.23102  1.42437

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            0.1893919  0.1042613   1.817  0.06984 .
genderMale             0.3908782  0.0467018   8.370 4.44e-16 ***
marrSingle             0.0507811  0.0557198   0.911  0.36251
ed                     0.0738640  0.0066154  11.165  < 2e-16 ***
ex                     0.0265297  0.0049741   5.334 1.42e-07 ***
I(ex^2)               -0.0003161  0.0001057  -2.990  0.00291 **
genderMale:marrSingle -0.1586452  0.0750830  -2.113  0.03506 *
---
Signif. codes:  0  ‘***’  0.001  ‘**’  0.01  ‘*’  0.05  ‘.’  0.1  ‘ ’  1

Residual standard error: 0.3959 on 543 degrees of freedom
Multiple R-Squared: 0.3547,Adjusted R-squared: 0.3476
F-statistic: 49.75 on 6 and 543 degrees of freedom,p-value:     0

> #Being married raises the wage for men by 13% but lowers it for women by 3%
> ###Exercise 4a (5 points):
> summary(lm(lnwage ~ union + gender + race + ed + ex + I(ex^2), data=cps78))
```

```
Call:
lm(formula = lnwage ~ union + gender + race + ed + ex + I(ex^2),
    data = cps78)

Residuals:
     Min       1Q   Median       3Q      Max
-2.41914 -0.23674  0.01682  0.21821  1.31584

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.1549723  0.1068589   1.450  0.14757
unionUnion   0.2071429  0.0368503   5.621 3.04e-08 ***
genderMale   0.3060477  0.0344415   8.886  < 2e-16 ***
raceNonwh   -0.1301175  0.0830156  -1.567  0.11761
raceOther    0.0271477  0.0688277   0.394  0.69342
ed           0.0746097  0.0066521  11.216  < 2e-16 ***
ex           0.0261914  0.0047174   5.552 4.43e-08 ***
I(ex^2)     -0.0003082  0.0001015  -3.035  0.00252 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3845 on 542 degrees of freedom
Multiple R-Squared: 0.3924,Adjusted R-squared: 0.3846
F-statistic: 50.01 on 7 and 542 degrees of freedom,p-value:     0

> exp(-0.1301175)
[1] 0.8779923

> #Being Hispanic lowers wages by 2.7%, byut being black lowers them
> #by 12.2 %

> #At what level of ex is lnwage maximized?
> #exeffect = 0.0261914 * ex -0.0003082 * ex^2
> #derivative =  0.0261914 - 2 * 0.0003082 * ex
> #derivative = 0 for ex=0.0261914/(2*0.0003082)

> 0.0261914/(2*0.0003082)
[1] 42.49091

> # age - ed - 6 =  42.49091
```



```
> # age = ed +  48.49091
> # for 8, 12, and 16 years of schooling the max earnings
> # are at ages 56.5, 60.5, and 64.5 years
> ##Exercise 4b (4 points) is a graph, not done here
> ##Exercise 4c (5 points)
> summary(lm(lnwage ~ gender + union + race + ed + ex + I(ex^2) + I(ed*ex), d

Call:
lm(formula = lnwage ~ gender + union + race + ed + ex + I(ex^2) +
    I(ed * ex), data = cps78)

Residuals:

     Min       1Q   Median       3Q      Max
-2.41207 -0.23922  0.01463  0.21645  1.32051

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0396495  0.1789073   0.222 0.824693
genderMale   0.3042639  0.0345241   8.813  < 2e-16 ***
unionUnion   0.2074045  0.0368638   5.626 2.96e-08 ***
raceNonwh   -0.1323898  0.0830908  -1.593 0.111673
raceOther    0.0319829  0.0691124   0.463 0.643718
ed           0.0824154  0.0117716   7.001 7.55e-12 ***
ex           0.0328854  0.0095716   3.436 0.000636 ***
I(ex^2)     -0.0003574  0.0001186  -3.013 0.002704 **
I(ed * ex)  -0.0003813  0.0004744  -0.804 0.421835
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3846 on 541 degrees of freedom
Multiple R-Squared: 0.3932,Adjusted R-squared: 0.3842
F-statistic: 43.81 on 8 and 541 degrees of freedom,p-value:     0
> #Maximum earnings ages must be computed as before

> ##Exercise 4d (4 points) not done here
> ##Exercise 4e (6 points) not done here
> ###Exercise 5a (3 points):
> #Naive approach to estimate impact of unionization on wages:
> summary(lm(lnwage ~ gender + union + race + ed + ex + I(ex^2), data=cps78))
```

```
Call:
lm(formula = lnwage ~ gender + union + race + ed + ex + I(ex^2),
    data = cps78)

Residuals:
     Min       1Q   Median       3Q      Max
-2.41914 -0.23674  0.01682  0.21821  1.31584

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.1549723  0.1068589   1.450  0.14757
genderMale   0.3060477  0.0344415   8.886  < 2e-16 ***
unionUnion   0.2071429  0.0368503   5.621 3.04e-08 ***
raceNonwh   -0.1301175  0.0830156  -1.567  0.11761
raceOther    0.0271477  0.0688277   0.394  0.69342
ed           0.0746097  0.0066521  11.216  < 2e-16 ***
ex           0.0261914  0.0047174   5.552 4.43e-08 ***
I(ex^2)     -0.0003082  0.0001015  -3.035  0.00252 **
---
Signif. codes:  0  ‘***’  0.001  ‘**’  0.01  ‘*’  0.05  ‘.’  0.1  ‘ ’  1

Residual standard error: 0.3845 on 542 degrees of freedom
Multiple R-Squared: 0.3924,Adjusted R-squared: 0.3846
F-statistic: 50.01 on 7 and 542 degrees of freedom,p-value:     0

> # What is wrong with the above?  It assumes that unions
> # only affect the intercept, everything else is the same
> ##Exercise 5b (2 points)

> tapply(lnwage, union, mean)
   Nonun    Union
1.600901 1.863137
> tapply(ed, union, mean)
   Nonun    Union
12.76178 12.02381
> table(gender, union)
        union
gender   Nonun Union
  Female   159    48
```



```
  Male     223   120
> table(race, union)
       union
race    Nonun Union
  Hisp     29     7
  Nonwh    35    22
  Other   318   139
> 7/(7+29)
[1] 0.1944444
> 22/(22+35)
[1] 0.3859649
> 139/(318+139)
[1] 0.3041575
> #19% of Hispanic, 39% of Nonwhite, and 30% of other (white) workers
> #in the sample are in unions
> ##Exercise 5c (3 points)
> summary(lm(lnwage ~ gender + race + ed + ex + I(ex^2), data=cps78, subset=u

Call:
lm(formula = lnwage ~ gender + race + ed + ex + I(ex^2), data = cps78,
    subset = union == "Union")

Residuals:
    Min      1Q  Median      3Q     Max
-2.3307 -0.1853  0.0160  0.2199  1.1992

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.9261456  0.2321964   3.989 0.000101 ***
genderMale   0.2239370  0.0684894   3.270 0.001317 **
raceNonwh   -0.3066717  0.1742287  -1.760 0.080278 .
raceOther   -0.0741660  0.1562131  -0.475 0.635591
ed           0.0399500  0.0138311   2.888 0.004405 **
ex           0.0313820  0.0098938   3.172 0.001814 **
I(ex^2)     -0.0004526  0.0002022  -2.239 0.026535 *
---
Signif. codes:  0  ‘***’  0.001  ‘**’  0.01  ‘*’  0.05  ‘.’  0.1  ‘ ’  1

Residual standard error: 0.3928 on 161 degrees of freedom
Multiple R-Squared: 0.2019,Adjusted R-squared: 0.1721
```

```
F-statistic: 6.787 on 6 and 161 degrees of freedom,p-value: 1.975e-06

> summary(lm(lnwage ~ gender + race + ed + ex + I(ex^2), data=cps78, subs

Call:
lm(formula = lnwage ~ gender + race + ed + ex + I(ex^2), data = cps78,
    subset = union == "Nonun")

Residuals:
     Min       1Q   Median       3Q      Max
-1.39107 -0.23775  0.01040  0.23337  1.29073

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.0095668  0.1193399  -0.080   0.9361
genderMale   0.3257661  0.0397961   8.186 4.22e-15 ***
raceNonwh   -0.0652018  0.0960570  -0.679   0.4977
raceOther    0.0444133  0.0761628   0.583   0.5602
ed           0.0852212  0.0075554  11.279  < 2e-16 ***
ex           0.0253813  0.0053710   4.726 3.25e-06 ***
I(ex^2)     -0.0002841  0.0001187  -2.392   0.0172 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3778 on 375 degrees of freedom
Multiple R-Squared: 0.4229,Adjusted R-squared: 0.4137
F-statistic:  45.8 on 6 and 375 degrees of freedom,p-value:     0

> #Are union-nonunion differences larger for females than males?
> #For this look at the intercepts for males and females in
> #the two regressions.  Say for white males and females:
> 0.9261456-0.0741660+0.2239370
[1] 1.075917
> 0.9261456-0.0741660
[1] 0.8519796
> -0.0095668+0.0444133+0.3257661
[1] 0.3606126
> -0.0095668+0.0444133
[1] 0.0348465
> 1.075917-0.3606126
```



```
[1] 0.7153044
> 0.8519796-0.0348465
[1] 0.8171331
>

>                #White Males     White Females
> #Union          1.075917        0.8519796
> #Nonunion       0.3606126       0.0348465
> #Difference     0.7153044       0.8171331
> #Difference is greater for women
> ###Exercise 6a (5 points)
> summary(lm(lnwage ~ gender + union + race + ed + ex + I(ex^2)))

Call:
lm(formula = lnwage ~ gender + union + race + ed + ex + I(ex^2))

Residuals:
     Min       1Q   Median       3Q      Max
-2.41914 -0.23674  0.01682  0.21821  1.31584

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.1549723  0.1068589   1.450  0.14757
genderMale   0.3060477  0.0344415   8.886  < 2e-16 ***
unionUnion   0.2071429  0.0368503   5.621 3.04e-08 ***
raceNonwh   -0.1301175  0.0830156  -1.567  0.11761
raceOther    0.0271477  0.0688277   0.394  0.69342
ed           0.0746097  0.0066521  11.216  < 2e-16 ***
ex           0.0261914  0.0047174   5.552 4.43e-08 ***
I(ex^2)     -0.0003082  0.0001015  -3.035  0.00252 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3845 on 542 degrees of freedom
Multiple R-Squared: 0.3924,Adjusted R-squared: 0.3846
F-statistic: 50.01 on 7 and 542 degrees of freedom,p-value:     0

> #To test whether Nonwh and Hisp have same intercept
> #one might generate a contrast matrix which collapses those
> #two and then run it and make an F-test
```

```
> #or make a contrast matrix which has this difference as one of
> #the dummies and use the t-test for that dummy
> ##Exercise 6b (2 points)
> table(race)
race
 Hisp Nonwh Other
   36    57   457
> tapply(lnwage, race, mean)
    Hisp    Nonwh    Other
1.529647 1.513404 1.713829
> tapply(lnwage, race, ed)
Error in get(x, envir, mode, inherits) : variable "ed" was not found
> tapply(ed, race, mean)
    Hisp    Nonwh    Other
10.30556 11.71930 12.81400
> table(gender, race)
        race
gender   Hisp Nonwh Other
  Female   12    28   167
  Male     24    29   290
> #Blacks, almost as many women than men, hispanic twice as many men,
> #Whites in between

>
> #Additional stuff:
> #There are two outliers in cps78 with wages of less than $1 per hour,
> #Both service workers, perhaps waitresses who did not report her tips?
> #What are the commands for extracting certain observations
> #by certain criteria and just print them?  The split command.
>
> #Interesting to do
> loess(lnwage ~ ed + ex, data=cps78)
> #loess is appropriate here because there are strong interation terms
> #How can one do loess after taking out the effects of gender for instan
> #Try the following, but I did not try it out yet:
> gam(lnwage ~ lo(ed,ex) + gender, data=cps78)
> #I should put more plotting commands in!
```

CHAPTER 23

The Mean Squared Error as an Initial Criterion of Precision

The question how "close" two random variables are to each other is a central concern in statistics. The goal of statistics is to find observed random variables which are "close" to the unobserved parameters or random outcomes of interest. These observed random variables are usually called "estimators" if the unobserved magnitude is nonrandom, and "predictors" if it is random. For *scalar* random variables we will use the mean squared error as a criterion for closeness. Its definition is $\mathrm{MSE}[\hat{\phi};\phi]$ (read it: mean squared error of $\hat{\phi}$ as an estimator or predictor, whatever the case may be, of ϕ):

$$\mathrm{MSE}[\hat{\phi};\phi] = \mathrm{E}[(\hat{\phi}-\phi)^2] \tag{23.0.1}$$

For our purposes, therefore, the estimator (or predictor) $\hat{\phi}$ of the unknown parameter (or unobserved random variable) ϕ is no worse than the alternative $\tilde{\phi}$ if $\mathrm{MSE}[\hat{\phi};\phi] \leq \mathrm{MSE}[\tilde{\phi};\phi]$. This is a criterion which can be applied before any observations are collected and actual estimations are made; it is an "initial" criterion regarding the expected average performance in a series of future trials (even though, in economics, usually only one trial is made).

23.1. Comparison of Two Vector Estimators

If one wants to compare two *vector* estimators, say $\hat{\boldsymbol{\phi}}$ and $\tilde{\boldsymbol{\phi}}$, it is often impossible to say which of two estimators is better. It may be the case that $\hat{\phi}_1$ is better than $\tilde{\phi}_1$ (in terms of MSE or some other criterion), but $\hat{\phi}_2$ is worse than $\tilde{\phi}_2$. And even if every component ϕ_i is estimated better by $\hat{\phi}_i$ than by $\tilde{\phi}_i$, certain linear combinations $\boldsymbol{t}^\top\boldsymbol{\phi}$ of the components of $\boldsymbol{\phi}$ may be estimated better by $\boldsymbol{t}^\top\tilde{\boldsymbol{\phi}}$ than by $\boldsymbol{t}^\top\hat{\boldsymbol{\phi}}$.

Problem 294. *2 points Construct an example of two vector estimators $\hat{\boldsymbol{\phi}}$ and $\tilde{\boldsymbol{\phi}}$ of the same random vector $\boldsymbol{\phi} = \begin{bmatrix}\phi_1 & \phi_2\end{bmatrix}^\top$, so that $\mathrm{MSE}[\hat{\phi}_i;\phi_i] < \mathrm{MSE}[\tilde{\phi}_i;\phi_i]$ for $i = 1,2$ but $\mathrm{MSE}[\hat{\phi}_1+\hat{\phi}_2;\phi_1+\phi_2] > \mathrm{MSE}[\tilde{\phi}_1+\tilde{\phi}_2;\phi_1+\phi_2]$. Hint: it is easiest to use an example in which all random variables are constants. Another hint: the geometric*

analog would be to find two vectors in a plane $\hat{\boldsymbol{\phi}}$ and $\tilde{\boldsymbol{\phi}}$. In each component (i.e., projection on the axes), $\hat{\boldsymbol{\phi}}$ is closer to the origin than $\tilde{\boldsymbol{\phi}}$. But in the projection on the diagonal, $\tilde{\boldsymbol{\phi}}$ is closer to the origin than $\hat{\boldsymbol{\phi}}$.

Answer. In the simplest counterexample, all variables involved are constants: $\boldsymbol{\phi} = \begin{bmatrix}0\\0\end{bmatrix}$, $\hat{\boldsymbol{\phi}} = \begin{bmatrix}1\\1\end{bmatrix}$, and $\tilde{\boldsymbol{\phi}} = \begin{bmatrix}-2\\2\end{bmatrix}$. □

One can only then say unambiguously that the vector $\hat{\boldsymbol{\phi}}$ is a no worse estimator than $\tilde{\boldsymbol{\phi}}$ if its MSE is smaller or equal *for every linear combination*. Theorem 23.1.1 will show that this is the case if and only if the *$\mathcal{MSE}$-matrix* of $\hat{\boldsymbol{\phi}}$ is smaller, by a nonnegative definite matrix, than that of $\tilde{\boldsymbol{\phi}}$. If this is so, then theorem 23.1.1 says that not only the MSE of all linear transformations, but also all other nonnegative definite quadratic loss functions involving these vectors (such as the trace of the $\mathcal{MSE}$-matrix, which is an often-used criterion) are minimized. In order to formulate and prove this, we first need a formal definition of the $\mathcal{MSE}$-matrix. We write $\mathcal{MSE}$ for the matrix and MSE for the scalar mean squared error. The $\mathcal{MSE}$-matrix of $\hat{\boldsymbol{\phi}}$ as an estimator of $\boldsymbol{\phi}$ is defined as

$$\mathcal{MSE}[\hat{\boldsymbol{\phi}};\boldsymbol{\phi}] = \mathcal{E}[(\hat{\boldsymbol{\phi}}-\boldsymbol{\phi})(\hat{\boldsymbol{\phi}}-\boldsymbol{\phi})^\top]\,. \tag{23.1.1}$$

Problem 295. *2 points Let $\boldsymbol{\theta}$ be a vector of possibly random parameters, and $\hat{\boldsymbol{\theta}}$ an estimator of $\boldsymbol{\theta}$. Show that*

$$\mathcal{MSE}[\hat{\boldsymbol{\theta}};\boldsymbol{\theta}] = \mathcal{V}[\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}] + (\mathcal{E}[\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}])(\mathcal{E}[\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}])^\top. \tag{23.1.2}$$

Don't assume the scalar result but make a proof that is good for vectors and scalars.

Answer. For any random vector $\boldsymbol{x}$ follows

$$\begin{aligned}\mathcal{E}[\boldsymbol{x}\boldsymbol{x}^\top] &= \mathcal{E}\left[(\boldsymbol{x}-\mathcal{E}[\boldsymbol{x}]+\mathcal{E}[\boldsymbol{x}])(\boldsymbol{x}-\mathcal{E}[\boldsymbol{x}]+\mathcal{E}[\boldsymbol{x}])^\top\right]\\ &= \mathcal{E}\left[(\boldsymbol{x}-\mathcal{E}[\boldsymbol{x}])(\boldsymbol{x}-\mathcal{E}[\boldsymbol{x}])^\top\right] - \mathcal{E}\left[(\boldsymbol{x}-\mathcal{E}[\boldsymbol{x}])\,\mathcal{E}[\boldsymbol{x}]^\top\right] - \mathcal{E}\left[\mathcal{E}[\boldsymbol{x}](\boldsymbol{x}-\mathcal{E}[\boldsymbol{x}])^\top\right] + \mathcal{E}\left[\mathcal{E}[\boldsymbol{x}]\,\mathcal{E}[\boldsymbol{x}]^\top\right]\\ &= \mathcal{V}[\boldsymbol{x}] - \boldsymbol{O} - \boldsymbol{O} + \mathcal{E}[\boldsymbol{x}]\,\mathcal{E}[\boldsymbol{x}]^\top.\end{aligned}$$

Setting $\boldsymbol{x} = \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}$ the statement follows. □

If $\boldsymbol{\theta}$ is nonrandom, formula (23.1.2) simplifies slightly, since in this case $\mathcal{V}[\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}] = \mathcal{V}[\hat{\boldsymbol{\theta}}]$. In this case, the $\mathcal{MSE}$ matrix is the covariance matrix plus the squared bias matrix. If $\boldsymbol{\theta}$ is nonrandom and in addition $\hat{\boldsymbol{\theta}}$ is unbiased, then the $\mathcal{MSE}$-matrix coincides with the covariance matrix.

Theorem 23.1.1. *Assume $\hat{\boldsymbol{\phi}}$ and $\tilde{\boldsymbol{\phi}}$ are two estimators of the parameter $\boldsymbol{\phi}$ (which is allowed to be random itself). Then conditions* (23.1.3), (23.1.4), *and*

(23.1.5) *are equivalent:*

(23.1.3) *For every constant vector* $\boldsymbol{t}$, $\quad \mathrm{MSE}[\boldsymbol{t}^\top\hat{\boldsymbol{\phi}};\boldsymbol{t}^\top\boldsymbol{\phi}] \leq \mathrm{MSE}[\boldsymbol{t}^\top\tilde{\boldsymbol{\phi}};\boldsymbol{t}^\top\boldsymbol{\phi}]$

(23.1.4) $\quad \mathcal{MSE}[\tilde{\boldsymbol{\phi}};\boldsymbol{\phi}] - \mathcal{MSE}[\hat{\boldsymbol{\phi}};\boldsymbol{\phi}]$ *is a nonnegative definite matrix*

(23.1.5) *For every nnd* $\boldsymbol{\Theta}$, $\quad \mathrm{E}\big[(\hat{\boldsymbol{\phi}}-\boldsymbol{\phi})^\top\boldsymbol{\Theta}(\hat{\boldsymbol{\phi}}-\boldsymbol{\phi})\big] \leq \mathrm{E}\big[(\tilde{\boldsymbol{\phi}}-\boldsymbol{\phi})^\top\boldsymbol{\Theta}(\tilde{\boldsymbol{\phi}}-\boldsymbol{\phi})\big]$.

Proof. Call $\mathcal{MSE}[\tilde{\boldsymbol{\phi}};\boldsymbol{\phi}] = \sigma^2\boldsymbol{\Xi}$ and $\mathcal{MSE}[\hat{\boldsymbol{\phi}};\boldsymbol{\phi}] = \sigma^2\boldsymbol{\Omega}$. To show that (23.1.3) implies (23.1.4), simply note that $\mathrm{MSE}[\boldsymbol{t}^\top\hat{\boldsymbol{\phi}};\boldsymbol{t}^\top\boldsymbol{\phi}] = \sigma^2\boldsymbol{t}^\top\boldsymbol{\Omega}\boldsymbol{t}$ and likewise $\mathrm{MSE}[\boldsymbol{t}^\top\tilde{\boldsymbol{\phi}};\boldsymbol{t}^\top\boldsymbol{\phi}]$ $\sigma^2\boldsymbol{t}^\top\boldsymbol{\Xi}\boldsymbol{t}$. Therefore (23.1.3) is equivalent to $\boldsymbol{t}^\top(\boldsymbol{\Xi}-\boldsymbol{\Omega})\boldsymbol{t} \geq 0$ for all $\boldsymbol{t}$, which is the defining property making $\boldsymbol{\Xi}-\boldsymbol{\Omega}$ nonnegative definite.

Here is the proof that (23.1.4) implies (23.1.5):

$$\mathrm{E}[(\hat{\boldsymbol{\phi}}-\boldsymbol{\phi})^\top\boldsymbol{\Theta}(\hat{\boldsymbol{\phi}}-\boldsymbol{\phi})] = \mathrm{E}[\mathrm{tr}\big((\hat{\boldsymbol{\phi}}-\boldsymbol{\phi})^\top\boldsymbol{\Theta}(\hat{\boldsymbol{\phi}}-\boldsymbol{\phi})\big)] =$$
$$= \mathrm{E}[\mathrm{tr}\big(\boldsymbol{\Theta}(\hat{\boldsymbol{\phi}}-\boldsymbol{\phi})(\hat{\boldsymbol{\phi}}-\boldsymbol{\phi})^\top\big)] = \mathrm{tr}\big(\boldsymbol{\Theta}\,\mathcal{E}[(\hat{\boldsymbol{\phi}}-\boldsymbol{\phi})(\hat{\boldsymbol{\phi}}-\boldsymbol{\phi})^\top]\big) = \sigma^2\,\mathrm{tr}\big(\boldsymbol{\Theta}\boldsymbol{\Omega}\big)$$

and in the same way

$$\mathrm{E}[(\tilde{\boldsymbol{\phi}}-\boldsymbol{\phi})^\top\boldsymbol{\Theta}(\tilde{\boldsymbol{\phi}}-\boldsymbol{\phi})] = \sigma^2\,\mathrm{tr}\big(\boldsymbol{\Theta}\boldsymbol{\Xi}\big).$$

The difference in the expected quadratic forms is therefore $\sigma^2\,\mathrm{tr}\big(\boldsymbol{\Theta}(\boldsymbol{\Xi}-\boldsymbol{\Omega})\big)$. By assumption, $\boldsymbol{\Xi}-\boldsymbol{\Omega}$ is nonnegative definite. Therefore, by theorem A.5.6 in the Mathematical Appendix, or by Problem 296 below, this trace is nonnegative.

To complete the proof, (23.1.5) has (23.1.3) as a special case if one sets $\boldsymbol{\Theta} = \boldsymbol{t}\boldsymbol{t}^\top$. □

Problem 296. *Show that if* $\boldsymbol{\Theta}$ *and* $\boldsymbol{\Sigma}$ *are symmetric and nonnegative definite, then* $\mathrm{tr}(\boldsymbol{\Theta}\boldsymbol{\Sigma}) \geq 0$. *You are allowed to use that* $\mathrm{tr}(\boldsymbol{A}\boldsymbol{B}) = \mathrm{tr}(\boldsymbol{B}\boldsymbol{A})$, *that the trace of a nonnegative definite matrix is* ≥ 0, *and Problem 129 (which is trivial).*

Answer. Write $\boldsymbol{\Theta} = \boldsymbol{R}\boldsymbol{R}^\top$; then $\mathrm{tr}(\boldsymbol{\Theta}\boldsymbol{\Sigma}) = \mathrm{tr}(\boldsymbol{R}\boldsymbol{R}^\top\boldsymbol{\Sigma}) = \mathrm{tr}(\boldsymbol{R}^\top\boldsymbol{\Sigma}\boldsymbol{R}) \geq 0$. □

Problem 297. *Consider two very simple-minded estimators of the unknown nonrandom parameter vector* $\boldsymbol{\phi} = \begin{bmatrix}\phi_1\\ \phi_2\end{bmatrix}$. *Neither of these estimators depends on any observations, they are constants. The first estimator is* $\hat{\boldsymbol{\phi}} = \left[\begin{smallmatrix}11\\11\end{smallmatrix}\right]$, *and the second is* $\tilde{\boldsymbol{\phi}} = \left[\begin{smallmatrix}12\\8\end{smallmatrix}\right]$.

• a. *2 points Compute the* $\mathcal{MSE}$*-matrices of these two estimators if the true value of the parameter vector is* $\boldsymbol{\phi} = \left[\begin{smallmatrix}10\\10\end{smallmatrix}\right]$. *For which estimator is the trace of the* $\mathcal{MSE}$ *matrix smaller?*

Answer. $\hat{\boldsymbol{\phi}}$ has smaller trace of the $\mathcal{MSE}$-matrix.

$$\hat{\boldsymbol{\phi}}-\boldsymbol{\phi} = \begin{bmatrix}1\\1\end{bmatrix}$$
$$\mathcal{MSE}[\hat{\boldsymbol{\phi}};\boldsymbol{\phi}] = \mathcal{E}[(\hat{\boldsymbol{\phi}}-\boldsymbol{\phi})(\hat{\boldsymbol{\phi}}-\boldsymbol{\phi})^\top]$$
$$= \mathcal{E}[\begin{bmatrix}1\\1\end{bmatrix}\begin{bmatrix}1 & 1\end{bmatrix}] = \mathcal{E}[\begin{bmatrix}1 & 1\\1 & 1\end{bmatrix}] = \begin{bmatrix}1 & 1\\1 & 1\end{bmatrix}$$
$$\tilde{\boldsymbol{\phi}}-\boldsymbol{\phi} = \begin{bmatrix}2\\-2\end{bmatrix}$$
$$\mathcal{MSE}[\tilde{\boldsymbol{\phi}};\boldsymbol{\phi}] = \begin{bmatrix}4 & -4\\-4 & 4\end{bmatrix}$$

Note that both $\mathcal{MSE}$-matrices are singular, i.e., both estimators allow an error-free look at certain linear combinations of the parameter vector. □

• b. *1 point Give two vectors* $\boldsymbol{g} = \left[\begin{smallmatrix}g_1\\g_2\end{smallmatrix}\right]$ *and* $\boldsymbol{h} = \left[\begin{smallmatrix}h_1\\h_2\end{smallmatrix}\right]$ *satisfying* $\mathrm{MSE}[\boldsymbol{g}^\top\hat{\boldsymbol{\phi}};\boldsymbol{g}^\top\boldsymbol{\phi}] < \mathrm{MSE}[\boldsymbol{g}^\top\tilde{\boldsymbol{\phi}};\boldsymbol{g}^\top\boldsymbol{\phi}]$ *and* $\mathrm{MSE}[\boldsymbol{h}^\top\hat{\boldsymbol{\phi}};\boldsymbol{h}^\top\boldsymbol{\phi}] > \mathrm{MSE}[\boldsymbol{h}^\top\tilde{\boldsymbol{\phi}};\boldsymbol{h}^\top\boldsymbol{\phi}]$ *(*$\boldsymbol{g}$ *and* $\boldsymbol{h}$ *are not unique; there are many possibilities).*

Answer. With $\boldsymbol{g} = \left[\begin{smallmatrix}1\\-1\end{smallmatrix}\right]$ and $\boldsymbol{h} = \left[\begin{smallmatrix}1\\1\end{smallmatrix}\right]$ for instance we get $\boldsymbol{g}^\top\hat{\boldsymbol{\phi}}-\boldsymbol{g}^\top\boldsymbol{\phi} = 0$, $\boldsymbol{g}^\top\tilde{\boldsymbol{\phi}}-\boldsymbol{g}^\top\boldsymbol{\phi} = 4$, $\boldsymbol{h}^\top\hat{\boldsymbol{\phi}};\boldsymbol{h}^\top\boldsymbol{\phi} = 2$, $\boldsymbol{h}^\top\tilde{\boldsymbol{\phi}};\boldsymbol{h}^\top\boldsymbol{\phi} = 0$, therefore $\mathrm{MSE}[\boldsymbol{g}^\top\hat{\boldsymbol{\phi}};\boldsymbol{g}^\top\boldsymbol{\phi}] = 0$, $\mathrm{MSE}[\boldsymbol{g}^\top\tilde{\boldsymbol{\phi}};\boldsymbol{g}^\top\boldsymbol{\phi}] = 16$, $\mathrm{MSE}[\boldsymbol{h}^\top\hat{\boldsymbol{\phi}};\boldsymbol{h}^\top\boldsymbol{\phi}] = 4$, $\mathrm{MSE}[\boldsymbol{h}^\top\tilde{\boldsymbol{\phi}};\boldsymbol{h}^\top\boldsymbol{\phi}] = 0$. An alternative way to compute this is e.g.

$$\mathcal{MSE}[\boldsymbol{h}^\top\tilde{\boldsymbol{\phi}};\boldsymbol{h}^\top\boldsymbol{\phi}] = \begin{bmatrix}1 & -1\end{bmatrix}\begin{bmatrix}4 & -4\\-4 & 4\end{bmatrix}\begin{bmatrix}1\\-1\end{bmatrix} = 16$$

□

• c. *1 point Show that neither* $\mathcal{MSE}[\hat{\boldsymbol{\phi}};\boldsymbol{\phi}] - \mathcal{MSE}[\tilde{\boldsymbol{\phi}};\boldsymbol{\phi}]$ *nor* $\mathcal{MSE}[\tilde{\boldsymbol{\phi}};\boldsymbol{\phi}] - \mathcal{MSE}[\hat{\boldsymbol{\phi}};\boldsymbol{\phi}]$ *is a nonnegative definite matrix. Hint: you are allowed to use the mathematical fact that if a matrix is nonnegative definite, then its determinant is nonnegative.*

Answer.

$$\mathcal{MSE}[\tilde{\boldsymbol{\phi}};\boldsymbol{\phi}] - \mathcal{MSE}[\hat{\boldsymbol{\phi}};\boldsymbol{\phi}] = \begin{bmatrix}3 & -5\\-5 & 3\end{bmatrix} \tag{23.1.6}$$

Its determinant is negative, and the determinant of its negative is also negative. □

CHAPTER 24

Sampling Properties of the Least Squares Estimator

The estimator $\hat{\boldsymbol{\beta}}$ was derived from a *geometric* argument, and everything which we showed so far are what [**DM93**, p. 3] calls its *numerical* as opposed to its *statistical* properties. But $\hat{\boldsymbol{\beta}}$ has also nice *statistical* or *sampling* properties. We are assuming right now the specification given in (18.1.3), in which $\boldsymbol{X}$ is an arbitrary matrix of full column rank, and we are not assuming that the errors must be Normally distributed. The assumption that $\boldsymbol{X}$ is nonrandom means that repeated samples are taken with the same $\boldsymbol{X}$-matrix. This is often true for experimental data, but not in econometrics. The sampling properties which we are really interested in are those where also the $\boldsymbol{X}$-matrix is random; we will derive those later. For this later derivation, the properties with fixed $\boldsymbol{X}$-matrix, which we are going to discuss presently, will be needed as an intermediate step. The assumption of fixed $\boldsymbol{X}$ is therefore a preliminary technical assumption, to be dropped later.

In order to know how good the estimator $\hat{\boldsymbol{\beta}}$ is, one needs the statistical properties of its "sampling error" $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$. This sampling error has the following formula:

$$\begin{aligned}\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\mathbf{y} - (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta} = \\ &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}) = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon}\end{aligned} \tag{24.0.7}$$

From (24.0.7) follows immediately that $\hat{\boldsymbol{\beta}}$ is unbiased, since $\mathcal{E}[(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon}] = \boldsymbol{o}$.

Unbiasedness does not make an estimator better, but many good estimators are unbiased, and it simplifies the math.

We will use the $\mathcal{MSE}$-matrix as a criterion for how good an estimator of a vector of unobserved parameters is. Chapter 23 gave some reasons why this is a sensible criterion (compare [**DM93**, Chapter 5.5]).

24.1. The Gauss Markov Theorem

Returning to the least squares estimator $\hat{\boldsymbol{\beta}}$, one obtains, using (24.0.7), that

$$\begin{aligned}\mathcal{MSE}[\hat{\boldsymbol{\beta}};\boldsymbol{\beta}] = \mathcal{E}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top] &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\,\mathcal{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top]\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1} = \\ &= \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}.\end{aligned} \tag{24.1.1}$$

This is a very simple formula. Its most interesting aspect is that this $\mathcal{MSE}$ matrix does not depend on the value of the true $\boldsymbol{\beta}$. In particular this means that it is *bounded* with respect to $\boldsymbol{\beta}$, which is important for someone who wants to be assured of a certain accuracy even in the worst possible situation.

Problem 298. *2 points Compute the $\mathcal{MSE}$-matrix $\mathcal{MSE}[\hat{\boldsymbol{\varepsilon}};\boldsymbol{\varepsilon}] = \mathcal{E}[(\hat{\boldsymbol{\varepsilon}} - \boldsymbol{\varepsilon})(\hat{\boldsymbol{\varepsilon}} - \boldsymbol{\varepsilon})^\top]$ of the residuals as predictors of the disturbances.*

Answer. Write $\hat{\boldsymbol{\varepsilon}} - \boldsymbol{\varepsilon} = \boldsymbol{M}\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon} = (\boldsymbol{M} - \boldsymbol{I})\boldsymbol{\varepsilon} = -\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon}$; therefore $\mathcal{MSE}[\hat{\boldsymbol{\varepsilon}};\boldsymbol{\varepsilon}] = \mathcal{E}[\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X} = \sigma^2\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$. Alternatively, start with $\hat{\boldsymbol{\varepsilon}} - \boldsymbol{\varepsilon} = \mathbf{y} - \hat{\mathbf{y}} - \boldsymbol{\varepsilon} = \boldsymbol{X}\boldsymbol{\beta} - \hat{\mathbf{y}} = \boldsymbol{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})$. This allows to use $\mathcal{MSE}[\hat{\boldsymbol{\varepsilon}};\boldsymbol{\varepsilon}] = \boldsymbol{X}\,\mathcal{MSE}[\hat{\boldsymbol{\beta}};\boldsymbol{\beta}]\boldsymbol{X}^\top = \sigma^2\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$. □

Problem 299. *2 points Let $\boldsymbol{\nu}$ be a random vector that is a linear transformation of $\mathbf{y}$, i.e., $\boldsymbol{\nu} = \boldsymbol{T}\mathbf{y}$ for some constant matrix $\boldsymbol{T}$. Furthermore $\boldsymbol{\nu}$ satisfies $\mathcal{E}[\boldsymbol{\nu}] = \boldsymbol{o}$. Show that from this follows $\boldsymbol{\nu} = \boldsymbol{T}\hat{\boldsymbol{\varepsilon}}$. (In other words, no other transformation of $\mathbf{y}$ with zero expected value is more "comprehensive" than $\boldsymbol{\varepsilon}$. However there are many other transformation of $\mathbf{y}$ with zero expected value which are as "comprehensive" as $\boldsymbol{\varepsilon}$).*

Answer. $\mathcal{E}[\boldsymbol{\nu}] = \boldsymbol{T}\boldsymbol{X}\boldsymbol{\beta}$ must be $\boldsymbol{o}$ whatever the value of $\boldsymbol{\beta}$. Therefore $\boldsymbol{T}\boldsymbol{X} = \boldsymbol{O}$, from which follows $\boldsymbol{T}\boldsymbol{M} = \boldsymbol{T}$. Since $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{M}\mathbf{y}$, this gives immediately $\boldsymbol{\nu} = \boldsymbol{T}\hat{\boldsymbol{\varepsilon}}$. (This is the statistical implication of the mathematical fact that $\boldsymbol{M}$ is a deficiency matrix of $\boldsymbol{X}$.) □

Problem 300. *2 points Show that $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\varepsilon}}$ are uncorrelated, i.e., $\operatorname{cov}[\hat{\beta}_i, \hat{\varepsilon}_j] = 0$ for all i, j. Defining the covariance matrix $\mathcal{C}[\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\varepsilon}}]$ as that matrix whose (i, j) element is $\operatorname{cov}[\hat{\beta}_i, \hat{\varepsilon}_j]$, this can also be written as $\mathcal{C}[\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\varepsilon}}] = \boldsymbol{O}$.* Hint: *The covariance matrix satisfies the rules $\mathcal{C}[\boldsymbol{A}\mathbf{y}, \boldsymbol{B}\mathbf{z}] = \boldsymbol{A}\,\mathcal{C}[\mathbf{y}, \mathbf{z}]\boldsymbol{B}^\top$ and $\mathcal{C}[\mathbf{y}, \mathbf{y}] = \mathcal{V}[\mathbf{y}]$. (Other rules for the covariance matrix, which will not be needed here, are $\mathcal{C}[\mathbf{z}, \mathbf{y}] = (\mathcal{C}[\mathbf{y}, \mathbf{z}])^\top$, $\mathcal{C}[\mathbf{x} + \mathbf{y}, \mathbf{z}] = \mathcal{C}[\mathbf{x}, \mathbf{z}] + \mathcal{C}[\mathbf{y}, \mathbf{z}]$, $\mathcal{C}[\mathbf{x}, \mathbf{y} + \mathbf{z}] = \mathcal{C}[\mathbf{x}, \mathbf{y}] + \mathcal{C}[\mathbf{x}, \mathbf{z}]$, and $\mathcal{C}[\mathbf{y}, \mathbf{c}] = \boldsymbol{O}$ if $\mathbf{c}$ is a vector of constants.)*

Answer. $\boldsymbol{A} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$ and $\boldsymbol{B} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$, therefore $\mathcal{C}[\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\varepsilon}}] = \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top) = \boldsymbol{O}$. □

Problem 301. *4 points Let $\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ be a regression model with intercept, in which the first column of $\boldsymbol{X}$ is the vector $\boldsymbol{\iota}$, and let $\hat{\boldsymbol{\beta}}$ the least squares estimator of*

$\boldsymbol{\beta}$. Show that the covariance matrix between $\bar{\mathsf{y}}$ and $\hat{\boldsymbol{\beta}}$, which is defined as the matrix (here consisting of one row only) that contains all the covariances

$$\mathcal{C}[\bar{\mathsf{y}}, \hat{\boldsymbol{\beta}}] \equiv \begin{bmatrix}\operatorname{cov}[\bar{\mathsf{y}}, \hat{\beta}_1] & \operatorname{cov}[\bar{\mathsf{y}}, \hat{\beta}_2] & \cdots & \operatorname{cov}[\bar{\mathsf{y}}, \hat{\beta}_k]\end{bmatrix} \tag{24.1.2}$$

has the following form: $\mathcal{C}[\bar{\mathsf{y}}, \hat{\boldsymbol{\beta}}] = \frac{\sigma^2}{n}\begin{bmatrix}1 & 0 & \cdots & 0\end{bmatrix}$ where n is the number of observations. Hint: *That the regression has an intercept term as first column of the $\boldsymbol{X}$-matrix means that $\boldsymbol{X}\boldsymbol{e}^{(1)} = \boldsymbol{\iota}$, where $\boldsymbol{e}^{(1)}$ is the unit vector having 1 in the first place and zeros elsewhere, and $\boldsymbol{\iota}$ is the vector which has ones everywhere.*

Answer. Write both $\bar{\mathsf{y}}$ and $\hat{\boldsymbol{\beta}}$ in terms of y, i.e., $\bar{\mathsf{y}} = \frac{1}{n}\boldsymbol{\iota}^\top\mathsf{y}$ and $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\mathsf{y}$. Therefore

$$\mathcal{C}[\bar{\mathsf{y}}, \hat{\boldsymbol{\beta}}] = \frac{1}{n}\boldsymbol{\iota}^\top\mathcal{V}[\mathsf{y}]\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1} = \frac{\sigma^2}{n}\boldsymbol{\iota}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1} = \frac{\sigma^2}{n}\boldsymbol{e}^{(1)\top}\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1} = \frac{\sigma^2}{n}\boldsymbol{e}^{(1)\top}. \tag{24.1.3}$$

□

Theorem 24.1.1. Gauss-Markov Theorem: *$\hat{\boldsymbol{\beta}}$ is the BLUE (Best Linear Unbiased Estimator) of $\boldsymbol{\beta}$ in the following vector sense: for every nonrandom coefficient vector $\boldsymbol{t}$, $\boldsymbol{t}^\top\hat{\boldsymbol{\beta}}$ is the scalar BLUE of $\boldsymbol{t}^\top\boldsymbol{\beta}$, i.e., every other linear unbiased estimator $\tilde{\phi} = \boldsymbol{a}^\top\mathsf{y}$ of $\phi = \boldsymbol{t}^\top\boldsymbol{\beta}$ has a bigger* MSE *than $\boldsymbol{t}^\top\hat{\boldsymbol{\beta}}$.*

Proof. Write the alternative linear estimator $\tilde{\phi} = \boldsymbol{a}^\top\mathsf{y}$ in the form

$$\tilde{\phi} = \big(\boldsymbol{t}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top + \boldsymbol{c}^\top\big)\mathsf{y} \tag{24.1.4}$$

then the sampling error is

$$\begin{aligned}\tilde{\phi} - \phi &= \big(\boldsymbol{t}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top + \boldsymbol{c}^\top\big)(\boldsymbol{X}\beta + \boldsymbol{\varepsilon}) - \boldsymbol{t}^\top\beta \\ &= \big(\boldsymbol{t}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top + \boldsymbol{c}^\top\big)\boldsymbol{\varepsilon} + \boldsymbol{c}^\top\boldsymbol{X}\beta.\end{aligned} \tag{24.1.5}$$

By assumption, the alternative estimator is unbiased, i.e., the expected value of this sampling error is zero regardless of the value of β. This is only possible if $\boldsymbol{c}^\top\boldsymbol{X} = \boldsymbol{o}^\top$. But then it follows

$$\begin{aligned}\operatorname{MSE}[\tilde{\phi}; \phi] &= \operatorname{E}[(\tilde{\phi} - \phi)^2] = \operatorname{E}[\big(\boldsymbol{t}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top + \boldsymbol{c}^\top\big)\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\big(\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{t} + \boldsymbol{c}\big)] = \\ &= \sigma^2\big(\boldsymbol{t}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top + \boldsymbol{c}^\top\big)\big(\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{t} + \boldsymbol{c}\big) = \sigma^2\boldsymbol{t}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{t} + \sigma^2\boldsymbol{c}^\top\boldsymbol{c},\end{aligned}$$

Here we needed again $\boldsymbol{c}^\top\boldsymbol{X} = \boldsymbol{o}^\top$. Clearly, this is minimized if $\boldsymbol{c} = \boldsymbol{o}$, in which case $\tilde{\phi} = \boldsymbol{t}^\top\hat{\boldsymbol{\beta}}$. □

Problem 302. *4 points Show: If $\tilde{\boldsymbol{\beta}}$ is a linear unbiased estimator of $\boldsymbol{\beta}$ and $\hat{\boldsymbol{\beta}}$ is the OLS estimator, then the difference of the $\mathcal{MSE}$-matrices $\mathcal{MSE}[\tilde{\boldsymbol{\beta}}; \boldsymbol{\beta}] - \mathcal{MSE}[\hat{\boldsymbol{\beta}}; \boldsymbol{\beta}]$ is nonnegative definite.*

Answer. (Compare [**DM93**, p. 159].) Any other linear estimator $\tilde{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ can be written as $\tilde{\boldsymbol{\beta}} = \big((\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top + \boldsymbol{C}\big)\mathsf{y}$. Its expected value is $\mathcal{E}[\tilde{\boldsymbol{\beta}}] = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{C}\boldsymbol{X}\boldsymbol{\beta}$. For $\tilde{\boldsymbol{\beta}}$ to be unbiased, regardless of the value of $\boldsymbol{\beta}$, $\boldsymbol{C}$ must satisfy $\boldsymbol{C}\boldsymbol{X} = \boldsymbol{O}$. But then it follows $\mathcal{MSE}[\tilde{\boldsymbol{\beta}}; \boldsymbol{\beta}] = \mathcal{V}[\tilde{\boldsymbol{\beta}}] = \sigma^2\big((\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top + \boldsymbol{C}\big)\big(\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1} + \boldsymbol{C}^\top\big) = \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1} + \sigma^2\boldsymbol{C}\boldsymbol{C}^\top$, i.e., it exceeds the $\mathcal{MSE}$-matrix of $\hat{\boldsymbol{\beta}}$ by a nonnegative definite matrix. □

24.2. Digression about Minimax Estimators

Theorem 24.1.1 is a somewhat puzzling property of the least squares estimator, since there is no reason in the world to restrict one's search for good estimators to unbiased estimators. An alternative and more enlightening characterization of $\hat{\boldsymbol{\beta}}$ does not use the concept of unbiasedness but that of a minimax estimator with respect to the MSE. For this I am proposing the following definition:

Definition 24.2.1. $\hat{\phi}$ is the linear minimax estimator of the scalar parameter ϕ with respect to the MSE if and only if for every other linear estimator $\tilde{\phi}$ there exists a value of the parameter vector $\boldsymbol{\beta}_0$ such that for all $\boldsymbol{\beta}_1$

$$\operatorname{MSE}[\tilde{\phi}; \phi|\boldsymbol{\beta} = \boldsymbol{\beta}_0] \geq \operatorname{MSE}[\hat{\phi}; \phi|\boldsymbol{\beta} = \boldsymbol{\beta}_1] \tag{24.2.1}$$

In other words, the worst that can happen if one uses any other $\tilde{\phi}$ is worse than the worst that can happen if one uses $\hat{\phi}$. Using this concept one can prove the following:

Theorem 24.2.2. *$\hat{\boldsymbol{\beta}}$ is a linear minimax estimator of the parameter vector $\boldsymbol{\beta}$ in the following sense: for every nonrandom coefficient vector $\boldsymbol{t}$, $\boldsymbol{t}^\top\hat{\boldsymbol{\beta}}$ is the linear minimax estimator of the scalar $\phi = \boldsymbol{t}^\top\boldsymbol{\beta}$ with respect to the* MSE. *I.e., for every other linear estimator $\tilde{\phi} = \boldsymbol{a}^\top\mathsf{y}$ of ϕ one can find a value $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ for which $\tilde{\phi}$ has a larger* MSE *than the largest possible* MSE *of $\boldsymbol{t}^\top\hat{\boldsymbol{\beta}}$.*

Proof: as in the proof of Theorem 24.1.1, write the alternative linear estimator $\tilde{\phi}$ in the form $\tilde{\phi} = \big(\boldsymbol{t}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top + \boldsymbol{c}^\top\big)\mathsf{y}$, so that the sampling error is given by (24.1.5). Then it follows

$$\operatorname{MSE}[\tilde{\phi}; \phi] = \operatorname{E}[(\tilde{\phi} - \phi)^2] = \operatorname{E}[\Big(\big(\boldsymbol{t}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top + \boldsymbol{c}^\top\big)\boldsymbol{\varepsilon} + \boldsymbol{c}^\top\boldsymbol{X}\beta\Big)\Big(\boldsymbol{\varepsilon}^\top\big(\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{t} + \boldsymbol{c}\big) + \beta^\top \tag{24.2.2}$$

$$= \sigma^2\big(\boldsymbol{t}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top + \boldsymbol{c}^\top\big)\big(\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{t} + \boldsymbol{c}\big) + \boldsymbol{c}^\top\boldsymbol{X}\boldsymbol{\beta}\boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{c} \tag{24.2.3}$$

Now there are two cases: if $\boldsymbol{c}^\top\boldsymbol{X} = \boldsymbol{o}^\top$, then $\operatorname{MSE}[\tilde{\phi}; \phi] = \sigma^2\boldsymbol{t}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{t} + \sigma^2\boldsymbol{c}^\top\boldsymbol{c}$. This does not depend on $\boldsymbol{\beta}$ and if $\boldsymbol{c} \neq \boldsymbol{o}$ then this MSE is larger than that for $\boldsymbol{c} = \boldsymbol{o}$. If $\boldsymbol{c}^\top\boldsymbol{X} \neq \boldsymbol{o}^\top$, then $\operatorname{MSE}[\tilde{\phi}; \phi]$ is unbounded, i.e., for any finite number ω one one can always find a $\boldsymbol{\beta}_0$ for which $\operatorname{MSE}[\tilde{\phi}; \phi] > \omega$. Since $\operatorname{MSE}[\hat{\phi}; \phi]$ is bounded, a $\boldsymbol{\beta}_0$ can be found that satisfies (24.2.1).

If we characterize the BLUE as a minimax estimator, we are using a consistent and unified principle. It is based on the concept of the MSE alone, not on a mixture between the concepts of unbiasedness and the MSE. This explains why the mathematical theory of the least squares estimator is so rich.

On the other hand, a minimax strategy is not a good estimation strategy. Nature is not the adversary of the researcher; it does not maliciously choose $\boldsymbol{\beta}$ in such a way that the researcher will be misled. This explains why the least squares principle, despite the beauty of its mathematical theory, does not give terribly good estimators (in fact, they are inadmissible, see the Section about the Stein rule below).

$\hat{\boldsymbol{\beta}}$ is therefore simultaneously the solution to two very different minimization problems. We will refer to it as the OLS estimate if we refer to its property of minimizing the sum of squared errors, and as the BLUE estimator if we think of it as the best linear unbiased estimator.

Note that even if σ^2 were known, one could not get a better linear unbiased estimator of $\boldsymbol{\beta}$.

24.3. Miscellaneous Properties of the BLUE

Problem 303.

• a. *1 point Instead of* (18.2.22) *one sometimes sees the formula*

$$\hat{\beta} = \frac{\sum(x_t - \bar{x})\mathsf{y}_t}{\sum(x_t - \bar{x})^2}. \tag{24.3.1}$$

for the slope parameter in the simple regression. Show that these formulas are mathematically equivalent.

Answer. Equivalence of (24.3.1) and (18.2.22) follows from $\sum(x_t - \bar{x}) = 0$ and therefore also $\bar{\mathsf{y}}\sum(x_t - \bar{x}) = 0$. Alternative proof, using matrix notation and the matrix $\boldsymbol{D}$ defined in Problem 189: (18.2.22) is $\frac{\boldsymbol{x}^\top \boldsymbol{D}^\top \boldsymbol{D}\mathsf{y}}{\boldsymbol{x}^\top \boldsymbol{D}^\top \boldsymbol{D}\boldsymbol{x}}$ and (24.3.1) is $\frac{\boldsymbol{x}^\top \boldsymbol{D}\mathsf{y}}{\boldsymbol{x}^\top \boldsymbol{D}^\top \boldsymbol{D}\boldsymbol{x}}$. They are equal because $\boldsymbol{D}$ is symmetric and idempotent.

□

• b. *1 point Show that*

$$\operatorname{var}[\hat{\beta}] = \frac{\sigma^2}{\sum(x_i - \bar{x})^2} \tag{24.3.2}$$

Answer. Write (24.3.1) as

$$\hat{\beta} = \frac{1}{\sum(x_t - \bar{x})^2}\sum(x_t - \bar{x})\mathsf{y}_t \qquad \Rightarrow \qquad \operatorname{var}[\hat{\beta}] = \frac{1}{\left(\sum(x_t - \bar{x})^2\right)^2}\sum(x_t - \bar{x})^2\sigma^2 \tag{24.3.3}$$

□

• c. *2 points Show that* $\operatorname{cov}[\hat{\beta}, \bar{\mathsf{y}}] = 0$.

Answer. This is a special case of problem 301, but it can be easily shown here separately:

$$\operatorname{cov}[\hat{\beta}, \bar{\mathsf{y}}] = \operatorname{cov}\left[\frac{\sum_s (x_s - \bar{x})\mathsf{y}_s}{\sum_t (x_t - \bar{x})^2}, \frac{1}{n}\sum_j \mathsf{y}_j\right] = \frac{1}{n\sum_t (x_t - \bar{x})^2}\operatorname{cov}\left[\sum_s (x_s - \bar{x})\mathsf{y}_s, \sum_j \mathsf{y}_j\right] =$$
$$= \frac{1}{n\sum_t (x_t - \bar{x})^2}\sum_s (x_s - \bar{x})\sigma^2 = 0.$$

□

• d. *2 points Using* (18.2.23) *show that*

$$\operatorname{var}[\hat{\alpha}] = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}\right) \tag{24.3.4}$$

Problem 304. *You have two data vectors x_i and y_i $(i = 1, \ldots, n)$, and the true model is*

$$\mathsf{y}_i = \beta x_i + \varepsilon_i \tag{24.3.5}$$

where x_i and ε_i satisfy the basic assumptions of the linear regression model. The least squares estimator for this model is

$$\tilde{\beta} = (\boldsymbol{x}^\top \boldsymbol{x})^{-1}\boldsymbol{x}^\top \mathsf{y} = \frac{\sum x_i \mathsf{y}_i}{\sum x_i^2} \tag{24.3.6}$$

• a. *1 point Is $\tilde{\beta}$ an unbiased estimator of β? (Proof is required.)*

Answer. First derive a nice expression for $\tilde{\beta} - \beta$:

$$\begin{aligned}
\tilde{\beta} - \beta &= \frac{\sum x_i \mathsf{y}_i}{\sum x_i^2} - \frac{\sum x_i^2 \beta}{\sum x_i^2} \\
&= \frac{\sum x_i(\mathsf{y}_i - x_i\beta)}{\sum x_i^2} \\
&= \frac{\sum x_i \varepsilon_i}{\sum x_i^2} \qquad \text{since} \quad \mathsf{y}_i = \beta x_i + \varepsilon_i \\
\operatorname{E}[\tilde{\beta} - \beta] &= \operatorname{E}\left[\frac{\sum x_i \varepsilon_i}{\sum x_i^2}\right] \\
&= \frac{\sum \operatorname{E}[x_i \varepsilon_i]}{\sum x_i^2} \\
&= \frac{\sum x_i \operatorname{E}[\varepsilon_i]}{\sum x_i^2} = 0 \qquad \text{since} \quad \operatorname{E}\varepsilon_i = 0.
\end{aligned}$$

□

• b. *2 points Derive the variance of $\tilde{\beta}$. (Show your work.)*

Answer.

$$
\begin{aligned}
\operatorname{var}\tilde\beta &= \mathrm{E}[\tilde\beta-\beta]^2 \\
&= \mathrm{E}\left(\frac{\sum x_i\varepsilon_i}{\sum x_i^2}\right)^2 \\
&= \frac{1}{(\sum x_i^2)^2}\,\mathrm{E}[\sum x_i\varepsilon_i]^2 \\
&= \frac{1}{(\sum x_i^2)^2}\left(\mathrm{E}\sum(x_i\varepsilon_i)^2 + 2\,\mathrm{E}\sum_{i<j}(x_i\varepsilon_i)(x_j\varepsilon_j)\right) \\
&= \frac{1}{(\sum x_i^2)^2}\sum \mathrm{E}[x_i\varepsilon_i]^2 \qquad \text{since the } \varepsilon_i\text{'s are uncorrelated, i.e., } \operatorname{cov}[\varepsilon_i,\varepsilon_j]=0 \text{ for } i\neq j \\
&= \frac{1}{(\sum x_i^2)^2}\sigma^2\sum x_i^2 \qquad \text{since all } \varepsilon_i \text{ have equal variance } \sigma^2 \\
&= \frac{\sigma^2}{\sum x_i^2}.
\end{aligned}
$$

□

Problem 305. *We still assume* (24.3.5) *is the true model. Consider an alternative estimator:*

$$
\hat\beta = \frac{\sum(x_i-\bar x)(y_i-\bar y)}{\sum(x_i-\bar x)^2} \tag{24.3.7}
$$

i.e., the estimator which would be the best linear unbiased estimator if the true model were (18.2.15).

• a. *2 points Is* $\hat\beta$ *still an unbiased estimator of* β *if* (24.3.5) *is the true model? (A short but rigorous argument may save you a lot of algebra here).*

Answer. One can argue it: $\hat\beta$ is unbiased for model (18.2.15) whatever the value of α or β, therefore also when $\alpha=0$, i.e., when the model is (24.3.5). But here is the pedestrian way:

$$
\begin{aligned}
\hat\beta &= \frac{\sum(x_i-\bar x)(y_i-\bar y)}{\sum(x_i-\bar x)^2} = \frac{\sum(x_i-\bar x)y_i}{\sum(x_i-\bar x)^2} \qquad \text{since} \quad \sum(x_i-\bar x)\bar y = 0 \\
&= \frac{\sum(x_i-\bar x)(\beta x_i+\varepsilon_i)}{\sum(x_i-\bar x)^2} \qquad \text{since} \quad y_i=\beta x_i+\varepsilon_i \\
&= \beta\frac{\sum(x_i-\bar x)x_i}{\sum(x_i-\bar x)^2} + \frac{\sum(x_i-\bar x)\varepsilon_i}{\sum(x_i-\bar x)^2} \\
&= \beta + \frac{\sum(x_i-\bar x)\varepsilon_i}{\sum(x_i-\bar x)^2} \qquad \text{since} \quad \sum(x_i-\bar x)x_i = \sum(x_i-\bar x)^2 \\
\mathrm{E}\,\hat\beta &= \mathrm{E}\,\beta + \mathrm{E}\,\frac{\sum(x_i-\bar x)\varepsilon_i}{\sum(x_i-\bar x)^2} \\
&= \beta + \frac{\sum(x_i-\bar x)\,\mathrm{E}\,\varepsilon_i}{\sum(x_i-\bar x)^2} = \beta \qquad \text{since } \mathrm{E}\,\varepsilon_i=0 \text{ for all } i\text{, i.e., } \hat\beta \text{ is unbiased.}
\end{aligned}
$$

□

• b. *2 points Derive the variance of* $\hat\beta$ *if* (24.3.5) *is the true model.*

Answer. One can again argue it: since the formula for $\operatorname{var}\hat\beta$ does not depend on what the true value of α is, it is the same formula.

$$
\operatorname{var}\hat\beta = \operatorname{var}\left(\beta + \frac{\sum(x_i-\bar x)\varepsilon_i}{\sum(x_i-\bar x)^2}\right) \tag{24.3.8}
$$

$$
= \operatorname{var}\left(\frac{\sum(x_i-\bar x)\varepsilon_i}{\sum(x_i-\bar x)^2}\right) \tag{24.3.9}
$$

$$
= \frac{\sum(x_i-\bar x)^2\operatorname{var}\varepsilon_i}{(\sum(x_i-\bar x)^2)^2} \qquad \text{since} \quad \operatorname{cov}[\varepsilon_i\varepsilon_j]=0 \tag{24.3.10}
$$

$$
= \frac{\sigma^2}{\sum(x_i-\bar x)^2}. \tag{24.3.11}
$$

□

• c. *1 point Still assuming* (24.3.5) *is the true model, would you prefer* $\hat\beta$ *or the* $\tilde\beta$ *from Problem 304 as an estimator of* β*?*

Answer. Since $\tilde\beta$ and $\hat\beta$ are both unbiased estimators, if (24.3.5) is the true model, the preferred estimator is the one with the smaller variance. As I will show, $\operatorname{var}\tilde\beta \le \operatorname{var}\hat\beta$ and, therefore, $\tilde\beta$ is preferred to $\hat\beta$. To show

$$
\operatorname{var}\hat\beta = \frac{\sigma^2}{\sum(x_i-\bar x)^2} \ge \frac{\sigma^2}{\sum x_i^2} = \operatorname{var}\tilde\beta \tag{24.3.12}
$$

one must show

$$
\sum(x_i-\bar x)^2 \le \sum x_i^2 \tag{24.3.13}
$$

which is a simple consequence of (12.1.1). Thus $\operatorname{var}\hat{\beta} \geq \operatorname{var}\tilde{\beta}$; the variances are equal only if $\bar{x} = 0$, i.e., if $\tilde{\beta} = \hat{\beta}$. □

PROBLEM 306. *Suppose the true model is* (18.2.15) *and the basic assumptions are satisfied.*

• a. *2 points In this situation,* $\tilde{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}$ *is generally a biased estimator of* β. *Show that its bias is*

$$\mathrm{E}[\tilde{\beta} - \beta] = \alpha \frac{n\bar{x}}{\sum x_i^2} \tag{24.3.14}$$

ANSWER. In situations like this it is always worth while to get a nice simple expression for the sampling error:

$$\tilde{\beta} - \beta = \frac{\sum x_i y_i}{\sum x_i^2} - \beta \tag{24.3.15}$$

$$= \frac{\sum x_i(\alpha + \beta x_i + \varepsilon_i)}{\sum x_i^2} - \beta \qquad \text{since } y_i = \alpha + \beta x_i + \varepsilon_i \tag{24.3.16}$$

$$= \alpha \frac{\sum x_i}{\sum x_i^2} + \beta \frac{\sum x_i^2}{\sum x_i^2} + \frac{\sum x_i \varepsilon_i}{\sum x_i^2} - \beta \tag{24.3.17}$$

$$= \alpha \frac{\sum x_i}{\sum x_i^2} + \frac{\sum x_i \varepsilon_i}{\sum x_i^2} \tag{24.3.18}$$

$$\mathrm{E}[\tilde{\beta} - \beta] = \mathrm{E}\,\alpha \frac{\sum x_i}{\sum x_i^2} + \mathrm{E}\,\frac{\sum x_i \varepsilon_i}{\sum x_i^2} \tag{24.3.19}$$

$$= \alpha \frac{\sum x_i}{\sum x_i^2} + \frac{\sum x_i\, \mathrm{E}\,\varepsilon_i}{\sum x_i^2} \tag{24.3.20}$$

$$= \alpha \frac{\sum x_i}{\sum x_i^2} + 0 = \alpha \frac{n\bar{x}}{\sum x_i^2} \tag{24.3.21}$$

This is $\neq 0$ unless $\bar{x} = 0$ or $\alpha = 0$. □

• b. *2 points Compute* $\operatorname{var}[\tilde{\beta}]$. *Is it greater or smaller than*

$$\frac{\sigma^2}{\sum (x_i - \bar{x})^2} \tag{24.3.22}$$

which is the variance of the OLS estimator in this model?

ANSWER.

$$\operatorname{var}\tilde{\beta} = \operatorname{var}\left[\frac{\sum x_i y_i}{\sum x_i^2}\right] \tag{24.3.23}$$

$$= \frac{1}{\left(\sum x_i^2\right)^2} \operatorname{var}\Big[\sum x_i y_i\Big] \tag{24.3.24}$$

$$= \frac{1}{\left(\sum x_i^2\right)^2} \sum x_i^2 \operatorname{var}[y_i] \tag{24.3.25}$$

$$= \frac{\sigma^2}{\left(\sum x_i^2\right)^2} \sum x_i^2 \quad \text{since all } y_i \text{ are uncorrelated and have equal variance } \sigma^2 \tag{24.3.26}$$

$$= \frac{\sigma^2}{\sum x_i^2}. \tag{24.3.27}$$

This variance is smaller or equal because $\sum x_i^2 \geq \sum (x_i - \bar{x})^2$. □

• c. *5 points Show that the* MSE *of* $\tilde{\beta}$ *is smaller than that of the OLS estimator if and only if the unknown true parameters* α *and* σ^2 *satisfy the equation*

$$\frac{\alpha^2}{\sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}\right)} < 1 \tag{24.3.28}$$

ANSWER. This implies some tedious algebra. Here it is important to set it up right.

$$\mathrm{MSE}[\tilde{\beta}; \beta] = \frac{\sigma^2}{\sum x_i^2} + \left(\frac{\alpha n \bar{x}}{\sum x_i^2}\right)^2 \leq \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$$

$$\left(\frac{\alpha n \bar{x}}{\sum x_i^2}\right)^2 \leq \frac{\sigma^2}{\sum (x_i - \bar{x})^2} - \frac{\sigma^2}{\sum x_i^2} = \frac{\sigma^2\left(\sum x_i^2 - \sum (x_i - \bar{x})^2\right)}{\sum (x_i - \bar{x})^2 \sum x_i^2}$$

$$= \frac{\sigma^2 n \bar{x}^2}{\sum (x_i - \bar{x})^2 \sum x_i^2}$$

$$\frac{\alpha^2 n}{\sum x_i^2} = \frac{\alpha^2}{\frac{1}{n}\sum (x_i - \bar{x})^2 + \bar{x}^2} \leq \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$$

$$\frac{\alpha^2}{\sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}\right)} \leq 1$$

Now look at this lefthand side; it is amazing and surprising that it is exactly the population equivalent of the F-test for testing $\alpha = 0$ in the regression *with* intercept. It can be estimated by replacing α^2 with $\hat{\boldsymbol{\alpha}}^2$ and σ^2 with s^2 (in the regression with intercept). Let's look at this statistic. If $\alpha = 0$ it has a F-distribution with 1 and $n - 2$ degrees of freedom. If $\alpha \neq 0$ it has what is called a noncentral distribution, and the only thing we needed to know so far was that it was likely to assume larger values than with $\alpha = 0$. This is why a small value of that statistic supported the hypothesis that $\alpha = 0$. But in the present case we are not testing whether $\alpha = 0$ but whether the constrained MSE is better than the unconstrained. This is the case of the above inequality holds, the limiting case being that it is an equality. If it is an equality, then the above statistic has a F distribution with noncentrality parameter 1/2. (Here all we need to know that: if $z \sim \mathsf{N}(\mu, 1)$ then

$z^2 \sim \chi^2_1$ with noncentrality parameter $\mu^2/2$. A noncentral F has a noncentral χ^2 in numerator and a central one in denominator.) The testing principle is therefore: compare the observed value with the upper α point of a F distribution with noncentrality parameter 1/2. This gives higher critical values than testing for $\alpha = 0$; i.e., one may reject that $\alpha = 0$ but not reject that the MSE of the contrained estimator is larger. This is as it should be. Compare [**Gre97**, 8.5.1 pp. 405–408] on this. □

From the Gauss-Markov theorem follows that for every nonrandom matrix $\boldsymbol{R}$, the BLUE of $\boldsymbol{\phi} = \boldsymbol{R}\boldsymbol{\beta}$ is $\hat{\boldsymbol{\phi}} = \boldsymbol{R}\hat{\boldsymbol{\beta}}$. Furthermore, the best linear unbiased predictor (BLUP) of $\boldsymbol{\varepsilon} = \mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}$ is the vector of residuals $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$.

PROBLEM 307. *Let $\tilde{\boldsymbol{\varepsilon}} = \boldsymbol{A}\mathbf{y}$ be a linear predictor of the disturbance vector $\boldsymbol{\varepsilon}$ in the model $\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim (\boldsymbol{o}, \sigma^2\boldsymbol{I})$.*

• a. *2 points Show that $\tilde{\boldsymbol{\varepsilon}}$ is unbiased, i.e., $\mathrm{E}[\tilde{\boldsymbol{\varepsilon}} - \boldsymbol{\varepsilon}] = \boldsymbol{o}$, regardless of the value of $\boldsymbol{\beta}$, if and only if $\boldsymbol{A}$ satisfies $\boldsymbol{A}\boldsymbol{X} = \boldsymbol{O}$.*

ANSWER. $\mathcal{E}[\boldsymbol{A}\mathbf{y} - \boldsymbol{\varepsilon}] = \mathcal{E}[\boldsymbol{A}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{A}\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}] = \boldsymbol{A}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{o} - \boldsymbol{o}$. This is $= \boldsymbol{o}$ for all $\boldsymbol{\beta}$ if and only if $\boldsymbol{A}\boldsymbol{X} = \boldsymbol{O}$ □

• b. *2 points Which unbiased linear predictor $\tilde{\boldsymbol{\varepsilon}} = \boldsymbol{A}\mathbf{y}$ of $\boldsymbol{\varepsilon}$ minimizes the $\mathcal{MSE}$-matrix $\mathcal{E}[(\tilde{\boldsymbol{\varepsilon}} - \boldsymbol{\varepsilon})(\tilde{\boldsymbol{\varepsilon}} - \boldsymbol{\varepsilon})^\top]$? Hint: Write $\boldsymbol{A} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top + \boldsymbol{C}$. What is the minimum value of this $\mathcal{MSE}$-matrix?*

ANSWER. Since $\boldsymbol{A}\boldsymbol{X} = \boldsymbol{O}$, the prediction error $\boldsymbol{A}\mathbf{y} - \boldsymbol{\varepsilon} = \boldsymbol{A}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{A}\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon} = (\boldsymbol{A} - \boldsymbol{I})\boldsymbol{\varepsilon}$; therefore one minimizes $\sigma^2(\boldsymbol{A} - \boldsymbol{I})(\boldsymbol{A} - \boldsymbol{I})^\top$ s. t. $\boldsymbol{A}\boldsymbol{X} = \boldsymbol{O}$. Using the hint, $\boldsymbol{C}$ must also satisfy $\boldsymbol{C}\boldsymbol{X} = \boldsymbol{O}$, and $(\boldsymbol{A} - \boldsymbol{I})(\boldsymbol{A} - \boldsymbol{I})^\top = (\boldsymbol{C} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top)(\boldsymbol{C}^\top - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top) = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top + \boldsymbol{C}\boldsymbol{C}^\top$, therefore one must set $\boldsymbol{C} = \boldsymbol{O}$. Minimum value is $\sigma^2\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$.

□

• c. *How does this best predictor relate to the OLS estimator $\hat{\boldsymbol{\beta}}$?*

ANSWER. It is equal to the residual vector $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$. □

PROBLEM 308. *This is a vector generalization of problem 198. Let $\hat{\boldsymbol{\beta}}$ the BLUE of $\boldsymbol{\beta}$ and $\tilde{\boldsymbol{\beta}}$ an arbitrary linear unbiased estimator of $\boldsymbol{\beta}$.*

• a. *2 points Show that $\mathcal{C}[\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}] = \boldsymbol{O}$.*

ANSWER. Say $\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{B}}\mathbf{y}$; unbiasedness means $\tilde{\boldsymbol{B}}\boldsymbol{X} = \boldsymbol{I}$. Therefore

$$\begin{aligned}
\mathcal{C}[\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}] &= \mathcal{C}[\big((\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top - \tilde{\boldsymbol{B}}\big)\mathbf{y}, (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\mathbf{y}] \\
&= \big((\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top - \tilde{\boldsymbol{B}}\big)\,\mathcal{V}[\mathbf{y}]\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1} \\
&= \sigma^2\big((\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top - \tilde{\boldsymbol{B}}\big)\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1} \\
&= \sigma^2\big((\boldsymbol{X}^\top\boldsymbol{X})^{-1} - (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\big) = \boldsymbol{O}.
\end{aligned}$$

□

• b. *2 points Show that $\mathcal{MSE}[\tilde{\boldsymbol{\beta}}; \boldsymbol{\beta}] = \mathcal{MSE}[\hat{\boldsymbol{\beta}}; \boldsymbol{\beta}] + \mathcal{V}[\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}]$*

ANSWER. Due to unbiasedness, $\mathcal{MSE} = \mathcal{V}$, and the decomposition $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} + (\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})$ is an uncorrelated sum. Here is more detail: $\mathcal{MSE}[\tilde{\boldsymbol{\beta}}; \boldsymbol{\beta}] = \mathcal{V}[\tilde{\boldsymbol{\beta}}] = \mathcal{V}[\hat{\boldsymbol{\beta}} + \tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}] = \mathcal{V}[\hat{\boldsymbol{\beta}}] + \mathcal{C}[\hat{\boldsymbol{\beta}}, \tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}] + \mathcal{C}[\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}] + \mathcal{V}[\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}]$ but the two $\mathcal{C}$-terms are the null matrices. □

PROBLEM 309. *3 points Given a simple regression $\mathrm{y}_t = \alpha + \beta x_t + \varepsilon_t$, where the ε_t are independent and identically distributed with mean μ and variance σ^2. Is it possible to consistently estimate all four parameters α, β, σ^2, and μ? If yes, explain how you would estimate them, and if no, what is the best you can do?*

ANSWER. Call $\tilde{\varepsilon}_t = \varepsilon_t - \mu$, then the equation reads $\mathrm{y}_t = \alpha + \mu + \beta x_t + \tilde{\varepsilon}_t$, with well behaved disturbances. Therefore one can estimate $\alpha + \mu$, β, and σ^2. This is also the best one can do; if $\alpha + \mu$ are equal, the y_t have the same joint distribution. □

PROBLEM 310. *3 points The model is $\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ but all rows of the $\boldsymbol{X}$-matrix are exactly equal. What can you do? Can you estimate $\boldsymbol{\beta}$? If not, are there any linear combinations of the components of $\boldsymbol{\beta}$ which you can estimate? Can you estimate σ^2?*

ANSWER. If all rows are equal, then each column is a multiple of $\boldsymbol{\iota}$. Therefore, if there are more than one column, none of the individual components of $\boldsymbol{\beta}$ can be estimated. But you can estimate $\boldsymbol{x}^\top\boldsymbol{\beta}$ (if $\boldsymbol{x}$ is one of the row vectors of $\boldsymbol{X}$) and you can estimate σ^2. □

PROBLEM 311. *This is* [**JHG+88**, 5.3.32]*: Consider the log-linear statistical model*

$$\mathrm{y}_t = \alpha x_t^\beta \exp \varepsilon_t = z_t \exp \varepsilon_t \tag{24.3.29}$$

with "well-behaved" disturbances ε_t. Here $z_t = \alpha x_t^\beta$ is the systematic portion of y_t, which depends on x_t. (This functional form is often used in models of demand and production.)

• a. *1 point Can this be estimated with the regression formalism?*

ANSWER. Yes, simply take logs:

$$\log \mathrm{y}_t = \log \alpha + \beta \log x_t + \varepsilon_t \tag{24.3.30}$$

□

• b. *1 point Show that the elasticity of the functional relationship between x_t and z_t*

$$\eta = \frac{\partial z_t / z_t}{\partial x_t / x_t} \tag{24.3.31}$$

does not depend on t, i.e., it is the same for all observations. Many authors talk about the elasticity of y_t with respect to x_t, but one should really only talk about the elasticity of z_t with respect to x_t, where z_t is the systematic part of y_t which can be estimated by $\hat{y}_t$.

Answer. The systematic functional relationship is $\log z_t = \log\alpha + \beta\log x_t$; therefore

$$\frac{\partial \log z_t}{\partial z_t} = \frac{1}{z_t} \tag{24.3.32}$$

which can be rewritten as

$$\frac{\partial z_t}{z_t} = \partial \log z_t; \tag{24.3.33}$$

The same can be done with x_t; therefore

$$\frac{\partial z_t/z_t}{\partial x_t/x_t} = \frac{\partial \log z_t}{\partial \log x_t} = \beta \tag{24.3.34}$$

What we just did was a tricky way to take a derivative. A less tricky way is:

$$\frac{\partial z_t}{\partial x_t} = \alpha\beta x_t^{\beta-1} = \beta z_t/x_t \tag{24.3.35}$$

Therefore

$$\frac{\partial z_t}{\partial x_t}\frac{x_t}{z_t} = \beta \tag{24.3.36}$$

□

Problem 312.

• a. *2 points What is the elasticity in the simple regression* $\mathsf{y}_t = \alpha + \beta x_t + \varepsilon_t$*?*

Answer.

$$\eta_t = \frac{\partial \mathsf{z}_t/\mathsf{z}_t}{\partial x_t/x_t} = \frac{\partial \mathsf{z}_t}{\partial x_t}\frac{x_t}{\mathsf{z}_t} = \frac{\beta x_t}{\mathsf{z}_t} = \frac{\beta x_t}{\alpha + \beta x_t} \tag{24.3.37}$$

This depends on the observation, and if one wants one number, a good way is to evaluate it at $\bar{x}$. □

• b. *Show that an estimate of this elasticity evaluated at* $\bar{x}$ *is* $h = \frac{\hat{\beta}\bar{x}}{\bar{\mathsf{y}}}$.

Answer. This comes from the fact that the fitted regression line goes through the point $\bar{x}, \bar{y}$. If one uses the other definition of elasticity, which Greene uses on p. 227 but no longer on p. 280, and which I think does not make much sense, one gets the same formula:

$$\eta_t = \frac{\partial \mathsf{y}_t/\mathsf{y}_t}{\partial x_t/x_t} = \frac{\partial \mathsf{y}_t}{\partial x_t}\frac{x_t}{\mathsf{y}_t} = \frac{\beta x_t}{\mathsf{y}_t} \tag{24.3.38}$$

This is different than (24.3.37), but if one evaluates it at the sample mean, both formulas give the same result $\frac{\hat{\beta}\bar{x}}{\bar{\mathsf{y}}}$. □

• c. *Show by the delta method that the estimator*

$$h = \frac{\hat{\beta}\bar{x}}{\bar{\mathsf{y}}} \tag{24.3.39}$$

of the elasticity in the simple regression model has the estimated asymptotic variance

$$\mathsf{s}^2 \begin{bmatrix} \frac{-h}{\bar{\mathsf{y}}} & \frac{\bar{x}(1-h)}{\bar{\mathsf{y}}} \end{bmatrix} \begin{bmatrix} 1 & \bar{x} \\ \bar{x} & \overline{x^2} \end{bmatrix}^{-1} \begin{bmatrix} \frac{-h}{\bar{\mathsf{y}}} \\ \frac{\bar{x}(1-h)}{\bar{\mathsf{y}}} \end{bmatrix} \tag{24.3.40}$$

• d. *Compare* [**Gre97**, example 6.20 on p. 280]. *Assume*

$$\frac{1}{n}(\boldsymbol{X}^\top\boldsymbol{X}) = \begin{bmatrix} 1 & \bar{x} \\ \bar{x} & \overline{x^2} \end{bmatrix} \to \boldsymbol{Q} = \begin{bmatrix} 1 & q \\ q & r \end{bmatrix} \tag{24.3.41}$$

where we assume for the sake of the argument that q *is known. The true elasticity of the underlying functional relationship, evaluated at* $\lim \bar{x}$*, is*

$$\eta = \frac{q\beta}{\alpha + q\beta} \tag{24.3.42}$$

Then

$$\mathsf{h} = \frac{q\hat{\beta}}{\hat{\alpha} + q\hat{\beta}} \tag{24.3.43}$$

is a consistent estimate for η*.*

A generalization of the log-linear model is the translog model, which is a second-order approximation to an unknown functional form, and which allows to model second-order effects such as elasticities of substitution etc. Used to model production, cost, and utility functions. Start with any function $v = f(u_1, \ldots, u_n)$ and make a second-order Taylor development around $\boldsymbol{u} = \boldsymbol{o}$:

$$v = f(\boldsymbol{o}) + \sum u_i \frac{\partial f}{\partial u_i}\Big|_{\boldsymbol{u}=\boldsymbol{o}} + \frac{1}{2}\sum_{i,j} u_i u_j \frac{\partial^2 f}{\partial u_i \partial u_j}\Big|_{\boldsymbol{u}=\boldsymbol{o}} \tag{24.3.44}$$

Now say $v = \log(\mathsf{y})$ and $u_i = \log(x_i)$, and the values of f and its derivatives at $\boldsymbol{o}$ are the coefficients to be estimated:

$$\log(\mathsf{y}) = \alpha + \sum \beta_i \log x_i + \frac{1}{2}\sum_{i,j} \gamma_{ij} \log x_i \log x_j + \varepsilon \tag{24.3.45}$$

Note that by Young's theorem it must be true that $\gamma_{kl} = \gamma_{lk}$.

The semi-log model is often used to model growth rates:

$$\log \mathsf{y}_t = \boldsymbol{x}_t^\top \boldsymbol{\beta} + \varepsilon_t \tag{24.3.46}$$

Here usually one of the columns of $\boldsymbol{X}$ is the time subscript t itself; [**Gre97**, p. 227] writes it as

$$\log \mathsf{y}_t = \boldsymbol{x}_t^\top \boldsymbol{\beta} + t\delta + \varepsilon_t \tag{24.3.47}$$

where δ is the autonomous growth rate. The logistic functional form is appropriate for adoption rates $0 \le \mathsf{y}_t \le 1$: the rate of adoption is slow at first, then rapid as the innovation gains popularity, then slow again as the market becomes saturated:

$$\mathsf{y}_t = \frac{\exp(\boldsymbol{x}_t^\top \boldsymbol{\beta} + t\delta + \varepsilon_t)}{1 + \exp(\boldsymbol{x}_t^\top \boldsymbol{\beta} + t\delta + \varepsilon_t)} \tag{24.3.48}$$

This can be linearized by the logit transformation:

$$\text{logit}(\mathsf{y}_t) = \log \frac{\mathsf{y}_t}{1-\mathsf{y}_t} = \boldsymbol{x}_t^\top \boldsymbol{\beta} + t\delta + \varepsilon_t \tag{24.3.49}$$

PROBLEM 313. *3 points Given a simple regression $\mathsf{y}_t = \alpha_t + \beta x_t$ which deviates from an ordinary regression in two ways: (1) There is no disturbance term. (2) The "constant term" α_t is random, i.e., in each time period t, the value of α_t is obtained by an independent drawing from a population with unknown mean μ and unknown variance σ^2. Is it possible to estimate all three parameters β, σ^2, and μ, and to "predict" each α_t? (Here I am using the term "prediction" for the estimation of a random parameter.) If yes, explain how you would estimate it, and if not, what is the best you can do?*

ANSWER. Call $\varepsilon_t = \alpha_t - \mu$, then the equation reads $\mathsf{y}_t = \mu + \beta x_t + \varepsilon_t$, with well behaved disturbances. Therefore one can estimate all the unknown parameters, and predict α_t by $\hat{\mu} + \varepsilon_t$. □

24.4. Estimation of the Variance

The formulas in this section use g-inverses (compare (A.3.1)) and are valid even if not all columns of $\boldsymbol{X}$ are linearly independent. q is the rank if $\boldsymbol{X}$. The proofs are not any more complicated than in the case that $\boldsymbol{X}$ has full rank, if one keeps in mind identity (A.3.3) and some other simple properties of g-inverses which are tacitly used at various places. Those readers who are only interested in the full-rank case should simply substitute $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$ for $(\boldsymbol{X}^\top\boldsymbol{X})^-$ and k for q (k is the number of columns of $\boldsymbol{X}$).

SSE, the attained minimum value of the Least Squares objective function, is a random variable too and we will now compute its expected value. It turns out that

$$\mathrm{E}[\mathsf{SSE}] = \sigma^2(n-q) \tag{24.4.1}$$

PROOF. $\mathsf{SSE} = \hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}}$, where $\hat{\boldsymbol{\varepsilon}} = \mathsf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = \mathsf{y} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^-\boldsymbol{X}^\top\mathsf{y} = \boldsymbol{M}\mathsf{y}$, with $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^-\boldsymbol{X}^\top$. From $\boldsymbol{M}\boldsymbol{X} = \boldsymbol{O}$ follows $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{M}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{M}\boldsymbol{\varepsilon}$. Since $\boldsymbol{M}$ is idempotent and symmetric, it follows $\hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}} = \boldsymbol{\varepsilon}^\top\boldsymbol{M}\boldsymbol{\varepsilon}$, therefore $\mathrm{E}[\hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}}] = \mathrm{E}[\operatorname{tr}\boldsymbol{\varepsilon}^\top\boldsymbol{M}\boldsymbol{\varepsilon}] = \mathrm{E}[\operatorname{tr}\boldsymbol{M}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top] = \sigma^2\operatorname{tr}\boldsymbol{M} = \sigma^2\operatorname{tr}(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^-\boldsymbol{X}^\top) = \sigma^2(n - \operatorname{tr}(\boldsymbol{X}^\top\boldsymbol{X})^-\boldsymbol{X}^\top\boldsymbol{X}) = \sigma^2(n-q)$. □

PROBLEM 314.

• a. *2 points Show that*

$$\mathsf{SSE} = \boldsymbol{\varepsilon}^\top\boldsymbol{M}\boldsymbol{\varepsilon} \qquad \textit{where} \qquad \boldsymbol{M} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^-\boldsymbol{X}^\top \tag{24.4.2}$$

ANSWER. $\mathsf{SSE} = \hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}}$, where $\hat{\boldsymbol{\varepsilon}} = \mathsf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = \mathsf{y} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^-\boldsymbol{X}^\top\mathsf{y} = \boldsymbol{M}\mathsf{y}$ where $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^-\boldsymbol{X}^\top$. From $\boldsymbol{M}\boldsymbol{X} = \boldsymbol{O}$ follows $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{M}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{M}\boldsymbol{\varepsilon}$. Since $\boldsymbol{M}$ is idempotent and symmetric, it follows $\hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}} = \boldsymbol{\varepsilon}^\top\boldsymbol{M}\boldsymbol{\varepsilon}$. □

• b. *1 point Is* SSE *observed? Is* $\boldsymbol{\varepsilon}$ *observed? Is* $\boldsymbol{M}$ *observed?*

• c. *3 points Under the usual assumption that* $\boldsymbol{X}$ *has full column rank, show that*

$$\mathrm{E}[\mathsf{SSE}] = \sigma^2(n-k) \tag{24.4.3}$$

ANSWER. $\mathrm{E}[\hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}}] = \mathrm{E}[\operatorname{tr}\boldsymbol{\varepsilon}^\top\boldsymbol{M}\boldsymbol{\varepsilon}] = \mathrm{E}[\operatorname{tr}\boldsymbol{M}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top] = \sigma^2\operatorname{tr}\boldsymbol{M} = \sigma^2\operatorname{tr}(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^-\boldsymbol{X}^\top) = \sigma^2(n - \operatorname{tr}(\boldsymbol{X}^\top\boldsymbol{X})^-\boldsymbol{X}^\top\boldsymbol{X}) = \sigma^2(n-k)$. □

PROBLEM 315. *As an alternative proof of* (24.4.3) *show that* $\mathsf{SSE} = \mathsf{y}^\top\boldsymbol{M}\mathsf{y}$ *and use theorem 9.2.1.*

From (24.4.3) follows that $\mathsf{SSE}/(n-q)$ is an unbiased estimate of σ^2. Although it is commonly suggested that $\mathsf{s}^2 = \mathsf{SSE}/(n-q)$ is an optimal estimator of σ^2, this is a fallacy. The question which estimator of σ^2 is best depends on the kurtosis of the distribution of the error terms. For instance, if the kurtosis is zero, which is the case when the error terms are normal, then a different scalar multiple of the SSE, namely, the Theil-Schweitzer estimator from [**TS61**]

$$\hat{\sigma}^2_{TS} = \frac{1}{n-q+2}\mathsf{y}^\top\boldsymbol{M}\mathsf{y} = \frac{1}{n-q+2}\sum_{i=1}^n \hat{\varepsilon}_i^2, \tag{24.4.4}$$

is biased but has lower MSE than s^2. Compare problem 191. The only thing one can say about s^2 is that it is a fairly good estimator which one can use when one does not know the kurtosis (but even in this case it is not the best one can do).

24.5. Mallow's Cp-Statistic as Estimator of the Mean Squared Error

PROBLEM 316. *We will compute here the $\mathcal{MSE}$-matrix of $\hat{\mathsf{y}}$ as an estimator of $\mathcal{E}[\mathsf{y}]$ in a regression which does not use the correct $\boldsymbol{X}$-matrix. For this we assume that $\mathsf{y} = \boldsymbol{\eta} + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim (\boldsymbol{o}, \sigma^2\boldsymbol{I})$. $\boldsymbol{\eta} = \mathcal{E}[\mathsf{y}]$ is an arbitrary vector of constants, and we do not assume that $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$ for some $\boldsymbol{\beta}$, i.e., we do not assume that $\boldsymbol{X}$ contains all the necessary explanatory variables. Regression of y on $\boldsymbol{X}$ gives the OLS estimator $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^-\boldsymbol{X}^\top\mathsf{y}$.*

• a. *2 points Show that the $\mathcal{MSE}$ matrix of $\hat{\mathsf{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ as estimator of $\boldsymbol{\eta}$ is*

$$\mathcal{MSE}[\boldsymbol{X}\hat{\boldsymbol{\beta}};\boldsymbol{\eta}] = \sigma^2\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^-\boldsymbol{X}^\top + \boldsymbol{M}\boldsymbol{\eta}\boldsymbol{\eta}^\top\boldsymbol{M} \tag{24.5.1}$$

where $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^-\boldsymbol{X}^\top$.

• b. *1 point Formula* (24.5.1) *for the $\mathcal{MSE}$ matrix depends on the unknown σ^2 and $\boldsymbol{\eta}$ and is therefore useless for estimation. If one cannot get an estimate of the whole $\mathcal{MSE}$ matrix, an often-used second best choice is its trace. Show that*

$$\operatorname{tr}\mathcal{MSE}[\boldsymbol{X}\hat{\boldsymbol{\beta}};\boldsymbol{\eta}] = \sigma^2 q + \boldsymbol{\eta}^\top\boldsymbol{M}\boldsymbol{\eta}. \tag{24.5.2}$$

where q is the rank of $\boldsymbol{X}$.

• c. *3 points If an unbiased estimator of the true σ^2 is available (call it s^2), then an unbiased estimator of the righthand side of* (24.5.2) *can be constructed using this* s^2 *and the* SSE *of the regression* $\mathsf{SSE} = \mathbf{y}^\top \boldsymbol{M}\mathbf{y}$. *Show that*

$$\mathrm{E}[\mathsf{SSE} - (n-2q)\mathsf{s}^2] = \sigma^2 q + \boldsymbol{\eta}^\top \boldsymbol{M}\boldsymbol{\eta}. \tag{24.5.3}$$

Hint: use equation (9.2.1). *If one does not have an unbiased estimator* s^2 *of* σ^2, *one usually gets such an estimator by regressing* $\mathbf{y}$ *on an* $\boldsymbol{X}$ *matrix which is so large that one can assume that it contains the true regressors.*

The statistic

$$\mathsf{C_p} = \frac{\mathsf{SSE}}{\mathsf{s}^2} + 2q - n \tag{24.5.4}$$

is called Mallow's $\mathsf{C_p}$ *statistic. It is a consistent estimator of* $\operatorname{tr}\mathcal{MSE}[\boldsymbol{X}\hat{\boldsymbol{\beta}};\boldsymbol{\eta}]/\sigma^2$. *If* $\boldsymbol{X}$ *contains all necessary variables, i.e.,* $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$ *for some* $\boldsymbol{\beta}$, *then* (24.5.2) *becomes* $\operatorname{tr}\mathcal{MSE}[\boldsymbol{X}\hat{\boldsymbol{\beta}};\boldsymbol{\eta}] = \sigma^2 q$, *i.e., in this case* $\mathsf{C_p}$ *should be close to* q. *Therefore the selection rule for regressions should be here to pick that regression for which the* $\mathsf{C_p}$*-value is closest to* q. *(This is an explanation; nothing to prove here.)*

If one therefore has several regressions and tries to decide which is the right one, it is recommended to plot C_p versus q for all regressions, and choose one for which this value is small and lies close to the diagonal. An example of this is given in problem 286.

24.6. Optimality of Variance Estimators

Regarding the estimator of σ^2, [**Ati62**] has the following result regarding minimum variance unbiased estimators:

THEOREM 24.6.1. *Assume* $\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ *where*

$$\mathrm{E}[\varepsilon_i] = 0 \tag{24.6.1}$$

$$\mathrm{E}[\varepsilon_i\varepsilon_j] = \sigma^2 \quad \textit{if } i = j \textit{ and 0 otherwise} \tag{24.6.2}$$

$$\mathrm{E}[\varepsilon_i\varepsilon_j\varepsilon_k] = \begin{cases} \sigma^3\gamma_1 & \textit{if } i = j = k \\ 0 & \textit{otherwise} \end{cases} \tag{24.6.3}$$

$$\mathrm{E}[\varepsilon_i\varepsilon_j\varepsilon_k\varepsilon_l] = \begin{cases} \sigma^4(\gamma_2+3) & \textit{if } i = j = k = l \\ \sigma^4 & \textit{if } i = j \neq k = l \textit{ or } i = k \neq j = l \textit{ or } i = l \neq j = k \\ 0 & \textit{otherwise.} \end{cases} \tag{24.6.4}$$

(Here γ_1 *is the skewness and* γ_2 *the kurtosis of* ε_i.*) This is for instance satisfied whenever all* ε_i *are independent drawings from the same population with* $\mathrm{E}[\varepsilon_i] = 0$ *and equal variances* $\operatorname{var}[\varepsilon_i] = \sigma^2$.

If either $\gamma_2 = 0$ *(which is for instance the case when* $\boldsymbol{\varepsilon}$ *is normally distributed), or if* $\boldsymbol{X}$ *is such that all diagonal elements of* $\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^-\boldsymbol{X}^\top$ *are equal, then the minimum* MSE *estimator in the class of unbiased estimators of* σ^2, *whatever the mean or dispersion matrix of* $\boldsymbol{\beta}$ *or its covariance matrix with* $\boldsymbol{\varepsilon}$ *may be, and which can be written in the form* $\mathbf{y}^\top\boldsymbol{A}\mathbf{y}$ *with a nonnegative definite* $\boldsymbol{A}$, *is*

$$s^2 = \frac{1}{n-q}\mathbf{y}^\top(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^-\boldsymbol{X}^\top)\mathbf{y}, \tag{24.6.5}$$

and

$$\mathrm{E}[(s^2-\sigma^2)^2] = \operatorname{var}[s^2] = \sigma^4\Big(\frac{2}{n-q} + \frac{\gamma_2}{n}\Big). \tag{24.6.6}$$

Proof: For notational convenience we will look at unbiased estimators of the form $\mathbf{y}^\top\boldsymbol{A}\mathbf{y}$ of $(n-q)\sigma^2$, and at the end divide by $n-q$. If $\boldsymbol{\beta}$ is nonrandom, then $\mathrm{E}[\mathbf{y}^\top\boldsymbol{A}\mathbf{y}] = \sigma^2\operatorname{tr}\boldsymbol{A} + \boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{A}\boldsymbol{X}\boldsymbol{\beta}$, and the estimator is unbiased iff $\boldsymbol{X}^\top\boldsymbol{A}\boldsymbol{X} = \boldsymbol{O}$ and $\operatorname{tr}\boldsymbol{A} = n-q$. Since $\boldsymbol{A}$ is assumed nonnegative definite, $\boldsymbol{X}^\top\boldsymbol{A}\boldsymbol{X} = \boldsymbol{O}$ is equivalent to $\boldsymbol{A}\boldsymbol{X} = \boldsymbol{O}$. From $\boldsymbol{A}\boldsymbol{X} = \boldsymbol{O}$ follows $\mathbf{y}^\top\boldsymbol{A}\mathbf{y} = \boldsymbol{\varepsilon}^\top\boldsymbol{A}\boldsymbol{\varepsilon}$, therefore the distribution of $\boldsymbol{\beta}$ no longer matters and (9.2.27) simplifies to

$$\operatorname{var}[\boldsymbol{\varepsilon}^\top\boldsymbol{A}\boldsymbol{\varepsilon}] = \sigma^4(\gamma_2\boldsymbol{a}^\top\boldsymbol{a} + 2\operatorname{tr}(\boldsymbol{A}^2)) \tag{24.6.7}$$

Now take an arbitrary nonnegative definite $\boldsymbol{A}$ with $\boldsymbol{A}\boldsymbol{X} = \boldsymbol{O}$ and $\operatorname{tr}\boldsymbol{A} = n-q$. Write it in the form $\boldsymbol{A} = \boldsymbol{M} + \boldsymbol{D}$, where $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^-\boldsymbol{X}^\top$, and correspondingly $\boldsymbol{a} = \boldsymbol{m} + \boldsymbol{d}$. The condition $\boldsymbol{A}\boldsymbol{X} = \boldsymbol{O}$ is equivalent to $\boldsymbol{A}\boldsymbol{M} = \boldsymbol{A}$ or, expressed in $\boldsymbol{D}$ instead of $\boldsymbol{A}$, $(\boldsymbol{M}+\boldsymbol{D})\boldsymbol{M} = \boldsymbol{M}+\boldsymbol{D}$, which simplifies to $\boldsymbol{D}\boldsymbol{M} = \boldsymbol{D}$. Furthermore, $\operatorname{tr}\boldsymbol{A} = n-q$ translates into $\operatorname{tr}\boldsymbol{D} = 0$. Hence $\boldsymbol{A}^2 = \boldsymbol{M} + 2\boldsymbol{D} + \boldsymbol{D}^2$, and $\operatorname{tr}\boldsymbol{A}^2 = n - q + \operatorname{tr}\boldsymbol{D}^2$. Plugged into (24.6.7) this gives

$$\operatorname{var}[\boldsymbol{\varepsilon}^\top\boldsymbol{A}\boldsymbol{\varepsilon}] = \sigma^4\Big(\gamma_2(\boldsymbol{m}^\top\boldsymbol{m} + 2\boldsymbol{m}^\top\boldsymbol{d} + \boldsymbol{d}^\top\boldsymbol{d}) + 2(n-q+\operatorname{tr}\boldsymbol{D}^2)\Big) \tag{24.6.8}$$

$$= \operatorname{var}[\boldsymbol{\varepsilon}^\top\boldsymbol{M}\boldsymbol{\varepsilon}] + \sigma^4\Big(\gamma_2(2\boldsymbol{m}^\top\boldsymbol{d} + \boldsymbol{d}^\top\boldsymbol{d}) + 2\operatorname{tr}\boldsymbol{D}^2\Big) \tag{24.6.9}$$

$$= \operatorname{var}[\boldsymbol{\varepsilon}^\top\boldsymbol{M}\boldsymbol{\varepsilon}] + \sigma^4\Big(\gamma_2\big(2\sum_i m_{ii}d_{ii} + \sum_i d_{ii}^2\big) + 2\sum_{i,j} d_{ij}^2\Big). \tag{24.6.10}$$

The minimization of this variance is easy if $\gamma_2 = 0$; then it follows $\boldsymbol{D} = \boldsymbol{O}$. The other case where one can easily do it is if $\boldsymbol{X}$ is such that the diagonal elements of $\boldsymbol{M}$ are all equal; then the first of the three summation terms is a multiple of $\operatorname{tr}\boldsymbol{D}$,

which is zero, and one obtains

$$\operatorname{var}[\boldsymbol{\varepsilon}^\top \boldsymbol{A}\boldsymbol{\varepsilon}] = \operatorname{var}[\boldsymbol{\varepsilon}^\top \boldsymbol{M}\boldsymbol{\varepsilon}] + \sigma^4\Big((\gamma_2+2)\sum_i d_{ii}^2 + 2\sum_{i,j:\, i\neq j} d_{ij}^2\Big). \tag{24.6.11}$$

Since $\gamma_2 + 2 \geq 0$ always, this is again minimized if $\boldsymbol{D} = \boldsymbol{O}$. If all diagonal elements of $\boldsymbol{M}$ are equal, they must have the value $(n-q)/n$, since $\operatorname{tr}\boldsymbol{M} = n-q$. therefore $\boldsymbol{m}^\top\boldsymbol{m} = (n-q)^2/n$, from which the formula for the MSE follows.

PROBLEM 317. *Show that, if $\boldsymbol{A}$ is not nnd, then $\boldsymbol{X}^\top\boldsymbol{A}\boldsymbol{X} = \boldsymbol{O}$ does not necessarily imply $\boldsymbol{A}\boldsymbol{X} = \boldsymbol{O}$.*

ANSWER. Counterexample: $\boldsymbol{A} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$ and $\boldsymbol{x} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$. □

CHAPTER 25

Variance Estimation: Should One Require Unbiasedness?

There is an imperfect analogy between linear estimation of the coefficients and quadratic estimation of the variance in the linear model. This chapter sorts out the principal commonalities and differences, a task obscured by the widespread but unwarranted imposition of the unbiasedness assumption. It is based on an unpublished paper co-authored with *Peter Ochshorn*.

We will work in the usual regression model

$$\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{25.0.12}$$

where $\mathbf{y}$ is a vector of n observations, $\boldsymbol{X}$ is nonstochastic with rank $r < n$, and the disturbance vector $\boldsymbol{\varepsilon}$ satisfies $\mathcal{E}[\boldsymbol{\varepsilon}] = \boldsymbol{o}$ and $\mathcal{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top] = \sigma^2\boldsymbol{I}$. The nonstochastic vector $\boldsymbol{\beta}$ and scalar $\sigma^2 > 0$ are the parameters to be estimated. The usual estimator of σ^2 is

$$\mathsf{s}^2 = \frac{1}{n-r}\mathbf{y}^\top\boldsymbol{M}\mathbf{y} = \frac{1}{n-r}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 \tag{25.0.13}$$

where $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^-\boldsymbol{X}^\top$ and $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{M}\mathbf{y}$. If $\boldsymbol{X}$ has full rank, then $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is the least squares estimator of $\boldsymbol{\beta}$. Just as $\hat{\boldsymbol{\beta}}$ is the best (minimum mean square error) linear unbiased estimator of $\boldsymbol{\beta}$, it has been shown in [**Ati62**], see also [**Seb77**, pp. 52/3], that under certain additional assumptions, s^2 is the best unbiased estimator of σ^2 which can be written in the form $\mathbf{y}^\top\boldsymbol{A}\mathbf{y}$ with a nonnegative definite $\boldsymbol{A}$. A precise formulation of these additional assumptions will be given below; they are, for instance, satisfied if $\boldsymbol{X} = \boldsymbol{\iota}$, the vector of ones, and the ε_i are i.i.d. But they are also satisfied for arbitrary $\boldsymbol{X}$ if $\boldsymbol{\varepsilon}$ is normally distributed. (In this last case, s^2 is best in the larger class of all unbiased estimators.)

This suggests an analogy between linear and quadratic estimation which is, however, by no means perfect. The results just cited pose the following puzzles:

- Why is s^2 not best nonnegative quadratic unbiased for arbitrary $\boldsymbol{X}$-matrix whenever the ε_i are i.i.d. with zero mean? What is the common logic behind the two disparate alternatives, that either restrictions on $\boldsymbol{X}$ or restrictions on the distribution of the ε_i can make s^2 optimal?
- It comes as a surprise that, again under the assumption of normality, a very simple modification of s^2, namely, the Theil-Schweitzer estimator from [**TS61**]

$$\hat{\sigma}^2 = \frac{1}{n-r+2}\mathbf{y}^\top\boldsymbol{M}\mathbf{y} = \frac{1}{n-r+2}\sum_{i=1}^{n}\hat{\varepsilon}_i^2, \tag{25.0.14}$$

  which is biased, has lower mean square error (MSE) than s^2.
- It is unclear why it is necessary to require ex ante that $\boldsymbol{A}$ is nonnegative definite. Wouldn't estimators which can yield negative values for σ^2 be automatically inferior to nonnegative ones?

We will show that these puzzles can be resolved if one replaces the requirement of unbiasedness by that of bounded MSE. (This is particularly satisfying since such a replacement is also called for in the case of linear estimators.) Then puzzle (2) disappears: the Theil-Schweitzer estimator is no longer an oddity but it is the best bounded MSE quadratic estimator of σ^2 when the kurtosis is zero. And puzzle (3) disappears as well: nonnegativity is only necessary because unbiasedness alone does not imply bounded MSE. Under this approach it becomes evident that there are two important disanalogies between linear and quadratic estimation: whereas the best bounded MSE linear estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is (a) unbiased and (b) does not depend on the nuisance parameter σ^2, the best quadratic bounded MSE estimator of σ^2 is (a) biased and (b) depends on a fourth-order nuisance parameter, the kurtosis of the disturbances. This, again, helps to dispel the false suggestiveness of puzzle (1). The main assumption is distributional. If the kurtosis is known, then the best nonnegative quadratic unbiased estimator exists. However it is uninteresting, since the (biased) best bounded MSE quadratic estimator is better. The class of unbiased estimators only then becomes interesting when the kurtosis is not known: for certain $\boldsymbol{X}$-matrices, the best nonnegative quadratic unbiased estimator does not depend on the kurtosis.

However even if the kurtosis is not known, this paper proposes to use as estimate of σ^2 the maximum value which one gets when one applies the best bonded mean squared error estimator for all possible values of the kurtosis.

25.1. Setting the Framework Straight

The assumption of unbiasedness has often been criticized. Despite its high-sounding name, there are no good reasons that one should confine one's search for good estimators to unbiased ones. Many good estimators are unbiased, but the property of unbiasedness has no bearing on how good an estimator is. In many cases

unbiased estimators do not exist or are not desirable. It is indeed surprising that the powerful building of least squares theory seems to rest on such a flimsy assumption as unbiasedness.

G. A. Barnard, in [**Bar63**], noted this and proposed to replace unbiasedness by bounded MSE, a requirement which can be justified by the researcher following an "insurance strategy": no bad surprises regarding the MSE of the estimator, whatever the value of the true $\boldsymbol{\beta}$. Barnard's suggestion has not found entrance into the textbooks—and indeed, since linear estimators in model (25.0.12) are unbiased if and only if they have bounded MSE, it might be considered an academic question.

It is usually not recognized that even in the linear case, the assumption of bounded MSE serves to unify the theory. Christensen's monograph [**Chr87**] treats, as we do here in chapter 27, best linear prediction on the basis of known first and second moments in parallel with the regression model. Both models have much in common, but there is one result which seems to set them apart: best linear predictors exist in one, but only best linear *unbiased* predictors in the other [**Chr87**, p. 226]. If one considers bounded MSE to be one of the basic assumptions, this seeming irregularity is easily explained: If the first and second moments are known, then *every* linear predictor has bounded MSE, while in the regression model only *unbiased* linear estimators do.

One might still argue that no real harm is done with the assumption of unbiasedness, because in the linear case, the best bounded MSE estimators or predictors turn out to be unbiased. This last defense of unbiasedness falls if one goes from linear to quadratic estimation. We will show that the best bounded MSE quadratic estimator is biased.

As in the the linear case, it is possible to derive these results without fully specifying the distributions involved. In order to compute the MSE of linear estimators, one needs to know the first and second moments of the disturbances, which is reflected in the usual assumption $\boldsymbol{\varepsilon} \sim (\boldsymbol{o}, \sigma^2\boldsymbol{I})$. For the MSE of quadratic estimators, one also needs information about the third and fourth moments. We will therefore derive optimal quadratic estimators of σ^2 based on the following assumptions regarding the first four moments, which are satisfied whenever the ε_i are independently identically distributed:

Assumption. A vector of n observations $\mathbf{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ is available, where ε_i satisfy

$$\mathrm{E}[\varepsilon_i] = 0 \tag{25.1.1}$$

$$\mathrm{E}[\varepsilon_i\varepsilon_j] = \begin{cases} \sigma^2 > 0 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} \tag{25.1.2}$$

$$\mathrm{E}[\varepsilon_i\varepsilon_j\varepsilon_k] = \begin{cases} \sigma^3\gamma_1 & \text{if } i = j = k \\ 0 & \text{otherwise} \end{cases} \tag{25.1.3}$$

$$\mathrm{E}[\varepsilon_i\varepsilon_j\varepsilon_k\varepsilon_l] = \begin{cases} \sigma^4(\gamma_2+3) & \text{if } i = j = k = l \\ \sigma^4 & \text{if } i = j \neq k = l \text{ or } i = k \neq j = l \text{ or } i = l \neq j = k \\ 0 & \text{otherwise.} \end{cases} \tag{25.1.4}$$

Here γ_1 is the skewness and γ_2 the kurtosis of ε_i. They are allowed to range within their natural limits

$$0 \leq \gamma_1^2 \leq \gamma_2 + 2. \tag{25.1.5}$$

Problem 318. *Show that the condition* (25.1.5), $\gamma_1^2 \leq \gamma_2 + 2$, *always holds.*

Answer.

$$(\sigma^3\gamma_1)^2 = (\mathrm{E}[\varepsilon^3])^2 = \big(\mathrm{cov}[\varepsilon, \varepsilon^2]\big)^2 \leq \mathrm{var}[\varepsilon]\,\mathrm{var}[\varepsilon^2] = \sigma^6(\gamma_2+2) \tag{25.1.6}$$

□

The concept of bounded MSE which is appropriate here requires the bound to be independent of the true value of $\boldsymbol{\beta}$, but it may depend on the "nuisance parameters" σ^2, γ_1, and γ_2:

Definition 25.1.1. The mean square error $\mathrm{E}[(\hat{\theta}-\theta)^2]$ of the estimator $\hat{\theta}$ of a scalar parameter θ in the linear model (25.0.12) will be said to be bounded (with respect to $\boldsymbol{\beta}$) if a finite number b exists with $\mathrm{E}[(\hat{\theta}-\theta)^2] \leq b$ regardless of the true value of $\boldsymbol{\beta}$. This bound b may depend on the known nonstochastic $\boldsymbol{X}$ and the distribution of $\boldsymbol{\varepsilon}$, but not on $\boldsymbol{\beta}$.

25.2. Derivation of the Best Bounded MSE Quadratic Estimator of the Variance

Theorem 25.2.1. *If the estimator $\tilde{\sigma}^2$ of σ^2 in the regression model* (25.0.12) *is quadratic, i.e., if it has the form $\tilde{\sigma}^2 = \mathbf{y}^\top\boldsymbol{A}\mathbf{y}$ with a symmetric $\boldsymbol{A}$, then its mean square error $\mathrm{E}[(\mathbf{y}^\top\boldsymbol{A}\mathbf{y}-\sigma^2)^2]$ is bounded (with respect to $\boldsymbol{\beta}$) if and only if $\boldsymbol{AX} = \boldsymbol{O}$.*

Proof: Clearly, the condition $\boldsymbol{AX} = \boldsymbol{O}$ is sufficient. It implies $\mathbf{y}^\top \boldsymbol{A}\mathbf{y} = \boldsymbol{\varepsilon}^\top \boldsymbol{A}\boldsymbol{\varepsilon}$, which therefore only depends on the distribution of $\boldsymbol{\varepsilon}$, not on the value of $\boldsymbol{\beta}$. To show necessity, note that bounded MSE means both bounded variance and bounded squared bias. The variance depends on skewness and kurtosis; writing $\boldsymbol{a}$ for the vector of diagonal elements of $\boldsymbol{A}$, it is

$$\operatorname{var}[\mathbf{y}^\top \boldsymbol{A}\mathbf{y}] = 4\sigma^2\boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{A}^2\boldsymbol{X}\boldsymbol{\beta} + 4\sigma^3\gamma_1\boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{a} + \sigma^4\Big(\gamma_2\boldsymbol{a}^\top\boldsymbol{a} + 2\operatorname{tr}(\boldsymbol{A}^2)\Big). \tag{25.2.1}$$

This formula can be found e.g. in [**Seb77**, pp. 14–16 and 52]. If $\boldsymbol{AX} \neq \boldsymbol{O}$, then a vector $\boldsymbol{\delta}$ exists with $\boldsymbol{\delta}^\top \boldsymbol{X}^\top \boldsymbol{A}^2\boldsymbol{X}\boldsymbol{\delta} > 0$; therefore, for the sequence $\boldsymbol{\beta} = j\boldsymbol{\delta}$, the variance is a quadratic polynomial in j, which is unbounded as $j \to \infty$.

The following ingredients are needed for the best bounded MSE quadratic estimator of σ^2:

THEOREM 25.2.2. *We will use the letter $\boldsymbol{\tau}$ to denote the vector whose ith component is the square of the ith residual $\tau_i = \hat{\varepsilon}_i^2$. Then*

$$\mathcal{E}[\boldsymbol{\tau}] = \sigma^2\boldsymbol{m} \tag{25.2.2}$$

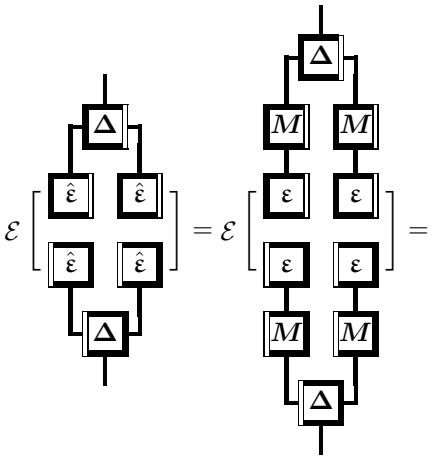
where $\boldsymbol{m}$ is the diagonal vector of $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^-\boldsymbol{X}^\top$. Furthermore,

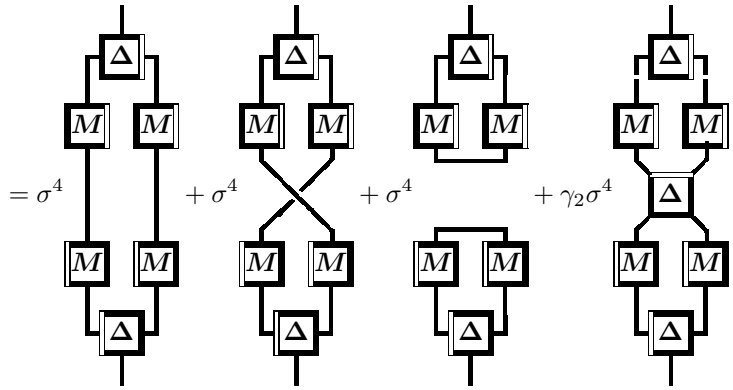
$$\mathcal{V}[\boldsymbol{\tau}] = \sigma^4\boldsymbol{\Omega} \qquad \text{where} \quad \boldsymbol{\Omega} = \gamma_2\boldsymbol{Q}^2 + 2\boldsymbol{Q} + \boldsymbol{m}\boldsymbol{m}^\top, \tag{25.2.3}$$

$\boldsymbol{Q}$ is the matrix with $q_{ij} = m_{ij}^2$, i.e., its elements are the squares of the elements of $\boldsymbol{M}$, and γ_2 is the kurtosis.

Here is a proof in tile notation: from (9.2.23) follows

(25.2.4)





These $\boldsymbol{m}$ and $\boldsymbol{\Omega}$ play an important and, to me, surprising role in the estimator of σ^2:

THEOREM 25.2.3. *The best bounded* MSE *quadratic estimator of σ^2 is*

$$\hat{\sigma}^2 = \boldsymbol{m}^\top\boldsymbol{\Omega}^-\boldsymbol{\tau} \tag{25.2.5}$$

where $\boldsymbol{m}$ and $\boldsymbol{\Omega}$ are defined as in Theorem 25.2.2. Other ways to write it are

$$\hat{\sigma}^2\mathbf{y}^\top\boldsymbol{M}\boldsymbol{\Lambda}\boldsymbol{M}\mathbf{y} = \sum_i \lambda_i\hat{\varepsilon}_i^2 \tag{25.2.6}$$

where $\boldsymbol{\lambda} = \boldsymbol{\Omega}^-\boldsymbol{m}$ or any other vector satisfying

$$\boldsymbol{\Omega}\boldsymbol{\lambda} = \boldsymbol{m}, \tag{25.2.7}$$

and $\boldsymbol{\Lambda}$ is the diagonal matrix with $\boldsymbol{\lambda}$ in the diagonal. The MSE *of estimator* (25.2.6) *is*

$$\mathrm{E}[(\hat{\sigma}^2 - \sigma^2)^2] = \sigma^4(1 - \boldsymbol{m}^\top\boldsymbol{\Omega}^-\boldsymbol{m}). \tag{25.2.8}$$

The estimator is negatively biased; its bias is

$$\mathrm{E}[\hat{\sigma}^2 - \sigma^2] = -\sigma^2(1 - \boldsymbol{m}^\top\boldsymbol{\Omega}^-\boldsymbol{m}). \tag{25.2.9}$$

The estimator (25.2.6) is therefore independent of the skewness of the disturbances, but it depends on their kurtosis. For zero kurtosis, it reduces to the Theil-Schweitzer estimator (25.0.14).

In order to prove theorem 25.2.3, one needs somewhat different matrix-algebraic tricks than those familiar from linear estimators. The problem at hand can be reduced to a minimum trace problem as defined in Rao [**Rao73**, pp. 65–66], and part of the following proof draws on a private communication of C. R. Rao regarding consistency of equation (1f.3.4) in [**Rao73**, p. 65].

Proof of theorem 25.2.3: Take an alternative estimator of the form $\tilde{\sigma}^2 = \mathbf{y}^\top \boldsymbol{A}\mathbf{y}$ where $\boldsymbol{A}$ is symmetric with $\boldsymbol{AX} = \boldsymbol{O}$. Since the MSE is variance plus squared bias, it follows, using (25.2.1) and $\mathrm{E}[\mathbf{y}^\top \boldsymbol{A}\mathbf{y}] = \sigma^2 \operatorname{tr} \boldsymbol{A} + \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{AX}\boldsymbol{\beta}$, that

$$\text{(25.2.10)} \qquad \mathrm{MSE} = \mathrm{E}[(\tilde{\sigma}^2 - \sigma^2)^2] = \sigma^4(\operatorname{tr} \boldsymbol{A} - 1)^2 + \sigma^4 \gamma_2 \boldsymbol{a}^\top \boldsymbol{a} + 2\sigma^4 \operatorname{tr}(\boldsymbol{A}^2)$$

where $\boldsymbol{a} = \operatorname{diag} \boldsymbol{A}$. We will first prove the following property of $\boldsymbol{a}$: every vector $\boldsymbol{u}$ with $\boldsymbol{Qu} = \boldsymbol{o}$ satisfies $\boldsymbol{a}^\top \boldsymbol{u} = 0$. Let $\boldsymbol{U}$ be the diagonal matrix with $\boldsymbol{u}$ in the diagonal, and note that $\boldsymbol{Qu} = \boldsymbol{o}$ implies $0 = \boldsymbol{u}^\top \boldsymbol{Qu}$. Writing $\boldsymbol{V} = \boldsymbol{MU}$ gives $v_{ij} = m_{ij} u_j$, therefore $\boldsymbol{u}^\top \boldsymbol{Qu} = \sum_{i,j} u_i q_{ij} u_j = \sum_{i,j} v_{ij} v_{ji} = \operatorname{tr}(\boldsymbol{MUMU})$. Since $\boldsymbol{M}^2 = \boldsymbol{M}$, $\operatorname{tr}(\boldsymbol{MUMU}) = 0$ implies $\operatorname{tr}(\boldsymbol{MUMUM}) = 0$, and since $\boldsymbol{M}$ is nonnegative definite, it follows $\boldsymbol{MUMUM} = \boldsymbol{O}$, and therefore already $\boldsymbol{MUM} = \boldsymbol{O}$. Since $\boldsymbol{AX} = \boldsymbol{O} = \boldsymbol{X}^\top \boldsymbol{A}$ implies $\boldsymbol{A} = \boldsymbol{MAM}$, one can write $\boldsymbol{a}^\top \boldsymbol{u} = \operatorname{tr}(\boldsymbol{AU}) = \operatorname{tr}(\boldsymbol{MAMU}) = \operatorname{tr}(\boldsymbol{AMUM}) = 0$.

This property of $\boldsymbol{a}$ can also be formulated as: there exists a vector $\boldsymbol{\lambda}$ with $\boldsymbol{a} = \boldsymbol{Q\lambda}$. Let $\boldsymbol{\Lambda}$ be the diagonal matrix with $\boldsymbol{\lambda}$ in the diagonal, and write $\boldsymbol{A} = \boldsymbol{M\Lambda M} + \boldsymbol{D}$. Then $\boldsymbol{DX} = \boldsymbol{O}$ and $\boldsymbol{D}$ has zeros in the diagonal, therefore $\operatorname{tr}(\boldsymbol{M\Lambda MD}) = \operatorname{tr}(\boldsymbol{\Lambda MDM}) = \operatorname{tr}(\boldsymbol{\Lambda D}) = 0$, since $\boldsymbol{\Lambda D}$ still has zeros in the diagonal. Therefore $\operatorname{tr}(\boldsymbol{A}^2) = \operatorname{tr}(\boldsymbol{M\Lambda M\Lambda}) + \operatorname{tr}(\boldsymbol{D}^2) = \boldsymbol{\lambda}^\top \boldsymbol{Q\lambda} + \operatorname{tr}(\boldsymbol{D}^2)$. Regarding $\boldsymbol{Q}$ observe that $\boldsymbol{m} = \operatorname{diag} \boldsymbol{M}$ can be written $\boldsymbol{m} = \boldsymbol{Q\iota}$, where $\boldsymbol{\iota}$ is the vector of ones, therefore $\operatorname{tr} \boldsymbol{A} = \boldsymbol{\iota}^\top \boldsymbol{Q\lambda} = \boldsymbol{m}^\top \boldsymbol{\lambda}$. Using all this in (25.2.10) gives

$$\text{(25.2.11)} \qquad \frac{1}{\sigma^4} \mathrm{MSE} = (\boldsymbol{m}^\top \boldsymbol{\lambda} - 1)^2 + \gamma_2 \boldsymbol{\lambda}^\top \boldsymbol{Q}^2 \boldsymbol{\lambda} + 2\boldsymbol{\lambda}^\top \boldsymbol{Q\lambda} + 2\operatorname{tr}(\boldsymbol{D}^2).$$

Define $\boldsymbol{\Sigma} = (\gamma_2 + 2)\boldsymbol{Q}^2 + 2(\boldsymbol{Q} - \boldsymbol{Q}^2)$. It is the sum of two nonnegative definite matrices: $\gamma_2 + 2 \geq 0$ by (25.1.5), and $\boldsymbol{Q} - \boldsymbol{Q}^2$ is nonnegative definite because $\boldsymbol{\lambda}^\top (\boldsymbol{Q} - \boldsymbol{Q}^2)\boldsymbol{\lambda}$ is the sum of the squares of the offdiagonal elements of $\boldsymbol{M\Lambda M}$. Therefore $\boldsymbol{\Sigma}$ is nonnegative definite and it follows $\boldsymbol{m} = (\boldsymbol{\Sigma} + \boldsymbol{mm}^\top)(\boldsymbol{\Sigma} + \boldsymbol{mm}^\top)^- \boldsymbol{m}$. (To see this, take any $\boldsymbol{P}$ with $\boldsymbol{\Sigma} = \boldsymbol{PP}^\top$ and apply the identity $\boldsymbol{T} = \boldsymbol{TT}^\top(\boldsymbol{TT}^\top)^- \boldsymbol{T}$, proof e.g. in [**Rao73**, p. 26], to the partitioned matrix $\boldsymbol{T} = \begin{bmatrix} \boldsymbol{P} & \boldsymbol{m} \end{bmatrix}$.)

Writing $\boldsymbol{\Omega} = \boldsymbol{\Sigma} + \boldsymbol{mm}^\top$, one verifies therefore

$$\text{(25.2.12)} \qquad \frac{1}{\sigma^4} \mathrm{MSE} = (\boldsymbol{\lambda} - \boldsymbol{\Omega}^- \boldsymbol{m})^\top \boldsymbol{\Omega}(\boldsymbol{\lambda} - \boldsymbol{\Omega}^- \boldsymbol{m}) - \boldsymbol{m}^\top \boldsymbol{\Omega}^- \boldsymbol{m} + 1 + 2\operatorname{tr}(\boldsymbol{D}^2).$$

Clearly, this is minimized by $\boldsymbol{D} = \boldsymbol{O}$ and any $\boldsymbol{\lambda}$ with $\boldsymbol{\Omega}(\boldsymbol{\lambda} - \boldsymbol{\Omega}^- \boldsymbol{m}) = \boldsymbol{o}$, which gives (25.2.7).

25.3. Unbiasedness Revisited

Unbiasedness of the estimator $\mathbf{y}^\top \boldsymbol{A}\mathbf{y}$ is equivalent to the two mathematical conditions $\operatorname{tr} \boldsymbol{A} = 1$ and $\boldsymbol{X}^\top \boldsymbol{AX} = \boldsymbol{O}$. This is not strong enough to ensure that the estimation error is a function of $\boldsymbol{\varepsilon}$ alone. In [**Hsu38**], the first treatment of best

quadratic estimation of σ^2, P. L. Hsu added therefore the condition that the MSE be independent of the value of $\boldsymbol{\beta}$.

But why should the data analyst be particulary interested in estimates whose MSE is independent of $\boldsymbol{\beta}$? The research following up on Hsu tried to get rid of this assumption again. C. R. Rao, in [**Rao52**], replaced independence of the MSE by the assumption that $\boldsymbol{A}$ be nonnegative definite. We argue that this was unfortunate, for the following two reasons:

- Although one can often read that it is "natural" to require $\boldsymbol{A}$ to be nonnegative definite (see for instance [**Ati62**, p. 84]), we disagree. Of course, one should expect the best estimator to be nonnegative, but is perplexing that one should have to *assume* it. We already noted this in puzzle (3) at the beginning.
- In the light of theorem 25.2.1, Hsu's additional condition is equivalent to the requirement of bounded MSE. It is therefore not as poorly motivated as it was at first assumed to be. Barnard's article [**Bar63**], arguing that this assumption is even in the linear case more meaningful than unbiasedness, appeared eleven years after Rao's [**Rao52**]. If one wanted to improve on Hsu's result, one should therefore discard the condition of unbiasedness, not that of bounded MSE.

Even the mathematical proof based on unbiasedness and nonnegative definiteness suggests that the condition $\boldsymbol{AX} = \boldsymbol{O}$, i.e., bounded MSE, is the more fundamental assumption. Nonnegative definitenes of $\boldsymbol{A}$ is used only once, in order to get from the condition $\boldsymbol{X}^\top \boldsymbol{AX} = \boldsymbol{O}$ implied by unbiasedness to $\boldsymbol{AX} = \boldsymbol{O}$. Unbiasedness and a nonnegative definite $\boldsymbol{A}$ together happen to imply bounded MSE, but neither condition separately should be considered "natural" in the present framework.

The foregoing discussion seems to be academic, since the best bounded MSE estimator depends on γ_2, which is rarely known. But it does not depend on it very much. I have not yet researched it fully, but it seems to be a concave function with a maximum somewhere. If one uses the estimate of σ^2 in order to assess the precision of some estimates, this maximum value may provide a conservative estimate which is still smaller than the unbiased estimate of σ^2. here these notes are still incomplete; I would like to know more about this maximum value, and it seems this would be the estimator which one should recommend.

If the requirement of unbiasedness has any redeeming qualities, they come from an unexpected yet remarkable fact. In some special cases one does not need to know the kurtosis if one restricts oneself to unbiased estimators of σ^2. In order to rederive this (known) result in our framework, we will first give a formula for Hsu's estimator. We obtain it from estimator (25.2.6) by multiplying it with the appropriate constant which makes it unbiased.

THEOREM 25.3.1. *The best bounded* MSE *quadratic unbiased estimator of* σ^2*, which is at the same time the best nonnegative quadratic unbiased estimator of* σ^2*, is*

$$\hat{\hat{\sigma}}^2 = \mathbf{y}^\top \boldsymbol{M}\boldsymbol{\Theta}\boldsymbol{M}\mathbf{y} = \sum_i \theta_i \hat{\varepsilon}_i^2 \tag{25.3.1}$$

where $\boldsymbol{\Theta}$ *is a diagonal matrix whose diagonal vector* $\boldsymbol{\theta}$ *satisfies the two conditions that*

$$\boldsymbol{\Omega}\boldsymbol{\theta} \quad \textit{is proportional to} \quad \boldsymbol{m}, \tag{25.3.2}$$

and that

$$\boldsymbol{m}^\top \boldsymbol{\theta} = 1 \tag{25.3.3}$$

(for instance one may use $\boldsymbol{\theta} = \boldsymbol{\lambda}\frac{1}{\boldsymbol{m}^\top\boldsymbol{\lambda}}$*.)* $\boldsymbol{M}$*,* $\boldsymbol{m}$*,* $\boldsymbol{\Omega}$*, and* $\boldsymbol{\lambda}$ *are the same as in theorem 25.2.3. The* MSE *of this estimator is*

$$\mathrm{E}[(\hat{\hat{\sigma}}^2 - \sigma^2)^2] = \sigma^4\left(\frac{1}{\boldsymbol{m}^\top\boldsymbol{\lambda}} - 1\right). \tag{25.3.4}$$

We omit the proof, which is very similar to that of theorem 25.2.3. In the general case, estimator (25.3.1) depends on the kurtosis, just as estimator (25.2.6) does. But if $\boldsymbol{X}$ is such that all diagonal elements of $\boldsymbol{M}$ are equal, a condition which Atiqullah in [**Ati62**] called "quadratically balanced," then it does not! Since $\operatorname{tr}\boldsymbol{M} = n - r$, equality of the diagonal elements implies $\boldsymbol{m} = \frac{n-r}{n}\boldsymbol{\iota}$. And since $\boldsymbol{m} = \boldsymbol{Q}\boldsymbol{\iota}$, any vector proportional to $\boldsymbol{\iota}$ satisfies (25.3.2), i.e., one can find solutions of (25.3.2) without knowing the kurtosis. (25.3.3) gives $\boldsymbol{\theta} = \boldsymbol{\iota}\frac{1}{n-r}$, i.e., the resulting estimator is none other than the unbiased s^2 defined in (25.0.13).

The property of unbiasedness which makes it so popular in the classroom—it is easy to check—gains here objective relevance. For the best nonnegative quadratic unbiased estimator one needs to know $\boldsymbol{\Omega}$ only up to a scalar factor, and in some special cases the unknown kurtosis merges into this arbitrary multiplicator.

25.4. Summary

If one replaces the requirement of unbiasedness by that of bounded MSE, one can not only unify some known results in linear estimation and prediction, but one also obtains a far-reaching analogy between linear estimation of $\boldsymbol{\beta}$ and quadratic estimation of σ^2. The most important dissimilarity is that, whereas one does not have to know the nuisance parameter σ^2 in order to write down the best linear bounded MSE estimator of $\boldsymbol{\beta}$, the best quadratic bounded MSE estimator of σ^2 depends on an additional fourth order nuisance parameter, namely, the kurtosis. In situations in which the kurtosis is known, one should consider the best quadratic bounded MSE estimator (25.2.6) of σ^2 to be the quadratic analog of the least squares

estimator $\hat{\boldsymbol{\beta}}$. It is a linear combination of the squared residuals, and if the kurtosis is zero, it specializes to the Theil-Schweitzer estimator (25.0.14). Regression computer packages, which require normality for large parts of their output, should therefore provide the Theil-Schweitzer estimate as a matter of course.

If the kurtosis is not known, one can always resort to s^2. It is unbiased and consistent, but does not have any optimality properties in the general case. If the design matrix is "quadratically balanced," s^2 can be justified better: in this case s^2 has minimum MSE in the class of nonnegative quadratic unbiased estimators (which is a subclass of all bounded MSE quadratic estimators).

The requirement of unbiasedness for the variance estimator in model (25.0.12) is therefore not as natural as is often assumed. Its main justification is that it may help to navigate around the unknown nuisance parameter "kurtosis."

CHAPTER 26

Nonspherical Positive Definite Covariance Matrix

The so-called "Generalized Least Squares" model specifies $\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim (\boldsymbol{o}, \sigma^2\boldsymbol{\Psi})$ where σ^2 is an unknown positive scalar, and $\boldsymbol{\Psi}$ is a *known* positive definite matrix.

This is simply the OLS model in disguise. To see this, we need a few more facts about positive definite matrices. $\boldsymbol{\Psi}$ is *nonnegative* definite if and only if a $\boldsymbol{Q}$ exists with $\boldsymbol{\Psi} = \boldsymbol{Q}\boldsymbol{Q}^\top$. If $\boldsymbol{\Psi}$ is *positive* definite, this $\boldsymbol{Q}$ can be chosen square and nonsingular. Then $\boldsymbol{P} = \boldsymbol{Q}^{-1}$ satisfies $\boldsymbol{P}^\top\boldsymbol{P}\boldsymbol{\Psi} = \boldsymbol{P}^\top\boldsymbol{P}\boldsymbol{Q}\boldsymbol{Q}^\top = \boldsymbol{I}$, i.e., $\boldsymbol{P}^\top\boldsymbol{P} = \boldsymbol{\Psi}^{-1}$, and also $\boldsymbol{P}\boldsymbol{\Psi}\boldsymbol{P}^\top = \boldsymbol{P}\boldsymbol{Q}\boldsymbol{Q}^\top\boldsymbol{P}^\top = \boldsymbol{I}$. Premultiplying the GLS model by $\boldsymbol{P}$ gives therefore a model whose disturbances have a spherical covariance matrix:

$$\boldsymbol{P}\mathbf{y} = \boldsymbol{P}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{P}\boldsymbol{\varepsilon} \qquad \boldsymbol{P}\boldsymbol{\varepsilon} \sim (\boldsymbol{o}, \sigma^2\boldsymbol{I}) \tag{26.0.1}$$

The OLS estimate of $\boldsymbol{\beta}$ in this transformed model is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{P}^\top\boldsymbol{P}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{P}^\top\boldsymbol{P}\mathbf{y} = (\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\mathbf{y}. \tag{26.0.2}$$

This $\hat{\boldsymbol{\beta}}$ is the BLUE of $\boldsymbol{\beta}$ in model (26.0.1), and since estimators which are linear in $\boldsymbol{P}\mathbf{y}$ are also linear in $\mathbf{y}$ and vice versa, $\hat{\boldsymbol{\beta}}$ is also the BLUE in the original GLS model.

Problem 319. *2 points Show that*

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\varepsilon} \tag{26.0.3}$$

and derive from this that $\hat{\boldsymbol{\beta}}$ *is unbiased and that* $\mathcal{MSE}[\hat{\boldsymbol{\beta}}; \boldsymbol{\beta}] = \sigma^2(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}$.

Answer. Proof of (26.0.3) is very similar to proof of (24.0.7). □

The objective function of the associated least squares problem is

$$\boldsymbol{\beta} = \hat{\boldsymbol{\beta}} \qquad \text{minimizes} \qquad (\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta})^\top\boldsymbol{\Psi}^{-1}(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}). \tag{26.0.4}$$

The normal equations are

$$\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\mathbf{y} \tag{26.0.5}$$

If $\boldsymbol{X}$ has full rank, then $\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X}$ is nonsingular, and the unique $\hat{\boldsymbol{\beta}}$ minimizing (26.0.4) is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\mathbf{y} \tag{26.0.6}$$

Problem 320. [**Seb77**, p. 386, 5] *Show that if* $\boldsymbol{\Psi}$ *is positive definite and* $\boldsymbol{X}$ *has full rank, then also* $\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X}$ *is positive definite. You are allowed to use, without proof, that the inverse of a positive definite matrix is also positive definite.*

Answer. From $\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X}\boldsymbol{a} = \boldsymbol{o}$ follows $\boldsymbol{a}^\top\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X}\boldsymbol{a} = 0$, and since $\boldsymbol{\Psi}^{-1}$ is positive definite, it follows $\boldsymbol{X}\boldsymbol{a} = \boldsymbol{o}$, and since $\boldsymbol{X}$ has full column rank, this implies $\boldsymbol{a} = \boldsymbol{o}$. □

Problem 321. *Show that* (26.0.5) *has always at least one solution, and that the general solution can be written as*

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-}\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\mathbf{y} + \boldsymbol{U}\boldsymbol{\gamma} \tag{26.0.7}$$

where $\boldsymbol{X} \perp \boldsymbol{U}$ *and* $\boldsymbol{\gamma}$ *is an arbitrary vector. Show furthermore that, if* $\hat{\boldsymbol{\beta}}$ *is a solution of* (26.0.5)*, and* $\boldsymbol{\beta}$ *is an arbitrary vector, then*

$$(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta})^\top\boldsymbol{\Psi}^{-1}(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}) = (\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top\boldsymbol{\Psi}^{-1}(\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^\top\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}). \tag{26.0.8}$$

Conclude from this that (26.0.5) *is a necessary and sufficient condition characterizing the values* $\hat{\boldsymbol{\beta}}$ *minimizing* (26.0.4).

Answer. One possible solution of (26.0.5) is $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-}\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\mathbf{y}$. Since the normal equations are consistent, (26.0.7) can be obtained from equation (A.4.1), using Problem 574. To prove (26.0.8), write (26.0.4) as $\big((\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) - \boldsymbol{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\big)^\top\boldsymbol{\Psi}^{-1}\big((\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) - \boldsymbol{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\big)$; since $\hat{\boldsymbol{\beta}}$ satisfies (26.0.5), the cross product terms disappear. Necessity of the normal equations: for any solution $\boldsymbol{\beta}$ of the minimization, $\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) = \boldsymbol{o}$. This together with (26.0.5) gives $\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\mathbf{y}$. □

The least squares objective function of the transformed model, which $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ minimizes, can be written

$$(\boldsymbol{P}\mathbf{y} - \boldsymbol{P}\boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{P}\mathbf{y} - \boldsymbol{P}\boldsymbol{X}\boldsymbol{\beta}) = (\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta})^\top\boldsymbol{\Psi}^{-1}(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}), \tag{26.0.9}$$

and whether one writes it in one form or the other, $1/(n-k)$ times the minimum value of that GLS objective function is still an unbiased estimate of σ^2.

Problem 322. *Show that the minimum value of the GLS objective function can be written in the form* $\mathbf{y}^\top\boldsymbol{M}\mathbf{y}$ *where* $\boldsymbol{M} = \boldsymbol{\Psi}^{-1} - \boldsymbol{\Psi}^{-1}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}$. *Does* $\boldsymbol{M}\boldsymbol{X} = \boldsymbol{O}$ *still hold? Does* $\boldsymbol{M}^2 = \boldsymbol{M}$ *or a similar simple identity still hold? Show that* $\boldsymbol{M}$ *is nonnegative definite. Show that* $\mathrm{E}[\mathbf{y}^\top\boldsymbol{M}\mathbf{y}] = (n-k)\sigma^2$.

Answer. In $(\mathbf{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^\top\boldsymbol{\Psi}^{-1}(\mathbf{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})$ plug in $\hat{\boldsymbol{\beta}}=(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\mathbf{y}$ and multiply out to get $\mathbf{y}^\top\boldsymbol{M}\mathbf{y}$. Yes, $\boldsymbol{M}\boldsymbol{X}=\boldsymbol{O}$ holds. $\boldsymbol{M}$ is no longer idempotent, but it satisfies $\boldsymbol{M}\boldsymbol{\Psi}\boldsymbol{M}=\boldsymbol{M}$. One way to show that it is nnd would be to use the first part of the question: for all $\boldsymbol{z}$, $\boldsymbol{z}^\top\boldsymbol{M}\boldsymbol{z}=(\boldsymbol{z}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\boldsymbol{z}-\boldsymbol{X}\hat{\boldsymbol{\beta}})$, and another way would be to use the second part of the question: $\boldsymbol{M}$ nnd because $\boldsymbol{M}\boldsymbol{\Psi}\boldsymbol{M}=\boldsymbol{M}$. To show expected value, show first that $\mathbf{y}^\top\boldsymbol{M}\mathbf{y}=\boldsymbol{\varepsilon}\boldsymbol{M}\boldsymbol{\varepsilon}$, and then use those tricks with the trace again. □

The simplest example of Generalized Least Squares is that where $\boldsymbol{\Psi}$ is diagonal (heteroskedastic data). In this case, the GLS objective function $(\mathbf{y}-\boldsymbol{X}\boldsymbol{\beta})^\top\boldsymbol{\Psi}^{-1}(\mathbf{y}-\boldsymbol{X}\boldsymbol{\beta})$ is simply a weighted least squares, with the weights being the inverses of the diagonal elements of $\boldsymbol{\Psi}$. This vector of inverse diagonal elements can be specified with the optional `weights` argument in R, see the help-file for `lm`. Heteroskedastic data arise for instance when each data point is an average over a different number of individuals.

If one runs OLS on the original instead of the transformed model, one gets an estimator, we will calle it here $\hat{\boldsymbol{\beta}}_{OLS}$, which is still unbiased. The estimator is usually also consistent, but no longer BLUE. This not only makes it less efficient than the GLS, but one also gets the wrong results if one relies on the standard computer printouts for significance tests etc. The estimate of σ^2 generated by this regression is now usually biased. How biased it is depends on the $\boldsymbol{X}$-matrix, but most often it seems biased upwards. The estimated standard errors in the regression printouts not only use the wrong s, but they also insert this wrong s into the wrong formula $\sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$ instead of $\sigma^2(\boldsymbol{X}\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}$ for $\mathcal{V}[\hat{\boldsymbol{\beta}}]$.

Problem 323. *In the generalized least squares model* $\mathbf{y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$ *with* $\boldsymbol{\varepsilon}\sim(\boldsymbol{o},\sigma^2\boldsymbol{\Psi})$, *the BLUE is*

$$\hat{\boldsymbol{\beta}}=(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\mathbf{y}. \tag{26.0.10}$$

We will write $\hat{\boldsymbol{\beta}}_{OLS}$ *for the ordinary least squares estimator*

$$\hat{\boldsymbol{\beta}}_{OLS}=(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\mathbf{y} \tag{26.0.11}$$

which has different properties now since we do not assume $\boldsymbol{\varepsilon}\sim(\boldsymbol{o},\sigma^2\boldsymbol{I})$ *but* $\boldsymbol{\varepsilon}\sim(\boldsymbol{o},\sigma^2\boldsymbol{\Psi})$.

- a. *1 point Is* $\hat{\boldsymbol{\beta}}_{OLS}$ *unbiased?*

- b. *2 points Show that, still under the assumption* $\boldsymbol{\varepsilon}\sim(\boldsymbol{o},\sigma^2\boldsymbol{\Psi})$, $\mathcal{V}[\hat{\boldsymbol{\beta}}_{OLS}]-\mathcal{V}[\hat{\boldsymbol{\beta}}]=\mathcal{V}[\hat{\boldsymbol{\beta}}_{OLS}-\hat{\boldsymbol{\beta}}]$. *(Write down the formulas for the left hand side and the right hand side and then show by matrix algebra that they are equal.) (This is what one should expect after Problem 198.) Since due to unbiasedness the covariance matrices are the* $\mathcal{MSE}$*-matrices, this shows that* $\mathcal{MSE}[\hat{\boldsymbol{\beta}}_{OLS};\boldsymbol{\beta}]-\mathcal{MSE}[\hat{\boldsymbol{\beta}};\boldsymbol{\beta}]$ *is nonnegative definite.*

Answer. Verify equality of the following two expressions for the differences in $\mathcal{MSE}$ matrices:

$$\mathcal{V}[\hat{\boldsymbol{\beta}}_{OLS}]-\mathcal{V}[\hat{\boldsymbol{\beta}}]=\sigma^2\left((\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Psi}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}-(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\right)=$$
$$=\sigma^2\left((\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top-(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\right)\boldsymbol{\Psi}\left(\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}-\boldsymbol{\Psi}^{-1}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\right)$$

□

Examples of GLS models are discussed in chapters 57 and 58.

CHAPTER 27

Best Linear Prediction

Best Linear Prediction is the second basic building block for the linear model, in addition to the OLS model. Instead of estimating a nonrandom parameter $\boldsymbol{\beta}$ about which no prior information is available, in the present situation one predicts a random variable $\boldsymbol{z}$ whose mean and covariance matrix are known. Most models to be discussed below are somewhere between these two extremes.

Christensen's [**Chr87**] is one of the few textbooks which treat best linear prediction on the basis of known first and second moments in parallel with the regression model. The two models have indeed so much in common that they should be treated together.

27.1. Minimum Mean Squared Error, Unbiasedness Not Required

Assume the expected values of the random vectors $\mathbf{y}$ and $\boldsymbol{z}$ are known, and their joint covariance matrix is known up to an unknown scalar factor $\sigma^2 > 0$. We will write this as

$$\begin{bmatrix}\mathbf{y}\\ \boldsymbol{z}\end{bmatrix} \sim \begin{bmatrix}\boldsymbol{\mu}\\ \boldsymbol{\nu}\end{bmatrix}, \sigma^2\begin{bmatrix}\boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}} & \boldsymbol{\Omega}_{\mathbf{y}z}\\ \boldsymbol{\Omega}_{z\mathbf{y}} & \boldsymbol{\Omega}_{zz}\end{bmatrix}, \qquad \sigma^2 > 0. \tag{27.1.1}$$

$\mathbf{y}$ is observed but $\boldsymbol{z}$ is not, and the goal is to predict $\boldsymbol{z}$ on the basis of the observation of $\mathbf{y}$.

There is a unique predictor of the form $\boldsymbol{z}^* = \boldsymbol{B}^*\mathbf{y}+\boldsymbol{b}^*$ (i.e., it is linear with a constant term, the technical term for this is "affine") with the following two properties: it is unbiased, and the prediction error is uncorrelated with $\mathbf{y}$, i.e.,

$$\mathcal{C}[\boldsymbol{z}^* - \boldsymbol{z}, \mathbf{y}] = \boldsymbol{O}. \tag{27.1.2}$$

The formulas for $\boldsymbol{B}^*$ and $\boldsymbol{b}^*$ are easily derived. Unbiasedness means $\boldsymbol{\nu} = \boldsymbol{B}^*\boldsymbol{\mu} + \boldsymbol{b}^*$, the predictor has therefore the form

$$\boldsymbol{z}^* = \boldsymbol{\nu} + \boldsymbol{B}^*(\mathbf{y} - \boldsymbol{\mu}). \tag{27.1.3}$$

Since

$$\boldsymbol{z}^* - \boldsymbol{z} = \boldsymbol{B}^*(\mathbf{y} - \boldsymbol{\mu}) - (\boldsymbol{z} - \boldsymbol{\nu}) = \begin{bmatrix}\boldsymbol{B}^* & -\boldsymbol{I}\end{bmatrix}\begin{bmatrix}\mathbf{y} - \boldsymbol{\mu}\\ \boldsymbol{z} - \boldsymbol{\nu}\end{bmatrix}, \tag{27.1.4}$$

the zero correlation condition (27.1.2) translates into

$$\boldsymbol{B}^*\boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}} = \boldsymbol{\Omega}_{z\mathbf{y}}, \tag{27.1.5}$$

which, due to equation (A.5.13) holds for $\boldsymbol{B}^* = \boldsymbol{\Omega}_{z\mathbf{y}}\boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}}^-$. Therefore the predictor

$$\boldsymbol{z}^* = \boldsymbol{\nu} + \boldsymbol{\Omega}_{z\mathbf{y}}\boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}}^-(\mathbf{y} - \boldsymbol{\mu}) \tag{27.1.6}$$

satisfies the two requirements.

Unbiasedness and condition (27.1.2) are sometimes interpreted to mean that $\boldsymbol{z}^*$ is an optimal predictor. Unbiasedness is often naively (but erroneously) considered to be a necessary condition for good estimators. And if the prediction error were correlated with the observed variable, the argument goes, then it would be possible to improve the prediction. Theorem 27.1.1 shows that despite the flaws in the argument, the result which it purports to show is indeed valid: $\boldsymbol{z}^*$ has the minimum $\mathcal{MSE}$ of all affine predictors, whether biased or not, of $\boldsymbol{z}$ on the basis of $\mathbf{y}$.

THEOREM 27.1.1. *In situation* (27.1.1), *the predictor* (27.1.6) *has, among all predictors of* $\boldsymbol{z}$ *which are affine functions of* $\mathbf{y}$, *the smallest* $\mathcal{MSE}$ *matrix. Its* $\mathcal{MSE}$ *matrix is*

$$\mathcal{MSE}[\boldsymbol{z}^*; \boldsymbol{z}] = \mathcal{E}[(\boldsymbol{z}^* - \boldsymbol{z})(\boldsymbol{z}^* - \boldsymbol{z})^\top] = \sigma^2(\boldsymbol{\Omega}_{zz} - \boldsymbol{\Omega}_{z\mathbf{y}}\boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}}^-\boldsymbol{\Omega}_{\mathbf{y}z}) = \sigma^2\boldsymbol{\Omega}_{zz.\mathbf{y}}. \tag{27.1.7}$$

PROOF. Look at any predictor of the form $\tilde{\boldsymbol{z}} = \tilde{\boldsymbol{B}}\mathbf{y}+\tilde{\boldsymbol{b}}$. Its bias is $\tilde{\boldsymbol{d}} = \mathcal{E}[\tilde{\boldsymbol{z}}-\boldsymbol{z}] = \tilde{\boldsymbol{B}}\boldsymbol{\mu} + \tilde{\boldsymbol{b}} - \boldsymbol{\nu}$, and by (23.1.2) one can write

$$\mathcal{E}[(\tilde{\boldsymbol{z}} - \boldsymbol{z})(\tilde{\boldsymbol{z}} - \boldsymbol{z})^\top] = \mathcal{V}[(\tilde{\boldsymbol{z}} - \boldsymbol{z})] + \tilde{\boldsymbol{d}}\tilde{\boldsymbol{d}}^\top \tag{27.1.8}$$

$$= \mathcal{V}\left[\begin{bmatrix}\tilde{\boldsymbol{B}} & -\boldsymbol{I}\end{bmatrix}\begin{bmatrix}\mathbf{y}\\ \boldsymbol{z}\end{bmatrix}\right] + \tilde{\boldsymbol{d}}\tilde{\boldsymbol{d}}^\top \tag{27.1.9}$$

$$= \sigma^2\begin{bmatrix}\tilde{\boldsymbol{B}} & -\boldsymbol{I}\end{bmatrix}\begin{bmatrix}\boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}} & \boldsymbol{\Omega}_{\mathbf{y}z}\\ \boldsymbol{\Omega}_{z\mathbf{y}} & \boldsymbol{\Omega}_{zz}\end{bmatrix}\begin{bmatrix}\tilde{\boldsymbol{B}}^\top\\ -\boldsymbol{I}\end{bmatrix} + \tilde{\boldsymbol{d}}\tilde{\boldsymbol{d}}^\top. \tag{27.1.10}$$

This $\mathcal{MSE}$-matrix is minimized if and only if $\boldsymbol{d}^* = \boldsymbol{o}$ and $\boldsymbol{B}^*$ satisfies (27.1.5). To see this, take any solution $\boldsymbol{B}^*$ of (27.1.5), and write $\tilde{\boldsymbol{B}} = \boldsymbol{B}^* + \tilde{\boldsymbol{D}}$. Since, due to theorem A.5.11, $\boldsymbol{\Omega}_{z\mathbf{y}} = \boldsymbol{\Omega}_{z\mathbf{y}}\boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}}^-\boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}}$, it follows $\boldsymbol{\Omega}_{z\mathbf{y}}\boldsymbol{B}^{*\top} = \boldsymbol{\Omega}_{z\mathbf{y}}\boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}}^-\boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}}\boldsymbol{B}^{*\top} = \boldsymbol{\Omega}_{z\mathbf{y}}\boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}}^-\boldsymbol{\Omega}_{\mathbf{y}z}$. Therefore

$$\mathcal{MSE}[\tilde{\boldsymbol{z}}; \boldsymbol{z}] = \sigma^2\begin{bmatrix}\boldsymbol{B}^* + \tilde{\boldsymbol{D}} & -\boldsymbol{I}\end{bmatrix}\begin{bmatrix}\boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}} & \boldsymbol{\Omega}_{\mathbf{y}z}\\ \boldsymbol{\Omega}_{z\mathbf{y}} & \boldsymbol{\Omega}_{zz}\end{bmatrix}\begin{bmatrix}\boldsymbol{B}^{*\top} + \tilde{\boldsymbol{D}}^\top\\ -\boldsymbol{I}\end{bmatrix} + \tilde{\boldsymbol{d}}\tilde{\boldsymbol{d}}^\top$$

$$= \sigma^2\begin{bmatrix}\boldsymbol{B}^* + \tilde{\boldsymbol{D}} & -\boldsymbol{I}\end{bmatrix}\begin{bmatrix}\boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}}\tilde{\boldsymbol{D}}^\top\\ -\boldsymbol{\Omega}_{zz.\mathbf{y}} + \boldsymbol{\Omega}_{z\mathbf{y}}\tilde{\boldsymbol{D}}^\top\end{bmatrix} + \tilde{\boldsymbol{d}}\tilde{\boldsymbol{d}}^\top \tag{27.1.11}$$

$$= \sigma^2(\boldsymbol{\Omega}_{zz.\mathbf{y}} + \tilde{\boldsymbol{D}}\boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}}\tilde{\boldsymbol{D}}^\top) + \tilde{\boldsymbol{d}}\tilde{\boldsymbol{d}}^\top. \tag{27.1.12}$$

The $\mathcal{MSE}$ matrix is therefore minimized (with minimum value $\sigma^2\boldsymbol{\Omega}_{\mathbf{z}\mathbf{z}.\mathbf{y}}$) if and only if $\tilde{\boldsymbol{d}} = \boldsymbol{o}$ and $\tilde{\boldsymbol{D}}\boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}} = \boldsymbol{O}$ which means that $\tilde{\boldsymbol{B}}$, along with $\boldsymbol{B}^*$, satisfies (27.1.5). □

PROBLEM 324. *Show that the solution of this minimum* MSE *problem is unique in the following sense: if $\boldsymbol{B}_1^*$ and $\boldsymbol{B}_2^*$ are two different solutions of* (27.1.5) *and $\boldsymbol{y}$ is any feasible observed value $\boldsymbol{y}$, plugged into equations* (27.1.3) *they will lead to the same predicted value $\boldsymbol{z}^*$.*

ANSWER. Comes from the fact that every feasible observed value of $\boldsymbol{y}$ can be written in the form $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}}\boldsymbol{q}$ for some $\boldsymbol{q}$, therefore $\boldsymbol{B}_i^*\boldsymbol{y} = \boldsymbol{B}_i^*\boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}}\boldsymbol{q} = \boldsymbol{\Omega}_{\mathbf{z}\mathbf{y}}\boldsymbol{q}$. □

The matrix $\boldsymbol{B}^*$ is also called the regression matrix of $\boldsymbol{z}$ on $\mathbf{y}$, and the unscaled covariance matrix has the form

$$\boldsymbol{\Omega} = \begin{bmatrix} \boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}} & \boldsymbol{\Omega}_{\mathbf{y}\mathbf{z}} \\ \boldsymbol{\Omega}_{\mathbf{z}\mathbf{y}} & \boldsymbol{\Omega}_{\mathbf{z}\mathbf{z}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}} & \boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}}\boldsymbol{X}^\top \\ \boldsymbol{X}\boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}} & \boldsymbol{X}\boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}}\boldsymbol{X}^\top + \boldsymbol{\Omega}_{\mathbf{z}\mathbf{z}.\mathbf{y}} \end{bmatrix} \tag{27.1.13}$$

Where we wrote here $\boldsymbol{B}^* = \boldsymbol{X}$ in order to make the analogy with regression clearer. A g-inverse is

$$\boldsymbol{\Omega}^- = \begin{bmatrix} \boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}}^- + \boldsymbol{X}^\top\boldsymbol{\Omega}_{\mathbf{z}\mathbf{z}.\mathbf{y}}^-\boldsymbol{X} & -\boldsymbol{X}^\top\boldsymbol{\Omega}_{\mathbf{z}\mathbf{z}.\mathbf{y}}^- \\ -\boldsymbol{X}^\top\boldsymbol{\Omega}_{\mathbf{z}\mathbf{z}.\mathbf{y}}^- & \boldsymbol{\Omega}_{\mathbf{z}\mathbf{z}.\mathbf{y}}^- \end{bmatrix} \tag{27.1.14}$$

and every g-inverse of the covariance matrix has a g-inverse of $\boldsymbol{\Omega}_{\mathbf{z}\mathbf{z}.\mathbf{y}}$ as its $\mathbf{z}\mathbf{z}$-partition. (Proof in Problem 592.)

If $\boldsymbol{\Omega} = \begin{bmatrix} \boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}} & \boldsymbol{\Omega}_{\mathbf{y}\mathbf{z}} \\ \boldsymbol{\Omega}_{\mathbf{z}\mathbf{y}} & \boldsymbol{\Omega}_{\mathbf{z}\mathbf{z}} \end{bmatrix}$ is nonsingular, 27.1.5 is also solved by $\boldsymbol{B}^* = -(\boldsymbol{\Omega}^{\mathbf{z}\mathbf{z}})^-\boldsymbol{\Omega}^{\mathbf{z}\mathbf{y}}$ where $\boldsymbol{\Omega}^{\mathbf{z}\mathbf{z}}$ and $\boldsymbol{\Omega}^{\mathbf{z}\mathbf{y}}$ are the corresponding partitions of the inverse $\boldsymbol{\Omega}^{-1}$. See Problem 592 for a proof. Therefore instead of 27.1.6 the predictor can also be written

$$\boldsymbol{z}^* = \boldsymbol{\nu} - \left(\boldsymbol{\Omega}^{\mathbf{z}\mathbf{z}}\right)^{-1}\boldsymbol{\Omega}^{\mathbf{z}\mathbf{y}}(\mathbf{y} - \boldsymbol{\mu}) \tag{27.1.15}$$

(note the minus sign) or

$$\boldsymbol{z}^* = \boldsymbol{\nu} - \boldsymbol{\Omega}_{\mathbf{z}\mathbf{z}.\mathbf{y}}\boldsymbol{\Omega}^{\mathbf{z}\mathbf{y}}(\mathbf{y} - \boldsymbol{\mu}). \tag{27.1.16}$$

PROBLEM 325. *This problem utilizes the concept of a bounded risk estimator, which is not yet explained very well in these notes. Assume $\mathbf{y}$, $\boldsymbol{z}$, $\boldsymbol{\mu}$, and $\boldsymbol{\nu}$ are jointly distributed random vectors. First assume $\boldsymbol{\nu}$ and $\boldsymbol{\mu}$ are observed, but $\mathbf{y}$ and $\boldsymbol{z}$ are not. Assume we know that in this case, the best linear bounded $\mathcal{MSE}$ predictor of $\mathbf{y}$ and $\boldsymbol{z}$ is $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$, with prediction errors distributed as follows:*

$$\begin{bmatrix} \mathbf{y} - \boldsymbol{\mu} \\ \boldsymbol{z} - \boldsymbol{\nu} \end{bmatrix} \sim \begin{bmatrix} \boldsymbol{o} \\ \boldsymbol{o} \end{bmatrix}, \sigma^2 \begin{bmatrix} \boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}} & \boldsymbol{\Omega}_{\mathbf{y}\mathbf{z}} \\ \boldsymbol{\Omega}_{\mathbf{z}\mathbf{y}} & \boldsymbol{\Omega}_{\mathbf{z}\mathbf{z}} \end{bmatrix}. \tag{27.1.17}$$

This is the initial information. Here it is unnecessary to specify the unconditional distributions of $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$, i.e., $\mathcal{E}[\boldsymbol{\mu}]$ and $\mathcal{E}[\boldsymbol{\nu}]$ as well as the joint covariance matrix of $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ are not needed, even if they are known.

Then in a second step assume that an observation of $\mathbf{y}$ becomes available, i.e., now $\mathbf{y}$, $\boldsymbol{\nu}$, and $\boldsymbol{\mu}$ are observed, but $\boldsymbol{z}$ still isn't. Then the predictor

$$\boldsymbol{z}^* = \boldsymbol{\nu} + \boldsymbol{\Omega}_{\mathbf{z}\mathbf{y}}\boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}}^-(\mathbf{y} - \boldsymbol{\mu}) \tag{27.1.18}$$

is the best linear bounded $\mathcal{MSE}$ predictor of $\boldsymbol{z}$ based on $\mathbf{y}$, $\boldsymbol{\mu}$, and $\boldsymbol{\nu}$.

• a. *Give special cases of this specification in which $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ are constant and $\mathbf{y}$ and $\boldsymbol{z}$ random, and one in which $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ and $\mathbf{y}$ are random and $\boldsymbol{z}$ is constant, and one in which $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ are random and $\mathbf{y}$ and $\boldsymbol{z}$ are constant.*

ANSWER. If $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ are constant, they are written $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$. From this follows $\boldsymbol{\mu} = \mathcal{E}[\mathbf{y}]$ and $\boldsymbol{\nu} = \mathcal{E}[\boldsymbol{z}]$ and $\sigma^2 \begin{bmatrix} \boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}} & \boldsymbol{\Omega}_{\mathbf{y}\boldsymbol{z}} \\ \boldsymbol{\Omega}_{\boldsymbol{z}\mathbf{y}} & \boldsymbol{\Omega}_{\boldsymbol{z}\boldsymbol{z}} \end{bmatrix} = \mathcal{V}[\begin{bmatrix} \mathbf{y} \\ rx \end{bmatrix}]$ and every linear predictor has bounded $\mathcal{MSE}$. Then the proof is as given earlier in this chapter. But an example in which $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ are not known constants but are observed random variables, and $\mathbf{y}$ is also a random variable but $\boldsymbol{z}$ is constant, is (28.0.26). Another example, in which $\mathbf{y}$ and $\boldsymbol{z}$ both are constants and $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ random, is constrained least squares (29.4.3). □

• b. *Prove equation 27.1.18.*

ANSWER. In this proof we allow all four $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ and $\mathbf{y}$ and $\boldsymbol{z}$ to be random. A linear predictor based on $\mathbf{y}$, $\boldsymbol{\mu}$, and $\boldsymbol{\nu}$ can be written as $\tilde{\boldsymbol{z}} = \boldsymbol{B}\mathbf{y} + \boldsymbol{C}\boldsymbol{\mu} + \boldsymbol{D}\boldsymbol{\nu} + \boldsymbol{d}$, therefore $\tilde{\boldsymbol{z}} - \boldsymbol{z} = \boldsymbol{B}(\mathbf{y} - \boldsymbol{\mu}) + (\boldsymbol{C} + \boldsymbol{B})\boldsymbol{\mu} + (\boldsymbol{D} - \boldsymbol{I})\boldsymbol{\nu} - (\boldsymbol{z} - \boldsymbol{\nu}) + \boldsymbol{d}$. $\mathcal{E}[\tilde{\boldsymbol{z}} - \boldsymbol{z}] = \boldsymbol{o} + (\boldsymbol{C} + \boldsymbol{B})\,\mathcal{E}[\boldsymbol{\mu}] + (\boldsymbol{D} - \boldsymbol{I})\,\mathcal{E}[\boldsymbol{\nu}] - \boldsymbol{o} + \boldsymbol{d}$. Assuming that $\mathcal{E}[\boldsymbol{\mu}]$ and $\mathcal{E}[\boldsymbol{\nu}]$ can be anything, the requirement of bounded $\mathcal{MSE}$ (or simply the requirement of unbiasedness, but this is not as elegant) gives $\boldsymbol{C} = -\boldsymbol{B}$ and $\boldsymbol{D} = \boldsymbol{I}$, therefore $\tilde{\boldsymbol{z}} = \boldsymbol{\nu} + \boldsymbol{B}(\mathbf{y} - \boldsymbol{\mu}) + \boldsymbol{d}$, and the estimation error is $\tilde{\boldsymbol{z}} - \boldsymbol{z} = \boldsymbol{B}(\mathbf{y} - \boldsymbol{\mu}) - (\boldsymbol{z} - \boldsymbol{\nu}) + \boldsymbol{d}$. Now continue as in the proof of theorem 27.1.1. I must still carry out this proof much more carefully! □

PROBLEM 326. *4 points According to* (27.1.2)*, the prediction error $\boldsymbol{z}^* - \boldsymbol{z}$ is uncorrelated with $\mathbf{y}$. If the distribution is such that the prediction error is even* independent *of $\mathbf{y}$ (as is the case if $\mathbf{y}$ and $\boldsymbol{z}$ are jointly normal), then $\boldsymbol{z}^*$ as defined in* (27.1.6) *is the conditional mean $\boldsymbol{z}^* = \mathcal{E}[\boldsymbol{z}|\mathbf{y}]$, and its $\mathcal{MSE}$-matrix as defined in* (27.1.7) *is the conditional variance $\mathcal{V}[\boldsymbol{z}|\mathbf{y}]$.*

ANSWER. From independence follows $\mathcal{E}[\boldsymbol{z}^* - \boldsymbol{z}|\mathbf{y}] = \mathcal{E}[\boldsymbol{z}^* - \boldsymbol{z}]$, and by the law of iterated expectations $\mathcal{E}[\boldsymbol{z}^* - \boldsymbol{z}] = \boldsymbol{o}$. Rewrite this as $\mathcal{E}[\boldsymbol{z}|\mathbf{y}] = \mathcal{E}[\boldsymbol{z}^*|\mathbf{y}]$. But since $\boldsymbol{z}^*$ is a function of $\mathbf{y}$, $\mathcal{E}[\boldsymbol{z}^*|\mathbf{y}] = \boldsymbol{z}^*$. Now the proof that the conditional dispersion matrix is the $\mathcal{MSE}$ matrix:

$$\begin{aligned} \mathcal{V}[\boldsymbol{z}|\mathbf{y}] &= \mathcal{E}[(\boldsymbol{z} - \mathcal{E}[\boldsymbol{z}|\mathbf{y}])(\boldsymbol{z} - \mathcal{E}[\boldsymbol{z}|\mathbf{y}])^\top|\mathbf{y}] = \mathcal{E}[(\boldsymbol{z} - \boldsymbol{z}^*)(\boldsymbol{z} - \boldsymbol{z}^*)^\top|\mathbf{y}] \\ &= \mathcal{E}[(\boldsymbol{z} - \boldsymbol{z}^*)(\boldsymbol{z} - \boldsymbol{z}^*)^\top] = \mathcal{MSE}[\boldsymbol{z}^*; \boldsymbol{z}]. \end{aligned} \tag{27.1.19}$$

□

PROBLEM 327. *Assume the expected values of $\mathbf{x}$, $\mathbf{y}$ and $\boldsymbol{z}$ are known, and their joint covariance matrix is known up to an unknown scalar factor $\sigma^2 > 0$.*

$$\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \\ \boldsymbol{z} \end{bmatrix} \sim \begin{bmatrix} \boldsymbol{\lambda} \\ \boldsymbol{\mu} \\ \boldsymbol{\nu} \end{bmatrix}, \sigma^2 \begin{bmatrix} \boldsymbol{\Omega}_{\mathbf{x}\mathbf{x}} & \boldsymbol{\Omega}_{\mathbf{x}\mathbf{y}} & \boldsymbol{\Omega}_{\mathbf{x}\boldsymbol{z}} \\ \boldsymbol{\Omega}_{\mathbf{x}\mathbf{y}}^\top & \boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}} & \boldsymbol{\Omega}_{\mathbf{y}\boldsymbol{z}} \\ \boldsymbol{\Omega}_{\mathbf{x}\boldsymbol{z}}^\top & \boldsymbol{\Omega}_{\mathbf{y}\boldsymbol{z}}^\top & \boldsymbol{\Omega}_{\boldsymbol{z}\boldsymbol{z}} \end{bmatrix}. \tag{27.1.20}$$

$\mathbf{x}$ *is the original information,* $\mathbf{y}$ *is additional information which becomes available, and* $\mathbf{z}$ *is the variable which we want to predict on the basis of this information.*

• a. *2 points Show that* $\mathbf{y}^* = \boldsymbol{\mu} + \boldsymbol{\Omega}_{\mathbf{xy}}^\top \boldsymbol{\Omega}_{\mathbf{xx}}^- (\mathbf{x} - \boldsymbol{\lambda})$ *is the best linear predictor of* $\mathbf{y}$ *and* $\mathbf{z}^* = \boldsymbol{\nu} + \boldsymbol{\Omega}_{\mathbf{xz}}^\top \boldsymbol{\Omega}_{\mathbf{xx}}^- (\mathbf{x} - \boldsymbol{\lambda})$ *the best linear predictor of* $\mathbf{z}$ *on the basis of the observation of* $\mathbf{x}$*, and that their joint* $\mathcal{MSE}$*-matrix is*

$$\mathcal{E}\left[\begin{bmatrix}\mathbf{y}^* - \mathbf{y} \\ \mathbf{z}^* - \mathbf{z}\end{bmatrix}\left[(\mathbf{y}^* - \mathbf{y})^\top \quad (\mathbf{z}^* - \mathbf{z})^\top\right]\right] = \sigma^2 \begin{bmatrix}\boldsymbol{\Omega}_{\mathbf{yy}} - \boldsymbol{\Omega}_{\mathbf{xy}}^\top \boldsymbol{\Omega}_{\mathbf{xx}}^- \boldsymbol{\Omega}_{\mathbf{xy}} & \boldsymbol{\Omega}_{\mathbf{yz}} - \boldsymbol{\Omega}_{\mathbf{xy}}^\top \boldsymbol{\Omega}_{\mathbf{xx}}^- \boldsymbol{\Omega}_{\mathbf{xz}} \\ \boldsymbol{\Omega}_{\mathbf{yz}}^\top - \boldsymbol{\Omega}_{\mathbf{xz}}^\top \boldsymbol{\Omega}_{\mathbf{xx}}^- \boldsymbol{\Omega}_{\mathbf{xy}} & \boldsymbol{\Omega}_{\mathbf{zz}} - \boldsymbol{\Omega}_{\mathbf{xz}}^\top \boldsymbol{\Omega}_{\mathbf{xx}}^- \boldsymbol{\Omega}_{\mathbf{xz}}\end{bmatrix}$$

which can also be written

$$= \sigma^2 \begin{bmatrix}\boldsymbol{\Omega}_{\mathbf{yy.x}} & \boldsymbol{\Omega}_{\mathbf{yz.x}} \\ \boldsymbol{\Omega}_{\mathbf{yz.x}}^\top & \boldsymbol{\Omega}_{\mathbf{zz.x}}\end{bmatrix}.$$

Answer. This part of the question is a simple application of the formulas derived earlier. For the $\mathcal{MSE}$-matrix you first get

$$\sigma^2\left(\begin{bmatrix}\boldsymbol{\Omega}_{\mathbf{yy}} & \boldsymbol{\Omega}_{\mathbf{yz}} \\ \boldsymbol{\Omega}_{\mathbf{yz}}^\top & \boldsymbol{\Omega}_{\mathbf{zz}}\end{bmatrix} - \begin{bmatrix}\boldsymbol{\Omega}_{\mathbf{xy}}^\top \\ \boldsymbol{\Omega}_{\mathbf{xz}}^\top\end{bmatrix} \boldsymbol{\Omega}_{\mathbf{xx}}^- \begin{bmatrix}\boldsymbol{\Omega}_{\mathbf{xy}} & \boldsymbol{\Omega}_{\mathbf{xz}}\end{bmatrix}\right)$$

□

• b. *5 points Show that the best linear predictor of* $\mathbf{z}$ *on the basis of the observations of* $\mathbf{x}$ and $\mathbf{y}$ *has the form*

$$\mathbf{z}^{**} = \mathbf{z}^* + \boldsymbol{\Omega}_{\mathbf{yz.x}}^\top \boldsymbol{\Omega}_{\mathbf{yy.x}}^- (\mathbf{y} - \mathbf{y}^*) \tag{27.1.21}$$

This is an important formula. All you need to compute $\mathbf{z}^{**}$ *is the best estimate* $\mathbf{z}^*$ *before the new information* $\mathbf{y}$ *became available, the best estimate* $\mathbf{y}^*$ *of that new information itself, and the joint* $\mathcal{MSE}$ *matrix of the two. The original data* $\mathbf{x}$ *and the covariance matrix* (27.1.20) *do not enter this formula.*

Answer. Follows from

$$\mathbf{z}^{**} = \boldsymbol{\nu} + \begin{bmatrix}\boldsymbol{\Omega}_{\mathbf{xz}}^\top & \boldsymbol{\Omega}_{\mathbf{yz}}^\top\end{bmatrix} \begin{bmatrix}\boldsymbol{\Omega}_{\mathbf{xx}} & \boldsymbol{\Omega}_{\mathbf{xy}} \\ \boldsymbol{\Omega}_{\mathbf{xy}}^\top & \boldsymbol{\Omega}_{\mathbf{yy}}\end{bmatrix}^- \begin{bmatrix}\mathbf{x} - \boldsymbol{\lambda} \\ \mathbf{y} - \boldsymbol{\mu}\end{bmatrix} =$$

Now apply (A.8.2):

$$= \boldsymbol{\nu} + \begin{bmatrix}\boldsymbol{\Omega}_{\mathbf{xz}}^\top & \boldsymbol{\Omega}_{\mathbf{yz}}^\top\end{bmatrix} \begin{bmatrix}\boldsymbol{\Omega}_{\mathbf{xx}}^- + \boldsymbol{\Omega}_{\mathbf{xx}}^- \boldsymbol{\Omega}_{\mathbf{xy}} \boldsymbol{\Omega}_{\mathbf{yy.x}}^- \boldsymbol{\Omega}_{\mathbf{xy}}^\top \boldsymbol{\Omega}_{\mathbf{xx}}^- & -\boldsymbol{\Omega}_{\mathbf{xx}}^- \boldsymbol{\Omega}_{\mathbf{xy}} \boldsymbol{\Omega}_{\mathbf{yy.x}}^- \\ -\boldsymbol{\Omega}_{\mathbf{yy.x}}^- \boldsymbol{\Omega}_{\mathbf{xy}}^\top \boldsymbol{\Omega}_{\mathbf{xx}}^- & \boldsymbol{\Omega}_{\mathbf{yy.x}}^-\end{bmatrix} \begin{bmatrix}\mathbf{x} - \boldsymbol{\lambda} \\ \mathbf{y} - \boldsymbol{\mu}\end{bmatrix} =$$

$$= \boldsymbol{\nu} + \begin{bmatrix}\boldsymbol{\Omega}_{\mathbf{xz}}^\top & \boldsymbol{\Omega}_{\mathbf{yz}}^\top\end{bmatrix} \begin{bmatrix}\boldsymbol{\Omega}_{\mathbf{xx}}^-(\mathbf{x} - \boldsymbol{\lambda}) + \boldsymbol{\Omega}_{\mathbf{xx}}^- \boldsymbol{\Omega}_{\mathbf{xy}} \boldsymbol{\Omega}_{\mathbf{yy.x}}^-(\mathbf{y}^* - \boldsymbol{\mu}) - \boldsymbol{\Omega}_{\mathbf{xx}}^- \boldsymbol{\Omega}_{\mathbf{xy}} \boldsymbol{\Omega}_{\mathbf{yy.x}}^-(\mathbf{y} - \boldsymbol{\mu}) \\ -\boldsymbol{\Omega}_{\mathbf{yy.x}}^-(\mathbf{y}^* - \boldsymbol{\mu}) + \boldsymbol{\Omega}_{\mathbf{yy.x}}^-(\mathbf{y} - \boldsymbol{\mu})\end{bmatrix} =$$

$$= \boldsymbol{\nu} + \begin{bmatrix}\boldsymbol{\Omega}_{\mathbf{xz}}^\top & \boldsymbol{\Omega}_{\mathbf{yz}}^\top\end{bmatrix} \begin{bmatrix}\boldsymbol{\Omega}_{\mathbf{xx}}^-(\mathbf{x} - \boldsymbol{\lambda}) - \boldsymbol{\Omega}_{\mathbf{xx}}^- \boldsymbol{\Omega}_{\mathbf{xy}} \boldsymbol{\Omega}_{\mathbf{yy.x}}^-(\mathbf{y} - \mathbf{y}^*) \\ +\boldsymbol{\Omega}_{\mathbf{yy.x}}^-(\mathbf{y} - \mathbf{y}^*)\end{bmatrix} =$$

$$= \boldsymbol{\nu} + \boldsymbol{\Omega}_{\mathbf{xz}}^\top \boldsymbol{\Omega}_{\mathbf{xx}}^-(\mathbf{x} - \boldsymbol{\lambda}) - \boldsymbol{\Omega}_{\mathbf{xz}}^\top \boldsymbol{\Omega}_{\mathbf{xx}}^- \boldsymbol{\Omega}_{\mathbf{xy}} \boldsymbol{\Omega}_{\mathbf{yy.x}}^-(\mathbf{y} - \mathbf{y}^*) + \boldsymbol{\Omega}_{\mathbf{yz}}^\top \boldsymbol{\Omega}_{\mathbf{yy.x}}^-(\mathbf{y} - \mathbf{y}^*) =$$

$$= \mathbf{z}^* + \left(\boldsymbol{\Omega}_{\mathbf{yz}}^\top - \boldsymbol{\Omega}_{\mathbf{xz}}^\top \boldsymbol{\Omega}_{\mathbf{xx}}^- \boldsymbol{\Omega}_{\mathbf{xy}}\right) \boldsymbol{\Omega}_{\mathbf{yy.x}}^-(\mathbf{y} - \mathbf{y}^*) = \mathbf{z}^* + \boldsymbol{\Omega}_{\mathbf{yz.x}}^\top \boldsymbol{\Omega}_{\mathbf{yy.x}}^-(\mathbf{y} - \mathbf{y}^*)$$

□

Problem 328. *Assume* $\mathbf{x}$*,* $\mathbf{y}$*, and* $\mathbf{z}$ *have a joint probability distribution, and the conditional expectation* $\mathcal{E}[\mathbf{z}|\mathbf{x}, \mathbf{y}] = \boldsymbol{\alpha}^* + \boldsymbol{A}^*\mathbf{x} + \boldsymbol{B}^*\mathbf{y}$ *is linear in* $\mathbf{x}$ *and* $\mathbf{y}$*.*

• a. *1 point Show that* $\mathcal{E}[\mathbf{z}|\mathbf{x}] = \boldsymbol{\alpha}^* + \boldsymbol{A}^*\mathbf{x} + \boldsymbol{B}^*\mathcal{E}[\mathbf{y}|\mathbf{x}]$*. Hint: you may use the law of iterated expectations in the following form:* $\mathcal{E}[\mathbf{z}|\mathbf{x}] = \mathcal{E}\left[\mathcal{E}[\mathbf{z}|\mathbf{x}, \mathbf{y}]\big|\mathbf{x}\right]$*.*

Answer. With this hint it is trivial: $\mathcal{E}[\mathbf{z}|\mathbf{x}] = \mathcal{E}\left[\boldsymbol{\alpha}^* + \boldsymbol{A}^*\mathbf{x} + \boldsymbol{B}^*\mathbf{y}\big|\mathbf{x}\right] = \boldsymbol{\alpha}^* + \boldsymbol{A}^*\mathbf{x} + \boldsymbol{B}^*\mathcal{E}[\mathbf{y}|\mathbf{x}]$. □

• b. *1 point The next three examples are from* [**CW99**, pp. 264/5]*: Assume* $\mathrm{E}[\mathrm{z}|\mathrm{x}, \mathrm{y}] = 1 + 2\mathrm{x} + 3\mathrm{y}$*,* x *and* y *are independent, and* $\mathrm{E}[\mathrm{y}] = 2$*. Compute* $\mathrm{E}[\mathrm{z}|\mathrm{x}]$*.*

Answer. According to the formula, $\mathrm{E}[\mathrm{z}|\mathrm{x}] = 1+2\mathrm{x}+3\mathrm{E}[\mathrm{y}|\mathrm{x}]$, but since x and y are independent, $\mathrm{E}[\mathrm{y}|\mathrm{x}] = \mathrm{E}[\mathrm{y}] = 2$; therefore $\mathrm{E}[\mathrm{z}|\mathrm{x}] = 7+2\mathrm{x}$. I.e., the slope is the same, but the intercept changes. □

• c. *1 point Assume again* $\mathrm{E}[\mathrm{z}|\mathrm{x}, \mathrm{y}] = 1 + 2\mathrm{x} + 3\mathrm{y}$*, but this time* x *and* y *are not independent but* $\mathrm{E}[\mathrm{y}|\mathrm{x}] = 2 - \mathrm{x}$*. Compute* $\mathrm{E}[\mathrm{z}|\mathrm{x}]$*.*

Answer. $\mathrm{E}[\mathrm{z}|\mathrm{x}] = 1 + 2\mathrm{x} + 3(2 - \mathrm{x}) = 7 - \mathrm{x}$. In this situation, both slope and intercept change, but it is still a linear relationship. □

• d. *1 point Again* $\mathrm{E}[\mathrm{z}|\mathrm{x}, \mathrm{y}] = 1 + 2\mathrm{x} + 3\mathrm{y}$*, and this time the relationship between* x *and* y *is nonlinear:* $\mathrm{E}[\mathrm{y}|\mathrm{x}] = 2 - e^{\mathrm{x}}$*. Compute* $\mathrm{E}[\mathrm{z}|\mathrm{x}]$*.*

Answer. $\mathrm{E}[\mathrm{z}|\mathrm{x}] = 1 + 2\mathrm{x} + 3(2 - e^{\mathrm{x}}) = 7 + 2\mathrm{x} - 3e^{\mathrm{x}}$. This time the marginal relationship between x and y is no longer linear. This is so despite the fact that, if all the variables are included, i.e., if both x and y are included, then the relationship is linear. □

• e. *1 point Assume* $\mathrm{E}[f(\mathrm{z})|\mathrm{x}, \mathrm{y}] = 1 + 2\mathrm{x} + 3\mathrm{y}$*, where* f *is a nonlinear function, and* $\mathrm{E}[\mathrm{y}|\mathrm{x}] = 2 - \mathrm{x}$*. Compute* $\mathrm{E}[f(\mathrm{z})|\mathrm{x}]$*.*

Answer. $\mathrm{E}[f(\mathrm{z})|\mathrm{x}] = 1 + 2\mathrm{x} + 3(2 - \mathrm{x}) = 7 - \mathrm{x}$. If one plots z against x and z, then the plots should be similar, though not identical, since the same transformation f will straighten them out. This is why the plots in the top row or right column of [**CW99**, p. 435] are so similar. □

Connection between prediction and inverse prediction: If $\mathbf{y}$ is observed and $\mathbf{z}$ is to be predicted, the BLUP is $\mathbf{z}^* - \boldsymbol{\nu} = \boldsymbol{B}^*(\mathbf{y} - \boldsymbol{\mu})$ where $\boldsymbol{B}^* = \boldsymbol{\Omega}_{\mathbf{zy}} \boldsymbol{\Omega}_{\mathbf{yy}}^-$. If $\mathbf{z}$ is observed and $\mathbf{y}$ is to be predicted, then the BLUP is $\mathbf{y}^* - \boldsymbol{\mu} = \boldsymbol{C}^*(\mathbf{z} - \boldsymbol{\nu})$ with $\boldsymbol{C}^* = \boldsymbol{\Omega}_{\mathbf{yz}} \boldsymbol{\Omega}_{\mathbf{zz}}^-$. $\boldsymbol{B}^*$ and $\boldsymbol{C}^*$ are connected by the formula

$$\boldsymbol{\Omega}_{\mathbf{yy}} \boldsymbol{B}^{*\top} = \boldsymbol{C}^* \boldsymbol{\Omega}_{\mathbf{zz}}. \tag{27.1.22}$$

This relationship can be used for graphical regression methods [**Coo98**, pp. 187/8]: If z is a scalar, it is much easier to determine the elements of $\boldsymbol{C}^*$ than those of $\boldsymbol{B}^*$. $\boldsymbol{C}^*$ consists of the regression slopes in the scatter plot of each of the observed variables against z. They can be read off easily from a scatterplot matrix. This works not only if the distribution is Normal, but also with arbitrary distributions as long as all conditional expectations between the explanatory variables are linear.

PROBLEM 329. *In order to make relationship* (27.1.22) *more intuitive, assume* x *and* ε *are Normally distributed and independent of each other, and* $\mathrm{E}[\varepsilon] = 0$. *Define* $\mathsf{y} = \alpha + \beta\mathsf{x} + \varepsilon$.

• a. *Show that* $\alpha + \beta\mathsf{x}$ *is the best linear predictor of* y *based on the observation of* x.

ANSWER. Follows from the fact that the predictor is unbiased and the prediction error is uncorrelated with x. □

• b. *Express* β *in terms of the variances and covariances of* x *and* y.

ANSWER. $\operatorname{cov}[\mathsf{x},\mathsf{y}] = \beta\operatorname{var}[\mathsf{x}]$, therefore $\beta = \frac{\operatorname{cov}[\mathsf{x},\mathsf{y}]}{\operatorname{var}[\mathsf{x}]}$ □

• c. *Since* x *and* y *are jointly normal, they can also be written* $\mathsf{x} = \gamma + \delta\mathsf{y} + \omega$ *where* ω *is independent of* y. *Express* δ *in terms of the variances and covariances of* x *and* y, *and show that* $\operatorname{var}[\mathsf{y}]\beta = \gamma\operatorname{var}[\mathsf{x}]$.

ANSWER. $\delta = \frac{\operatorname{cov}[\mathsf{x},\mathsf{y}]}{\operatorname{var}[\mathsf{y}]}$. □

• d. *Now let us extend the model a little: assume* x_1, x_2, *and* ε *are Normally distributed and independent of each other, and* $\mathrm{E}[\varepsilon] = 0$. *Define* $\mathsf{y} = \alpha + \beta_1\mathsf{x}_1 + \beta_2\mathsf{x}_2 + \varepsilon$. *Again express* β_1 *and* β_2 *in terms of variances and covariances of* x_1, x_2, *and* y.

ANSWER. Since x_1 and x_2 are independent, one gets the same formulas as in the univariate case: from $\operatorname{cov}[\mathsf{x}_1,\mathsf{y}] = \beta_1\operatorname{var}[\mathsf{x}_1]$ and $\operatorname{cov}[\mathsf{x}_2,\mathsf{y}] = \beta_2\operatorname{var}[\mathsf{x}_2]$ follows $\beta_1 = \frac{\operatorname{cov}[\mathsf{x}_1,\mathsf{y}]}{\operatorname{var}[\mathsf{x}_1]}$ and $\beta_2 = \frac{\operatorname{cov}[\mathsf{x}_2,\mathsf{y}]}{\operatorname{var}[\mathsf{x}_2]}$. □

• e. *Since* x_1 *and* y *are jointly normal, they can also be written* $\mathsf{x}_1 = \gamma_1 + \delta_1\mathsf{y} + \omega_1$, *where* ω_1 *is independent of* y. *Likewise,* $\mathsf{x}_2 = \gamma_2 + \delta_2\mathsf{y} + \omega_2$, *where* ω_2 *is independent of* y. *Express* δ_1 *and* δ_2 *in terms of the variances and covariances of* x_1, x_2, *and* y, *and show that*

$$\begin{bmatrix}\delta_1\\ \delta_2\end{bmatrix}\operatorname{var}[\mathsf{y}] = \begin{bmatrix}\operatorname{var}[\mathsf{x}_1] & 0\\ 0 & \operatorname{var}[\mathsf{x}_2]\end{bmatrix}\begin{bmatrix}\beta_1\\ \beta_2\end{bmatrix} \tag{27.1.23}$$

This is (27.1.22) *in the present situation.*

ANSWER. $\delta_1 = \frac{\operatorname{cov}[\mathsf{x}_1,\mathsf{y}]}{\operatorname{var}[\mathsf{y}]}$ and $\delta_2 = \frac{\operatorname{cov}[\mathsf{x}_2,\mathsf{y}]}{\operatorname{var}[\mathsf{y}]}$. □

27.2. The Associated Least Squares Problem

For every estimation problem there is an associated "least squares" problem. In the present situation, $\boldsymbol{z}^*$ is that value which, together with the given observation $\boldsymbol{y}$, "blends best" into the population defined by $\boldsymbol{\mu}$, $\boldsymbol{\nu}$ and the dispersion matrix $\boldsymbol{\Omega}$, in the following sense: Given the observed value $\boldsymbol{y}$, the vector $\boldsymbol{z}^* = \boldsymbol{\nu} + \boldsymbol{\Omega}_{zy}\boldsymbol{\Omega}_{yy}^{-}(\boldsymbol{y}-\boldsymbol{\mu})$

is that value $\boldsymbol{z}$ for which $\begin{bmatrix}\boldsymbol{y}\\ \boldsymbol{z}\end{bmatrix}$ has smallest Mahalanobis distance from the population defined by the mean vector $\begin{bmatrix}\boldsymbol{\mu}\\ \boldsymbol{\nu}\end{bmatrix}$ and the covariance matrix $\sigma^2\begin{bmatrix}\boldsymbol{\Omega}_{yy} & \boldsymbol{\Omega}_{yz}\\ \boldsymbol{\Omega}_{zy} & \boldsymbol{\Omega}_{zz}\end{bmatrix}$.

In the case of singular $\boldsymbol{\Omega}_{zz}$, it is only necessary to minimize among those $\boldsymbol{z}$ which have finite distance from the population, i.e., which can be written in the form $\boldsymbol{z} = \boldsymbol{\nu} + \boldsymbol{\Omega}_{zz}\boldsymbol{q}$ for some $\boldsymbol{q}$. We will also write $r = \operatorname{rank}\begin{bmatrix}\boldsymbol{\Omega}_{yy} & \boldsymbol{\Omega}_{yz}\\ \boldsymbol{\Omega}_{zy} & \boldsymbol{\Omega}_{zz}\end{bmatrix}$. Therefore, $\boldsymbol{z}^*$ solves the following "least squares problem:"

$$\boxed{\boldsymbol{z} = \boldsymbol{z}^* \quad \text{min.}\ \frac{1}{r\sigma^2}\begin{bmatrix}\boldsymbol{y}-\boldsymbol{\mu}\\ \boldsymbol{z}-\boldsymbol{\nu}\end{bmatrix}^{\top}\begin{bmatrix}\boldsymbol{\Omega}_{yy} & \boldsymbol{\Omega}_{yz}\\ \boldsymbol{\Omega}_{zy} & \boldsymbol{\Omega}_{zz}\end{bmatrix}^{-}\begin{bmatrix}\boldsymbol{y}-\boldsymbol{\mu}\\ \boldsymbol{z}-\boldsymbol{\nu}\end{bmatrix}\ \text{s.t.}\ \boldsymbol{z} = \boldsymbol{\nu} + \boldsymbol{\Omega}_{zz}\boldsymbol{q}\ \text{for some}\ \boldsymbol{q}.} \tag{27.2.1}$$

To prove this, use (A.8.2) to invert the dispersion matrix:

$$\begin{bmatrix}\boldsymbol{\Omega}_{yy} & \boldsymbol{\Omega}_{yz}\\ \boldsymbol{\Omega}_{zy} & \boldsymbol{\Omega}_{zz}\end{bmatrix}^{-} = \begin{bmatrix}\boldsymbol{\Omega}_{yy}^{-} + \boldsymbol{\Omega}_{yy}^{-}\boldsymbol{\Omega}_{yz}\boldsymbol{\Omega}_{zz.y}^{-}\boldsymbol{\Omega}_{zy}\boldsymbol{\Omega}_{yy}^{-} & -\boldsymbol{\Omega}_{yy}^{-}\boldsymbol{\Omega}_{yz}\boldsymbol{\Omega}_{zz.y}^{-}\\ -\boldsymbol{\Omega}_{zz.y}^{-}\boldsymbol{\Omega}_{zy}\boldsymbol{\Omega}_{yy}^{-} & \boldsymbol{\Omega}_{zz.y}^{-}\end{bmatrix}. \tag{27.2.2}$$

If one plugs $\boldsymbol{z} = \boldsymbol{z}^*$ into this objective function, one obtains a very simple expression:

$$(\boldsymbol{y}-\boldsymbol{\mu})^{\top}\begin{bmatrix}\boldsymbol{I} & \boldsymbol{\Omega}_{yy}^{-}\boldsymbol{\Omega}_{yz}\end{bmatrix}\begin{bmatrix}\boldsymbol{\Omega}_{yy}^{-} + \boldsymbol{\Omega}_{yy}^{-}\boldsymbol{\Omega}_{yz}\boldsymbol{\Omega}_{zz.y}^{-}\boldsymbol{\Omega}_{zy}\boldsymbol{\Omega}_{yy}^{-} & -\boldsymbol{\Omega}_{yy}^{-}\boldsymbol{\Omega}_{yz}\boldsymbol{\Omega}_{zz.y}^{-}\\ -\boldsymbol{\Omega}_{zz.y}^{-}\boldsymbol{\Omega}_{zy}\boldsymbol{\Omega}_{yy}^{-} & \boldsymbol{\Omega}_{zz.y}^{-}\end{bmatrix}\begin{bmatrix}\boldsymbol{I}\\ \boldsymbol{\Omega}_{zy}\boldsymbol{\Omega}_{yy}^{-}\end{bmatrix}(\boldsymbol{y}-\boldsymbol{\mu}) = \tag{27.2.3}$$

$$= (\boldsymbol{y}-\boldsymbol{\mu})^{\top}\boldsymbol{\Omega}_{yy}^{-}(\boldsymbol{y}-\boldsymbol{\mu}). \tag{27.2.4}$$

Now take any $\boldsymbol{z}$ of the form $\boldsymbol{z} = \boldsymbol{\nu} + \boldsymbol{\Omega}_{zz}\boldsymbol{q}$ for some $\boldsymbol{q}$ and write it in the form $\boldsymbol{z} = \boldsymbol{z}^* + \boldsymbol{\Omega}_{zz}\boldsymbol{d}$, i.e.,

$$\begin{bmatrix}\boldsymbol{y}-\boldsymbol{\mu}\\ \boldsymbol{z}-\boldsymbol{\nu}\end{bmatrix} = \begin{bmatrix}\boldsymbol{y}-\boldsymbol{\mu}\\ \boldsymbol{z}^*-\boldsymbol{\nu}\end{bmatrix} + \begin{bmatrix}\boldsymbol{o}\\ \boldsymbol{\Omega}_{zz}\boldsymbol{d}\end{bmatrix}.$$

Then the cross product terms in the objective function disappear:

$$\begin{bmatrix}\boldsymbol{o}^{\top} & \boldsymbol{d}^{\top}\boldsymbol{\Omega}_{zz}\end{bmatrix}\begin{bmatrix}\boldsymbol{\Omega}_{yy}^{-} + \boldsymbol{\Omega}_{yy}^{-}\boldsymbol{\Omega}_{yz}\boldsymbol{\Omega}_{zz.y}^{-}\boldsymbol{\Omega}_{zy}\boldsymbol{\Omega}_{yy}^{-} & -\boldsymbol{\Omega}_{yy}^{-}\boldsymbol{\Omega}_{yz}\boldsymbol{\Omega}_{zz.y}^{-}\\ -\boldsymbol{\Omega}_{zz.y}^{-}\boldsymbol{\Omega}_{zy}\boldsymbol{\Omega}_{yy}^{-} & \boldsymbol{\Omega}_{zz.y}^{-}\end{bmatrix}\begin{bmatrix}\boldsymbol{I}\\ \boldsymbol{\Omega}_{zy}\boldsymbol{\Omega}_{yy}^{-}\end{bmatrix}(\boldsymbol{y}-\boldsymbol{\mu}) = \tag{27.2.5}$$

$$= \begin{bmatrix}\boldsymbol{o}^{\top} & \boldsymbol{d}^{\top}\boldsymbol{\Omega}_{zz}\end{bmatrix}\begin{bmatrix}\boldsymbol{\Omega}_{yy}^{-}\\ \boldsymbol{O}\end{bmatrix}(\boldsymbol{y}-\boldsymbol{\mu}) = 0$$

Therefore this gives a larger value of the objective function.

PROBLEM 330. *Use problem 579 for an alternative proof of this.*

From (27.2.1) follows that $\boldsymbol{z}^*$ is the mode of the normal density function, and since the mode is the mean, this is an alternative proof, in the case of nonsingular covariance matrix, when the density exists, that $\boldsymbol{z}^*$ is the normal conditional mean.

27.3. Prediction of Future Observations in the Regression Model

For a moment let us go back to the model $\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$ with spherically distributed disturbances $\boldsymbol{\varepsilon} \sim (\boldsymbol{o}, \sigma^2\boldsymbol{I})$. This time, our goal is not to estimate $\boldsymbol{\beta}$, but the situation is the following: For a new set of observations of the explanatory variables $\boldsymbol{X}_0$ the values of the dependent variable $\mathbf{y}_0 = \boldsymbol{X}_0\boldsymbol{\beta} + \boldsymbol{\varepsilon}_0$ have not yet been observed and we want to predict them. The obvious predictor is $\mathbf{y}_0^* = \boldsymbol{X}_0\hat{\boldsymbol{\beta}} = \boldsymbol{X}_0(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\mathbf{y}$.

Since

$$\mathbf{y}_0^* - \mathbf{y}_0 = \boldsymbol{X}_0(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\mathbf{y} - \mathbf{y}_0 = \tag{27.3.1}$$

$$= \boldsymbol{X}_0(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{X}_0(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon}-\boldsymbol{X}_0\boldsymbol{\beta}-\boldsymbol{\varepsilon}_0 = \boldsymbol{X}_0(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon}-\boldsymbol{\varepsilon}_0$$

one sees that $\mathrm{E}[\mathbf{y}_0^* - \mathbf{y}_0] = \boldsymbol{o}$, i.e., it is an unbiased predictor. And since $\boldsymbol{\varepsilon}$ and $\boldsymbol{\varepsilon}_0$ are uncorrelated, one obtains

$$\mathcal{MSE}[\mathbf{y}_0^*; \mathbf{y}_0] = \mathcal{V}[\mathbf{y}_0^* - \mathbf{y}_0] = \mathcal{V}[\boldsymbol{X}_0(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon}] + \mathcal{V}[\boldsymbol{\varepsilon}_0] \tag{27.3.2}$$

$$= \sigma^2(\boldsymbol{X}_0(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}_0^\top + \boldsymbol{I}). \tag{27.3.3}$$

Problem 331 shows that this is the Best Linear Unbiased Predictor (BLUP) of $\mathbf{y}_0$ on the basis of $\mathbf{y}$.

Problem 331. *The prediction problem in the Ordinary Least Squares model can be formulated as follows:*

$$\begin{bmatrix}\mathbf{y}\\ \mathbf{y}_0\end{bmatrix} = \begin{bmatrix}\boldsymbol{X}\\ \boldsymbol{X}_0\end{bmatrix}\boldsymbol{\beta} + \begin{bmatrix}\boldsymbol{\varepsilon}\\ \boldsymbol{\varepsilon}_0\end{bmatrix} \qquad \mathcal{E}\begin{bmatrix}\boldsymbol{\varepsilon}\\ \boldsymbol{\varepsilon}_0\end{bmatrix} = \begin{bmatrix}\boldsymbol{o}\\ \boldsymbol{o}\end{bmatrix} \qquad \mathcal{V}\begin{bmatrix}\boldsymbol{\varepsilon}\\ \boldsymbol{\varepsilon}_0\end{bmatrix} = \sigma^2\begin{bmatrix}\boldsymbol{I} & \boldsymbol{O}\\ \boldsymbol{O} & \boldsymbol{I}\end{bmatrix}. \tag{27.3.4}$$

$\boldsymbol{X}$ and $\boldsymbol{X}_0$ are known, $\mathbf{y}$ is observed, $\mathbf{y}_0$ is not observed.

• a. *4 points Show that $\mathbf{y}_0^* = \boldsymbol{X}_0\hat{\boldsymbol{\beta}}$ is the Best Linear Unbiased Predictor (BLUP) of $\mathbf{y}_0$ on the basis of $\mathbf{y}$, where $\hat{\boldsymbol{\beta}}$ is the OLS estimate in the model $\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.*

Answer. Take any other predictor $\tilde{\mathbf{y}}_0 = \tilde{\boldsymbol{B}}\mathbf{y}$ and write $\tilde{\boldsymbol{B}} = \boldsymbol{X}_0(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top + \boldsymbol{D}$. Unbiasedness means $\mathcal{E}[\tilde{\mathbf{y}}_0 - \mathbf{y}_0] = \boldsymbol{X}_0(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{D}\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}_0\boldsymbol{\beta} = \boldsymbol{o}$, from which follows $\boldsymbol{D}\boldsymbol{X} = \boldsymbol{O}$. Because of unbiasedness we know $\mathcal{MSE}[\tilde{\mathbf{y}}_0; \mathbf{y}_0] = \mathcal{V}[\tilde{\mathbf{y}}_0 - \mathbf{y}_0]$. Since the prediction error can be written $\tilde{\mathbf{y}}_0 - \mathbf{y} = \begin{bmatrix}\boldsymbol{X}_0(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top + \boldsymbol{D} & -\boldsymbol{I}\end{bmatrix}\begin{bmatrix}\mathbf{y}\\ \mathbf{y}_0\end{bmatrix}$, one obtains

$$\begin{aligned}
\mathcal{V}[\tilde{\mathbf{y}}_0 - \mathbf{y}_0] &= \begin{bmatrix}\boldsymbol{X}_0(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top + \boldsymbol{D} & -\boldsymbol{I}\end{bmatrix}\mathcal{V}\begin{bmatrix}\mathbf{y}\\ \mathbf{y}_0\end{bmatrix}\begin{bmatrix}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}_0^\top + \boldsymbol{D}^\top\\ -\boldsymbol{I}\end{bmatrix}\\
&= \sigma^2\begin{bmatrix}\boldsymbol{X}_0(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top + \boldsymbol{D} & -\boldsymbol{I}\end{bmatrix}\begin{bmatrix}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}_0^\top + \boldsymbol{D}^\top\\ -\boldsymbol{I}\end{bmatrix}\\
&= \sigma^2\big(\boldsymbol{X}_0(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top + \boldsymbol{D}\big)\big(\boldsymbol{X}_0(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top + \boldsymbol{D}\big)^\top + \sigma^2\boldsymbol{I}\\
&= \sigma^2\big(\boldsymbol{X}_0(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}_0^\top + \boldsymbol{D}\boldsymbol{D}^\top + \boldsymbol{I}\big).
\end{aligned}$$

This is smallest for $\boldsymbol{D} = \boldsymbol{O}$. □

• b. *2 points From our formulation of the Gauss-Markov theorem in Theorem 24.1.1 it is obvious that the same $\mathbf{y}_0^* = \boldsymbol{X}_0\hat{\boldsymbol{\beta}}$ is also the Best Linear Unbiased Estimator of $\boldsymbol{X}_0\boldsymbol{\beta}$, which is the expected value of $\mathbf{y}_0$. You are not required to reprove this here, but you are asked to compute $\mathcal{MSE}[\boldsymbol{X}_0\hat{\boldsymbol{\beta}}; \boldsymbol{X}_0\boldsymbol{\beta}]$ and compare it with $\mathcal{MSE}[\mathbf{y}_0^*; \mathbf{y}_0]$. Can you explain the difference?*

Answer. Estimation error and $\mathcal{MSE}$ are

$$\boldsymbol{X}_0\hat{\boldsymbol{\beta}} - \boldsymbol{X}_0\boldsymbol{\beta} = \boldsymbol{X}_0(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \boldsymbol{X}_0(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon} \quad \text{due to (??)}$$

$$\mathcal{MSE}[\boldsymbol{X}_0\hat{\boldsymbol{\beta}}; \boldsymbol{X}_0\boldsymbol{\beta}] = \mathcal{V}[\boldsymbol{X}_0\hat{\boldsymbol{\beta}} - \boldsymbol{X}_0\boldsymbol{\beta}] = \mathcal{V}[\boldsymbol{X}_0(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon}] = \sigma^2\boldsymbol{X}_0(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}_0^\top.$$

It differs from the prediction $\mathcal{MSE}$ matrix by $\sigma^2\boldsymbol{I}$, which is the uncertainty about the value of the new disturbance $\boldsymbol{\varepsilon}_0$ about which the data have no information. □

[**Gre97**, p. 369] has an enlightening formula showing how the prediction intervals increase if one goes away from the center of the data.

Now let us look at the prediction problem in the Generalized Least Squares model

$$\begin{bmatrix}\mathbf{y}\\ \mathbf{y}_0\end{bmatrix} = \begin{bmatrix}\boldsymbol{X}\\ \boldsymbol{X}_0\end{bmatrix}\boldsymbol{\beta} + \begin{bmatrix}\boldsymbol{\varepsilon}\\ \boldsymbol{\varepsilon}_0\end{bmatrix} \qquad \mathcal{E}\begin{bmatrix}\boldsymbol{\varepsilon}\\ \boldsymbol{\varepsilon}_0\end{bmatrix} = \begin{bmatrix}\boldsymbol{o}\\ \boldsymbol{o}\end{bmatrix} \qquad \mathcal{V}\begin{bmatrix}\boldsymbol{\varepsilon}\\ \boldsymbol{\varepsilon}_0\end{bmatrix} = \sigma^2\begin{bmatrix}\boldsymbol{\Psi} & \boldsymbol{C}\\ \boldsymbol{C}^\top & \boldsymbol{\Psi}_0\end{bmatrix}. \tag{27.3.5}$$

$\boldsymbol{X}$ and $\boldsymbol{X}_0$ are known, $\mathbf{y}$ is observed, $\mathbf{y}_0$ is not observed, and we assume $\boldsymbol{\Psi}$ is *positive* definite. If $\boldsymbol{C} = \boldsymbol{O}$, the BLUP of $\mathbf{y}_0$ is $\boldsymbol{X}_0\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is the BLUE in the model $\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. In other words, all new disturbances are simply predicted by zero. If past and future disturbances are correlated, this predictor is no longer optimal.

In [**JHG+88**, pp. 343–346] it is proved that the best linear unbiased predictor of $\mathbf{y}_0$ is

$$\mathbf{y}_0^* = \boldsymbol{X}_0\hat{\boldsymbol{\beta}} + \boldsymbol{C}^\top\boldsymbol{\Psi}^{-1}(\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}). \tag{27.3.6}$$

where $\hat{\boldsymbol{\beta}}$ is the generalized least squares estimator of $\boldsymbol{\beta}$, and that its $\mathcal{MSE}$-matrix $\mathcal{MSE}[\mathbf{y}_0^*; \mathbf{y}_0]$ is

$$\sigma^2\Big(\boldsymbol{\Psi}_0 - \boldsymbol{C}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{C} + (\boldsymbol{X}_0 - \boldsymbol{C}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X}_0^\top - \boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{C})\Big). \tag{27.3.7}$$

Problem 332. *Derive the formula for the $\mathcal{MSE}$ matrix from the formula of the predictor, and compute the joint $\mathcal{MSE}$ matrix for the predicted values and the parameter vector.*

Answer. The prediction error is, using (26.0.3),

$$\mathfrak{y}_0^* - \mathfrak{y}_0 = \boldsymbol{X}_0\hat{\boldsymbol{\beta}} - \boldsymbol{X}_0\boldsymbol{\beta} + \boldsymbol{X}_0\boldsymbol{\beta} - \mathfrak{y}_0 + \boldsymbol{C}^\top\boldsymbol{\Psi}^{-1}(\mathfrak{y} - \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \tag{27.3.8}$$

$$= \boldsymbol{X}_0(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \boldsymbol{\varepsilon}_0 + \boldsymbol{C}^\top\boldsymbol{\Psi}^{-1}(\boldsymbol{\varepsilon} - \boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})) \tag{27.3.9}$$

$$= \boldsymbol{C}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\varepsilon} + (\boldsymbol{X}_0 - \boldsymbol{C}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \boldsymbol{\varepsilon}_0 \tag{27.3.10}$$

$$= \begin{bmatrix}\boldsymbol{C}^\top\boldsymbol{\Psi}^{-1} + (\boldsymbol{X}_0 - \boldsymbol{C}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1} & -\boldsymbol{I}\end{bmatrix}\begin{bmatrix}\boldsymbol{\varepsilon}\\ \boldsymbol{\varepsilon}_0\end{bmatrix} \tag{27.3.11}$$

The $\mathcal{MSE}$-matrix is therefore $\mathcal{E}[(\mathfrak{y}_0^* - \mathfrak{y}_0)(\mathfrak{y}_0^* - \mathfrak{y}_0)^\top] =$

$$= \sigma^2\begin{bmatrix}\boldsymbol{C}^\top\boldsymbol{\Psi}^{-1} + (\boldsymbol{X}_0 - \boldsymbol{C}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1} & -\boldsymbol{I}\end{bmatrix}\begin{bmatrix}\boldsymbol{\Psi} & \boldsymbol{C}\\ \boldsymbol{C}^\top & \boldsymbol{\Psi}_0\end{bmatrix}\begin{bmatrix}\boldsymbol{\Psi}^{-1}\boldsymbol{C} + \boldsymbol{\Psi}^{-1}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X}_0^\top - \boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{C})\\ -\boldsymbol{I}\end{bmatrix} \tag{27.3.12}$$

and the joint $\mathcal{MSE}$ matrix with the sampling error of the parameter vector $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ is

$$\sigma^2\begin{bmatrix}\boldsymbol{C}^\top\boldsymbol{\Psi}^{-1} + (\boldsymbol{X}_0 - \boldsymbol{C}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1} & -\boldsymbol{I}\\ (\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1} & \boldsymbol{O}\end{bmatrix} \tag{27.3.13}$$

$$\begin{bmatrix}\boldsymbol{\Psi} & \boldsymbol{C}\\ \boldsymbol{C}^\top & \boldsymbol{\Psi}_0\end{bmatrix}\begin{bmatrix}\boldsymbol{\Psi}^{-1}\boldsymbol{C} + \boldsymbol{\Psi}^{-1}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X}_0^\top - \boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{C}) & \boldsymbol{\Psi}^{-1}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\\ -\boldsymbol{I} & \boldsymbol{O}\end{bmatrix} =$$

$$= \sigma^2\begin{bmatrix}\boldsymbol{C}^\top\boldsymbol{\Psi}^{-1} + (\boldsymbol{X}_0 - \boldsymbol{C}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1} & -\boldsymbol{I}\\ (\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1} & \boldsymbol{O}\end{bmatrix} \tag{27.3.14}$$

$$\begin{bmatrix}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X}_0^\top - \boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{C}) & \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\\ \boldsymbol{C}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{C} + \boldsymbol{C}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X}_0^\top - \boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{C}) - \boldsymbol{\Psi}_0 & \boldsymbol{C}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\end{bmatrix}$$

If one multiplies this out, one gets

$$\begin{bmatrix}\boldsymbol{\Psi}_0 - \boldsymbol{C}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{C} + (\boldsymbol{X}_0 - \boldsymbol{C}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X}_0^\top - \boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{C}) & (\boldsymbol{X}_0 - \boldsymbol{C}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\\ (\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X}_0^\top - \boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{C}) & (\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\end{bmatrix} \tag{27.3.15}$$

The upper left diagonal element is as claimed in (27.3.7). □

The strategy of the proof given in ITPE is similar to the strategy used to obtain the GLS results, namely, to transform the data in such a way that the disturbances are well behaved. Both data vectors $\mathfrak{y}$ and $\mathfrak{y}_0$ will be transformed, but this transformation must have the following additional property: the transformed $\mathfrak{y}$ must be a function of $\mathfrak{y}$ alone, not of $\mathfrak{y}_0$. Once such a transformation is found, it is easy to predict the transformed $\mathfrak{y}_0$ on the basis of the transformed $\mathfrak{y}$, and from this one also obtains a prediction of $\mathfrak{y}_0$ on the basis of $\mathfrak{y}$.

Here is some heuristics in order to understand formula (27.3.6). Assume for a moment that $\boldsymbol{\beta}$ was known. Then you can apply theorem **??** to the model

$$\begin{bmatrix}\mathfrak{y}\\ \mathfrak{y}_0\end{bmatrix} \sim \begin{bmatrix}\boldsymbol{X}\boldsymbol{\beta}\\ \boldsymbol{X}_0\boldsymbol{\beta}\end{bmatrix}, \sigma^2\begin{bmatrix}\boldsymbol{\Psi} & \boldsymbol{C}\\ \boldsymbol{C}^\top & \boldsymbol{\Psi}_0\end{bmatrix} \tag{27.3.16}$$

to get $\mathfrak{y}_0^* = \boldsymbol{X}_0\boldsymbol{\beta} + \boldsymbol{C}^\top\boldsymbol{\Psi}^{-1}(\mathfrak{y} - \boldsymbol{X}\boldsymbol{\beta})$ as best linear predictor of $\mathfrak{y}_0$ on the basis of $\mathfrak{y}$. According to theorem **??**, its $\mathcal{MSE}$ matrix is $\sigma^2(\boldsymbol{\Psi}_0 - \boldsymbol{C}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{C})$. Since $\boldsymbol{\beta}$ is not known, replace it by $\hat{\boldsymbol{\beta}}$, which gives exactly (27.3.6). This adds $\mathcal{MSE}[\boldsymbol{X}_0\hat{\boldsymbol{\beta}} + \boldsymbol{C}^\top\boldsymbol{\Psi}^{-1}(\mathfrak{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}); \boldsymbol{X}_0\boldsymbol{\beta} + \boldsymbol{C}^\top\boldsymbol{\Psi}^{-1}(\mathfrak{y} - \boldsymbol{X}\boldsymbol{\beta})]$ to the $\mathcal{MSE}$-matrix, which gives (27.3.7).

Problem 333. *Show that*

$$\mathcal{MSE}[\boldsymbol{X}_0\hat{\boldsymbol{\beta}} + \boldsymbol{C}^\top\boldsymbol{\Psi}^{-1}(\mathfrak{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}); \boldsymbol{X}_0\boldsymbol{\beta} + \boldsymbol{C}^\top\boldsymbol{\Psi}^{-1}(\mathfrak{y} - \boldsymbol{X}\boldsymbol{\beta})] = \tag{27.3.17}$$
$$= \sigma^2(\boldsymbol{X}_0 - \boldsymbol{C}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X}_0^\top - \boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{C}).$$

Answer. What is predicted is a random variable, therefore the $\mathcal{MSE}$ matrix is the covariance matrix of the prediction error. The prediction error is $(\boldsymbol{X}_0 - \boldsymbol{C}^\top\boldsymbol{\Psi}^{-1})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$, its covariance matrix is therefore $\sigma^2(\boldsymbol{X}_0 - \boldsymbol{C}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X}_0^\top - \boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{C})$. □

Problem 334. *In the following we work with partitioned matrices. Given the model*

$$\begin{bmatrix}\mathfrak{y}\\ \mathfrak{y}_0\end{bmatrix} = \begin{bmatrix}\boldsymbol{X}\\ \boldsymbol{X}_0\end{bmatrix}\boldsymbol{\beta} + \begin{bmatrix}\boldsymbol{\varepsilon}\\ \boldsymbol{\varepsilon}_0\end{bmatrix} \qquad \mathrm{E}[\begin{bmatrix}\boldsymbol{\varepsilon}\\ \boldsymbol{\varepsilon}_0\end{bmatrix}] = \begin{bmatrix}\boldsymbol{o}\\ \boldsymbol{o}\end{bmatrix} \qquad \mathcal{V}[\begin{bmatrix}\boldsymbol{\varepsilon}\\ \boldsymbol{\varepsilon}_0\end{bmatrix}] = \sigma^2\begin{bmatrix}\boldsymbol{\Psi} & \boldsymbol{C}\\ \boldsymbol{C}^\top & \boldsymbol{\Psi}_0\end{bmatrix}. \tag{27.3.18}$$

$\boldsymbol{X}$ has full rank. $\mathfrak{y}$ is observed, $\mathfrak{y}_0$ is not observed. $\boldsymbol{C}$ is not the null matrix.

• a. *Someone predicts $\mathfrak{y}_0$ by $\mathfrak{y}_0^* = \boldsymbol{X}_0\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\mathfrak{y}$ is the BLUE of $\boldsymbol{\beta}$. Is this predictor unbiased?*

Answer. Yes, since $\mathcal{E}[\mathfrak{y}_0] = \boldsymbol{X}_0\boldsymbol{\beta}$, and $\mathcal{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$. □

• b. *Compute the $\mathcal{MSE}$ matrix $\mathcal{MSE}[\boldsymbol{X}_0\hat{\boldsymbol{\beta}}; \mathfrak{y}_0]$ of this predictor.* Hint: *For any matrix $\boldsymbol{B}$, the difference $\boldsymbol{B}\mathfrak{y} - \mathfrak{y}_0$ can be written in the form $\begin{bmatrix}\boldsymbol{B} & -\boldsymbol{I}\end{bmatrix}\begin{bmatrix}\mathfrak{y}\\ \mathfrak{y}_0\end{bmatrix}$.* Hint: *For an unbiased predictor (or estimator), the $\mathcal{MSE}$ matrix is the covariance matrix of the prediction (or estimation) error.*

Answer.

$$\mathrm{E}[(\boldsymbol{B}\mathfrak{y} - \mathfrak{y}_0)(\boldsymbol{B}\mathfrak{y} - \mathfrak{y}_0)^\top] = \mathcal{V}[\boldsymbol{B}\mathfrak{y} - \mathfrak{y}_0] \tag{27.3.19}$$

$$= \mathcal{V}\Big[\begin{bmatrix}\boldsymbol{B} & -\boldsymbol{I}\end{bmatrix}\begin{bmatrix}\mathfrak{y}\\ \mathfrak{y}_0\end{bmatrix}\Big] \tag{27.3.20}$$

$$= \sigma^2\begin{bmatrix}\boldsymbol{B} & -\boldsymbol{I}\end{bmatrix}\begin{bmatrix}\boldsymbol{\Psi} & \boldsymbol{C}\\ \boldsymbol{C}^\top & \boldsymbol{\Psi}_0\end{bmatrix}\begin{bmatrix}\boldsymbol{B}^\top\\ -\boldsymbol{I}\end{bmatrix} \tag{27.3.21}$$

$$= \sigma^2\big(\boldsymbol{B}\boldsymbol{\Psi}\boldsymbol{B}^\top - \boldsymbol{C}^\top\boldsymbol{B}^\top - \boldsymbol{C}\boldsymbol{B} + \boldsymbol{\Psi}_0\big). \tag{27.3.22}$$

Now one must use $\boldsymbol{B} = \boldsymbol{X}_0(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}$. One ends up with

$$\mathcal{MSE}[\boldsymbol{X}_0\hat{\boldsymbol{\beta}}; \mathfrak{y}_0] = \sigma^2\Big(\boldsymbol{X}_0(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}_0^\top - \boldsymbol{C}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}_0^\top - \boldsymbol{X}_0(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{C} + \tag{27.3.23}$$

□

• c. *Compare its $\mathcal{MSE}$-matrix with formula* (27.3.7). *Is the difference nonnegative definite?*

Answer. To compare it with the minimum $\mathcal{MSE}$ matrix, it can also be written as

(27.3.24)

$$\mathcal{MSE}[\boldsymbol{X}_0\hat{\boldsymbol{\beta}};\mathbf{y}_0] = \sigma^2\Big(\boldsymbol{\Psi}_0+(\boldsymbol{X}_0-\boldsymbol{C}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X}_0^\top-\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{C})-\boldsymbol{C}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-}$$

i.e., it exceeds the minimum $\mathcal{MSE}$ matrix by $\boldsymbol{C}^\top(\boldsymbol{\Psi}^{-1}-\boldsymbol{\Psi}^{-1}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1})\boldsymbol{C}$. This is nnd because the matrix in parentheses is $\boldsymbol{M}=\boldsymbol{M}\boldsymbol{\Psi}\boldsymbol{M}$, refer here to Problem 322. □

CHAPTER 28

Updating of Estimates When More Observations become Available

The theory of the linear model often deals with pairs of models which are nested in each other, one model either having more data or more stringent parameter restrictions than the other. We will discuss such nested models in three forms: in the remainder of the present chapter 28 we will see how estimates must be updated when more observations become available, in chapter 29 how the imposition of a linear constraint affects the parameter estimates, and in chapter 30 what happens if one adds more regressors.

Assume you have already computed the BLUE $\hat{\boldsymbol{\beta}}$ on the basis of the observations $\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, and afterwards additional data $\mathbf{y}_0 = \boldsymbol{X}_0\boldsymbol{\beta} + \boldsymbol{\varepsilon}_0$ become available. Then $\hat{\boldsymbol{\beta}}$ can be updated using the following principles:

Before the new observations became available, the information given in the original dataset not only allowed to estimate $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}$, but also yielded a prediction $\mathbf{y}_0^* = \boldsymbol{X}_0\hat{\boldsymbol{\beta}}$ of the additional data. The estimation error $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ and the prediction error $\mathbf{y}_0^* - \mathbf{y}_0$ are unobserved, but we know their expected values (the zero vectors), and we also know their joint covariance matrix up to the unknown factor σ^2. After the additional data have become available, we can compute the actual value of the prediction error $\mathbf{y}_0^* - \mathbf{y}_0$. This allows us to also get a better idea of the actual value of the estimation error, and therefore we can get a better estimator of $\boldsymbol{\beta}$. The following steps are involved:

(1) Make the best prediction $\mathbf{y}_0^*$ of the new data $\mathbf{y}_0$ based on $\mathbf{y}$.

(2) Compute the joint covariance matrix of the prediction error $\mathbf{y}_0^* - \mathbf{y}_0$ of the new data by the old (which is observed) and the sampling error in the old regression $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ (which is unobserved).

(3) Use the formula for best linear prediction (**??**) to get a predictor $\boldsymbol{z}^*$ of $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$.

(4) Then $\hat{\hat{\boldsymbol{\beta}}} = \hat{\boldsymbol{\beta}} - \boldsymbol{z}^*$ is the BLUE of $\boldsymbol{\beta}$ based on the joint observations $\mathbf{y}$ and $\mathbf{y}_0$.

(5) The sum of squared errors of the updated model minus that of the basic model is the standardized prediction error $\mathsf{SSE}^* - \mathsf{SSE} = (\mathbf{y}_0^* - \mathbf{y}_0)^\top \boldsymbol{\Omega}^{-1}(\mathbf{y}_0^* - \mathbf{y}_0)$ where $\mathsf{SSE}^* = (\mathbf{y} - \boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}})^\top(\mathbf{y} - \boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}})$ $\mathcal{V}[\mathbf{y}_0^* - \mathbf{y}_0] = \sigma^2\boldsymbol{\Omega}$.

In the case of *one* additional observation and spherical covariance matrix, this procedure yields the following formulas:

Problem 335. *Assume $\hat{\boldsymbol{\beta}}$ is the BLUE on the basis of the observation $\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, and a new observation $\mathsf{y}_0 = \boldsymbol{x}_0^\top\boldsymbol{\beta} + \varepsilon_0$ becomes available. Show that the updated estimator has the form*

$$\hat{\hat{\boldsymbol{\beta}}} = \hat{\boldsymbol{\beta}} + (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_0\frac{\mathsf{y}_0 - \boldsymbol{x}_0^\top\hat{\boldsymbol{\beta}}}{1 + \boldsymbol{x}_0^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_0}. \tag{28.0.25}$$

Answer. Set it up as follows:

$$\begin{bmatrix}\mathsf{y}_0 - \boldsymbol{x}_0^\top\hat{\boldsymbol{\beta}}\\ \boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\end{bmatrix} \sim \begin{bmatrix}0\\ \boldsymbol{o}\end{bmatrix}, \sigma^2\begin{bmatrix}\boldsymbol{x}_0^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_0 + 1 & \boldsymbol{x}_0^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\\ (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_0 & (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\end{bmatrix} \tag{28.0.26}$$

and use (27.1.18). By the way, if the covariance matrix is not spherical but is $\begin{bmatrix}\boldsymbol{\Psi} & \boldsymbol{c}\\ \boldsymbol{c}^\top & \psi_0\end{bmatrix}$ we get from (27.3.6)

$$\mathsf{y}_0^* = \boldsymbol{x}_0^\top\hat{\boldsymbol{\beta}} + \boldsymbol{c}^\top\boldsymbol{\Psi}^{-1}(\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \tag{28.0.27}$$

and from (27.3.15)

$$\begin{bmatrix}\mathsf{y}_0 - \mathsf{y}_0^*\\ \boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\end{bmatrix} \sim \begin{bmatrix}0\\ \boldsymbol{o}\end{bmatrix}, \sigma^2 \tag{28.0.28}$$

$$\begin{bmatrix}\psi_0 - \boldsymbol{c}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{c} + (\boldsymbol{x}_0^\top - \boldsymbol{c}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{x}_0 - \boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{c}) & (\boldsymbol{x}_0^\top - \boldsymbol{c}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\\ (\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{x}_0 - \boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{c}) & (\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\end{bmatrix}$$

□

• a. *Show that the residual $\hat{\hat{\varepsilon}}_0$ from the full regression is the following nonrandom multiple of the "predictive" residual $\mathsf{y}_0 - \boldsymbol{x}_0^\top\hat{\boldsymbol{\beta}}$:*

$$\hat{\hat{\varepsilon}}_0 = \mathsf{y}_0 - \boldsymbol{x}_0^\top\hat{\hat{\boldsymbol{\beta}}} = \frac{1}{1 + \boldsymbol{x}_0^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_0}(\mathsf{y}_0 - \boldsymbol{x}_0^\top\hat{\boldsymbol{\beta}}) \tag{28.0.29}$$

Interestingly, this is the predictive residual divided by its relative variance (to standardize it one would have to divide it by its relative standard deviation). Compare this with (31.2.9).

Answer. (28.0.29) can either be derived from (28.0.25), or from the following alternative application of the updating principle: All the information which the old observations have for the estimate of $\boldsymbol{x}_0^\top\boldsymbol{\beta}$ is contained in $\hat{\mathsf{y}}_0 = \boldsymbol{x}_0^\top\hat{\boldsymbol{\beta}}$. The information which the updated regression, which

includes the additional observation, has about $\boldsymbol{x}_0^\top\boldsymbol{\beta}$ can therefore be represented by the following two "observations":

$$\begin{bmatrix}\hat{\mathsf{y}}_0\\ \mathsf{y}_0\end{bmatrix} = \begin{bmatrix}1\\1\end{bmatrix}\boldsymbol{x}_0^\top\boldsymbol{\beta} + \begin{bmatrix}\delta_1\\ \delta_2\end{bmatrix} \qquad \begin{bmatrix}\delta_1\\ \delta_2\end{bmatrix} \sim \left(\begin{bmatrix}0\\0\end{bmatrix}, \sigma^2\begin{bmatrix}\boldsymbol{x}_0^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_0 & 0\\ 0 & 1\end{bmatrix}\right) \tag{28.0.30}$$

This is a regression model with two observations and one unknown parameter, $\boldsymbol{x}_0^\top\boldsymbol{\beta}$, which has a nonspherical error covariance matrix. The formula for the BLUE of $\boldsymbol{x}_0^\top\boldsymbol{\beta}$ in model (28.0.30) is

(28.0.31)

$$\hat{\mathsf{y}}_0 = \left(\begin{bmatrix}1 & 1\end{bmatrix}\begin{bmatrix}\boldsymbol{x}_0^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_0 & 0\\ 0 & 1\end{bmatrix}^{-1}\begin{bmatrix}1\\1\end{bmatrix}\right)^{-1}\begin{bmatrix}1 & 1\end{bmatrix}\begin{bmatrix}\boldsymbol{x}_0^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_0 & 0\\ 0 & 1\end{bmatrix}^{-1}\begin{bmatrix}\hat{\mathsf{y}}_0\\ \mathsf{y}_0\end{bmatrix}$$

$$= \frac{1}{1+\frac{1}{\boldsymbol{x}_0^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_0}}\left(\frac{\hat{\mathsf{y}}_0}{\boldsymbol{x}_0^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_0} + \mathsf{y}_0\right) \tag{28.0.32}$$

$$= \frac{1}{1+\boldsymbol{x}_0^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_0}(\hat{\mathsf{y}}_0 + \boldsymbol{x}_0^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_0\mathsf{y}_0). \tag{28.0.33}$$

Now subtract (28.0.33) from y_0 to get (28.0.29).

□

Using (28.0.29), one can write (28.0.25) as

$$\hat{\hat{\boldsymbol{\beta}}} = \hat{\boldsymbol{\beta}} + (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_0\hat{\hat{\varepsilon}}_0 \tag{28.0.34}$$

Later, in (32.4.1), one will see that it can also be written in the form

$$\hat{\hat{\boldsymbol{\beta}}} = \hat{\boldsymbol{\beta}} + (\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{x}_0(\mathsf{y}_0 - \boldsymbol{x}_0^\top\hat{\boldsymbol{\beta}}) \tag{28.0.35}$$

where $\boldsymbol{Z} = \begin{bmatrix}\boldsymbol{X}\\ \boldsymbol{x}_0^\top\end{bmatrix}$.

Problem 336. *Show the following fact which is point (5) in the above updating principle in this special case: If one takes the squares of the standardized predictive residuals, one gets the difference of the* SSE *for the regression with and without the additional observation* y_0

$$\mathsf{SSE}^* - \mathsf{SSE} = \frac{(\mathsf{y}_0 - \boldsymbol{x}_0^\top\hat{\boldsymbol{\beta}})^2}{1+\boldsymbol{x}_0^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_0} \tag{28.0.36}$$

Answer. The sum of squared errors in the old regression is $\mathsf{SSE} = (\mathsf{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\mathsf{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})$; the sum of squared errors in the updated regression is $\mathsf{SSE}^* = (\mathsf{y}-\boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}})^\top(\mathsf{y}-\boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}}) + \hat{\hat{\varepsilon}}_0{}^2$. From (28.0.34) follows

$$\mathsf{y}-\boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}} = \mathsf{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_0\hat{\hat{\varepsilon}}_0. \tag{28.0.37}$$

If one squares this, the cross product terms fall away: $(\mathsf{y}-\boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}})^\top(\mathsf{y}-\boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}}) = (\mathsf{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\mathsf{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}) + \hat{\hat{\varepsilon}}_0\boldsymbol{x}_0^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_0\hat{\hat{\varepsilon}}_0$. Adding $\hat{\hat{\varepsilon}}_0{}^2$ to both sides gives $\mathsf{SSE}^* = \mathsf{SSE} + \hat{\hat{\varepsilon}}_0{}^2(1+\boldsymbol{x}_0^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_0)$. Now use (28.0.29) to get (28.0.36). □

CHAPTER 29

Constrained Least Squares

One of the assumptions for the linear model was that nothing is known about the true value of $\boldsymbol{\beta}$. Any k-vector $\boldsymbol{\gamma}$ is a possible candidate for the value of $\boldsymbol{\beta}$. We used this assumption e.g. when we concluded that an unbiased estimator $\tilde{\boldsymbol{B}}\mathbf{y}$ of $\boldsymbol{\beta}$ must satisfy $\tilde{\boldsymbol{B}}\boldsymbol{X} = \boldsymbol{I}$. Now we will modify this assumption and assume we know that the true value $\boldsymbol{\beta}$ satisfies the linear constraint $\boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{u}$. To fix notation, assume $\boldsymbol{y}$ be a $n \times 1$ vector, $\boldsymbol{u}$ a $i \times 1$ vector, $\boldsymbol{X}$ a $n \times k$ matrix, and $\boldsymbol{R}$ a $i \times k$ matrix. In addition to our usual assumption that all columns of $\boldsymbol{X}$ are linearly independent (i.e., $\boldsymbol{X}$ has full column rank) we will also make the assumption that all *rows* of $\boldsymbol{R}$ are linearly independent (which is called: $\boldsymbol{R}$ has full row rank). In other words, the matrix of constraints $\boldsymbol{R}$ does not include "redundant" constraints which are linear combinations of the other constraints.

29.1. Building the Constraint into the Model

Problem 337. *Given a regression with a constant term and two explanatory variables which we will call x and z, i.e.,*

$$\mathrm{y}_t = \alpha + \beta x_t + \gamma z_t + \varepsilon_t \tag{29.1.1}$$

• a. *1 point How will you estimate β and γ if it is known that $\beta = \gamma$?*

Answer. Write

$$\mathrm{y}_t = \alpha + \beta(x_t + z_t) + \varepsilon_t \tag{29.1.2}$$

□

• b. *1 point How will you estimate β and γ if it is known that $\beta + \gamma = 1$?*

Answer. Setting $\gamma = 1 - \beta$ gives the regression

$$\mathrm{y}_t - z_t = \alpha + \beta(x_t - z_t) + \varepsilon_t \tag{29.1.3}$$

□

• c. *3 points Go back to a. If you add the original z as an additional regressor into the modified regression incorporating the constraint $\beta = \gamma$, then the coefficient of z is no longer an estimate of the original γ, but of a new parameter δ which is a*

linear combination of α, β, and γ. Compute this linear combination, i.e., express δ in terms of α, β, and γ. Remark (no proof required): this regression is equivalent to (29.1.1)*, and it allows you to* test *the constraint.*

Answer. It you add z as additional regressor into (29.1.2), you get $\mathrm{y}_t = \alpha + \beta(x_t + z_t) + \delta z_t + \varepsilon_t$. Now substitute the right hand side from (29.1.1) for y to get $\alpha + \beta x_t + \gamma z_t + \varepsilon_t = \alpha + \beta(x_t + z_t) + \delta z_t + \varepsilon_t$. Cancelling out gives $\gamma z_t = \beta z_t + \delta z_t$, in other words, $\gamma = \beta + \delta$. In this regression, therefore, the coefficient of z is split into the sum of two terms, the first term is the value it should be if the constraint were satisfied, and the other term is the difference from that. □

• d. *2 points Now do the same thing with the modified regression from part b which incorporates the constraint $\beta + \gamma = 1$: include the original z as an additional regressor and determine the meaning of the coefficient of z.*

What Problem 337 suggests is true in general: every constrained Least Squares problem can be reduced to an equivalent unconstrained Least Squares problem with fewer explanatory variables. Indeed, one can consider *every* least squares problem to be "constrained" because the assumption $\mathcal{E}[\mathbf{y}] = \boldsymbol{X}\boldsymbol{\beta}$ for some $\boldsymbol{\beta}$ is equivalent to a linear constraint on $\mathcal{E}[\mathbf{y}]$. The decision not to include certain explanatory variables in the regression can be considered the decision to set certain elements of $\boldsymbol{\beta}$ zero, which is the imposition of a constraint. If one writes a certain regression model as a constrained version of some other regression model, this simply means that one is interested in the relationship between two nested regressions.

Problem 273 is another example here.

29.2. Conversion of an Arbitrary Constraint into a Zero Constraint

This section, which is nothing but the matrix version of Problem 337, follows [**DM93**, pp. 16–19]. By reordering the elements of $\boldsymbol{\beta}$ one can write the constraint $\boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{u}$ in the form

$$\begin{bmatrix}\boldsymbol{R}_1 & \boldsymbol{R}_2\end{bmatrix}\begin{bmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\end{bmatrix} \equiv \boldsymbol{R}_1\boldsymbol{\beta}_1 + \boldsymbol{R}_2\boldsymbol{\beta}_2 = \boldsymbol{u} \tag{29.2.1}$$

where $\boldsymbol{R}_1$ is a nonsingular $i \times i$ matrix. Why can that be done? The rank of $\boldsymbol{R}$ is i, i.e., all the rows are linearly independent. Since row rank is equal to column rank, there are also i linearly independent columns. Use those for $\boldsymbol{R}_1$. Using this same partition, the original regression can be written

$$\mathbf{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon} \tag{29.2.2}$$

Now one can solve (29.2.1) for $\boldsymbol{\beta}_1$ to get

$$\boldsymbol{\beta}_1 = \boldsymbol{R}_1^{-1}\boldsymbol{u} - \boldsymbol{R}_1^{-1}\boldsymbol{R}_2\boldsymbol{\beta}_2 \tag{29.2.3}$$

Plug (29.2.3) into (29.2.2) and rearrange to get a regression which is equivalent to the constrained regression:

$$\boldsymbol{y} - \boldsymbol{X}_1\boldsymbol{R}_1^{-1}\boldsymbol{u} = (\boldsymbol{X}_2 - \boldsymbol{X}_1\boldsymbol{R}_1^{-1}\boldsymbol{R}_2)\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon} \tag{29.2.4}$$

or

$$\boldsymbol{y}^* = \boldsymbol{Z}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon} \tag{29.2.5}$$

One more thing is noteworthy here: if we add $\boldsymbol{X}_1$ as additional regressors into (29.2.5), we get a regression that is equivalent to (29.2.2). To see this, define the difference between the left hand side and right hand side of (29.2.3) as $\boldsymbol{\gamma}_1 = \boldsymbol{\beta}_1 - \boldsymbol{R}_1^{-1}\boldsymbol{u} + \boldsymbol{R}_1^{-1}\boldsymbol{R}_2\boldsymbol{\beta}_2$; then the constraint (29.2.1) is equivalent to the "zero constraint" $\boldsymbol{\gamma}_1 = \boldsymbol{o}$, and the regression

$$\boldsymbol{y} - \boldsymbol{X}_1\boldsymbol{R}_1^{-1}\boldsymbol{u} = (\boldsymbol{X}_2 - \boldsymbol{X}_1\boldsymbol{R}_1^{-1}\boldsymbol{R}_2)\boldsymbol{\beta}_2 + \boldsymbol{X}_1(\boldsymbol{\beta}_1 - \boldsymbol{R}_1^{-1}\boldsymbol{u} + \boldsymbol{R}_1^{-1}\boldsymbol{R}_2\boldsymbol{\beta}_2) + \boldsymbol{\varepsilon} \tag{29.2.6}$$

is equivalent to the original regression (29.2.2). (29.2.6) can also be written as

$$\boldsymbol{y}^* = \boldsymbol{Z}_2\boldsymbol{\beta}_2 + \boldsymbol{X}_1\boldsymbol{\gamma}_1 + \boldsymbol{\varepsilon} \tag{29.2.7}$$

The coefficient of $\boldsymbol{X}_1$, if it is added back into (29.2.5), is therefore $\boldsymbol{\gamma}_1$.

Problem 338. [**DM93**] *assert on p. 17, middle, that*

$$\mathrm{R}[\boldsymbol{X}_1, \boldsymbol{Z}_2] = \mathrm{R}[\boldsymbol{X}_1, \boldsymbol{X}_2]. \tag{29.2.8}$$

where $\boldsymbol{Z}_2 = \boldsymbol{X}_2 - \boldsymbol{X}_1\boldsymbol{R}_1^{-1}\boldsymbol{R}_2$. *Give a proof.*

Answer. We have to show

$$\{\boldsymbol{z}\colon \boldsymbol{z} = \boldsymbol{X}_1\boldsymbol{\gamma} + \boldsymbol{X}_2\boldsymbol{\delta}\} = \{\boldsymbol{z}\colon \boldsymbol{z} = \boldsymbol{X}_1\boldsymbol{\alpha} + \boldsymbol{Z}_2\boldsymbol{\beta}\} \tag{29.2.9}$$

First $\subset$: given $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$ we need a $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ with

$$\boldsymbol{X}_1\boldsymbol{\gamma} + \boldsymbol{X}_2\boldsymbol{\delta} = \boldsymbol{X}_1\boldsymbol{\alpha} + (\boldsymbol{X}_2 - \boldsymbol{X}_1\boldsymbol{R}_1^{-1}\boldsymbol{R}_2)\boldsymbol{\beta} \tag{29.2.10}$$

This can be accomplished with $\boldsymbol{\beta} = \boldsymbol{\delta}$ and $\boldsymbol{\alpha} = \boldsymbol{\gamma} + \boldsymbol{R}_1^{-1}\boldsymbol{R}_2\boldsymbol{\delta}$. The other side is even more trivial: given $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, multiplying out the right side of (29.2.10) gives $\boldsymbol{X}_1\boldsymbol{\alpha} + \boldsymbol{X}_2\boldsymbol{\beta} - \boldsymbol{X}_1\boldsymbol{R}_1^{-1}\boldsymbol{R}_2\boldsymbol{\beta}$, i.e., $\boldsymbol{\delta} = \boldsymbol{\beta}$ and $\boldsymbol{\gamma} = \boldsymbol{\alpha} - \boldsymbol{R}_1^{-1}\boldsymbol{R}_2\boldsymbol{\beta}$. □

29.3. Lagrange Approach to Constrained Least Squares

The constrained least squares estimator is that $k \times 1$ vector $\boldsymbol{\beta} = \hat{\tilde{\boldsymbol{\beta}}}$ which minimizes $\mathsf{SSE} = (\mathsf{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\mathsf{y} - \boldsymbol{X}\boldsymbol{\beta})$ subject to the linear constraint $\boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{u}$.

Again, we assume that $\boldsymbol{X}$ has full column and $\boldsymbol{R}$ full row rank.

The Lagrange approach to constrained least squares, which we follow here, is given in [**Gre97**, Section 7.3 on pp. 341/2], also [**DM93**, pp. 90/1]:

The Constrained Least Squares problem can be solved with the help of the "Lagrange function," which is a function of the $k \times 1$ vector $\boldsymbol{\beta}$ and an additional $i \times 1$ vector $\boldsymbol{\lambda}$ of "Lagrange multipliers":

$$\mathsf{L}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = (\mathsf{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\mathsf{y} - \boldsymbol{X}\boldsymbol{\beta}) + (\boldsymbol{R}\boldsymbol{\beta} - \boldsymbol{u})^\top\boldsymbol{\lambda} \tag{29.3.1}$$

$\boldsymbol{\lambda}$ can be considered a vector of "penalties" for violating the constraint. For every possible value of $\boldsymbol{\lambda}$ one computes that $\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}$ which minimizes L for that $\boldsymbol{\lambda}$ (This is an unconstrained minimization problem.) It will turn out that for one of the values $\boldsymbol{\lambda} = \boldsymbol{\lambda}^*$, the corresponding $\boldsymbol{\beta} = \hat{\tilde{\boldsymbol{\beta}}}$ satisfies the constraint. This $\hat{\tilde{\boldsymbol{\beta}}}$ is the solution of the constrained minimization problem we are looking for.

Problem 339. *4 points Show the following: If* $\boldsymbol{\beta} = \hat{\tilde{\boldsymbol{\beta}}}$ *is the unconstrained minimum argument of the Lagrange function*

$$\mathsf{L}(\boldsymbol{\beta}, \boldsymbol{\lambda}^*) = (\mathsf{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\mathsf{y} - \boldsymbol{X}\boldsymbol{\beta}) + (\boldsymbol{R}\boldsymbol{\beta} - \boldsymbol{u})^\top\boldsymbol{\lambda}^* \tag{29.3.2}$$

for some fixed value $\boldsymbol{\lambda}^*$, *and if at the same time* $\hat{\tilde{\boldsymbol{\beta}}}$ *satisfies* $\boldsymbol{R}\hat{\tilde{\boldsymbol{\beta}}} = \boldsymbol{u}$, *then* $\boldsymbol{\beta} = \hat{\tilde{\boldsymbol{\beta}}}$ *minimizes* $(\mathsf{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\mathsf{y} - \boldsymbol{X}\boldsymbol{\beta})$ *subject to the constraint* $\boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{u}$.

Answer. Since $\hat{\tilde{\boldsymbol{\beta}}}$ minimizes the Lagrange function, we know that

$$(\mathsf{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}})^\top(\mathsf{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}) + (\boldsymbol{R}\tilde{\boldsymbol{\beta}} - \boldsymbol{u})^\top\boldsymbol{\lambda}^* \geq (\mathsf{y} - \boldsymbol{X}\hat{\tilde{\boldsymbol{\beta}}})^\top(\mathsf{y} - \boldsymbol{X}\hat{\tilde{\boldsymbol{\beta}}}) + (\boldsymbol{R}\hat{\tilde{\boldsymbol{\beta}}} - \boldsymbol{u})^\top\boldsymbol{\lambda}^* \tag{29.3.3}$$

for all $\tilde{\boldsymbol{\beta}}$. Since by assumption, $\hat{\tilde{\boldsymbol{\beta}}}$ also satisfies the constraint, this simplifies to:

$$(\mathsf{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}})^\top(\mathsf{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}) + (\boldsymbol{R}\tilde{\boldsymbol{\beta}} - \boldsymbol{u})^\top\boldsymbol{\lambda}^* \geq (\mathsf{y} - \boldsymbol{X}\hat{\tilde{\boldsymbol{\beta}}})^\top(\mathsf{y} - \boldsymbol{X}\hat{\tilde{\boldsymbol{\beta}}}). \tag{29.3.4}$$

This is still true for all $\tilde{\boldsymbol{\beta}}$. If we only look at those $\tilde{\boldsymbol{\beta}}$ which satisfy the constraint, we get

$$(\mathsf{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}})^\top(\mathsf{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}) \geq (\mathsf{y} - \boldsymbol{X}\hat{\tilde{\boldsymbol{\beta}}})^\top(\mathsf{y} - \boldsymbol{X}\hat{\tilde{\boldsymbol{\beta}}}). \tag{29.3.5}$$

This means, $\hat{\tilde{\boldsymbol{\beta}}}$ is the constrained minimum argument.

□

Instead of imposing the constraint itself, one imposes a penalty function which has such a form that the agents will "voluntarily" heed the constraint. This is a familiar principle in neoclassical economics: instead of restricting pollution to a certain level, tax the polluters so much that they will voluntarily stay within the desired level.

The proof which follows now not only derives the formula for $\hat{\tilde{\boldsymbol{\beta}}}$ but also shows that there is always a $\boldsymbol{\lambda}^*$ for which $\hat{\tilde{\boldsymbol{\beta}}}$ satisfies $\boldsymbol{R}\hat{\tilde{\boldsymbol{\beta}}} = \boldsymbol{u}$.

Problem 340. *2 points Use the simple matrix differentiation rules* $\partial(\boldsymbol{w}^\top\boldsymbol{\beta})/\partial\boldsymbol{\beta}^\top = \boldsymbol{w}^\top$ *and* $\partial(\boldsymbol{\beta}^\top\boldsymbol{M}\boldsymbol{\beta})/\partial\boldsymbol{\beta}^\top = 2\boldsymbol{\beta}^\top\boldsymbol{M}$ *to compute* $\partial\mathsf{L}/\partial\boldsymbol{\beta}^\top$ *where*

$$\mathsf{L}(\boldsymbol{\beta}) = (\mathsf{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\mathsf{y} - \boldsymbol{X}\boldsymbol{\beta}) + (\boldsymbol{R}\boldsymbol{\beta} - \boldsymbol{u})^\top\boldsymbol{\lambda} \tag{29.3.6}$$

Answer. Write the objective function as $\mathbf{y}^\top\mathbf{y} - 2\mathbf{y}^\top\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\lambda}^\top\boldsymbol{R}\boldsymbol{\beta} - \boldsymbol{\lambda}^\top\boldsymbol{u}$ to get (29.3.7). □

Our goal is to find a $\hat{\hat{\boldsymbol{\beta}}}$ and a $\boldsymbol{\lambda}^*$ so that (a) $\boldsymbol{\beta} = \hat{\hat{\boldsymbol{\beta}}}$ minimizes $\mathsf{L}(\boldsymbol{\beta}, \boldsymbol{\lambda}^*)$ and (b) $\boldsymbol{R}\hat{\hat{\boldsymbol{\beta}}} = \boldsymbol{u}$. In other words, $\hat{\hat{\boldsymbol{\beta}}}$ and $\boldsymbol{\lambda}^*$ together satisfy the following two conditions: (a) they must satisfy the first order condition for the unconstrained minimization of L with respect to $\boldsymbol{\beta}$, i.e., $\hat{\hat{\boldsymbol{\beta}}}$ must annul

$$\partial\mathsf{L}/\partial\boldsymbol{\beta}^\top = -2\mathbf{y}^\top\boldsymbol{X} + 2\boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{X} + \boldsymbol{\lambda}^{*\top}\boldsymbol{R}, \tag{29.3.7}$$

and (b) $\hat{\hat{\boldsymbol{\beta}}}$ must satisfy the constraint (29.3.9).

(29.3.7) and (29.3.9) are two linear matrix equations which can indeed be solved for $\hat{\hat{\boldsymbol{\beta}}}$ and $\boldsymbol{\lambda}^*$. I wrote (29.3.7) as a row vector, because the Jacobian of a scalar function is a row vector, but it is usually written as a column vector. Since this conventional notation is arithmetically a little simpler here, we will replace (29.3.7) with its transpose (29.3.8). Our starting point is therefore

$$2\boldsymbol{X}^\top\boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}} = 2\boldsymbol{X}^\top\mathbf{y} - \boldsymbol{R}^\top\boldsymbol{\lambda}^* \tag{29.3.8}$$

$$\boldsymbol{R}\hat{\hat{\boldsymbol{\beta}}} - \boldsymbol{u} = \boldsymbol{o} \tag{29.3.9}$$

Some textbook treatments have an extra factor 2 in front of $\boldsymbol{\lambda}^*$, which makes the math slightly smoother, but which has the disadvantage that the Lagrange multiplier can no longer be interpreted as the "shadow price" for violating the constraint.

Solve (29.3.8) for $\hat{\hat{\boldsymbol{\beta}}}$ to get that $\hat{\hat{\boldsymbol{\beta}}}$ which minimizes L for any given $\boldsymbol{\lambda}^*$:

$$\hat{\hat{\boldsymbol{\beta}}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\mathbf{y} - \frac{1}{2}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\boldsymbol{\lambda}^* = \hat{\boldsymbol{\beta}} - \frac{1}{2}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\boldsymbol{\lambda}^* \tag{29.3.10}$$

Here $\hat{\boldsymbol{\beta}}$ on the right hand side is the *unconstrained* OLS estimate. Plug this formula for $\hat{\hat{\boldsymbol{\beta}}}$ into (29.3.9) in order to determine that value of $\boldsymbol{\lambda}^*$ for which the corresponding $\hat{\hat{\boldsymbol{\beta}}}$ satisfies the constraint:

$$\boldsymbol{R}\hat{\boldsymbol{\beta}} - \frac{1}{2}\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\boldsymbol{\lambda}^* - \boldsymbol{u} = \boldsymbol{o}. \tag{29.3.11}$$

Since $\boldsymbol{R}$ has full row rank and $\boldsymbol{X}$ full column rank, $\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top$ has an inverse (Problem 341). Therefore one can solve for $\boldsymbol{\lambda}^*$:

$$\boldsymbol{\lambda}^* = 2\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{u}) \tag{29.3.12}$$

If one substitutes this $\boldsymbol{\lambda}^*$ back into (29.3.10), one gets the formula for the constrained least squares estimator:

$$\hat{\hat{\boldsymbol{\beta}}} = \hat{\boldsymbol{\beta}} - (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{u}). \tag{29.3.13}$$

Problem 341. *If $\boldsymbol{R}$ has full row rank and $\boldsymbol{X}$ full column rank, show that $\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top$ has an inverse.*

Answer. Since it is nonnegative definite we have to show that it is positive definite. $\boldsymbol{b}^\top\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\boldsymbol{b} = 0$ implies $\boldsymbol{b}^\top\boldsymbol{R} = \boldsymbol{o}^\top$ because $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$ is positive definite, and this implies $\boldsymbol{b} = \boldsymbol{o}$ because $\boldsymbol{R}$ has full row rank. □

Problem 342. *Assume $\boldsymbol{\varepsilon} \sim (\boldsymbol{o}, \sigma^2\boldsymbol{\Psi})$ with a nonsingular $\boldsymbol{\Psi}$ and show: If one minimizes $\mathsf{SSE} = (\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta})^\top\boldsymbol{\Psi}^{-1}(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta})$ subject to the linear constraint $\boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{u}$, the formula for the minimum argument $\hat{\hat{\boldsymbol{\beta}}}$ is the following modification of* (29.3.13)*:*

$$\hat{\hat{\boldsymbol{\beta}}} = \hat{\boldsymbol{\beta}} - (\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{u}) \tag{29.3.14}$$

where $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\mathbf{y}$. This formula is given in [**JHG+88**, (11.2.38) on p. 457]. *Remark, which you are not asked to prove: this is the best linear unbiased estimator if $\boldsymbol{\varepsilon} \sim (\boldsymbol{o}, \sigma^2\boldsymbol{\Psi})$ among all linear estimators which are unbiased whenever the true $\boldsymbol{\beta}$ satisfies the constraint $\boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{u}$.)*

Answer. Lagrange function is

$$\begin{aligned}\mathsf{L}(\boldsymbol{\beta}, \boldsymbol{\lambda}) &= (\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta})^\top\boldsymbol{\Psi}^{-1}(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}) + (\boldsymbol{R}\boldsymbol{\beta} - \boldsymbol{u})^\top\boldsymbol{\lambda} \\ &= \mathbf{y}^\top\mathbf{y} - 2\mathbf{y}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\lambda}^\top\boldsymbol{R}\boldsymbol{\beta} - \boldsymbol{\lambda}^\top\boldsymbol{u}\end{aligned}$$

Jacobian is

$$\partial\mathsf{L}/\partial\boldsymbol{\beta}^\top = -2\mathbf{y}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X} + 2\boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X} + \boldsymbol{\lambda}^\top\boldsymbol{R},$$

Transposing and setting it zero gives

$$2\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}} = 2\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\mathbf{y} - \boldsymbol{R}^\top\boldsymbol{\lambda}^* \tag{29.3.15}$$

Solve (29.3.15) for $\hat{\hat{\boldsymbol{\beta}}}$:

$$\hat{\hat{\boldsymbol{\beta}}} = (\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\mathbf{y} - \frac{1}{2}(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{R}^\top\boldsymbol{\lambda}^* = \hat{\boldsymbol{\beta}} - \frac{1}{2}(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{R}^\top\boldsymbol{\lambda}^* \tag{29.3.16}$$

Here $\hat{\boldsymbol{\beta}}$ is the unconstrained GLS estimate. Plug $\hat{\hat{\boldsymbol{\beta}}}$ into the constraint (29.3.9):

$$\boldsymbol{R}\hat{\boldsymbol{\beta}} - \frac{1}{2}\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{R}^\top\boldsymbol{\lambda}^* - \boldsymbol{u} = \boldsymbol{o}. \tag{29.3.17}$$

Since $\boldsymbol{R}$ has full row rank and $\boldsymbol{X}$ full column rank and $\boldsymbol{\Psi}$ is nonsingular, $\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{R}^\top$ still has an inverse. Therefore

$$\boldsymbol{\lambda}^* = 2\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{u}) \tag{29.3.18}$$

Now substitute this $\boldsymbol{\lambda}^*$ back into (29.3.16):

$$\hat{\hat{\boldsymbol{\beta}}} = \hat{\boldsymbol{\beta}} - (\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{u}). \tag{29.3.19}$$

□

29.4. Constrained Least Squares as the Nesting of Two Simpler Models

The imposition of a constraint can also be considered the addition of new information: a certain linear transformation of $\boldsymbol{\beta}$, namely, $\boldsymbol{R\beta}$, is observed without error.

PROBLEM 343. *Assume the random* $\boldsymbol{\beta} \sim (\hat{\boldsymbol{\beta}}, \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1})$ *is unobserved, but one observes* $\boldsymbol{R\beta} = \mathbf{u}$.

• a. *2 points Compute the best linear predictor of* $\boldsymbol{\beta}$ *on the basis of the observation* $\mathbf{u}$. Hint: *First write down the joint means and covariance matrix of* $\mathbf{u}$ *and* $\boldsymbol{\beta}$.

ANSWER.

$$\begin{bmatrix}\mathbf{u}\\ \boldsymbol{\beta}\end{bmatrix} \sim \left(\begin{bmatrix}\boldsymbol{R\hat{\beta}}\\ \hat{\boldsymbol{\beta}}\end{bmatrix}, \sigma^2\begin{bmatrix}\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top & \boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\\ (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top & (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\end{bmatrix}\right). \tag{29.4.1}$$

Therefore application of formula (**??**) gives

$$\boldsymbol{\beta}^* = \hat{\boldsymbol{\beta}} + (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\left(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\right)^{-1}(\mathbf{u} - \boldsymbol{R\hat{\beta}}). \tag{29.4.2}$$

□

• b. *1 point Look at the formula for the predictor you just derived. Have you seen this formula before? Describe the situation in which this formula is valid as a BLUE-formula, and compare the situation with the situation here.*

ANSWER. Of course, constrained least squares. But in contrained least squares, $\boldsymbol{\beta}$ is nonrandom and $\hat{\boldsymbol{\beta}}$ is random, while here it is the other way round. □

In the unconstrained OLS model, i.e., before the "observation" of $\boldsymbol{u} = \boldsymbol{R\beta}$, the best bounded $\mathcal{MSE}$ estimators of $\boldsymbol{u}$ and $\boldsymbol{\beta}$ are $\boldsymbol{R\hat{\beta}}$ and $\hat{\boldsymbol{\beta}}$, with the sampling errors having the following means and variances:

$$\begin{bmatrix}\boldsymbol{u} - \boldsymbol{R\hat{\beta}}\\ \boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\end{bmatrix} \sim \left(\begin{bmatrix}\boldsymbol{o}\\ \boldsymbol{o}\end{bmatrix}, \sigma^2\begin{bmatrix}\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top & \boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\\ (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top & (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\end{bmatrix}\right) \tag{29.4.3}$$

After the observation of $\boldsymbol{u}$ we can therefore apply (27.1.18) to get exactly equation (29.3.13) for $\hat{\hat{\boldsymbol{\beta}}}$. This is probably the easiest way to derive this equation, but it derives constrained least squares by the minimization of the $\mathcal{MSE}$-matrix, not by the least squares problem.

29.5. Solution by Quadratic Decomposition

An alternative purely algebraic solution method for this constrained minimization problem rewrites the OLS objective function in such a way that one sees immediately what the constrained minimum value is.

Start with the decomposition (18.2.12) which can be used to show optimality of the OLS estimate:

$$(\mathbf{y} - \boldsymbol{X\beta})^\top(\mathbf{y} - \boldsymbol{X\beta}) = (\mathbf{y} - \boldsymbol{X\hat{\beta}})^\top(\mathbf{y} - \boldsymbol{X\hat{\beta}}) + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^\top\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}).$$

Split the second term again, using $\hat{\boldsymbol{\beta}} - \hat{\hat{\boldsymbol{\beta}}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\left(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\right)^{-1}(\boldsymbol{R\hat{\beta}} - \boldsymbol{u})$:

$$\begin{aligned}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^\top\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) &= \left(\boldsymbol{\beta} - \hat{\hat{\boldsymbol{\beta}}} - (\hat{\boldsymbol{\beta}} - \hat{\hat{\boldsymbol{\beta}}})\right)^\top\boldsymbol{X}^\top\boldsymbol{X}\left(\boldsymbol{\beta} - \hat{\hat{\boldsymbol{\beta}}} - (\hat{\boldsymbol{\beta}} - \hat{\hat{\boldsymbol{\beta}}})\right)\\ &= (\boldsymbol{\beta} - \hat{\hat{\boldsymbol{\beta}}})^\top\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{\beta} - \hat{\hat{\boldsymbol{\beta}}})\\ &\quad - 2(\boldsymbol{\beta} - \hat{\hat{\boldsymbol{\beta}}})^\top\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\left(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\right)^{-1}(\boldsymbol{R\hat{\beta}} - \boldsymbol{u})\\ &\quad + (\hat{\boldsymbol{\beta}} - \hat{\hat{\boldsymbol{\beta}}})^\top\boldsymbol{X}^\top\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \hat{\hat{\boldsymbol{\beta}}}).\end{aligned}$$

The cross product terms can be simplified to $-2(\boldsymbol{R\beta} - \boldsymbol{u})^\top\left(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\right)^{-1}(\boldsymbol{R\hat{\beta}} - \boldsymbol{u})$, and the last term is $(\boldsymbol{R\hat{\beta}} - \boldsymbol{u})^\top\left(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\right)^{-1}(\boldsymbol{R\hat{\beta}} - \boldsymbol{u})$. Therefore the objective function for an arbitrary $\boldsymbol{\beta}$ can be written as

$$\begin{aligned}(\mathbf{y} - \boldsymbol{X\beta})^\top(\mathbf{y} - \boldsymbol{X\beta}) &= (\mathbf{y} - \boldsymbol{X\hat{\beta}})^\top(\mathbf{y} - \boldsymbol{X\hat{\beta}})\\ &\quad + (\boldsymbol{\beta} - \hat{\hat{\boldsymbol{\beta}}})^\top\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{\beta} - \hat{\hat{\boldsymbol{\beta}}})\\ &\quad - 2(\boldsymbol{R\beta} - \boldsymbol{u})^\top\left(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\right)^{-1}(\boldsymbol{R\hat{\beta}} - \boldsymbol{u})\\ &\quad + (\boldsymbol{R\hat{\beta}} - \boldsymbol{u})^\top\left(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\right)^{-1}(\boldsymbol{R\hat{\beta}} - \boldsymbol{u})\end{aligned}$$

The first and last terms do not depend on $\boldsymbol{\beta}$ at all; the third term is zero whenever $\boldsymbol{\beta}$ satisfies $\boldsymbol{R\beta} = \boldsymbol{u}$; and the second term is minimized if and only if $\boldsymbol{\beta} = \hat{\hat{\boldsymbol{\beta}}}$, in which case it also takes the value zero.

29.6. Sampling Properties of Constrained Least Squares

Again, this variant of the least squares principle leads to estimators with desirable sampling properties. Note that $\hat{\hat{\boldsymbol{\beta}}}$ is an affine function of $\mathbf{y}$. We will compute $\mathcal{E}[\hat{\hat{\boldsymbol{\beta}}} - \boldsymbol{\beta}]$ and $\mathcal{MSE}[\hat{\hat{\boldsymbol{\beta}}}; \boldsymbol{\beta}]$ not only in the case that the true $\boldsymbol{\beta}$ satisfies $\boldsymbol{R\beta} = \boldsymbol{u}$, but also in the case that it does not. For this, let us first get a suitable representation of the

sampling error:

$$\begin{aligned}\hat{\hat{\boldsymbol{\beta}}}-\boldsymbol{\beta}&=(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})+(\hat{\hat{\boldsymbol{\beta}}}-\hat{\boldsymbol{\beta}})=\\ &=(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})-(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}\boldsymbol{R}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\\ &\quad-(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\boldsymbol{\beta}-\boldsymbol{u}).\end{aligned}\tag{29.6.1}$$

The last term is zero if $\boldsymbol{\beta}$ satisfies the constraint. Now use (24.0.7) twice to get

$$\hat{\hat{\boldsymbol{\beta}}}-\boldsymbol{\beta}=\boldsymbol{W}\boldsymbol{X}^\top\boldsymbol{\varepsilon}-(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\boldsymbol{\beta}-\boldsymbol{u})\tag{29.6.2}$$

where

$$\boldsymbol{W}=(\boldsymbol{X}^\top\boldsymbol{X})^{-1}-(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}.\tag{29.6.3}$$

If $\boldsymbol{\beta}$ satisfies the constraint, (29.6.2) simplifies to $\hat{\hat{\boldsymbol{\beta}}}-\boldsymbol{\beta}=\boldsymbol{W}\boldsymbol{X}^\top\boldsymbol{\varepsilon}$. In this case, therefore, $\hat{\hat{\boldsymbol{\beta}}}$ is unbiased and $\mathcal{MSE}[\hat{\hat{\boldsymbol{\beta}}};\boldsymbol{\beta}]=\sigma^2\boldsymbol{W}$ (Problem 344). Since $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}-\boldsymbol{W}$ is nonnegative definite, $\mathcal{MSE}[\hat{\hat{\boldsymbol{\beta}}};\boldsymbol{\beta}]$ is smaller than $\mathcal{MSE}[\hat{\boldsymbol{\beta}};\boldsymbol{\beta}]$ by a nonnegative definite matrix. This should be expected, since $\hat{\hat{\boldsymbol{\beta}}}$ uses more information than $\hat{\boldsymbol{\beta}}$.

Problem 344.

- a. *Show that $\boldsymbol{W}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{W}=\boldsymbol{W}$ (i.e., $\boldsymbol{X}^\top\boldsymbol{X}$ is a g-inverse of $\boldsymbol{W}$).*

Answer. This is a tedious matrix multiplication. □

- b. *Use this to show that $\mathcal{MSE}[\hat{\hat{\boldsymbol{\beta}}};\boldsymbol{\beta}]=\sigma^2\boldsymbol{W}$.*

(Without proof:) The Gauss-Markov theorem can be extended here as follows: the constrained least squares estimator is the best linear unbiased estimator among all linear (or, more precisely, affine) estimators which are unbiased whenever the true $\boldsymbol{\beta}$ satisfies the constraint $\boldsymbol{R}\boldsymbol{\beta}=\boldsymbol{u}$. Note that there are more estimators which are unbiased whenever the true $\boldsymbol{\beta}$ satisfies the constraint than there are estimators which are unbiased for all $\boldsymbol{\beta}$.

If $\boldsymbol{R}\boldsymbol{\beta}\neq\boldsymbol{u}$, then $\hat{\hat{\boldsymbol{\beta}}}$ is biased. Its bias is

$$\mathcal{E}[\hat{\hat{\boldsymbol{\beta}}}-\boldsymbol{\beta}]=-(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\boldsymbol{\beta}-\boldsymbol{u}).\tag{29.6.4}$$

Due to the decomposition (23.1.2) of the $\mathcal{MSE}$ matrix into dispersion matrix plus squared bias, it follows

$$\begin{aligned}\mathcal{MSE}[\hat{\hat{\boldsymbol{\beta}}};\boldsymbol{\beta}]&=\sigma^2\boldsymbol{W}+\\ &\quad+(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\boldsymbol{\beta}-\boldsymbol{u})\cdot\\ &\quad\cdot(\boldsymbol{R}\boldsymbol{\beta}-\boldsymbol{u})^\top\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}.\end{aligned}\tag{29.6.5}$$

Even if the true parameter does not satisfy the constraint, it is still possible that the constrained least squares estimator has a better $\mathcal{MSE}$ matrix than the unconstrained one. This is the case if and only if the true parameter values $\boldsymbol{\beta}$ and σ^2 satisfy

$$(\boldsymbol{R}\boldsymbol{\beta}-\boldsymbol{u})^\top\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\boldsymbol{\beta}-\boldsymbol{u})\leq\sigma^2.\tag{29.6.6}$$

This equation, which is the same as [**Gre97**, (8-27) on p. 406], is an interesting result, because the obvious estimate of the lefthand side in (29.6.6) is i times the value of the F-test statistic for the hypothesis $\boldsymbol{R}\boldsymbol{\beta}=\boldsymbol{u}$. To test for this, one has to use the *noncentral* F-test with parameters i, $n-k$, and $1/2$.

Problem 345. *2 points This Problem motivates Equation* (29.6.6). *If $\hat{\hat{\boldsymbol{\beta}}}$ is a better estimator of $\boldsymbol{\beta}$ than $\hat{\boldsymbol{\beta}}$, then $\boldsymbol{R}\hat{\hat{\boldsymbol{\beta}}}=\boldsymbol{u}$ is also a better estimator of $\boldsymbol{R}\boldsymbol{\beta}$ than $\boldsymbol{R}\hat{\boldsymbol{\beta}}$. Show that this latter condition is not only necessary but already sufficient, i.e., if $\mathcal{MSE}[\boldsymbol{R}\hat{\boldsymbol{\beta}};\boldsymbol{R}\boldsymbol{\beta}]-\mathcal{MSE}[\boldsymbol{u};\boldsymbol{R}\boldsymbol{\beta}]$ is nonnegative definite then $\boldsymbol{\beta}$ and σ^2 satisfy* (29.6.6). *You are allowed to use, without proof, theorem A.5.9 in the mathematical Appendix.*

Answer. We have to show

$$\sigma^2\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top-(\boldsymbol{R}\boldsymbol{\beta}-\boldsymbol{u})(\boldsymbol{R}\boldsymbol{\beta}-\boldsymbol{u})^\top\tag{29.6.7}$$

is nonnegative definite. Since $\boldsymbol{\Omega}=\sigma^2\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top$ has an inverse, theorem A.5.9 immediately leads to (29.6.6). □

29.7. Estimation of the Variance in Constrained OLS

Next we will compute the expected value of the minimum value of the constrained OLS objective funtion, i.e., $\mathrm{E}[\hat{\hat{\boldsymbol{\varepsilon}}}^\top\hat{\hat{\boldsymbol{\varepsilon}}}]$ where $\hat{\hat{\boldsymbol{\varepsilon}}}=\boldsymbol{y}-\boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}}$, again without necessarily making the assumption that $\boldsymbol{R}\boldsymbol{\beta}=\boldsymbol{u}$:

$$\hat{\hat{\boldsymbol{\varepsilon}}}=\boldsymbol{y}-\boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}}=\hat{\boldsymbol{\varepsilon}}+\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{u}).\tag{29.7.1}$$

Since $\boldsymbol{X}^\top\hat{\boldsymbol{\varepsilon}}=\boldsymbol{o}$, it follows

$$\hat{\hat{\boldsymbol{\varepsilon}}}^\top\hat{\hat{\boldsymbol{\varepsilon}}}=\hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}}+(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{u})^\top\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{u}).\tag{29.7.2}$$

Now note that $\mathcal{E}[\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{u}]=\boldsymbol{R}\boldsymbol{\beta}-\boldsymbol{u}$ and $\mathcal{V}[\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{u}]=\sigma^2\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top$. Therefore use (9.2.1) in theorem 9.2.1 and $\operatorname{tr}\Big(\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}\Big)=i$ to get

$$\text{(29.7.3)}\quad \mathrm{E}[(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{u})^\top\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{u})]= \\ =\sigma^2 i+(\boldsymbol{R}\boldsymbol{\beta}-\boldsymbol{u})^\top\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\boldsymbol{\beta}-\boldsymbol{u})$$

Since $\mathrm{E}[\hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}}]=\sigma^2(n-k)$, it follows

$$\text{(29.7.4)}\qquad \mathrm{E}[\hat{\hat{\boldsymbol{\varepsilon}}}^\top\hat{\hat{\boldsymbol{\varepsilon}}}]=\sigma^2(n+i-k)+(\boldsymbol{R}\boldsymbol{\beta}-\boldsymbol{u})^\top\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\boldsymbol{\beta}-\boldsymbol{u}).$$

In other words, $\hat{\hat{\boldsymbol{\varepsilon}}}^\top\hat{\hat{\boldsymbol{\varepsilon}}}/(n+i-k)$ is an unbiased estimator of σ^2 if the constraint holds, and it is biased upwards if the constraint does not hold. The adjustment of the degrees of freedom is what one should expect: a regression with k explanatory variables and i constraints can always be rewritten as a regression with $k-i$ different explanatory variables (see Section 29.2), and the distribution of the SSE does not depend on the values taken by the explanatory variables at all, only on how many there are. The unbiased estimate of σ^2 is therefore

$$\text{(29.7.5)}\qquad \hat{\hat{\sigma}}^2=\hat{\hat{\boldsymbol{\varepsilon}}}^\top\hat{\hat{\boldsymbol{\varepsilon}}}/(n+i-k)$$

Here is some geometric intuition: $\mathbf{y}=\boldsymbol{X}\hat{\boldsymbol{\beta}}+\hat{\boldsymbol{\varepsilon}}$ is an orthogonal decomposition, since $\hat{\boldsymbol{\varepsilon}}$ is orthogonal to all columns of $\boldsymbol{X}$. From orthogonality follows $\mathbf{y}^\top\mathbf{y}=\hat{\boldsymbol{\beta}}^\top\boldsymbol{X}^\top\boldsymbol{X}\hat{\boldsymbol{\beta}}+\hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}}$. If one splits up $\mathbf{y}=\boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}}+\hat{\hat{\boldsymbol{\varepsilon}}}$, one should expect this to be orthogonal as well. But this is only the case if $\boldsymbol{u}=\boldsymbol{o}$. If $\boldsymbol{u}\neq\boldsymbol{o}$, one first has to shift the origin of the coordinate system to a point which can be written in the form $\boldsymbol{X}\boldsymbol{\beta}_0$ where $\boldsymbol{\beta}_0$ satisfies the constraint:

Problem 346. *3 points Assume $\hat{\hat{\boldsymbol{\beta}}}$ is the constrained least squares estimate, and $\boldsymbol{\beta}_0$ is any vector satisfying $\boldsymbol{R}\boldsymbol{\beta}_0=\boldsymbol{u}$. Show that in the decomposition*

$$\text{(29.7.6)}\qquad \mathbf{y}-\boldsymbol{X}\boldsymbol{\beta}_0=\boldsymbol{X}(\hat{\hat{\boldsymbol{\beta}}}-\boldsymbol{\beta}_0)+\hat{\hat{\boldsymbol{\varepsilon}}}$$

the two vectors on the righthand side are orthogonal.

Answer. We have to show $(\hat{\hat{\boldsymbol{\beta}}}-\boldsymbol{\beta}_0)^\top\boldsymbol{X}^\top\hat{\hat{\boldsymbol{\varepsilon}}}=0$. Since $\hat{\hat{\boldsymbol{\varepsilon}}}=\mathbf{y}-\boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}}=\mathbf{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}+\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\hat{\hat{\boldsymbol{\beta}}})=\hat{\boldsymbol{\varepsilon}}+\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\hat{\hat{\boldsymbol{\beta}}})$, and we already know that $\boldsymbol{X}^\top\hat{\boldsymbol{\varepsilon}}=\boldsymbol{o}$, it is necessary and sufficient to show that $(\hat{\hat{\boldsymbol{\beta}}}-\boldsymbol{\beta}_0)^\top\boldsymbol{X}^\top\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\hat{\hat{\boldsymbol{\beta}}})=0$. By (29.3.13),

$$(\hat{\hat{\boldsymbol{\beta}}}-\boldsymbol{\beta}_0)^\top\boldsymbol{X}^\top\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\hat{\hat{\boldsymbol{\beta}}})=(\hat{\hat{\boldsymbol{\beta}}}-\boldsymbol{\beta}_0)^\top\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{u}) \\ =(\boldsymbol{u}-\boldsymbol{u})^\top\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{u})=0.$$

□

If $\boldsymbol{u}=\boldsymbol{o}$, then one has two orthogonal decompositions: $\mathbf{y}=\hat{\mathbf{y}}+\hat{\boldsymbol{\varepsilon}}$, and $\mathbf{y}=\hat{\hat{\mathbf{y}}}+\hat{\hat{\boldsymbol{\varepsilon}}}$. And if one connects the footpoints of these two orthogonal decompositions, one obtains an orthogonal decomposition into three parts:

Problem 347. *Assume $\hat{\hat{\boldsymbol{\beta}}}$ is the constrained least squares estimator subject to the constraint $\boldsymbol{R}\boldsymbol{\beta}=\boldsymbol{o}$, and $\hat{\boldsymbol{\beta}}$ is the unconstrained least squares estimator.*

• a. *1 point With the usual notation $\hat{\mathbf{y}}=\boldsymbol{X}\hat{\boldsymbol{\beta}}$ and $\hat{\hat{\mathbf{y}}}=\boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}}$, show that*

$$\text{(29.7.7)}\qquad \mathbf{y}=\hat{\hat{\mathbf{y}}}+(\hat{\mathbf{y}}-\hat{\hat{\mathbf{y}}})+\hat{\boldsymbol{\varepsilon}}$$

Point out these vectors in the `reggeom` *simulation.*

Answer. In the `reggeom`-simulation, $\mathbf{y}$ is the purple line; $\boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}}$ is the red line starting at the origin, one could also call it $\hat{\hat{\mathbf{y}}}$; $\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\hat{\hat{\boldsymbol{\beta}}})=\hat{\mathbf{y}}-\hat{\hat{\mathbf{y}}}$ is the light blue line, and $\hat{\boldsymbol{\varepsilon}}$ is the green line which does not start at the origin. In other words: if one projects $\mathbf{y}$ on a plane, and also on a line in that plane, and then connects the footpoints of these two projections, one obtains a zig-zag line with two right angles. □

• b. *4 points Show that in* (29.7.7) *the three vectors $\hat{\hat{\mathbf{y}}}$, $\hat{\mathbf{y}}-\hat{\hat{\mathbf{y}}}$, and $\hat{\boldsymbol{\varepsilon}}$ are orthogonal. You are allowed to use, without proof, formula* (29.3.13)*:*

Answer. One has to verify that the scalar products of the three vectors on the right hand side of (29.7.7) are zero. $\hat{\hat{\mathbf{y}}}^\top\hat{\boldsymbol{\varepsilon}}=\hat{\hat{\boldsymbol{\beta}}}^\top\boldsymbol{X}^\top\hat{\boldsymbol{\varepsilon}}=0$ and $(\hat{\mathbf{y}}-\hat{\hat{\mathbf{y}}})^\top\hat{\boldsymbol{\varepsilon}}=(\hat{\boldsymbol{\beta}}-\hat{\hat{\boldsymbol{\beta}}})^\top\boldsymbol{X}^\top\hat{\boldsymbol{\varepsilon}}=0$ follow from $\boldsymbol{X}^\top\hat{\boldsymbol{\varepsilon}}=\boldsymbol{o}$; geometrically on can simply say that $\hat{\mathbf{y}}$ and $\hat{\hat{\mathbf{y}}}$ are in the space spanned by the columns of $\boldsymbol{X}$, and $\hat{\boldsymbol{\varepsilon}}$ is orthogonal to that space. Finally, using (29.3.13) for $\hat{\boldsymbol{\beta}}-\hat{\hat{\boldsymbol{\beta}}}$,

$$\hat{\hat{\mathbf{y}}}^\top(\hat{\mathbf{y}}-\hat{\hat{\mathbf{y}}})=\hat{\hat{\boldsymbol{\beta}}}^\top\boldsymbol{X}^\top\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\hat{\hat{\boldsymbol{\beta}}})= \\ =\hat{\hat{\boldsymbol{\beta}}}^\top\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}\boldsymbol{R}\hat{\boldsymbol{\beta}}= \\ =\hat{\hat{\boldsymbol{\beta}}}^\top\boldsymbol{R}^\top\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}\boldsymbol{R}\hat{\boldsymbol{\beta}}=0$$

because $\hat{\hat{\boldsymbol{\beta}}}$ satisfies the constraint $\boldsymbol{R}\hat{\hat{\boldsymbol{\beta}}}=\boldsymbol{o}$, hence $\hat{\hat{\boldsymbol{\beta}}}^\top\boldsymbol{R}^\top=\boldsymbol{o}^\top$. □

Problem 348.

• a. *3 points In the model $\mathbf{y}=\boldsymbol{\beta}+\boldsymbol{\varepsilon}$, where $\mathbf{y}$ is a $n\times 1$ vector, and $\boldsymbol{\varepsilon}\sim(\boldsymbol{o},\sigma^2\boldsymbol{I})$, subject to the constraint $\boldsymbol{\iota}^\top\boldsymbol{\beta}=0$, compute $\hat{\hat{\boldsymbol{\beta}}}$, $\hat{\hat{\boldsymbol{\varepsilon}}}$, and the unbiased estimate $\hat{\hat{\sigma}}^2$. Give general formulas and the numerical results for the case $\mathbf{y}^\top=\begin{bmatrix}-1 & 0 & 1 & 2\end{bmatrix}$. All you need to do is evaluate the appropriate formulas and correctly count the number of degrees of freedom.*

Answer. The unconstrained least squares estimate of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}}=\mathbf{y}$, and since $\boldsymbol{X}=\boldsymbol{I}$, $\boldsymbol{R}=\boldsymbol{\iota}^\top$, and $\boldsymbol{u}=0$, the constrained LSE has the form $\hat{\hat{\boldsymbol{\beta}}}=\mathbf{y}-\boldsymbol{\iota}(\boldsymbol{\iota}^\top\boldsymbol{\iota})^{-1}(\boldsymbol{\iota}^\top\mathbf{y})=\mathbf{y}-\boldsymbol{\iota}\bar{\mathbf{y}}$ by (29.3.13). If $\mathbf{y}^\top=[-1,0,1,2]$ this gives $\hat{\hat{\boldsymbol{\beta}}}^\top=[-1.5,-0.5,0.5,1.5]$. The residuals in the constrained model are

therefore $\hat{\hat{\varepsilon}} = \iota\bar{y}$, i.e., $\hat{\hat{\varepsilon}} = [0.5, 0.5, 0.5, 0.5]$. Since one has n observations, n parameters and 1 constraint, the number of degrees of freedom is 1. Therefore $\hat{\hat{\sigma}}^2 = \hat{\hat{\varepsilon}}^\top \hat{\hat{\varepsilon}}/1 = n\bar{y}^2$ which is $= 1$ in our case. □

• b. *1 point Can you think of a practical situation in which this model might be appropriate?*

Answer. This can occur if one measures data which theoretically add to zero, and the measurement errors are independent and have equal standard deviations. □

• c. *2 points Check your results against a SAS printout (or do it in any other statistical package) with the data vector* $\mathbf{y}^\top = [-1\ 0\ 1\ 2]$. *Here are the sas commands:*

```
data zeromean;
input y x1 x2 x3 x4;
cards;
-1 1 0 0 0
 0 0 1 0 0
 1 0 0 1 0
 2 0 0 0 1
;
proc reg;
model y= x1 x2 x3 x4 /
noint;
restrict x1+x2+x3+x4=0;
output out=zerout
residual=ehat;
run;
proc print data=zerout;
run;
```

Problem 349. *Least squares estimates of the coefficients of a linear regression model often have signs that are regarded by the researcher to be 'wrong'. In an effort to obtain the 'right' signs, the researcher may be tempted to drop statistically insignificant variables from the equation.* [**Lea75**] *showed that such attempts necessarily fail: there can be no change in sign of any coefficient which is more significant than the coefficient of the omitted variable. The present exercise shows this, using a different proof than Leamer's. You will need the formula for the constrained least squares estimator subject to one linear constraint* $\boldsymbol{r}^\top\boldsymbol{\beta} = u$, *which is*

$$\hat{\hat{\boldsymbol{\beta}}} = \hat{\boldsymbol{\beta}} - \boldsymbol{V}\boldsymbol{r}\big(\boldsymbol{r}^\top\boldsymbol{V}\boldsymbol{r}\big)^{-1}(\boldsymbol{r}^\top\hat{\boldsymbol{\beta}} - u). \tag{29.7.8}$$

where $\boldsymbol{V} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}$.

• a. *In order to assess the sensitivity of the estimate of any linear combination of the elements of* $\boldsymbol{\beta}$, $\phi = \boldsymbol{t}^\top\boldsymbol{\beta}$, *due to imposition of the constraint, it makes sense to divide the change* $\boldsymbol{t}^\top\hat{\boldsymbol{\beta}} - \boldsymbol{t}^\top\hat{\hat{\boldsymbol{\beta}}}$ *by the standard deviation of* $\boldsymbol{t}^\top\hat{\boldsymbol{\beta}}$, *i.e., to look at*

$$\frac{\boldsymbol{t}^\top(\hat{\boldsymbol{\beta}} - \hat{\hat{\boldsymbol{\beta}}})}{\sigma\sqrt{\boldsymbol{t}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{t}}}. \tag{29.7.9}$$

Such a standardization allows you to compare the sensitivity of different linear combinations. Show that that linear combination of the elements of $\hat{\boldsymbol{\beta}}$ *which is affected most if one imposes the constraint* $\boldsymbol{r}^\top\boldsymbol{\beta} = u$ *is the constraint* $\boldsymbol{t} = \boldsymbol{r}$ *itself. If this value is small, then no other linear combination of the elements of* $\hat{\boldsymbol{\beta}}$ *will be affected much by the imposition of the constraint either.*

Answer. Using (29.7.8) and equation (32.4.1) one obtains

$$\max_{\boldsymbol{t}} \frac{(\boldsymbol{t}^\top(\hat{\boldsymbol{\beta}} - \hat{\hat{\boldsymbol{\beta}}}))^2}{\sigma^2\boldsymbol{t}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{t}} = \frac{(\hat{\boldsymbol{\beta}} - \hat{\hat{\boldsymbol{\beta}}})^\top\boldsymbol{X}^\top\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \hat{\hat{\boldsymbol{\beta}}})}{\sigma^2} =$$

$$= \frac{(\boldsymbol{r}^\top\hat{\boldsymbol{\beta}} - u)^\top\big(\boldsymbol{r}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{r}\big)^{-1}(\boldsymbol{r}^\top\hat{\boldsymbol{\beta}} - u)}{\sigma^2} = \frac{(\boldsymbol{r}^\top\hat{\boldsymbol{\beta}} - u)^2}{\sigma^2\boldsymbol{r}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{r}}$$

□

29.8. Inequality Restrictions

With linear inequality restrictions, it makes sense to have $\boldsymbol{R}$ of deficient rank, these are like two different half planes in the same plane, and the restrictions define a quarter plane, or a triangle, etc.

One obvious approach would be: compute the unrestricted estimator, see what restrictions it violates, and apply these restrictions with equality. But this equality restricted estimator may then suddenly violate other restrictions.

One brute force approach would be: impose all combinations of restrictions and see if the so partially restricted parameter satisfies the other restrictions too; and among those that do, choose the one with the lowest SSE.

[**Gre97**, 8.5.3 on pp. 411/12] has good discussion. The inequality restricted estimator is biased, unless the true parameter value satisfies all inequality restrictions with equality. It is always a mixture between the unbiased $\hat{\boldsymbol{\beta}}$ and some restricted estimator which is biased if this condition does not hold.

Its variance is always smaller than that of $\hat{\boldsymbol{\beta}}$ but, incredibly, its MSE will sometimes be larger than that of $\hat{\boldsymbol{\beta}}$. Don't understand how this comes about.

29.9. Application: Biased Estimators and Pre-Test Estimators

The formulas about Constrained Least Squares which were just derived suggest that it is sometimes advantageous (in terms of MSE) to impose constraints even if they do not really hold. In other words, one should not put all explanatory variables into a regression which have an influence, but only the main ones. A logical extension of this idea is the common practice of first testing whether some variables have significant influence and dropping the variables if they do not. These so-called pre-test estimators are very common. [**DM93**, Chapter 3.7, pp. 94–98] says something about them. Pre-test estimation this seems a good procedure, but the graph regarding MSE shows it is not: the pre-test estimator never has lowest MSE, and it has highest MSE exactly in the area where it is most likely to be applied.

CHAPTER 30

Additional Regressors

A good detailed explanation of the topics covered in this chapter is [**DM93**, pp. 19–24]. [**DM93**] use the addition of variables as their main paradigm for going from a more restrictive to a less restrictive model.

In this chapter, the usual regression model is given in the form

$$\mathbf{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon} = \begin{bmatrix}\boldsymbol{X}_1 & \boldsymbol{X}_2\end{bmatrix}\begin{bmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\end{bmatrix} + \boldsymbol{\varepsilon} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim (\boldsymbol{o}, \sigma^2\boldsymbol{I}) \tag{30.0.1}$$

where $\boldsymbol{X} = \begin{bmatrix}\boldsymbol{X}_1 & \boldsymbol{X}_2\end{bmatrix}$ has full column rank, and the coefficient vector is $\boldsymbol{\beta} = \begin{bmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\end{bmatrix}$.

We take a sequential approach to this regression. First we regress $\mathbf{y}$ on $\boldsymbol{X}_1$ alone, which gives the regression coefficient $\hat{\hat{\boldsymbol{\beta}}}_1$. This by itself is an inconsistent estimator of $\boldsymbol{\beta}_1$, but we will use it as a stepping stone towards the full regression. We make use of the information gained by the regression on $\boldsymbol{X}_1$ in our computation of the full regression. Such a sequential approach may be appropriate in the following situations:

- If regression on $\boldsymbol{X}_1$ is much simpler than the combined regression, for instance if $\boldsymbol{X}_1$ contains dummy or trend variables, and the dataset is large. Example: model (64.3.4).
- If we want to fit the regressors in $\boldsymbol{X}_2$ by graphical methods and those in $\boldsymbol{X}_1$ by analytical methods (added variable plots).
- If we need an estimate of $\boldsymbol{\beta}_2$ but are not interested in an estimate of $\boldsymbol{\beta}_1$.
- If we want to test the joint significance of the regressors in $\boldsymbol{X}_2$, while $\boldsymbol{X}_1$ consists of regressors not being tested.

If one regresses $\mathbf{y}$ on $\boldsymbol{X}_1$, one gets $\mathbf{y} = \boldsymbol{X}_1\hat{\hat{\boldsymbol{\beta}}}_1 + \hat{\hat{\boldsymbol{\varepsilon}}}$. Of course, $\hat{\hat{\boldsymbol{\beta}}}_1$ is an inconsistent estimator of $\boldsymbol{\beta}_1$, since some explanatory variables are left out. And $\hat{\hat{\boldsymbol{\varepsilon}}}$ is orthogonal to $\boldsymbol{X}_1$ but not to $\boldsymbol{X}_2$.

The iterative "backfitting" method proceeds from here as follows: it regresses $\hat{\hat{\boldsymbol{\varepsilon}}}$ on $\boldsymbol{X}_2$, which gives another residual, which is again orthogonal on $\boldsymbol{X}_2$ but no longer orthogonal on $\boldsymbol{X}_1$. Then this new residual is regressed on $\boldsymbol{X}_1$ again, etc.

Problem 350. *The purpose of this Problem is to get a graphical intuition of the issues in sequential regression. Make sure the stand-alone program* `xgobi` *is installed on your computer (in Debian GNU-Linux do* `apt-get install xgobi`*), and the* R*-interface* `xgobi` *is installed (the* R*-command is simply* `install.packages("xgobi")`*, or, on a Debian system the preferred argument is* `install.packages("xgobi", lib = "/usr/lib/R/library")`*). You have to give the commands* `library(xgobi)` *and then* `reggeom()`*. This produces a graph in the* `XGobi` *window which looks like* [**DM93**, Figure 3b on p. 22]*. If you switch from the* `XYPlot` *view to the* `Rotation` *view, you will see the same lines rotating 3-dimensionally, and you can interact with this graph. You will see that this graph shows the dependent variable* $\mathbf{y}$*, the regression of* $\mathbf{y}$ *on* $\boldsymbol{x}_1$*, and the regression of* $\mathbf{y}$ *on* $\boldsymbol{x}_1$ *and* $\boldsymbol{x}_2$*.*

- a. *1 point In order to show that you have correctly identified which line is* $\mathbf{y}$*, please answer the following two questions: Which color is* $\mathbf{y}$*: red, yellow, light blue, dark blue, green, purple, or white? If it is yellow, also answer the question: Is it that yellow line which is in part covered by a red line, or is it the other one? If it is red, green, or dark blue, also answer the question: Does it start at the origin or not?*

- b. *1 point Now answer the same two questions about* $\boldsymbol{x}_1$*.*

- c. *1 point Now answer the same two questions about* $\boldsymbol{x}_2$*.*

- d. *1 point Now answer the same two questions about* $\hat{\hat{\boldsymbol{\varepsilon}}}$*, the residual in the regression of* $\mathbf{y}$ *on* $\boldsymbol{x}_1$*.*

- e. *Now assume* $\boldsymbol{x}_1$ *is the vector of ones. The* R^2 *of this regression is a ratio of the squared lengths of two of the lines in the regression. Which lines?*

- f. *2 points If one regresses* $\hat{\hat{\boldsymbol{\varepsilon}}}$ *on* $\boldsymbol{x}_2$*, one gets a decomposition* $\hat{\hat{\boldsymbol{\varepsilon}}} = \mathbf{h} + \mathbf{k}$*, where* $\mathbf{h}$ *is a multiple of* $\boldsymbol{x}_2$ *and* $\mathbf{k}$ *orthogonal to* $\boldsymbol{x}_2$*. This is the next step in the backfitting algorithm. Draw this decomposition into the diagram. The points are already invisibly present. Therefore you should use the line editor to connect the points. You may want to increase the magnification scale of the figure for this. (In my version of* `XGobi`*, I often lose lines if I try to add more lines. This seems to be a bug which will probably be fixed eventually.) Which label does the corner point of the decomposition have? Make a geometric argument that the new residual* $\mathbf{k}$ *is no longer orthogonal to* $\boldsymbol{x}_2$*.*

- g. *1 point The next step in the backfitting procedure is to regress* $\mathbf{k}$ *on* $\boldsymbol{x}_1$*. The corner point for this decomposition is again invisibly in the animation. Identify the two endpoints of the residual in this regression. Hint: the* R*-command* `example(reggeom)` *produces a modified version of the animation in which the backfitting prodecure is highlighted. The successive residuals which are used as regressors*

are drawn in dark blue, and the quickly improving approximations to the fitted value are connected by a red zig-zag line.

• h. *1 point The diagram contains the points for two more backfitting steps. Identify the endpoints of both residuals.*

• i. *2 points Of the five cornerpoints obtained by simple regressions, c, p, q, r, and s, three lie on one straight line, and the other two on a different straight line, with the intersection of these straight lines being the corner point in the multiple regression of* $\mathbf{y}$ *on* $\boldsymbol{x}_1$ *and* $\boldsymbol{x}_2$*. Which three points are on the same line, and how can these two lines be characterized?*

• j. *1 point Of the lines cp, pq, qr, and rs, two are parallel to* $\boldsymbol{x}_1$*, and two parallel to* $\boldsymbol{x}_2$*. Which two are parallel to* $\boldsymbol{x}_1$*?*

• k. *1 point Draw in the regression of* $\mathbf{y}$ *on* $\boldsymbol{x}_2$*.*

• l. *3 points Which two variables are plotted against each other in an added-variable plot for* $\boldsymbol{x}_2$*?*

Here are the coordinates of some of the points in this animation:

$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\mathbf{y}$	$\hat{\mathbf{y}}$	$\hat{\hat{\mathbf{y}}}$
5	-1	3	3	3
0	4	3	3	0
0	0	4	0	0

In the dataset which R submits to `XGobi`, all coordinates are multiplied by 1156, which has the effect that all the points included in the animation have integer coordinates.

Problem 351. *2 points How do you know that the decomposition* $\begin{bmatrix}3\\3\\4\end{bmatrix} = \begin{bmatrix}3\\0\\0\end{bmatrix} + \begin{bmatrix}0\\3\\4\end{bmatrix}$ *is* $\mathbf{y} = \hat{\hat{\mathbf{y}}} + \hat{\hat{\boldsymbol{\varepsilon}}}$ *in the regression of* $\mathbf{y} = \begin{bmatrix}3\\3\\4\end{bmatrix}$ *on* $\boldsymbol{x}_1 = \begin{bmatrix}5\\0\\0\end{bmatrix}$*?*

Answer. Besides the equation $\mathbf{y} = \hat{\hat{\mathbf{y}}} + \hat{\hat{\boldsymbol{\varepsilon}}}$ we have to check two things: (1) $\hat{\hat{\mathbf{y}}}$ is a linear combination of all the explanatory variables (here: is a multiple of $\boldsymbol{x}_1$), and (2) $\hat{\hat{\boldsymbol{\varepsilon}}}$ is orthogonal to all explanatory variables. Compare Problem **??**. □

Problem 352. *3 points In the same way, check that the decomposition* $\begin{bmatrix}3\\3\\4\end{bmatrix} = \begin{bmatrix}3\\3\\0\end{bmatrix} + \begin{bmatrix}0\\0\\4\end{bmatrix}$ *is* $\mathbf{y} = \hat{\mathbf{y}} + \boldsymbol{\varepsilon}$ *in the regression of* $\mathbf{y} = \begin{bmatrix}3\\3\\4\end{bmatrix}$ *on* $\boldsymbol{x}_1 = \begin{bmatrix}5\\0\\0\end{bmatrix}$ *and* $\boldsymbol{x}_2 = \begin{bmatrix}-1\\4\\0\end{bmatrix}$.

Answer. Besides the equation $\mathbf{y} = \hat{\hat{\mathbf{y}}} + \hat{\hat{\boldsymbol{\varepsilon}}}$ we have to check two things: (1) $\hat{\hat{\mathbf{y}}}$ is a linear combination of all the explanatory variables. Since both $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ have zero as third coordinate, and they are linearly independent, they span the whole plane, therefore $\hat{\mathbf{y}}$, which also has the third coordinate zero, is their linear combination. (2) $\hat{\hat{\boldsymbol{\varepsilon}}}$ is orthogonal to both explanatory variables because its only nonzero coordinate is the third. □

The residuals $\hat{\hat{\boldsymbol{\varepsilon}}}$ in the regression on $\boldsymbol{x}_1$ are $\mathbf{y} - \hat{\hat{\mathbf{y}}} = \begin{bmatrix}3\\3\\4\end{bmatrix} - \begin{bmatrix}3\\0\\0\end{bmatrix} = \begin{bmatrix}0\\3\\4\end{bmatrix}$. This vector is clearly orthogonal to $\boldsymbol{x}_1 = \begin{bmatrix}3\\0\\0\end{bmatrix}$. Now let us regress $\hat{\hat{\boldsymbol{\varepsilon}}} = \begin{bmatrix}0\\3\\4\end{bmatrix}$ on $\boldsymbol{x}_2 = \begin{bmatrix}-1\\4\\0\end{bmatrix}$. Say $\boldsymbol{h}$ is the vector of fitted values and $\boldsymbol{k}$ the residual vector in this regression. We saw in problem 350 that this is the next step in backfitting, but $\boldsymbol{k}$ is not the same as the residual vector $\hat{\boldsymbol{\varepsilon}}$ in the multiple regression, because $\boldsymbol{k}$ is not orthogonal to $\boldsymbol{x}_1$. In order to get the correct residual in the joint regression and also the correct coefficient of $\boldsymbol{x}_2$, one must regress $\hat{\hat{\boldsymbol{\varepsilon}}}$ only on that part of $\boldsymbol{x}_2$ which is orthogonal to $\boldsymbol{x}_1$. This regressor is the dark blue line starting at the origin.

In formulas: One gets the correct $\hat{\boldsymbol{\varepsilon}}$ and $\hat{\boldsymbol{\beta}}_2$ by regressingx $\hat{\hat{\boldsymbol{\varepsilon}}} = \boldsymbol{M}_1\mathbf{y}$ not on $\boldsymbol{X}_2$ but on $\boldsymbol{M}_1\boldsymbol{X}_2$, where $\boldsymbol{M}_1 = \boldsymbol{I} - \boldsymbol{X}_1(\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top$ is the matrix which forms the residuals under the regression on $\boldsymbol{X}_1$. In other words, one has to remove the influence of $\boldsymbol{X}_1$ not only from the dependent but also the independent variables. Instead of regressing the residuals $\hat{\hat{\boldsymbol{\varepsilon}}} = \boldsymbol{M}_1\mathbf{y}$ on $\boldsymbol{X}_2$, one has to regress them on what is new about $\boldsymbol{X}_2$ after we know $\boldsymbol{X}_1$, i.e., on what remains of $\boldsymbol{X}_2$ after taking out the effect of $\boldsymbol{X}_1$, which is $\boldsymbol{M}_1\boldsymbol{X}_2$. The regression which gets the correct $\hat{\boldsymbol{\beta}}_2$ is therefore

$$\boldsymbol{M}_1\mathbf{y} = \boldsymbol{M}_1\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 + \hat{\boldsymbol{\varepsilon}} \tag{30.0.2}$$

In formulas, the correct $\hat{\boldsymbol{\beta}}_2$ is

$$\hat{\boldsymbol{\beta}}_2 = (\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top\boldsymbol{M}_1\mathbf{y}. \tag{30.0.3}$$

This regression also yields the correct covariance matrix. (The only thing which is not right is the number of degrees of freedom). The regression is therefore fully representative of the additional effect of $\boldsymbol{x}_2$, and the plot of $\hat{\hat{\boldsymbol{\varepsilon}}}$ against $\boldsymbol{M}_1\boldsymbol{X}_2$ with the fitted line drawn (which has the correct slope $\hat{\boldsymbol{\beta}}_2$) is called the "added variable plot" for $\boldsymbol{X}_2$. [**CW99**, pp. 244–246] has a good discussion of added variable plots.

Problem 353. *2 points Show that in the model* (30.0.1)*, the estimator* $\hat{\boldsymbol{\beta}}_2 = (\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top\boldsymbol{M}_1\mathbf{y}$ *is unbiased. Compute* $\mathcal{MSE}[\hat{\boldsymbol{\beta}}_2;\boldsymbol{\beta}_2]$.

Answer. $\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2 = (\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top\boldsymbol{M}_1(\boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}) - \boldsymbol{\beta}_2 = (\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{\varepsilon}$; therefore $\mathcal{MSE}[\hat{\boldsymbol{\beta}}_2;\boldsymbol{\beta}_2] = \sigma^2(\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{M}_1^\top\boldsymbol{X}_2(\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1} = \sigma^2(\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}$. □

In order to get an estimate of $\hat{\boldsymbol{\beta}}_1$, one can again do what seems intuitive, namely, regress $\mathbf{y} - \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2$ on $\boldsymbol{X}_1$. This gives

$$\hat{\boldsymbol{\beta}}_1 = (\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top(\mathbf{y} - \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2). \tag{30.0.4}$$

This regression also gives the right residuals, but not the right estimates of the covariance matrix.

PROBLEM 354. *The three Figures in* [**DM93**, p. 22] *can be seen in* `XGobi` *if you use the instructions in Problem 350. The purple line represents the dependent variable* $\mathbf{y}$*, and the two yellow lines the explanatory variables* $\boldsymbol{x}_1$ *and* $\boldsymbol{x}_2$*. (*$\boldsymbol{x}_1$ *is the one which is in part red.) The two green lines represent the unconstrained regression* $\mathbf{y} = \hat{\mathbf{y}} + \hat{\boldsymbol{\varepsilon}}$*, and the two red lines the constrained regression* $\mathbf{y} = \hat{\hat{\mathbf{y}}} + \hat{\hat{\boldsymbol{\varepsilon}}}$ *where* $\mathbf{y}$ *is only regressed on* $\boldsymbol{x}_1$*. The two dark blue lines, barely visible against the dark blue background, represent the regression of* $\boldsymbol{x}_2$ *on* $\boldsymbol{x}_1$*.*

• a. *The first diagram which* `XGobi` *shows on startup is* [**DM93**, diagram (b) on p. 22]*. Go into the* `Rotation` *view and rotate the diagram in such a way that the view is* [**DM93**, Figure (a)]*. You may want to delete the two white lines, since they are not shown in Figure (a).*

• b. *Make a geometric argument that the light blue line, which represents* $\hat{\mathbf{y}} - \hat{\hat{\mathbf{y}}} = \boldsymbol{X}(\hat{\boldsymbol{\beta}} - \hat{\hat{\boldsymbol{\beta}}})$*, is orthogonal on the green line* $\hat{\boldsymbol{\varepsilon}}$ *(this is the green line which ends at the point* $\mathbf{y}$*, i.e., not the green line which starts at the origin).*

ANSWER. The light blue line lies in the plane spanned by $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, and $\hat{\boldsymbol{\varepsilon}}$ is orthogonal to this plane. □

• c. *Make a geometric argument that the light blue line is also orthogonal to the red line* $\hat{\hat{\mathbf{y}}}$ *emanating from the origin.*

ANSWER. This is a little trickier. The red line $\hat{\hat{\boldsymbol{\varepsilon}}}$ is orthogonal to $\boldsymbol{x}_1$, and the green line $\hat{\boldsymbol{\varepsilon}}$ is also orthogonal to $\boldsymbol{x}_1$. Together, $\hat{\boldsymbol{\varepsilon}}$ and $\hat{\hat{\boldsymbol{\varepsilon}}}$ span therefore the plane orthogonal to $\boldsymbol{x}_1$. Since the light blue line lies in the plane spanned by $\hat{\boldsymbol{\varepsilon}}$ and $\hat{\hat{\boldsymbol{\varepsilon}}}$, it is orthogonal to $\boldsymbol{x}_1$. □

Question 354 shows that the decomposition $\mathbf{y} = \hat{\hat{\mathbf{y}}} + (\hat{\mathbf{y}} - \hat{\hat{\mathbf{y}}}) + \hat{\boldsymbol{\varepsilon}}$ is orthogonal, i.e., all 3 vectors $\hat{\hat{\mathbf{y}}}$, $\hat{\mathbf{y}} - \hat{\hat{\mathbf{y}}}$, and $\hat{\boldsymbol{\varepsilon}}$ are orthogonal to each other. This is (29.7.6) in the special case that $\boldsymbol{u} = \boldsymbol{o}$ and therefore $\boldsymbol{\beta}_0 = \boldsymbol{o}$.

One can use this same animation also to show the following: If you first project the purple line on the plane spanned by the yellow lines, you get the green line in the plane. If you then project that green line on $\boldsymbol{x}_1$, which is a subspace of the plane, then you get the red section of the yellow line. This is the same result as if you had projected the purple line directly on $\boldsymbol{x}_1$. A matrix-algebraic proof of this fact is given in (A.6.3).

The same animation allows us to verify the following:

- In the regression of $\mathbf{y}$ on $\boldsymbol{x}_1$, the coefficient is $\hat{\hat{\beta}}_1$, and the residual is $\hat{\hat{\boldsymbol{\varepsilon}}}$.
- In the regression of $\mathbf{y}$ on $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, the coefficients are $\hat{\beta}_1$, $\hat{\beta}_2$, and the residual is $\hat{\boldsymbol{\varepsilon}}$.
- In the regression of $\mathbf{y}$ on $\boldsymbol{x}_1$ and $\boldsymbol{M}_1\boldsymbol{x}_2$, the coefficients are $\hat{\hat{\beta}}_1$, $\hat{\beta}_2$, and the residual is $\hat{\boldsymbol{\varepsilon}}$. The residual is $\hat{\boldsymbol{\varepsilon}}$ because the space spanned by the regressors is the same as in the regression on $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, and $\hat{\boldsymbol{\varepsilon}}$ only depends on that space.
- In the regression of $\mathbf{y}$ on $\boldsymbol{M}_1\boldsymbol{x}_2$, the coefficient is $\hat{\beta}_2$, because the regressor I am leaving out is orthogonal to $\boldsymbol{M}_1\boldsymbol{x}_2$. The residual contains the contribution of the left-out variable, i.e., it is $\hat{\boldsymbol{\varepsilon}} + \hat{\beta}_1\boldsymbol{x}_1$.
- But in the regression of $\hat{\hat{\boldsymbol{\varepsilon}}} = \boldsymbol{M}_1\mathbf{y}$ on $\boldsymbol{M}_1\boldsymbol{x}_2$, the coefficient is $\hat{\beta}_2$ and the residual $\hat{\boldsymbol{\varepsilon}}$.

This last statement is (30.0.3).

Now let us turn to proving all this mathematically. The "brute force" proof, i.e., the proof which is conceptually simplest but has to plow through some tedious mathematics, uses (18.2.4) with partitioned matrix inverses. For this we need (30.0.5).

PROBLEM 355. *4 points This is a simplified version of question 593. Show the following, by multiplying* $\boldsymbol{X}^\top\boldsymbol{X}$ *with its alleged inverse: If* $\boldsymbol{X} = \begin{bmatrix}\boldsymbol{X}_1 & \boldsymbol{X}_2\end{bmatrix}$ *has full column rank, then* $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$ *is the following partitioned matrix:*

(30.0.5)
$$\begin{bmatrix}\boldsymbol{X}_1^\top\boldsymbol{X}_1 & \boldsymbol{X}_1^\top\boldsymbol{X}_2\\ \boldsymbol{X}_2^\top\boldsymbol{X}_1 & \boldsymbol{X}_2^\top\boldsymbol{X}_2\end{bmatrix}^{-1} = \begin{bmatrix}(\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1} + \boldsymbol{K}_1^\top\boldsymbol{X}_2(\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top\boldsymbol{K}_1 & -\boldsymbol{K}_1^\top\boldsymbol{X}_2(\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}\\ -(\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top\boldsymbol{K}_1 & (\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\end{bmatrix}$$

where $\boldsymbol{M}_1 = \boldsymbol{I} - \boldsymbol{X}_1(\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top$ *and* $\boldsymbol{K}_1 = \boldsymbol{X}_1(\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1}$*.*

From (30.0.5) one sees that the covariance matrix in regression (30.0.3) is the lower left partition of the covariance matrix in the full regression (30.0.1).

PROBLEM 356. *6 points Use the usual formula* $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\mathbf{y}$ *together with* (30.0.5) *to prove* (30.0.3) *and* (30.0.4)*.*

ANSWER. (18.2.4) reads here

(30.0.6)
$$\begin{bmatrix}\hat{\beta}_1\\ \hat{\beta}_2\end{bmatrix} = \begin{bmatrix}(\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1} + \boldsymbol{K}_1^\top\boldsymbol{X}_2(\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top\boldsymbol{K}_1 & -\boldsymbol{K}_1^\top\boldsymbol{X}_2(\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\\ -(\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top\boldsymbol{K}_1 & (\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\end{bmatrix}\begin{bmatrix}\boldsymbol{X}_1^\top\mathbf{y}\\ \boldsymbol{X}_2^\top\mathbf{y}\end{bmatrix}$$

Since $\boldsymbol{M}_1 = \boldsymbol{I} - \boldsymbol{K}_1\boldsymbol{X}_1^\top$, one can simplify

(30.0.7)
$$\hat{\boldsymbol{\beta}}_2 = -(\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top\boldsymbol{K}_1\boldsymbol{X}_1^\top\mathfrak{y} + (\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top\mathfrak{y}$$

(30.0.8)
$$= (\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top\boldsymbol{M}\mathfrak{y}$$

(30.0.9)
$$\hat{\boldsymbol{\beta}}_1 = (\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top\mathfrak{y} + \boldsymbol{K}_1^\top\boldsymbol{X}_2(\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top\boldsymbol{K}_1\boldsymbol{X}_1^\top\mathfrak{y} - \boldsymbol{K}_1^\top\boldsymbol{X}_2(\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top\mathfrak{y}$$

(30.0.10)
$$= \boldsymbol{K}_1^\top\mathfrak{y} - \boldsymbol{K}_1^\top\boldsymbol{X}_2(\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top(\boldsymbol{I} - \boldsymbol{K}_1\boldsymbol{X}_1^\top)\mathfrak{y}$$

(30.0.11)
$$= \boldsymbol{K}_1^\top\mathfrak{y} - \boldsymbol{K}_1^\top\boldsymbol{X}_2(\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top\boldsymbol{M}_1\mathfrak{y}$$

(30.0.12)
$$= \boldsymbol{K}_1^\top(\mathfrak{y} - \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2)$$

□

[**Gre97**, pp. 245–7] follow a different proof strategy: he solves the partitioned normal equations

$$\begin{bmatrix}\boldsymbol{X}_1^\top\boldsymbol{X}_1 & \boldsymbol{X}_1^\top\boldsymbol{X}_2\\ \boldsymbol{X}_2^\top\boldsymbol{X}_1 & \boldsymbol{X}_2^\top\boldsymbol{X}_2\end{bmatrix}\begin{bmatrix}\hat{\boldsymbol{\beta}}_1\\ \hat{\boldsymbol{\beta}}_2\end{bmatrix}\begin{bmatrix}\boldsymbol{X}_1^\top\mathfrak{y}\\ \boldsymbol{X}_2^\top\mathfrak{y}\end{bmatrix} \tag{30.0.13}$$

directly, without going through the inverse. A third proof strategy, used by [**Seb77**, pp. 65–72], is followed in Problems 358 and 359.

Problem 357. *5 points* [**Gre97**, problem 18 on p. 326]. *The following matrix gives the slope in the simple regression of the column variable on the row variable:*

(30.0.14)

$\boldsymbol{y}$	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	
1	0.03	0.36	$\boldsymbol{y}$
0.4	1	0.3	$\boldsymbol{x}_1$
1.2	0.075	1	$\boldsymbol{x}_2$

For example, if $\boldsymbol{y}$ is regressed on $\boldsymbol{x}_1$, the slope is 0.4, but if $\boldsymbol{x}_1$ is regressed on $\boldsymbol{y}$, the slope is 0.03. All variables have zero means, so the constant terms in all regressions are zero. What are the two slope coefficients in the multiple regression of $\boldsymbol{y}$ on $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$? Hint: Use the partitioned normal equation as given in [**Gre97**, p. 245] *in the special case when each of the partitions of $\boldsymbol{X}$ has only one colum.*

Answer.

$$\begin{bmatrix}\boldsymbol{x}_1^\top\boldsymbol{x}_1 & \boldsymbol{x}_1^\top\boldsymbol{x}_2\\ \boldsymbol{x}_2^\top\boldsymbol{x}_1 & \boldsymbol{x}_2^\top\boldsymbol{x}_2\end{bmatrix}\begin{bmatrix}\hat{\beta}_1\\ \hat{\beta}_2\end{bmatrix} = \begin{bmatrix}\boldsymbol{x}_1^\top\mathfrak{y}\\ \boldsymbol{x}_2^\top\mathfrak{y}\end{bmatrix} \tag{30.0.15}$$

The first row reads

$$\hat{\beta}_1 + (\boldsymbol{x}_1^\top\boldsymbol{x}_1)^{-1}\boldsymbol{x}_1^\top\boldsymbol{x}_2\hat{\beta}_2 = (\boldsymbol{x}_1^\top\boldsymbol{x}_1)^{-1}\boldsymbol{x}_1^\top\boldsymbol{y} \tag{30.0.16}$$

which is the upper line of [**Gre97**, (6.24) on p, 245], and in our numbers this is $\hat{\beta}_1 = 0.4 - 0.3\hat{\beta}_2$. The second row reads

$$(\boldsymbol{x}_2^\top\boldsymbol{x}_2)^{-1}\boldsymbol{x}_2^\top\boldsymbol{x}_1\hat{\beta}_1 + \hat{\beta}_2 = (\boldsymbol{x}_2^\top\boldsymbol{x}_2)^{-1}\boldsymbol{x}_2^\top\boldsymbol{y} \tag{30.0.17}$$

or in our numbers $0.075\hat{\beta}_2 + \hat{\beta}_2 = 1.2$. Plugging in the formula for $\hat{\beta}_1$ gives $0.075 \cdot 0.4 - 0.075 \cdot 0.3\hat{\beta}_2 + \hat{\beta}_2 = 1.2$. This gives $\hat{\beta}_2 = 1.17/0.9775 = 1.196931 = 1.2$ roughly, and $\hat{\beta}_1 = 0.4 - 0.36 = 0.0409207 = 0.041$ roughly. □

Problem 358. *Derive* (30.0.3) *and* (30.0.4) *from the first order conditions for minimizing*

$$(\mathfrak{y} - \boldsymbol{X}_1\boldsymbol{\beta}_1 - \boldsymbol{X}_2\boldsymbol{\beta}_2)^\top(\mathfrak{y} - \boldsymbol{X}_1\boldsymbol{\beta}_1 - \boldsymbol{X}_2\boldsymbol{\beta}_2). \tag{30.0.18}$$

Answer. Start by writing down the OLS objective function for the full model. Perhaps we can use the more sophisticated matrix differentiation rules?

(30.0.19)
$$(\mathfrak{y}-\boldsymbol{X}_1\boldsymbol{\beta}_1-\boldsymbol{X}_2\boldsymbol{\beta}_2)^\top(\mathfrak{y}-\boldsymbol{X}_1\boldsymbol{\beta}_1-\boldsymbol{X}_2\boldsymbol{\beta}_2) = \mathfrak{y}^\top\mathfrak{y}+\boldsymbol{\beta}_1^\top\boldsymbol{X}_1^\top\boldsymbol{X}_1\boldsymbol{\beta}_1+\boldsymbol{\beta}_2^\top\boldsymbol{X}_2^\top\boldsymbol{X}_2\boldsymbol{\beta}_2-2\mathfrak{y}^\top\boldsymbol{X}_1\boldsymbol{\beta}_1-2\mathfrak{y}^\top\boldsymbol{X}_2\boldsymbol{\beta}_2+2\boldsymbol{\beta}_2^\top\boldsymbol{X}$$

Taking partial derivatives with respect to $\boldsymbol{\beta}_1^\top$ and $\boldsymbol{\beta}_2^\top$ gives

(30.0.20)
$$2\boldsymbol{\beta}_1^\top\boldsymbol{X}_1^\top\boldsymbol{X}_1 - 2\mathfrak{y}^\top\boldsymbol{X}_1 + 2\boldsymbol{\beta}_2^\top\boldsymbol{X}_2^\top\boldsymbol{X}_1 \quad \text{or, transposed} \quad 2\boldsymbol{X}_1^\top\boldsymbol{X}_1\boldsymbol{\beta}_1 - 2\boldsymbol{X}_1^\top\mathfrak{y} + 2\boldsymbol{X}_1^\top\boldsymbol{X}_2\boldsymbol{\beta}_2$$

(30.0.21)
$$2\boldsymbol{\beta}_2^\top\boldsymbol{X}_2^\top\boldsymbol{X}_2 - 2\mathfrak{y}^\top\boldsymbol{X}_2^\top + 2\boldsymbol{\beta}_1^\top\boldsymbol{X}_1^\top\boldsymbol{X}_2 \quad \text{or, transposed} \quad 2\boldsymbol{X}_2^\top\boldsymbol{X}_2\boldsymbol{\beta}_2 - 2\boldsymbol{X}_2^\top\mathfrak{y} + 2\boldsymbol{X}_2^\top\boldsymbol{X}_1\boldsymbol{\beta}_1$$

Setting them zero and replacing $\boldsymbol{\beta}_1$ by $\hat{\boldsymbol{\beta}}_1$ and $\boldsymbol{\beta}_2$ by $\hat{\boldsymbol{\beta}}_2$ gives

$$\boldsymbol{X}_1^\top\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 = \boldsymbol{X}_1^\top(\mathfrak{y} - \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2) \tag{30.0.22}$$

$$\boldsymbol{X}_2^\top\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 = \boldsymbol{X}_2^\top(\mathfrak{y} - \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1). \tag{30.0.23}$$

Premultiply (30.0.22) by $\boldsymbol{X}_1(\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1}$:

$$\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 = \boldsymbol{X}_1(\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top(\mathfrak{y} - \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2). \tag{30.0.24}$$

Plug this into (30.0.23):

$$\boldsymbol{X}_2^\top\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 = \boldsymbol{X}_2^\top\Big(\mathfrak{y} - \boldsymbol{X}_1(\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top\mathfrak{y} + \boldsymbol{X}_1(\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2\Big) \tag{30.0.25}$$

$$\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 = \boldsymbol{X}_2^\top\boldsymbol{M}_1\mathfrak{y}. \tag{30.0.26}$$

(30.0.26) is the normal equation of the regression of $\boldsymbol{M}_1\mathfrak{y}$ on $\boldsymbol{M}_1\boldsymbol{X}_2$; it immediately implies (30.0.3). Once $\hat{\boldsymbol{\beta}}_2$ is known, (30.0.22) is the normal equation of the regression of $\mathfrak{y} - \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2$ on $\boldsymbol{X}_1$, which gives (30.0.4). □

Problem 359. *Using* (30.0.3) *and* (30.0.4) *show that the residuals in regression* (30.0.1) *are identical to those in the regression of $\boldsymbol{M}_1\mathfrak{y}$ on $\boldsymbol{M}_1\boldsymbol{X}_2$.*

Answer.

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 - \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 \tag{30.0.27}$$

$$= \mathbf{y} - \boldsymbol{X}_1(\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top(\mathbf{y} - \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2) - \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 \tag{30.0.28}$$

$$= \boldsymbol{M}_1\mathbf{y} - \boldsymbol{M}_1\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2. \tag{30.0.29}$$

□

Problem 360. *The following problem derives one of the main formulas for adding regressors, following* [**DM93**, pp. 19–24]. *We are working in model* (30.0.1).

• a. *1 point Show that, if $\boldsymbol{X}$ has full column rank, then $\boldsymbol{X}^\top\boldsymbol{X}$, $\boldsymbol{X}_1^\top\boldsymbol{X}_1$, and $\boldsymbol{X}_2^\top\boldsymbol{X}_2$ are nonsingular. Hint: A matrix $\boldsymbol{X}$ has full column rank if $\boldsymbol{X}\boldsymbol{a} = \boldsymbol{o}$ implies $\boldsymbol{a} = \boldsymbol{o}$.*

Answer. From $\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{a} = \boldsymbol{o}$ follows $\boldsymbol{a}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{a} = 0$ which can also be written $\|\boldsymbol{X}\boldsymbol{a}\| = 0$. Therefore $\boldsymbol{X}\boldsymbol{a} = \boldsymbol{o}$, and since the columns are linearly independent, it follows $\boldsymbol{a} = \boldsymbol{o}$. $\boldsymbol{X}_1^\top\boldsymbol{X}_1$ and $\boldsymbol{X}_2^\top\boldsymbol{X}_2$ are nonsingular because, along with $\boldsymbol{X}$, also $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ have full column rank. □

• b. *1 point Define $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$ and $\boldsymbol{M}_1 = \boldsymbol{I} - \boldsymbol{X}_1(\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top$. Show that both $\boldsymbol{M}$ and $\boldsymbol{M}_1$ are projection matrices. (Give the definition of a projection matrix.) Which spaces do they project on? Which space is bigger?*

Answer. A projection matrix is symmetric and idempotent. That $\boldsymbol{M}\boldsymbol{M} = \boldsymbol{M}$ is easily verified. $\boldsymbol{M}$ projects on the orthogonal complement of the column space of $\boldsymbol{X}$, and $\boldsymbol{M}_1$ on that of $\boldsymbol{X}_1$. I.e., $\boldsymbol{M}_1$ projects on the larger space. □

• c. *2 points Prove that $\boldsymbol{M}_1\boldsymbol{M} = \boldsymbol{M}$ and that $\boldsymbol{M}\boldsymbol{X}_1 = \boldsymbol{O}$ as well as $\boldsymbol{M}\boldsymbol{X}_2 = \boldsymbol{O}$. You will need each these equationse below. What is their geometric meaning?*

Answer. $\boldsymbol{X}_1 = \begin{bmatrix}\boldsymbol{X}_1 & \boldsymbol{X}_2\end{bmatrix}\begin{bmatrix}\boldsymbol{I}\\ \boldsymbol{O}\end{bmatrix} = \boldsymbol{X}\boldsymbol{A}$, say. Therefore $\boldsymbol{M}_1\boldsymbol{M} = (\boldsymbol{I} - \boldsymbol{X}\boldsymbol{A}(\boldsymbol{A}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{A})^{-1}\boldsymbol{A}^\top\boldsymbol{X}^\top)$ $\boldsymbol{M}$ because $\boldsymbol{X}^\top\boldsymbol{M} = \boldsymbol{O}$. Geometrically this means that the space on which $\boldsymbol{M}$ projects is a subspace of the space on which $\boldsymbol{M}_1$ projects. To show that $\boldsymbol{M}\boldsymbol{X}_2 = \boldsymbol{O}$ note that $\boldsymbol{X}_2$ can be written in the form $\boldsymbol{X}_2 = \boldsymbol{X}\boldsymbol{B}$, too; this time, $\boldsymbol{B} = \begin{bmatrix}\boldsymbol{O}\\ \boldsymbol{I}\end{bmatrix}$. $\boldsymbol{M}\boldsymbol{X}_2 = \boldsymbol{O}$ means geometrically that $\boldsymbol{M}$ projects on a space that is orthogonal to all columns of $\boldsymbol{X}_2$. □

• d. *2 points Show that $\boldsymbol{M}_1\boldsymbol{X}_2$ has full column rank.*

Answer. If $\boldsymbol{M}_1\boldsymbol{X}_2\boldsymbol{b} = \boldsymbol{o}$, then $\boldsymbol{X}_2\boldsymbol{b} = \boldsymbol{X}_1\boldsymbol{a}$ for some $\boldsymbol{a}$. We showed this in Problem 234. Therefore $\begin{bmatrix}\boldsymbol{X}_1 & \boldsymbol{X}_2\end{bmatrix}\begin{bmatrix}-\boldsymbol{a}\\ \boldsymbol{b}\end{bmatrix} = \begin{bmatrix}\boldsymbol{o}\\ \boldsymbol{o}\end{bmatrix}$, and since $\begin{bmatrix}\boldsymbol{X}_1 & \boldsymbol{X}_2\end{bmatrix}$ has full column rank, it follows $\begin{bmatrix}-\boldsymbol{a}\\ \boldsymbol{b}\end{bmatrix} = \begin{bmatrix}\boldsymbol{o}\\ \boldsymbol{o}\end{bmatrix}$, in particular $\boldsymbol{b} = \boldsymbol{o}$. □

• e. *1 point Here is some more notation: the regression of $\mathbf{y}$ on $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ can also be represented by the equation*

$$\mathbf{y} = \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 + \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 + \hat{\boldsymbol{\varepsilon}} \tag{30.0.30}$$

The difference between (30.0.1) *and* (30.0.30) *is that* (30.0.30) *contains the parameter estimates, not their true values, and the residuals, not the true disturbances. Explain the difference between residuals and disturbances, and between the fitted regression line and the true regression line.*

• f. *1 point Verify that premultiplication of* (30.0.30) *by $\boldsymbol{M}_1$ gives*

$$\boldsymbol{M}_1\mathbf{y} = \boldsymbol{M}_1\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 + \hat{\boldsymbol{\varepsilon}} \tag{30.0.31}$$

Answer. We need $\boldsymbol{M}_1\boldsymbol{X}_1 = \boldsymbol{O}$ and $\boldsymbol{M}_1\hat{\boldsymbol{\varepsilon}} = \boldsymbol{M}_1\boldsymbol{M}\mathbf{y} = \boldsymbol{M}\mathbf{y} = \hat{\boldsymbol{\varepsilon}}$ (or this can also besseen because $\boldsymbol{X}_1^\top\hat{\boldsymbol{\varepsilon}} = \boldsymbol{o}$). □

• g. *2 points Prove that* (30.0.31) *is the fit which one gets if one regresses $\boldsymbol{M}_1\mathbf{y}$ on $\boldsymbol{M}_1\boldsymbol{X}_2$. In other words, if one runs OLS with dependent variable $\boldsymbol{M}_1\mathbf{y}$ and explanatory variables $\boldsymbol{M}_1\boldsymbol{X}_2$, one gets the same $\hat{\boldsymbol{\beta}}_2$ and $\hat{\boldsymbol{\varepsilon}}$ as in* (30.0.31)*, which are the same $\hat{\boldsymbol{\beta}}_2$ and $\hat{\boldsymbol{\varepsilon}}$ as in the complete regression* (30.0.30).

Answer. According to Problem **??** we have to check $\boldsymbol{X}_2^\top\boldsymbol{M}_1\hat{\boldsymbol{\varepsilon}} = \boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{M}\mathbf{y} = \boldsymbol{X}_2^\top\boldsymbol{M}\mathbf{y} = \boldsymbol{O}\mathbf{y} = \boldsymbol{o}$. □

• h. *1 point Show that $\mathcal{V}[\hat{\boldsymbol{\beta}}_2] = (\boldsymbol{X}_2\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}$. Are the variance estimates and confidence intervals valid, which the computer automatically prints out if one regresses $\boldsymbol{M}_1\mathbf{y}$ on $\boldsymbol{M}_1\boldsymbol{X}_2$?*

Answer. Yes except for the number of degrees of freedom. □

• i. *4 points If one premultiplies* (30.0.1) *by $\boldsymbol{M}_1$, one obtains*

$$\boldsymbol{M}_1\mathbf{y} = \boldsymbol{M}_1\boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{M}_1\boldsymbol{\varepsilon}, \qquad \boldsymbol{M}_1\boldsymbol{\varepsilon} \sim (\boldsymbol{o}, \sigma^2\boldsymbol{M}_1) \tag{30.0.32}$$

Although the covariance matrix of the disturbance $\boldsymbol{M}_1\boldsymbol{\varepsilon}$ in (30.0.32) *is no longer spherical, show that nevertheless the $\hat{\boldsymbol{\beta}}_2$ obtained by running OLS on* (30.0.32) *is the BLUE of $\boldsymbol{\beta}_2$ based on the information given in* (30.0.32) *(i.e., assuming that $\boldsymbol{M}_1\mathbf{y}$ and $\boldsymbol{M}_1\boldsymbol{X}_2$ are known, but not necessarily $\boldsymbol{M}_1$, $\mathbf{y}$, and $\boldsymbol{X}_2$ separately). Hint: this proof is almost identical to the proof that for spherically distributed disturbances the OLS is BLUE (e.g. given in* [**DM93**, p. 159]*), but you have to add some $\boldsymbol{M}_1$'s to your formulas.*

Answer. Any other linear estimator $\tilde{\boldsymbol{\gamma}}$ of $\boldsymbol{\beta}_2$ can be written as $\tilde{\boldsymbol{\gamma}} = \big((\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top + \boldsymbol{C}\big)\boldsymbol{M}_1\mathbf{y}$. Its expected value is $\mathrm{E}[\tilde{\boldsymbol{\gamma}}] = (\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{C}\boldsymbol{M}_1\boldsymbol{X}_2\boldsymbol{\beta}_2$. For $\tilde{\boldsymbol{\gamma}}$ to be unbiased, regardless of the value of $\boldsymbol{\beta}_2$, $\boldsymbol{C}$ must satisfy $\boldsymbol{C}\boldsymbol{M}_1\boldsymbol{X}_2 = \boldsymbol{O}$. From this follows $\mathcal{MSE}[\tilde{\boldsymbol{\gamma}};\boldsymbol{\beta}_2] = \mathcal{V}[\tilde{\boldsymbol{\gamma}}] = \sigma^2\big((\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top + \boldsymbol{C}\big)\boldsymbol{M}_1\big(\boldsymbol{X}_2(\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1} + \boldsymbol{C}^\top\big) = \sigma^2(\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1} + \sigma^2\boldsymbol{C}\boldsymbol{M}_1\boldsymbol{C}^\top$, i.e., it exceeds the $\mathcal{MSE}$-matrix of $\hat{\boldsymbol{\beta}}$ by a nonnegative definite matrix. Is it unique? The *formula*

for the BLUE is not unique, since one can add any $\boldsymbol{C}$ with $\boldsymbol{C}\boldsymbol{M}_1\boldsymbol{C}^\top = \boldsymbol{O}$ or equivalently $\boldsymbol{C}\boldsymbol{M}_1 = \boldsymbol{O}$ or $\boldsymbol{C} = \boldsymbol{A}\boldsymbol{X}$ for some $\boldsymbol{A}$. However such a $\boldsymbol{C}$ applied to a dependent variable of the form $\boldsymbol{M}_1\mathsf{y}$ will give the null vector, therefore the *values* of the BLUE for those values of y which are possible are indeed unique.

□

• j. *1 point Once $\hat{\boldsymbol{\beta}}_2$ is known, one can move it to the left hand side in* (30.0.30) *to get*

$$\mathsf{y} - \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 = \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 + \hat{\boldsymbol{\varepsilon}} \tag{30.0.33}$$

Prove that one gets the right values of $\hat{\boldsymbol{\beta}}_1$ and of $\hat{\boldsymbol{\varepsilon}}$ if one regresses $\mathsf{y} - \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2$ on $\boldsymbol{X}_1$.

Answer. The simplest answer just observes that $\boldsymbol{X}_1^\top\hat{\boldsymbol{\varepsilon}} = \boldsymbol{o}$. Or: The normal equation for this pseudo-regression is $\boldsymbol{X}_1^\top\mathsf{y} - \boldsymbol{X}_1^\top\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 = \boldsymbol{X}_1^\top\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1$, which holds due to the normal equation for the full model. □

• k. *1 point Does* (30.0.33) *also give the right covariance matrix for $\hat{\boldsymbol{\beta}}_1$?*

Answer. No, since $\mathsf{y} - \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2$ has a different covariance matrix than $\sigma^2\boldsymbol{I}$. □

This following Problems gives some applications of the results in Problem 360. You are allowed to use the results of Problem 360 without proof.

Problem 361. *Assume your regression involves an intercept, i.e., the matrix of regressors is $\begin{bmatrix}\boldsymbol{\iota} & \boldsymbol{X}\end{bmatrix}$, where $\boldsymbol{X}$ is the matrix of the "true" explanatory variables with no vector of ones built in, and $\boldsymbol{\iota}$ the vector of ones. The regression can therefore be written*

$$\mathsf{y} = \boldsymbol{\iota}\alpha + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{30.0.34}$$

• a. *1 point Show that the OLS estimate of the slope parameters $\boldsymbol{\beta}$ can be obtained by regressing $\underline{\mathsf{y}}$ on $\underline{\boldsymbol{X}}$ without intercept, where $\underline{\mathsf{y}}$ and $\underline{\boldsymbol{X}}$ are the variables with their means taken out, i.e., $\underline{\mathsf{y}} = \boldsymbol{D}\mathsf{y}$ and $\underline{\boldsymbol{X}} = \boldsymbol{D}\boldsymbol{X}$, with $\boldsymbol{D} = \boldsymbol{I} - \frac{1}{n}\boldsymbol{\iota}\boldsymbol{\iota}^\top$.*

Answer. This is called the "sweeping out of means." It follows immediately from (30.0.3). This is the usual procedure to do regression with a constant term: in simple regression $\mathsf{y}_i = \alpha + \beta x_i + \varepsilon_i$, (30.0.3) is equation (18.2.22):

$$\hat{\beta} = \frac{\sum(x_i - \bar{x})(\mathsf{y}_i - \bar{\mathsf{y}})}{\sum(x_i - \bar{x})^2}. \tag{30.0.35}$$

□

• b. *Show that the OLS estimate of the intercept is $\hat{\alpha} = \bar{\mathsf{y}} - \bar{\boldsymbol{x}}^\top\hat{\boldsymbol{\beta}}$ where $\bar{\boldsymbol{x}}^\top$ is the row vector of column means of $\boldsymbol{X}$, i.e., $\bar{\boldsymbol{x}}^\top = \frac{1}{n}\boldsymbol{\iota}^\top\boldsymbol{X}$.*

Answer. This is exactly (30.0.4). Here is a more specific argument: The intercept $\hat{\alpha}$ is obtained by regressing $\mathsf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ on $\boldsymbol{\iota}$. The normal equation for this second regression is $\boldsymbol{\iota}^\top\mathsf{y} - \boldsymbol{\iota}^\top\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{\iota}^\top\boldsymbol{\iota}\hat{\alpha}$. If $\bar{\mathsf{y}}$ is the mean of y, and $\bar{\boldsymbol{x}}^\top$ the row vector consisting of means of the colums of $\boldsymbol{X}$, then this gives $\bar{\mathsf{y}} = \bar{\boldsymbol{x}}^\top\hat{\boldsymbol{\beta}} + \hat{\alpha}$. In the case of simple regression, this was derived earlier as formula (18.2.23). □

• c. *2 points Show that $\mathcal{MSE}[\hat{\boldsymbol{\beta}};\boldsymbol{\beta}] = \sigma^2(\underline{\boldsymbol{X}}^\top\underline{\boldsymbol{X}})^{-1}$. (Use the formula for $\hat{\boldsymbol{\beta}}$.)*

Answer. Since

$$\begin{bmatrix}\boldsymbol{\iota}^\top \\ \boldsymbol{X}^\top\end{bmatrix}\begin{bmatrix}\boldsymbol{\iota} & \boldsymbol{X}\end{bmatrix} = \begin{bmatrix} n & n\bar{\boldsymbol{x}}^\top \\ \bar{\boldsymbol{x}}n & \boldsymbol{X}^\top\boldsymbol{X}\end{bmatrix}, \tag{30.0.36}$$

it follows by Problem 593

$$\left(\begin{bmatrix}\boldsymbol{\iota}^\top \\ \boldsymbol{X}^\top\end{bmatrix}\begin{bmatrix}\boldsymbol{\iota} & \boldsymbol{X}\end{bmatrix}\right)^{-1} = \begin{bmatrix} 1/n + \bar{\boldsymbol{x}}^\top(\underline{\boldsymbol{X}}^\top\underline{\boldsymbol{X}})^{-1}\bar{\boldsymbol{x}} & -\bar{\boldsymbol{x}}^\top(\underline{\boldsymbol{X}}^\top\underline{\boldsymbol{X}})^{-1} \\ -(\underline{\boldsymbol{X}}^\top\underline{\boldsymbol{X}})^{-1}\bar{\boldsymbol{x}} & (\underline{\boldsymbol{X}}^\top\underline{\boldsymbol{X}})^{-1}\end{bmatrix} \tag{30.0.37}$$

In other words, one simply does as if the actual regressors had been the data with their means removed, and then takes the inverse of that design matrix. The only place where on has to be careful is the number of degrees of freedom. See also Seber [**Seb77**, section 11.7] about centering and scaling the data. □

• d. *3 points Show that $\hat{\mathsf{y}} - \boldsymbol{\iota}\bar{\mathsf{y}} = \underline{\boldsymbol{X}}\hat{\boldsymbol{\beta}}$.*

Answer. First note that $\boldsymbol{X} = \underline{\boldsymbol{X}} + \frac{1}{n}\boldsymbol{\iota}\boldsymbol{\iota}^\top\boldsymbol{X} = \underline{\boldsymbol{X}} + \boldsymbol{\iota}\bar{\boldsymbol{x}}^\top$ where $\bar{\boldsymbol{x}}^\top$ is the row vector of means of $\boldsymbol{X}$. By definition, $\hat{\mathsf{y}} = \boldsymbol{\iota}\hat{\alpha} + \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{\iota}\hat{\alpha} + \underline{\boldsymbol{X}}\hat{\boldsymbol{\beta}} + \boldsymbol{\iota}\bar{\boldsymbol{x}}^\top\hat{\boldsymbol{\beta}} = \boldsymbol{\iota}(\hat{\alpha} + \bar{\boldsymbol{x}}^\top\hat{\boldsymbol{\beta}}) + \underline{\boldsymbol{X}}\hat{\boldsymbol{\beta}} = \boldsymbol{\iota}\bar{\mathsf{y}} + \underline{\boldsymbol{X}}\hat{\boldsymbol{\beta}}$. □

• e. *2 points Show that* $\mathsf{R}^2 = \dfrac{\underline{\mathsf{y}}^\top\underline{\boldsymbol{X}}(\underline{\boldsymbol{X}}^\top\underline{\boldsymbol{X}})^{-1}\underline{\boldsymbol{X}}^\top\underline{\mathsf{y}}}{\underline{\mathsf{y}}^\top\underline{\mathsf{y}}}$

Answer.

$$R^2 = \frac{(\hat{\mathsf{y}} - \bar{\mathsf{y}}\boldsymbol{\iota})^\top(\hat{\mathsf{y}} - \bar{\mathsf{y}}\boldsymbol{\iota})}{\underline{\mathsf{y}}^\top\underline{\mathsf{y}}} = \frac{\hat{\boldsymbol{\beta}}^\top\underline{\boldsymbol{X}}^\top\underline{\boldsymbol{X}}\hat{\boldsymbol{\beta}}}{\underline{\mathsf{y}}^\top\underline{\mathsf{y}}} \tag{30.0.38}$$

and now plugging in the formula for $\hat{\boldsymbol{\beta}}$ the result follows. □

• f. *3 points Now, split once more $\underline{\boldsymbol{X}} = \begin{bmatrix}\underline{\boldsymbol{X}}_1 & \underline{\boldsymbol{x}}_2\end{bmatrix}$ where the second partition $\underline{\boldsymbol{x}}_2$ consists of one column only, and $\underline{\boldsymbol{X}}$ is, as above, the $\boldsymbol{X}$ matrix with the column means taken out. Conformably, $\hat{\boldsymbol{\beta}} = \begin{bmatrix}\hat{\boldsymbol{\beta}}_1 \\ \hat{\beta}_2\end{bmatrix}$. Show that*

$$\operatorname{var}[\hat{\beta}_2] = \frac{\sigma^2}{\underline{\boldsymbol{x}}^\top\underline{\boldsymbol{x}}}\frac{1}{(1 - R^2_{2.})} \tag{30.0.39}$$

where $R^2_{2.}$ is the R^2 in the regression of $\boldsymbol{x}_2$ on all other variables in $\boldsymbol{X}$. This is in [**Gre97**, (9.3) on p. 421]. *Hint: you should first show that* $\operatorname{var}[\hat{\beta}_2] = \sigma^2/\underline{\boldsymbol{x}}_2^\top\underline{\boldsymbol{M}}_1\underline{\boldsymbol{x}}_2$ *where* $\underline{\boldsymbol{M}}_1 = \boldsymbol{I} - \underline{\boldsymbol{X}}_1(\underline{\boldsymbol{X}}_1^\top\underline{\boldsymbol{X}}_1)^{-1}\underline{\boldsymbol{X}}_1^\top$. *Here is an interpretation of* (30.0.39) *which you don't have to prove: $\sigma^2/\underline{\boldsymbol{x}}^\top\underline{\boldsymbol{x}}$ is the variance in a simple regression with a constant*

term and $\boldsymbol{x}_2$ as the only explanatory variable, and $1/(1-R^2_{2\cdot})$ is called the variance inflation factor.

Answer. Note that we are not talking about the variance of the constant term but that of all the other terms.

$$\underline{\boldsymbol{x}}_2^\top \underline{\boldsymbol{M}}_1 \underline{\boldsymbol{x}}_2 = \underline{\boldsymbol{x}}_2^\top \underline{\boldsymbol{x}}_2 + \underline{\boldsymbol{x}}_2^\top \underline{\boldsymbol{X}}_1 (\underline{\boldsymbol{X}}_1^\top \underline{\boldsymbol{X}}_1)^{-1} \underline{\boldsymbol{X}}_1^\top \underline{\boldsymbol{x}}_2 = \underline{\boldsymbol{x}}_2^\top \underline{\boldsymbol{x}}_2 \left(1 + \frac{\underline{\boldsymbol{x}}_2^\top \underline{\boldsymbol{X}}_1 (\underline{\boldsymbol{X}}_1^\top \underline{\boldsymbol{X}}_1)^{-1} \underline{\boldsymbol{X}}_1^\top \underline{\boldsymbol{x}}_2}{\underline{\boldsymbol{x}}_2^\top \underline{\boldsymbol{x}}_2}\right) \tag{30.0.40}$$

and since the fraction is $R^2_{2\cdot}$, i.e., it is the R^2 in the regression of $\underline{\boldsymbol{x}}_2$ on all other variables in $\underline{\boldsymbol{X}}$, we get the result. □

30.1. Selection of Regressors

One problem often arising in practical statistical work is to select, from a given pool of regressors, those regressions with one, two, etc. regressors which have the best fit. Such routines are used very often in practical work, but they have been somewhat shunned in the econometrics text books, since these methods are considered suspect in the classical frequentist regression approach. This is a nice example of theory-practice inconsistency due to the weakness of this classical approach. Recently, some very good books about this subject have appeared, for instance [**Mil90**] and [**MT98**].

This is a discussion of the `leaps` procedure in `Splus`. Its purpose is to select, from a given pool of regressors, those regressions with one, two, etc. regressors which have the best fit. One may also ask for the sets of size `nbest` of best regressions.

It uses a procedure proposed in a paper titled "Regression by Leaps and Bounds" [**FW74**]. As the title says, the SSE of regression with more variables are used as *bounds* for the SSE of regressions with some of those variables omitted, in order to be able to *leap* ahead in the list and not to have to examine all regressions.

The paper computes the SSE of these regressions by sweeping. As discussed above, this is a procedure which can do or undo regressions. If you start with the sum of squares and cross products matrix, sweeping will introduce regressors, and if you start with the inverse of the SSCP matrix, sweeping will remove regressors.

Look at the example with five regressors in table 2 of that article. The regressors are ordered, perhaps by value of `t` statistic, in such a way that the most promising ones come first. Regressions on subsets containing the first four regressors will be built up in the "Product Traverse" by *introducing* regressors one by one. By contrast, all regressions on subsets of variables which contain the fifth variable will be gained by *eliminating* regressors from the full set. Why this inverse procedure? The hope is that certain regressions with more variables (among them the "least promising" fifth variable) have such a high SSE that it will become superfluous to run the regressions with subsets of these variables, since one knows that they cannot be better than other regressions already performed.

The idea is therefore to compute some good regression with few variables and some bad regressions with many variables; and if the bad regressions with many variables have a too large SSE compared with the good regressions, one does not have to examine the subsets of these bad regressions. The precise implementation is more a piece of engineering than mathematics. Let's just go through their example. Stage 0 must always be done. In it, the following regressions are performed:

Product Traverse		Inverse Traverse	
Regressors	SSE	Regressors	SSE
1	668	2345	660
12	615	1345	605
123	612	1245	596
1234	592	1235	596

Note that in the product traverse the procedure is lexicographical, i.e., the lowest placed regressors are introduced first, since they promise to have the lowest SSE. In the inverse traverse, the regressions on all four-variable sets which include the fifth variable are generated. Our excerpt of table 2 does not show how these regressions were computed; from the full table one can see that the two regressions shown in each row are generated by a sweep on the same pivot index. In the "product traverse," the source matrix of each sweep operation is the result of the previous regression. For the inverse traverse, the source is in each of these four regressions the same, namely, the inverse of the SSCP matrix, but different regressors are eliminated by the sweep.

Now we are at the beginning of stage 1. Is it necessary to perform the sweep which generates regression 124? No other regression will be derived from 124, therefore we only have to look at regression 124 itself, not any subsets of these three variables. It would not necessary to perform this regression if the regression with variables 1245 (596) had a higher SSE than the best three-variable regression run so far, which is 123, whose SSE is 612. Since 596 is not higher than 612, we must run the regression. It gives

Product Traverse		Inverse Traverse	
Regressors	SSE	Regressors	SSE
124	615	125	597

First regression of stage 2. It is only necessary to run the regression on 13 if the best two-variable regression so far is not better than the regression 1345 (605). The only two-variable regression we have so far is that on 12 (615), and since $615 > 605$, we have to perform sweep 13.

Product Traverse		Inverse Traverse	
Regressors	SSE	Regressors	SSE
13	641	145	618

Second regression in stage 2. Is it necessary to run the regression on 134? The SSE of 1345 is 605, and the best three-variable regression so far is 125 with 519, therefore it is not necessary to run 134, which necessarily must give a SSE higher than 605. (If one does run it, its SSE is indeed 612).

Stage 3: do we have to perform the sweep 14? The best two-variable regression right now is 12 (615), which is better than the 618 of 145, therefore no point in running 14.

By continuing this procedure one can gain substantial advantages over methods in which all regressions must be performed. In the given example, only 7 out of the 15 possible regressions must be run.

CHAPTER 31

Residuals: Standardized, Predictive, "Studentized"

31.1. Three Decisions about Plotting Residuals

After running a regression it is always advisable to look at the residuals. Here one has to make three decisions.

The first decision is whether to look at the ordinary residuals

$$\hat{\varepsilon}_i = \mathsf{y}_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}} \tag{31.1.1}$$

($\boldsymbol{x}_i^\top$ is the ith row of $\boldsymbol{X}$), or the "predictive" residuals, which are the residuals computed using the OLS estimate of $\boldsymbol{\beta}$ gained from all the other data except the data point where the residual is taken. If one writes $\hat{\boldsymbol{\beta}}(i)$ for the OLS estimate without the ith observation, the defining equation for the ith predictive residual, which we call $\hat{\varepsilon}_i(i)$, is

$$\hat{\varepsilon}_i(i) = \mathsf{y}_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}(i). \tag{31.1.2}$$

The second decision is whether to standardize the residuals or not, i.e., whether to divide them by their estimated standard deviations or not. Since $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{M}\mathsf{y}$, the variance of the ith ordinary residual is

$$\operatorname{var}[\hat{\varepsilon}_i] = \sigma^2 m_{ii} = \sigma^2(1 - h_{ii}), \tag{31.1.3}$$

and regarding the predictive residuals it will be shown below, see (31.2.9), that

$$\operatorname{var}[\hat{\varepsilon}_i(i)] = \frac{\sigma^2}{m_{ii}} = \frac{\sigma^2}{1 - h_{ii}}. \tag{31.1.4}$$

Here

$$h_{ii} = \boldsymbol{x}_i^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_i. \tag{31.1.5}$$

(Note that $\boldsymbol{x}_i$ is the ith *row* of $\boldsymbol{X}$ written as a *column* vector.) h_{ii} is the ith diagonal element of the "hat matrix" $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$, the projector on the column space of $\boldsymbol{X}$. This projector is called "hat matrix" because $\hat{\mathsf{y}} = \boldsymbol{H}\mathsf{y}$, i.e., $\boldsymbol{H}$ puts the "hat" on y.

Problem 362. *2 points Show that the ith diagonal element of the "hat matrix" $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$ is $\boldsymbol{x}_i^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{x}_i$ where $\boldsymbol{x}_i$ is the ith row of $\boldsymbol{X}$ written as a column vector.*

Answer. In terms of $\boldsymbol{e}_i$, the n-vector with 1 on the ith place and 0 everywhere else, $\boldsymbol{x}_i = \boldsymbol{X}^\top \boldsymbol{e}_i$, and the ith diagonal element of the hat matrix is $\boldsymbol{e}_i^\top \boldsymbol{H}\boldsymbol{e}_i = \boldsymbol{e}_i^\top \boldsymbol{X}_i(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{e}_i = \boldsymbol{x}_i^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{x}_i$. □

Problem 363. *2 points The variance of the ith disturbance is σ^2. Is the variance of the ith residual bigger than σ^2, smaller than σ^2, or equal to σ^2? (Before doing the math, first argue in words what you would expect it to be.) What about the variance of the predictive residual? Prove your answers mathematically. You are allowed to use* (31.2.9) *without proof.*

Answer. Here is only the math part of the answer: $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{M}\mathsf{y}$. Since $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{H}$ is idempotent and symmetric, we get $\mathcal{V}[\boldsymbol{M}\mathsf{y}] = \sigma^2\boldsymbol{M}$, in particular this means $\operatorname{var}[\hat{\varepsilon}_i] = \sigma^2 m_{ii}$ where m_{ii} is the ith diagonal elements of $\boldsymbol{M}$. Then $m_{ii} = 1 - h_{ii}$. Since all diagonal elements of projection matrices are between 0 and 1, the answer is: the variances of the ordinary residuals cannot be bigger than σ^2. Regarding predictive residuals, if we plug $m_{ii} = 1 - h_{ii}$ into (31.2.9) it becomes

$$\hat{\varepsilon}_i(i) = \frac{1}{m_{ii}}\hat{\varepsilon}_i \qquad \text{therefore} \qquad \operatorname{var}[\hat{\varepsilon}_i(i)] = \frac{1}{m_{ii}^2}\sigma^2 m_{ii} = \frac{\sigma^2}{m_{ii}} \tag{31.1.6}$$

which is bigger than σ^2. □

Problem 364. *Decide in the following situations whether you want predictive residuals or ordinary residuals, and whether you want them standardized or not.*

• a. *1 point You are looking at the residuals in order to check whether the associated data points are outliers and do perhaps not belong into the model.*

Answer. Here one should use the predictive residuals. If the ith observation is an outlier which should not be in the regression, then one should not use it when running the regression. Its inclusion may have a strong influence on the regression result, and therefore the residual may not be as conspicuous. One should standardize them. □

• b. *1 point You are looking at the residuals in order to assess whether there is heteroskedasticity.*

Answer. Here you want them standardized, but there is no reason to use the predictive residuals. Ordinary residuals are a little more precise than predictive residuals because they are based on more observations. □

• c. *1 point You are looking at the residuals in order to assess whether the disturbances are autocorrelated.*

Answer. Same answer as for b. □

• d. *1 point You are looking at the residuals in order to assess whether the disturbances are normally distributed.*

Answer. In my view, one should make a normal QQ-plot of standardized residuals, but one should not use the predictive residuals. To see why, let us first look at the distribution of the standardized residuals before division by s. Each $\hat{\varepsilon}_i/\sqrt{1-h_{ii}}$ is normally distributed with mean zero and standard deviation σ. (But different such residuals are not independent.) If one takes a QQ-plot of those residuals against the normal distribution, one will get in the limit a straight line with slope σ. If one divides every residual by s, the slope will be close to 1, but one will again get something approximating a straight line. The fact that s is random does not affect the relation of the residuals to each other, and this relation is what determines whether or not the QQ-plot approximates a straight line.

But Belsley, Kuh, and Welsch on [**BKW80**, p. 43] draw a normal probability plot of the studentized, not the standardized, residuals. They give no justification for their choice. I think it is the wrong choice.

□

• e. *1 point Is there any situation in which you do* not *want to standardize the residuals?*

Answer. Standardization is a mathematical procedure which is justified when certain conditions hold. But there is no guarantee that these conditions acutally hold, and in order to get a more immediate impression of the fit of the curve one may want to look at the unstandardized residuals. □

The third decision is how to plot the residuals. Never do it against $\mathbf{y}$. Either do it against the predicted $\hat{\mathbf{y}}$, or make several plots against all the columns of the $\boldsymbol{X}$-matrix.

In time series, also a plot of the residuals against time is called for.

Another option are the partial residual plots, see about this also (30.0.2). Say $\hat{\boldsymbol{\beta}}[h]$ is the estimated parameter vector, which is estimated with the full model, but after estimation we drop the h-th parameter, and $\boldsymbol{X}[h]$ is the $\boldsymbol{X}$-matrix without the hth column, and $\boldsymbol{x}_h$ is the hth column of the $\boldsymbol{X}$-matrix. Then by (30.0.4), the estimate of the hth slope parameter is the same as that in the simple regression of $\mathbf{y}-\boldsymbol{X}[h]\hat{\boldsymbol{\beta}}[h]$ on $\boldsymbol{x}_h$. The plot of $\mathbf{y}-\boldsymbol{X}[h]\hat{\boldsymbol{\beta}}[h]$ against $\boldsymbol{x}_h$ is called the hth partial residual plot.

To understand this better, start out with a regression $\mathrm{y}_i = \alpha + \beta x_i + \gamma z_i + \varepsilon_i$; which gives you the fitted values $\mathrm{y}_i = \hat{\alpha}+\hat{\beta}x_i+\hat{\gamma}z_i+\hat{\varepsilon}_i$. Now if you regress $\mathrm{y}_i-\hat{\alpha}-\hat{\beta}x_i$ on x_i and z_i then the intercept will be zero and the estimated coefficient of x_i will be zero, and the estimated coefficient of z_i will be $\hat{\gamma}$, and the residuals will be $\hat{\varepsilon}_i$. The plot of $\mathrm{y}_i-\hat{\alpha}-\hat{\beta}x_i$ versus z_i is the partial residuals plot for z.

31.2. Relationship between Ordinary and Predictive Residuals

In equation (31.1.2), the ith predictive residuals was defined in terms of $\hat{\boldsymbol{\beta}}(i)$, the parameter estimate from the regression of $\mathbf{y}$ on $\boldsymbol{X}$ with the ith observation left out. We will show now that there is a very simple mathematical relationship between the ith predictive residual and the ith ordinary residual, namely, equation (31.2.9).

(It is therefore not necessary to run n different regressions to get the n predictive residuals.)

We will write $\mathbf{y}(i)$ for the $\mathbf{y}$ vector with the ith element deleted, and $\boldsymbol{X}(i)$ is the matrix $\boldsymbol{X}$ with the ith row deleted.

Problem 365. *2 points Show that*

$$\boldsymbol{X}(i)^\top\boldsymbol{X}(i) = \boldsymbol{X}^\top\boldsymbol{X} - \boldsymbol{x}_i\boldsymbol{x}_i^\top \tag{31.2.1}$$

$$\boldsymbol{X}(i)^\top\mathbf{y}(i) = \boldsymbol{X}^\top\mathbf{y} - \boldsymbol{x}_i\mathrm{y}_i. \tag{31.2.2}$$

Answer. Write (31.2.2) as $\boldsymbol{X}^\top\mathbf{y} = \boldsymbol{X}(i)^\top\mathbf{y}(i) + \boldsymbol{x}_i\mathrm{y}_i$, and observe that with our definition of $\boldsymbol{x}_i$ as column vectors representing the rows of $\boldsymbol{X}$, $\boldsymbol{X}^\top = \begin{bmatrix}\boldsymbol{x}_1 & \cdots & \boldsymbol{x}_n\end{bmatrix}$. Therefore

$$\boldsymbol{X}^\top\mathbf{y} = \begin{bmatrix}\boldsymbol{x}_1 & \dots & \boldsymbol{x}_n\end{bmatrix}\begin{bmatrix}\mathrm{y}_1\\ \vdots\\ \mathrm{y}_n\end{bmatrix} = \boldsymbol{x}_1\mathrm{y}_1 + \cdots + \boldsymbol{x}_n\mathrm{y}_n. \tag{31.2.3}$$

□

An important stepping stone towards the proof of (31.2.9) is equation (31.2.8), which gives a relationship between h_{ii} and

$$h_{ii}(i) = \boldsymbol{x}_i^\top(\boldsymbol{X}(i)^\top\boldsymbol{X}(i))^{-1}\boldsymbol{x}_i. \tag{31.2.4}$$

$\hat{\mathrm{y}}_i(i) = \boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}(i)$ has variance $\sigma^2 h_{ii}(i)$. The following problems give the steps necessary to prove (31.2.8). We begin with a simplified version of theorem A.8.2 in the Mathematical Appendix:

Theorem 31.2.1. *Let $\boldsymbol{A}$ be a nonsingular $k\times k$ matrix, $\delta\neq 0$ a scalar, and $\boldsymbol{b}$ a $k\times 1$ vector with $\boldsymbol{b}^\top\boldsymbol{A}^{-1}\boldsymbol{b}+\delta\neq 0$. Then*

$$\left(\boldsymbol{A}+\frac{\boldsymbol{b}\boldsymbol{b}^\top}{\delta}\right)^{-1} = \boldsymbol{A}^{-1} - \frac{\boldsymbol{A}^{-1}\boldsymbol{b}\boldsymbol{b}^\top\boldsymbol{A}^{-1}}{\delta+\boldsymbol{b}^\top\boldsymbol{A}^{-1}\boldsymbol{b}}. \tag{31.2.5}$$

Problem 366. *Prove* (31.2.5) *by showing that the product of the matrix with its alleged inverse is the unit matrix.*

Problem 367. *As an application of* (31.2.5) *show that*

$$(\boldsymbol{X}^\top\boldsymbol{X})^{-1} + \frac{(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}}{1-h_{ii}} \quad \textit{is the inverse of} \quad \boldsymbol{X}(i)^\top\boldsymbol{X}(i). \tag{31.2.6}$$

Answer. This is (31.2.5), or (A.8.20), with $\boldsymbol{A}=\boldsymbol{X}^\top\boldsymbol{X}$, $\boldsymbol{b}=\boldsymbol{x}_i$, and $\delta=-1$.

□

PROBLEM 368. *Using* (31.2.6) *show that*

$$(\boldsymbol{X}(i)^\top \boldsymbol{X}(i))^{-1}\boldsymbol{x}_i = \frac{1}{1-h_{ii}}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{x}_i, \tag{31.2.7}$$

and using (31.2.7) *show that* $h_{ii}(i)$ *is related to* h_{ii} *by the equation*

$$1 + h_{ii}(i) = \frac{1}{1-h_{ii}} \tag{31.2.8}$$

[**Gre97**, (9-37) on p. 445] *was apparently not aware of this relationship.*

PROBLEM 369. *Prove the following mathematical relationship between predictive residuals and ordinary residuals:*

$$\hat{\varepsilon}_i(i) = \frac{1}{1-h_{ii}}\hat{\varepsilon}_i \tag{31.2.9}$$

which is the same as (28.0.29)*, only in a different notation.*

ANSWER. For this we have to apply the above mathematical tools. With the help of (31.2.7) (transpose it!) and (31.2.2), (31.1.2) becomes

$$\begin{aligned}
\hat{\varepsilon}_i(i) &= \mathsf{y}_i - \boldsymbol{x}_i^\top(\boldsymbol{X}(i)^\top \boldsymbol{X}(i))^{-1}\boldsymbol{X}(i)^\top \mathbf{y}(i) \\
&= \mathsf{y}_i - \frac{1}{1-h_{ii}}\boldsymbol{x}_i^\top(\boldsymbol{X}^\top \boldsymbol{X})^{-1}(\boldsymbol{X}^\top \mathbf{y} - \boldsymbol{x}_i\mathsf{y}_i) \\
&= \mathsf{y}_i - \frac{1}{1-h_{ii}}\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}} + \frac{1}{1-h_{ii}}\boldsymbol{x}_i^\top(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{x}_i\mathsf{y}_i \\
&= \mathsf{y}_i\left(1 + \frac{h_{ii}}{1-h_{ii}}\right) - \frac{1}{1-h_{ii}}\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}} \\
&= \frac{1}{1-h_{ii}}(\mathsf{y}_i - \boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}})
\end{aligned}$$

This is a little tedious but simplifies extremely nicely at the end. □

The relationship (31.2.9) is so simple because the estimation of $\eta_i = \boldsymbol{x}_i^\top\boldsymbol{\beta}$ can be done in two steps. First collect the information which the $n-1$ observations other than the ith contribute to the estimation of $\eta_i = \boldsymbol{x}_i^\top\boldsymbol{\beta}$ is contained in $\hat{\mathsf{y}}_i(i)$. The information from all observations except the ith can be written as

$$\hat{\mathsf{y}}_i(i) = \eta_i + \delta_i \qquad \delta_i \sim (0, \sigma^2 h_{ii}(i)) \tag{31.2.10}$$

Here δ_i is the "sampling error" or "estimation error" $\hat{\mathsf{y}}_i(i) - \eta_i$ from the regression of $\mathbf{y}(i)$ on $\boldsymbol{X}(i)$. If we combine this compound "observation" with the ith observation y_i, we get

$$\begin{bmatrix}\hat{\mathsf{y}}_i(i)\\ \mathsf{y}_i\end{bmatrix} = \begin{bmatrix}1\\1\end{bmatrix}\eta_i + \begin{bmatrix}\delta_i\\ \varepsilon_i\end{bmatrix} \qquad \begin{bmatrix}\delta_i\\ \varepsilon_i\end{bmatrix} \sim \left(\begin{bmatrix}0\\0\end{bmatrix}, \sigma^2\begin{bmatrix}h_{ii}(i) & 0\\ 0 & 1\end{bmatrix}\right) \tag{31.2.11}$$

This is a regression model similar to model (18.1.1), but this time with a nonspherical covariance matrix.

PROBLEM 370. *Show that the BLUE of* η_i *in model* (31.2.11) *is*

$$\hat{\mathsf{y}}_i = (1-h_{ii})\hat{\mathsf{y}}_i(i) + h_{ii}\mathsf{y}_i = \hat{\mathsf{y}}_i(i) + h_{ii}\hat{\varepsilon}_i(i) \tag{31.2.12}$$

Hint: apply (31.2.8)*. Use this to prove* (31.2.9)*.*

ANSWER. As shown in problem 206, the BLUE in this situation is the weighted average of the observations with the weights proportional to the inverses of the variances. I.e., the first observation has weight

$$\frac{1/h_{ii}(i)}{1/h_{ii}(i)+1} = \frac{1}{1+h_{ii}(i)} = 1 - h_{ii}. \tag{31.2.13}$$

Since the sum of the weights must be 1, the weight of the second observation is h_{ii}.

Here is an alternative solution, using formula (26.0.2) for the BLUE, which reads here

$$\begin{aligned}
\hat{\mathsf{y}}_i &= \left(\begin{bmatrix}1 & 1\end{bmatrix}\begin{bmatrix}\frac{h_{ii}}{1-h_{ii}} & 0\\ 0 & 1\end{bmatrix}^{-1}\begin{bmatrix}1\\1\end{bmatrix}\right)^{-1}\begin{bmatrix}1 & 1\end{bmatrix}\begin{bmatrix}\frac{h_{ii}}{1-h_{ii}} & 0\\ 0 & 1\end{bmatrix}^{-1}\begin{bmatrix}\hat{\mathsf{y}}_i(i)\\ \mathsf{y}_i\end{bmatrix} = \\
&= h_{ii}\begin{bmatrix}1 & 1\end{bmatrix}\begin{bmatrix}\frac{1-h_{ii}}{h_{ii}} & 0\\ 0 & 1\end{bmatrix}\begin{bmatrix}\hat{\mathsf{y}}_i(i)\\ \mathsf{y}_i\end{bmatrix} = (1-h_{ii})\hat{\mathsf{y}}_i(i) + h_{ii}\mathsf{y}_i.
\end{aligned}$$

Now subtract this last formula from y_i to get $\mathsf{y}_i - \hat{\mathsf{y}}_i = (1-h_{ii})(\mathsf{y}_i - \hat{\mathsf{y}}_i(i))$, which is (31.2.9). □

31.3. Standardization

In this section we will show that the standardized predictive residual is what is sometimes called the "studentized" residual. It is recommended not to use the term "studentized residual" but say "standardized predictive residual" instead.

The standardization of the ordinary residuals has two steps: every $\hat{\varepsilon}_i$ is divided by its "relative" standard deviation $\sqrt{1-h_{ii}}$, and then by s, an estimate of σ, the standard deviation of the true disturbances. In formulas,

$$\text{the } i\text{th standardized ordinary residual} = \frac{\hat{\varepsilon}_i}{\mathsf{s}\sqrt{1-h_{ii}}}. \tag{31.3.1}$$

Standardization of the ith *predictive* residual has the same two steps: first divide the predictive residual (31.2.9) by the relative standard deviation, and then divide by $\mathsf{s}(i)$. But a look at formula (31.2.9) shows that the ordinary and the predictive residual differ only by a nonrandom factor. Therefore the first step of the standardization yields exactly the same result whether one starts with an ordinary or a predictive residual. Standardized predictive residuals differ therefore from standardized ordinary residuals only in the second step:

$$\text{the } i\text{th standardized predictive residual} = \frac{\hat{\varepsilon}_i}{\mathsf{s}(i)\sqrt{1-h_{ii}}}. \tag{31.3.2}$$

Note that equation (31.3.2) writes the standardized predictive residual as a function of the *ordinary* residual, not the predictive residual. The standardized predictive residual is sometimes called the "studentized" residual.

PROBLEM 371. *3 points The ith predictive residual has the formula*

$$\hat{\varepsilon}_i(i) = \frac{1}{1-h_{ii}}\hat{\varepsilon}_i \tag{31.3.3}$$

You do not have to prove this formula, but you are asked to derive the standard deviation of $\hat{\varepsilon}_i(i)$, and to derive from it a formula for the standardized ith predictive residual.

This similarity between these two formulas has lead to widespread confusion. Even [**BKW80**] seem to have been unaware of the significance of "studentization"; they do not work with the concept of predictive residuals at all.

The standardized predictive residuals have a t-distribution, because they are a normally distributed variable divided by an independent χ^2 over its degrees of freedom. (But note that the joint distribution of all standardized predictive residuals is *not* a multivariate t.) Therefore one can use the quantiles of the t-distribution to judge, from the size of these residuals, whether one has an extreme observation or not.

PROBLEM 372. *Following* [**DM93**, p. 34], *we will use* (30.0.3) *and the other formulas regarding additional regressors to prove the following: If you add a dummy variable which has the value 1 for the ith observation and the value 0 for all other observations to your regression, then the coefficient estimate of this dummy is the ith predictive residual, and the coefficient estimate of the other parameters after inclusion of this dummy is equal to $\hat{\boldsymbol{\beta}}(i)$. To fix notation (and without loss of generality), assume the ith observation is the last observation, i.e., $i = n$, and put the dummy variable first in the regression:*

$$\begin{bmatrix} \mathbf{y}(n) \\ \mathbf{y}_n \end{bmatrix} = \begin{bmatrix} \boldsymbol{o} & \boldsymbol{X}(n) \\ 1 & \boldsymbol{x}_n^\top \end{bmatrix} \begin{bmatrix} \alpha \\ \boldsymbol{\beta} \end{bmatrix} + \begin{bmatrix} \hat{\boldsymbol{\varepsilon}}(i) \\ \hat{\varepsilon}_n \end{bmatrix} \qquad \text{or} \quad \mathbf{y} = \begin{bmatrix} \boldsymbol{e}_n & \boldsymbol{X} \end{bmatrix} \begin{bmatrix} \alpha \\ \boldsymbol{\beta} \end{bmatrix} + \boldsymbol{\varepsilon} \tag{31.3.4}$$

• a. *2 points With the definition $\boldsymbol{X}_1 = \boldsymbol{e}_n = \begin{bmatrix} \boldsymbol{o} \\ 1 \end{bmatrix}$, write $\boldsymbol{M}_1 = \boldsymbol{I} - \boldsymbol{X}_1(\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top$ as a 2×2 partitioned matrix.*

ANSWER.

$$\boldsymbol{M}_1 = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{o} \\ \boldsymbol{o}^\top & 1 \end{bmatrix} - \begin{bmatrix} \boldsymbol{o} \\ 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{o}^\top & 1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{o} \\ \boldsymbol{o}^\top & 0 \end{bmatrix}; \qquad \begin{bmatrix} \boldsymbol{I} & \boldsymbol{o} \\ \boldsymbol{o}^\top & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{z}(i) \\ z_i \end{bmatrix} = \begin{bmatrix} \boldsymbol{z}(i) \\ 0 \end{bmatrix} \tag{31.3.5}$$

i.e., $\boldsymbol{M}_1$ simply annulls the last element. □

• b. *2 points Either show mathematically, perhaps by evaluating $(\boldsymbol{X}_2^\top \boldsymbol{M}_1 \boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top \boldsymbol{M}$ or give a good heuristic argument (as* [**DM93**] *do), that regressing $\boldsymbol{M}_1\mathbf{y}$ on $\boldsymbol{M}_1\boldsymbol{X}$ gives the same parameter estimate as regressing $\mathbf{y}$ on $\boldsymbol{X}$ with the nth observation dropped.*

ANSWER. (30.0.2) reads here

$$\begin{bmatrix} \mathbf{y}(n) \\ 0 \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}(n) \\ \boldsymbol{o}^\top \end{bmatrix} \hat{\boldsymbol{\beta}}(i) + \begin{bmatrix} \hat{\boldsymbol{\varepsilon}}(i) \\ 0 \end{bmatrix} \tag{31.3.6}$$

in other words, the estimate of $\boldsymbol{\beta}$ is indeed $\hat{\boldsymbol{\beta}}(i)$, and the first $n-1$ elements of the residual are indeed the residuals one gets in the regression without the ith observation. This is so ugly because the singularity shows here in the zeros of the last row, usually it does not show so much. But this way one also sees that it gives zero as the last residual, and this is what one needs to know!

To have a mathematical proof that the last row with zeros does not affect the estimate, evaluate (30.0.3)

$$\begin{aligned} \hat{\boldsymbol{\beta}}_2 &= (\boldsymbol{X}_2^\top \boldsymbol{M}_1 \boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top \boldsymbol{M}_1 \mathbf{y} \\ &= \left(\begin{bmatrix} \boldsymbol{X}(n)^\top & \boldsymbol{x}_n \end{bmatrix} \begin{bmatrix} \boldsymbol{I} & \boldsymbol{o} \\ \boldsymbol{o}^\top & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{X}(n) \\ \boldsymbol{x}_n^\top \end{bmatrix} \right)^{-1} \begin{bmatrix} \boldsymbol{X}(n)^\top & \boldsymbol{x}_n \end{bmatrix} \begin{bmatrix} \boldsymbol{I} & \boldsymbol{o} \\ \boldsymbol{o}^\top & 0 \end{bmatrix} \begin{bmatrix} \mathbf{y}(n) \\ \mathbf{y}_n \end{bmatrix} \\ &= (\boldsymbol{X}(n)^\top \boldsymbol{X}(n))^{-1}\boldsymbol{X}(n)^\top \mathbf{y}(n) = \hat{\boldsymbol{\beta}}(n) \end{aligned}$$

□

• c. *2 points Use the fact that the residuals in the regression of $\boldsymbol{M}_1\mathbf{y}$ on $\boldsymbol{M}_1\boldsymbol{X}$ are the same as the residuals in the full regression* (31.3.4) *to show that $\hat{\alpha}$ is the nth predictive residual.*

ANSWER. $\hat{\alpha}$ is obtained from that last row, which reads $\mathbf{y}_n = \hat{\alpha} + \boldsymbol{x}_n^\top \hat{\boldsymbol{\beta}}(i)$, i.e., $\hat{\alpha}$ is the predictive residual. □

• d. *2 points Use* (30.0.3) *with $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ interchanged to get a formula for $\hat{\alpha}$.*

ANSWER. $\hat{\alpha} = (\boldsymbol{X}_1^\top \boldsymbol{M}\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top \boldsymbol{M}\mathbf{y} = \frac{1}{m_{nn}}\hat{\varepsilon}_n = \frac{1}{1-h_{nn}}\hat{\varepsilon}_n$, here $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$. □

• e. *2 points From* (30.0.4) *follows that also $\hat{\boldsymbol{\beta}}_2 = (\boldsymbol{X}_2^\top \boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top(\mathbf{y} - \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1)$. Use this to prove*

$$\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}(i) = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{x}_i \hat{\varepsilon}_i \frac{1}{1-h_{ii}} \tag{31.3.7}$$

which is [**DM93**, equation (1.40) on p. 33].

ANSWER. For this we also need to show that one gets the right $\hat{\boldsymbol{\beta}}(i)$ if one regresses $\mathbf{y} - \boldsymbol{e}_n\hat{\alpha}$, or, in other words $\mathbf{y} - \boldsymbol{e}_n\hat{\varepsilon}_n(n)$, on $\boldsymbol{X}$. In other words, $\hat{\boldsymbol{\beta}}(n) = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top(\mathbf{y} - \boldsymbol{e}_n\hat{\varepsilon}_n(n))$, which is exactly (32.4.1). □

CHAPTER 32

Regression Diagnostics

"Regression Diagnostics" can either concentrate on observations or on variables. Regarding observations, it looks for outliers or influential data in the dataset. Regarding variables, it checks whether there are highly collinear variables, or it keeps track of how much each variable contributes to the $\mathcal{MSE}$ of the regression. Collinearity is discussed in [**DM93**, 6.3] and [**Gre97**, 9.2]. Regression diagnostics needs five to ten times more computer resources than the regression itself, and often relies on graphics, therefore it has only recently become part of the standard procedures.

PROBLEM 373. *1 point Define multicollinearity.*

- a. *2 points What are the symptoms of multicollinearity?*
- b. *2 points How can one detect multicollinearity?*
- c. *2 points How can one remedy multicollinearity?*

32.1. Missing Observations

First case: data on $\mathbf{y}$ are missing. If you use a least squares predictor then this will not give any change in the estimates and although the computer will think it is more efficient it isn't.

What other schemes are there? Filling in the missing $\mathbf{y}$ by the arithmetic mean of the observed $\mathbf{y}$ does not give an unbiased estimator.

General conclusion: in a single-equation context, filling in missing $\mathbf{y}$ not a good idea.

Now missing values in the $\boldsymbol{X}$-matrix.

If there is only one regressor and a constant term, then the zero order filling in of $\bar{x}$ "results in no changes and is equivalent with dropping the incomplete data."

The alternative: filling it with zeros and adding a dummy for the data with missing observation amounts to exactly the same thing.

The only case where filling in missing data makes sense is: if you have multiple regression and you can predict the missing data in the $\boldsymbol{X}$ matrix from the other data in the $\boldsymbol{X}$ matrix.

32.2. Grouped Data

If single observations are replaced by arithmetic means of groups of observations, then the error variances vary with the size of the group. If one takes this into consideration, GLS still has good properties, although having the original data is of course more efficient.

32.3. Influential Observations and Outliers

The following discussion focuses on diagnostics regarding *observations*. To be more precise, we will investigate how each single observation affects the fit established by the other data. (One may also ask how the addition of any *two* observations affects the fit, etc.)

32.3.1. The "Leverage". The ith diagonal element h_{ii} of the "hat matrix" is called the "leverage" of the ith observation. The leverage satisfies the following identity

$$\hat{\mathsf{y}}_i = (1 - h_{ii})\hat{\mathsf{y}}_i(i) + h_{ii}\mathsf{y}_i \tag{32.3.1}$$

h_{ii} is therefore is the weight which y_i has in the least squares estimate $\hat{\mathsf{y}}_i$ of $\eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$, compared with all other observations, which contribute to $\hat{\mathsf{y}}_i$ through $\hat{\mathsf{y}}_i(i)$. The larger this weight, the more strongly this one observation will influence the estimate of η_i (and if the estimate of η_i is affected, then other parameter estimates may be affected too).

PROBLEM 374. *3 points Explain the meanings of all the terms in equation* (32.3.1) *and use that equation to explain why h_{ii} is called the "leverage" of the ith observation. Is every observation with high leverage also "influential" (in the sense that its removal would greatly change the regression estimates)?*

ANSWER. $\hat{\mathsf{y}}_i$ is the fitted value for the ith observation, i.e., it is the BLUE of η_i, of the expected value of the ith observation. It is a weighted average of two quantities: the actual observation y_i (which has η_i as expected value), and $\hat{\mathsf{y}}_i(i)$, which is the BLUE of η_i based on all the other observations except the ith. The weight of the ith observation in this weighted average is called the "leverage" of the ith observation. The sum of all leverages is always k, the number of parameters in the regression. If the leverage of one individual point is much greater than k/n, then this point has much more influence on its own fitted value than one should expect just based on the number of observations,

Leverage is not the same as influence; if an observation has high leverage, but by accident the observed value y_i is very close to $\hat{\mathsf{y}}_i(i)$, then removal of this observation will not change the regression results much. Leverage is potential influence. Leverage does not depend on any of the observations, one only needs the $\boldsymbol{X}$ matrix to compute it. □

Those observations whose $\boldsymbol{x}$-values are away from the other observations have "leverage" and can therefore potentially influence the regression results more than the

others. h_{ii} serves as a measure of this distance. Note that h_{ii} only depends on the $\boldsymbol{X}$-matrix, not on $\mathbf{y}$, i.e., points may have a high leverage but not be influential, because the associated $\mathbf{y}_i$ blends well into the fit established by the other data. However, regardless of the observed value of $\mathbf{y}$, observations with high leverage always affect the covariance matrix of $\hat{\boldsymbol{\beta}}$.

$$h_{ii} = \frac{\det(\boldsymbol{X}^\top\boldsymbol{X}) - \det(\boldsymbol{X}(i)^\top\boldsymbol{X}(i))}{\det(\boldsymbol{X}^\top\boldsymbol{X})}, \tag{32.3.2}$$

where $\boldsymbol{X}(i)$ is the $\boldsymbol{X}$-matrix without the ith observation.

PROBLEM 375. *Prove equation* (32.3.2).

ANSWER. Since $\boldsymbol{X}^\top(i)\boldsymbol{X}(i) = \boldsymbol{X}^\top\boldsymbol{X} - \boldsymbol{x}_i\boldsymbol{x}_i^\top$, use theorem A.7.3 with $\boldsymbol{W} = \boldsymbol{X}^\top\boldsymbol{X}$, $\alpha = -1$, and $\boldsymbol{d} = \boldsymbol{x}_i$. □

PROBLEM 376. *Prove the following facts about the diagonal elements of the so-called "hat matrix"* $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$, *which has its name because* $\boldsymbol{H}\mathbf{y} = \hat{\mathbf{y}}$, *i.e., it puts the hat on* $\mathbf{y}$.

• a. *1 point* $\boldsymbol{H}$ *is a projection matrix, i.e., it is symmetric and idempotent.*

ANSWER. Symmetry follows from the laws for the transposes of products: $\boldsymbol{H}^\top = (\boldsymbol{ABC})^\top = \boldsymbol{C}^\top\boldsymbol{B}^\top\boldsymbol{A}^\top = \boldsymbol{H}$ where $\boldsymbol{A} = \boldsymbol{X}$, $\boldsymbol{B} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}$ which is symmetric, and $\boldsymbol{C} = \boldsymbol{X}^\top$. Idempotency $\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$. □

• b. *1 point Prove that a symmetric idempotent matrix is nonnegative definite.*

ANSWER. If $\boldsymbol{H}$ is symmetric and idempotent, then for arbitrary $\boldsymbol{g}$, $\boldsymbol{g}^\top\boldsymbol{H}\boldsymbol{g} = \boldsymbol{g}^\top\boldsymbol{H}^\top\boldsymbol{H}\boldsymbol{g} = \|\boldsymbol{H}\boldsymbol{g}\|^2 \geq 0$. But $\boldsymbol{g}^\top\boldsymbol{H}\boldsymbol{g} \geq 0$ for all $\boldsymbol{g}$ is the criterion which makes $\boldsymbol{H}$ nonnegative definite. □

• c. *2 points Show that*

$$0 \leq h_{ii} \leq 1 \tag{32.3.3}$$

ANSWER. If $\boldsymbol{e}_i$ is the vector with a 1 on the ith place and zeros everywhere else, then $\boldsymbol{e}_i^\top\boldsymbol{H}\boldsymbol{e}_i = h_{ii}$. From $\boldsymbol{H}$ nonnegative definite follows therefore that $h_{ii} \geq 0$. $h_{ii} \leq 1$ follows because $\boldsymbol{I} - \boldsymbol{H}$ is symmetric and idempotent (and therefore nonnegative definite) as well: it is the projection on the orthogonal complement. □

• d. *2 points Show: the average value of the* h_{ii} *is* $\sum h_{ii}/n = k/n$, *where* k *is the number of columns of* $\boldsymbol{X}$. *(Hint: for this you must compute the trace* $\operatorname{tr}\boldsymbol{H}$.*)*

ANSWER. The average can be written as

$$\tfrac{1}{n}\operatorname{tr}(\boldsymbol{H}) = \tfrac{1}{n}\operatorname{tr}(\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top) = \tfrac{1}{n}\operatorname{tr}(\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}) = \tfrac{1}{n}\operatorname{tr}(\boldsymbol{I}_k) = \tfrac{k}{n}.$$

Here we used $\operatorname{tr}\boldsymbol{BC} = \operatorname{tr}\boldsymbol{CB}$ (Theorem A.1.2). □

• e. *1 point Show that* $\frac{1}{n}\boldsymbol{\iota}\boldsymbol{\iota}^\top$ *is a projection matrix. Here* $\boldsymbol{\iota}$ *is the* n*-vector of ones.*

• f. *2 points Show: If the regression has a constant term, then* $\boldsymbol{H} - \frac{1}{n}\boldsymbol{\iota}\boldsymbol{\iota}^\top$ *is a projection matrix.*

ANSWER. If $\boldsymbol{\iota}$, the vector of ones, is one of the columns of $\boldsymbol{X}$ (or a linear combination of these columns), this means there is a vector $\boldsymbol{a}$ with $\boldsymbol{\iota} = \boldsymbol{X}\boldsymbol{a}$. From this follows $\boldsymbol{H}\boldsymbol{\iota}\boldsymbol{\iota}^\top = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{a}\boldsymbol{\iota}^\top = \boldsymbol{X}\boldsymbol{a}\boldsymbol{\iota}^\top = \boldsymbol{\iota}\boldsymbol{\iota}^\top$. One can use this to show that $\boldsymbol{H} - \frac{1}{n}\boldsymbol{\iota}\boldsymbol{\iota}^\top$ is idempotent: $(\boldsymbol{H} - \frac{1}{n}\boldsymbol{\iota}\boldsymbol{\iota}^\top)(\boldsymbol{H} - \frac{1}{n}\boldsymbol{\iota}\boldsymbol{\iota}^\top) = \boldsymbol{H}\boldsymbol{H} - \boldsymbol{H}\frac{1}{n}\boldsymbol{\iota}\boldsymbol{\iota}^\top - \frac{1}{n}\boldsymbol{\iota}\boldsymbol{\iota}^\top\boldsymbol{H} + \frac{1}{n}\boldsymbol{\iota}\boldsymbol{\iota}^\top\frac{1}{n}\boldsymbol{\iota}\boldsymbol{\iota}^\top = \boldsymbol{H} - \frac{1}{n}\boldsymbol{\iota}\boldsymbol{\iota}^\top - \frac{1}{n}\boldsymbol{\iota}\boldsymbol{\iota}^\top + \frac{1}{n}\boldsymbol{\iota}\boldsymbol{\iota}^\top = \boldsymbol{H} - \frac{1}{n}\boldsymbol{\iota}\boldsymbol{\iota}^\top$. □

• g. *1 point Show: If the regression has a constant term, then one can sharpen inequality* (32.3.3) *to* $1/n \leq h_{ii} \leq 1$.

ANSWER. $\boldsymbol{H} - \boldsymbol{\iota}\boldsymbol{\iota}^\top/n$ is a projection matrix, therefore nonnegative definite, therefore its diagonal elements $h_{ii} - 1/n$ are nonnegative. □

• h. *3 points Why is* h_{ii} *called the "leverage" of the* i*th observation? To get full points, you must give a really good verbal explanation.*

ANSWER. Use equation (31.2.12). Effect on any other linear combination of $\hat{\boldsymbol{\beta}}$ is less than the effect on $\hat{\mathbf{y}}_i$. Distinguish from influence. Leverage depends only on $\boldsymbol{X}$ matrix, not on $\mathbf{y}$. □

h_{ii} is closely related to the test statistic testing whether the $\boldsymbol{x}_i$ comes from the same multivariate normal distribution as the other rows of the $\boldsymbol{X}$-matrix. Belsley, Kuh, and Welsch [**BKW80**, p. 17] say those observations i with $h_{ii} > 2k/n$, i.e., more than twice the average, should be considered as "leverage points" which might deserve some attention.

32.4. Sensitivity of Estimates to Omission of One Observation

The most straightforward approach to sensitivity analysis is to see how the estimates of the parameters of interest are affected if one leaves out the ith observation. In the case of linear regression, it is not necessary for this to run n different regressions, but one can derive simple formulas for the changes in the parameters of interest. Interestingly, the various sensitivity measures to be discussed below only depend on the two quantities h_{ii} and $\hat{\varepsilon}_i$.

32.4.1. Changes in the Least Squares Estimate. Define $\hat{\boldsymbol{\beta}}(i)$ to be the OLS estimate computed without the ith observation, and $\hat{\varepsilon}_i(i) = \frac{1}{1-h_{ii}}\hat{\varepsilon}_i$ the ith predictive residual. Then

$$\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}(i) = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_i\hat{\varepsilon}_i(i) \tag{32.4.1}$$

PROBLEM 377. *Show* (32.4.1) *by methods very similar to the proof of* (31.2.9)

Answer. Here is this brute-force proof, I think from [**BKW80**]: Let $\mathbf{y}(i)$ be the $\mathbf{y}$ vector with the ith observation deleted. As shown in Problem 365, $\boldsymbol{X}^\top(i)\mathbf{y}(i) = \boldsymbol{X}^\top\mathbf{y} - \boldsymbol{x}_i\mathrm{y}_i$. Therefore by (31.2.6)

$$\hat{\boldsymbol{\beta}}(i) = (\boldsymbol{X}^\top(i)\boldsymbol{X}(i))^{-1}\boldsymbol{X}^\top(i)\mathbf{y}(i) = \left((\boldsymbol{X}^\top\boldsymbol{X})^{-1} + \frac{(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}}{1-h_{ii}}\right)\left(\boldsymbol{X}^\top\mathbf{y} - \boldsymbol{x}_i\mathrm{y}_i\right) =$$
$$= \hat{\boldsymbol{\beta}} - (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_i\mathrm{y}_i + \frac{1}{1-h_{ii}}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}} - \frac{h_{ii}}{1-h_{ii}}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_i\mathrm{y}_i$$
$$= \hat{\boldsymbol{\beta}} - \frac{1}{1-h_{ii}}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_i\mathrm{y}_i + \frac{1}{1-h_{ii}}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - \frac{1}{1-h_{ii}}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_i\hat{\varepsilon}_i$$

□

To understand (32.4.1), note the following fact which is interesting in its own right: $\hat{\boldsymbol{\beta}}(i)$, which is defined as the OLS estimator if one drops the ith observation, can also be obtained as the OLS estimator if one replaces the ith observation by the prediction of the ith observation on the basis of all other observations, i.e., by $\hat{\mathrm{y}}_i(i)$. Writing $\mathbf{y}((i))$ for the vector $\mathbf{y}$ whose ith observation has been replaced in this way, one obtains

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\mathbf{y}; \qquad \hat{\boldsymbol{\beta}}(i) = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\mathbf{y}((i)). \tag{32.4.2}$$

Since $\mathbf{y} - \mathbf{y}((i)) = \boldsymbol{e}_i\hat{\varepsilon}_i(i)$ and $\boldsymbol{x}_i = \boldsymbol{X}^\top\boldsymbol{e}_i$ (32.4.1) follows.

The quantities h_{ii}, $\hat{\boldsymbol{\beta}}(i)-\hat{\boldsymbol{\beta}}$, and $\mathrm{s}^2(i)$ are computed by the R-function `lm.influence`. Compare [**CH93**, pp. 129–131].

32.4.2. Scaled Measures of Sensitivity. In order to assess the sensitivity of the estimate of any linear combination of the elements of $\boldsymbol{\beta}$, $\phi = \boldsymbol{t}^\top\boldsymbol{\beta}$, it makes sense to divide the change in $\boldsymbol{t}^\top\hat{\boldsymbol{\beta}}$ due to omission of the ith observation by the standard deviation of $\boldsymbol{t}^\top\hat{\boldsymbol{\beta}}$, i.e., to look at

$$\frac{\boldsymbol{t}^\top(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}(i))}{\sigma\sqrt{\boldsymbol{t}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{t}}}. \tag{32.4.3}$$

Such a standardization makes it possible to compare the sensitivity of different linear combinations, and to ask: Which linear combination of the elements of $\hat{\boldsymbol{\beta}}$ is affected most if one drops the ith observation? Interestingly and, in hindsight, perhaps not surprisingly, the linear combination which is most sensitive to the addition of the ith observation, is $\boldsymbol{t} = \boldsymbol{x}_i$.

For a mathematical proof we need the following inequality, which is nothing but the Cauchy-Schwartz inequality in disguise:

Theorem 32.4.1. *If $\boldsymbol{\Omega}$ is positive definite symmetric, then*

$$\max_{\boldsymbol{g}} \frac{(\boldsymbol{g}^\top\boldsymbol{x})^2}{\boldsymbol{g}^\top\boldsymbol{\Omega}\boldsymbol{g}} = \boldsymbol{x}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{x}. \tag{32.4.4}$$

If the denominator in the fraction on the lefthand side is zero, then $\boldsymbol{g} = \boldsymbol{o}$ and therefore the numerator is necessarily zero as well. In this case, the fraction itself should be considered zero.

Proof: As in the derivation of the BLUE with nonsperical covariance matrix, pick a nonsingular $\boldsymbol{Q}$ with $\boldsymbol{\Omega} = \boldsymbol{Q}\boldsymbol{Q}^\top$, and define $\boldsymbol{P} = \boldsymbol{Q}^{-1}$. Then it follows $\boldsymbol{P}\boldsymbol{\Omega}\boldsymbol{P}^\top = \boldsymbol{I}$. Define $\boldsymbol{y} = \boldsymbol{P}\boldsymbol{x}$ and $\boldsymbol{h} = \boldsymbol{Q}^\top\boldsymbol{g}$. Then $\boldsymbol{h}^\top\boldsymbol{y} = \boldsymbol{g}^\top\boldsymbol{x}$, $\boldsymbol{h}^\top\boldsymbol{h} = \boldsymbol{g}^\top\boldsymbol{\Omega}\boldsymbol{g}$, and $\boldsymbol{y}^\top\boldsymbol{y} = \boldsymbol{x}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{x}$. Therefore (32.4.4) follows from the Cauchy-Schwartz inequality $(\boldsymbol{h}^\top\boldsymbol{y})^2 \le (\boldsymbol{h}^\top\boldsymbol{h})(\boldsymbol{y}^\top\boldsymbol{y})$.

Using Theorem 32.4.1 and equation (32.4.1) one obtains

$$\max_{\boldsymbol{t}} \frac{(\boldsymbol{t}^\top(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}(i)))^2}{\sigma^2\boldsymbol{t}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{t}} = \frac{1}{\sigma^2}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}(i))^\top\boldsymbol{X}^\top\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}(i)) = \tag{32.4.5}$$
$$= \frac{1}{\sigma^2}\boldsymbol{x}_i^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_i\hat{\varepsilon}_i^2(i) = \frac{h_{ii}}{\sigma^2}\hat{\varepsilon}_i^2(i)$$

Now we will show that the linear combination which attains this maximum, i.e., which is most sensitive to the addition of the ith observation, is $\boldsymbol{t} = \boldsymbol{x}_i$. If one premultiplies (32.4.1) by $\boldsymbol{x}_i^\top$ one obtains

$$\hat{\mathrm{y}}_i - \hat{\mathrm{y}}_i(i) = \boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}} - \boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}(i) = \frac{h_{ii}}{1-h_{ii}}\hat{\varepsilon}_i = h_{ii}\hat{\varepsilon}_i(i) \tag{32.4.6}$$

If one divides (32.4.6) by the standard deviation of $\hat{\mathrm{y}}_i$, i.e., if one applies the construction (32.4.3), one obtains

$$\frac{\hat{\mathrm{y}}_i - \hat{\mathrm{y}}_i(i)}{\sigma\sqrt{h_{ii}}} = \frac{\sqrt{h_{ii}}}{\sigma}\hat{\varepsilon}_i(i) = \frac{\sqrt{h_{ii}}}{\sigma(1-h_{ii})}\hat{\varepsilon}_i \tag{32.4.7}$$

If $\hat{\mathrm{y}}_i$ changes only little (compared with the standard deviation of $\hat{\mathrm{y}}_i$) if the ith observation is removed, then no other linear combination of the elements of $\hat{\boldsymbol{\beta}}$ will be affected much by the omission of this observation either.

The righthand side of (32.4.7), with σ estimated by $\mathrm{s}(i)$, is called by [**BKW80**] and many others DFFITS (which stands for DiFference in FIT, Standardized). If one takes its square, divides it by k, and estimates σ^2 by s^2 (which is more consistent than using $\mathrm{s}^2(i)$, since one standardizes by the standard deviation of $\boldsymbol{t}^\top\hat{\boldsymbol{\beta}}$ and not by that of $\boldsymbol{t}^\top\hat{\boldsymbol{\beta}}(i)$), one obtains Cook's distance [**Coo77**]. (32.4.5) gives an equation for Cook's distance in terms of $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}(i)$:

(32.4.8)
$$\text{Cook's distance} = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}(i))^\top\boldsymbol{X}^\top\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}(i))}{k\mathrm{s}^2} = \frac{h_{ii}}{k\mathrm{s}^2}\hat{\varepsilon}_i^2(i) = \frac{h_{ii}}{k\mathrm{s}^2(1-h_{ii})^2}\hat{\varepsilon}_i^2$$

Problem 378. *Can you think of a situation in which an observation has a small residual but a large "influence" as measured by Cook's distance?*

Answer. Assume "all observations are clustered near each other while the solitary odd observation lies a way out" as Kmenta wrote in [**Kme86**, p. 426]. If the observation happens to lie on the regression line, then it can be discovered by its influence on the variance-covariance matrix (32.3.2), i.e., in this case only the h_{ii} count. □

Problem 379. *The following is the example given in* [**Coo77**]. *In* R, *the command* `data(longley)` *makes the data frame* `longley` *available, which has the famous Longley-data, a standard example for a highly multicollinear dataset. These data are also available on the web at* `www.econ.utah.edu/ehrbar/data/longley.txt`. `attach(longley)` *makes the individual variables available as* R*-objects.*

• a. *3 points Look at the data in a scatterplot matrix and explain what you see. Later we will see that one of the observations is in the regression much more influential than the rest. Can you see from the scatterplot matrix which observation that might be?*

Answer. In linux, you first have to give the command `x11()` in order to make the graphics window available. In windows, this is not necessary. It is important to display the data in a reasonable order, therefore instead of `pairs(longley)` you should do something like `attach(longley)` and then `pairs(cbind(Year, Population, Employed, Unemployed, Armed.Forces, GNP, GNP.deflator))`. Put `Year` first, so that all variables are plotted against `Year` on the horizontal axis.

Population vs. year is a very smooth line.

Population vs GNP also quite smooth.

You see the huge increase in the armed forced in 1951 due to the Korean War, which led to a (temporary) drop in unemployment and a (not so temporary) jump in the GNP deflator.

Otherwise the unemployed show the stop-and-go scenario of the fifties.

unemployed is not correlated with anything.

One should expect a strong negative correlation between employed and unemployed, but this is not the case. □

• b. *4 points Run a regression of the model* `Employed ~ GNP.deflator + GNP + Unemployed + Armed.Forces + Population + Year` *and discuss the result.*

Answer. To fit a regression run `longley.fit <- lm(Employed ~ GNP + Unemployed + Armed.Force + Population + Year)`. You can see the regression results by typing `summary(longley.fit)`.

Armed forces and unemployed are significant and have negative sign, as expected.

GNP and Population are insignificant and have negative sign too, this is not expected. GNP, Population and Year are highly collinear.

□

• c. *3 points Make plots of the ordinary residuals and the standardized residuals against time. How do they differ? In* R, *the commands are* `plot(Year, residuals(long type="h", ylab="Ordinary Residuals in Longley Regression")`. *In order to get the next plot in a different graphics window, so that you can compare them, do now either* `x11()` *in linux or* `windows()` *in windows, and then* `plot(Year, rstandard(longley.fit), type="h", ylab="Standardized Residuals in Longley Regression")`.

Answer. You see that the standardized residuals at the edge of the dataset are bigger than the ordinary residuals. The datapoints at the edge are better able to attract the regression plane than those in the middle, therefore the ordinary residuals are "too small." Standardization corrects for this. □

• d. *4 points Make plots of the predictive residuals. Apparently there is no special command in* R *to do this, therefore you should use formula* (31.2.9). *Also plot the standardized predictive residuals, and compare them.*

Answer. The predictive residuals are `plot(Year, residuals(longley.fit)/(1-hatvalues(longley.fit type="h", ylab="Predictive Residuals in Longley Regression")`. The standardized predictive residuals are often called studentized residuals, `plot(Year, rstudent(longley.fit), type="h", ylab="Standardized predictive Residuals in Longley Regression")`.

A comparison shows an opposite effect as with the ordinary residuals: the predictive residuals at the edge of the dataset are too *large*, and standardization corrects this.

Specific results: standardized predictive residual in 1950 smaller than that in 1962, but predictive residual in 1950 is very close to 1962.

standardized predictive residual in 1951 smaller than that in 1956, but predictive residual in 1951 is larger than in 1956.

Largest predictive residual is 1951, but largest standardized predictive residual is 1956.

□

• e. *3 points Make a plot of the leverage, i.e., the h_{ii}-values, using* `plot(Year, hatvalues(longley.fit), type="h", ylab="Leverage in Longley Regression")`, *and explain what leverage means.*

• f. *3 points One observation is much more influential than the others; which is it? First look at the plots for the residuals, then look also at the plot for leverage, and try to guess which is the most influential observation. Then do it the right way. Can you give reasons based on your prior knowledge about the time period involved why an observation in that year might be influential?*

Answer. The "right" way is to use Cook's distance: `plot(Year, cooks.distance(longley.fit), type="h", ylab="Cook's Distance in Longley Regression")`

One sees that 1951 towers above all others. It does not have highest leverage, but it has second-highest, and a bigger residual than the point with the highest leverage.

1951 has the largest distance of .61. The second largest is the last observation in the dataset, 1962, with a distance of .47, and the others have .24 or less. Cook says: removal of 1951 point will move the least squares estimate to the edge of a 35% confidence region around $\hat{\boldsymbol{\beta}}$. This point is probably so influential because 1951 was the first full year of the Korean war. One would not be able to detect this point from the ordinary residuals, standardized or not! The predictive residuals are a little better; their maximum is at 1951, but several other residuals are almost as large. 1951 is so influential because it has an extremely high hat-value, and one of the highest values for the ordinary residuals! □

At the end don't forget to `detach(longley)` *if you have attached it before.*

32.4.3. Changes in the Sum of Squared Errors. For the computation of $\mathsf{s}^2(i)$ from the regression results one can take advantage of the following simple relationship between the SSE for the regression with and without the ith observation:

$$\mathsf{SSE} - \mathsf{SSE}(i) = \frac{\hat{\varepsilon}_i^2}{1-h_{ii}} \tag{32.4.9}$$

Problem 380. *Use* (32.4.9) *to derive the following formula for* $\mathsf{s}^2(i)$*:*

$$\mathsf{s}^2(i) = \frac{1}{n-k-1}\Big((n-k)\mathsf{s}^2 - \frac{\hat{\varepsilon}_i^2}{1-h_{ii}}\Big) \tag{32.4.10}$$

Answer. This merely involves re-writing SSE and $\mathsf{SSE}(i)$ in terms of s^2 and $\mathsf{s}^2(i)$.

$$\mathsf{s}^2(i) = \frac{\mathsf{SSE}(i)}{n-1-k} = \frac{1}{n-k-1}\Big(\mathsf{SSE} - \frac{\hat{\varepsilon}_i^2}{1-h_{ii}}\Big) \tag{32.4.11}$$

□

Proof of equation (32.4.9)*:*

$$\begin{aligned}
\mathsf{SSE}(i) &= \sum_{j:\ j\neq i} (\mathsf{y}_j - \boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}(i))^2 = \sum_{j:\ j\neq i} \big(\mathsf{y}_j - \boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}} - \boldsymbol{x}_j^\top(\hat{\boldsymbol{\beta}}(i) - \hat{\boldsymbol{\beta}})\big)^2 \\
&= \sum_{j:\ j\neq i} \Big(\hat{\varepsilon}_j + \frac{h_{ji}}{1-h_{ii}}\hat{\varepsilon}_i\Big)^2 \\
&= \sum_j \Big(\hat{\varepsilon}_j + \frac{h_{ji}}{1-h_{ii}}\hat{\varepsilon}_i\Big)^2 - \Big(\frac{1}{1-h_{ii}}\hat{\varepsilon}_i\Big)^2 \\
&= \sum_j \hat{\varepsilon}_j^2 + \frac{2\hat{\varepsilon}_i}{1-h_{ii}}\sum_j h_{ij}\hat{\varepsilon}_j + \Big(\frac{\hat{\varepsilon}_i}{1-h_{ii}}\Big)^2 \sum_j h_{ji}^2 - \Big(\frac{\hat{\varepsilon}_i}{1-h_{ii}}\Big)^2
\end{aligned}$$

In the last line the first term is SSE. The second term is zero because $\boldsymbol{H}\hat{\boldsymbol{\varepsilon}} = \boldsymbol{o}$. Furthermore, $h_{ii} = \sum_j h_{ji}^2$ because $\boldsymbol{H}$ is symmetric and idempotent, therefore the sum of the last two items is $-\hat{\varepsilon}_i^2/(1-h_{ii})$.

Note that every single relationship we have derived so far is a function of $\hat{\varepsilon}_i$ and h_{ii}.

Problem 381. *3 points What are the main concepts used in modern "Regression Diagnostics"? Can it be characterized to be a careful look at the residuals, or does it have elements which cannot be inferred from the residuals alone?*

Answer. Leverage (sometimes it is called "potential") is something which cannot be inferred from the residuals, it does not depend on y at all. □

Problem 382. *An observation in a linear regression model is "influential" if its omission causes large changes to the regression results. Discuss how you would ascertain in practice whether a given observation is influential or not.*

• a. *What is meant by leverage? Does high leverage necessarily imply that an observation is influential?*

Answer. Leverage is potential influence. It only depends of $\boldsymbol{X}$, not on y. It is the distance of the observation from the center of gravity of all observations. Whether this is actual influence depends on the y-values. □

• b. *How are the concepts of leverage and influence affected by sample size?*

• c. *What steps would you take when alerted to the presence of an influential observation?*

Answer. Make sure you know whether the results you rely on are affected if that influential observation is dropped. Try to find out why this observation is influential (e.g. in the Longley data the observations in the year when the Korean War started are influential). □

• d. *What is a "predictive residual" and how does it differ from an ordinary residual?*

• e. *Discuss situations in which one would want to deal with the "predictive" residuals rather than the ordinary residuals, and situations in which one would want residuals standardized versus situations in which it would be preferable to have the unstandardized residuals.*

Problem 383. *6 points Describe what you would do to ascertain that a regression you ran is correctly specified?*

Answer. Economic theory behind that regression, size and sign of coefficients, plot residuals versus predicted values, time, and every independent variable, run all tests: F-test, t-tests, R^2, DW, portmanteau test, forecasting, multicollinearity, influence statistics, overfitting to see if other variables are significant, try to defeat the result by using alternative variables, divide time period into subperiods in order to see if parameters are constant over time, pre-test specification assumptions. □

CHAPTER 33

Regression Graphics

The "regression" referred to in the title of this chapter is not necessarily *linear* regression. The *population* regression can be defined as follows: The random scalar y and the random vector $\mathbf{x}$ have a joint distribution, and we want to know how the conditional distribution of $\mathsf{y}|\mathbf{x} = \boldsymbol{x}$ depends on the value $\boldsymbol{x}$. The distributions themselves are not known, but we have datasets and we use graphical means to estimate the distributions from these datasets.

PROBLEM 384. *Someone said on an email list about statistics: if you cannot* see *an effect in the data, then there is no use trying to estimate it. Right or wrong?*

ANSWER. One argument one might give is the curse of dimensionality. Also higher moments of the distribution, kurtosis etc., cannot be seen very cleary with the plain eye. □

33.1. Scatterplot Matrices

One common graphical method to explore a dataset is to make a scatter plot of each data series against each other and arrange these plots in a matrix. In R, the `pairs` function does this. Scatterplot matrices should be produced in the preliminary stages of the investigation, but the researcher should not think he or she is done after having looked at the scatterplot matrices.

In the construction of scatter plot matrices, it is good practice to change the signs of some of the variables in order to make all correlations positive if this is possible.

[**BT99**, pp. 17–20] gives a good example of what kinds of things can be seen from looking at scatterplot matrices. The data for this book are available at http://biometrics.a

PROBLEM 385. *5 points Which inferences about the datasets can you draw from looking at the scatterplot matrix in* [**BT99**, Exhibit 3.2, p. 14]*?*

ANSWER. The discussion on [**BT99**, p. 19?] distinguishes three categories. First the univariate phenomena:

- yield is more concentrated for local genotypes (•) than for imports (∘);
- the converse is true for protein % but not as pronounced;
- oil % and seed size are lower for local genotypes (•); regarding seed size, the heaviest • is ligher than the lightest ∘;

- height and lodging are greater for local genotypes.

Bivariate phenomena are either within-group or between-group phenomena or both.:

- negative relationship of protein % and oil % (both within • and ∘);
- positive relationship of oil % and seed size (both within • and ∘ and also between these groups);
- negative relationship, between groups, of seed size and height;
- positive relationship of height and lodging (within ∘ and between groups);
- negative relationship of oil % and lodging (between groups and possibly within •);
- negative relationship of seed size and lodging (between groups);
- positive relationship of height and lodging (between groups).

The between group pehnomena are, of course, not due to an interaction between the groups, but they are the consequence of univariate phenomena. As a third category, the authors point out unusual individual points:

- 1 high ∘ for yield;
- 1 high • (still lower than all the ∘s) for seed size;
- 1 low ∘ for lodging;
- 1 low • for protein % and oil % in combination.

□

[**Coo98**, Figure 2.8 on p. 29] shows a scatterplot matrix of the "horse mussel" data, originally from [**Cam89**]. This graph is also available at `www.stat.umn.edu/RegGraph/graphics/Figure_2.8.gif`. Horse mussels, (Atrinia), were sampled from the Marlborough Sounds. The five variables are L = Shell length in mm, W = Shell width in mm, H = Shell height in mm, S = Shell mass in g, and M = Muscle mass in g. M is the part of the mussel that is edible.

PROBLEM 386. *3 points In the mussel data set, M is the "response" (according to* [**Coo98**]*). Is it justified to call this variable the "response" and the other variables the explanatory variables, and if so, how would you argue for it?*

ANSWER. This is one of the issues which is not sufficiently discussed in the literature. It would be justified if the dimensions and weight of the shell were exogenous to the weight of the edible part of the mussel. I.e., if the mussel first grows the shell, and then it fills this shell wish muscle, and the dimensions of the shell affect how big the muscle can grow, but the muscle itself does not have an influence on the dimensions of the shell. If this is the case, then it makes sense to look at the distribution of M conditionally on the other variables, i.e., ask the question: given certain weights and dimensions of the shell, what is the nature of the mechanism by which the muscle grows inside this shell. But if muscle and shell grow together, both affected by the same variables (temperature, nutrition, daylight, etc.), then the conditional distribution is not informative. In this case, the joint distribution is of interest. □

In order to get this dataset into R, you simply say `data(mussels)`, after having said `library(ecmet)`. Then you need the command `pairs(mussels)` to get the scatterplot matrix. Also interesting is `pairs(log(mussels))`, especially since the log transformation is appropriate if one explains volume and weight by length, height, and width.

The scatter plot of M versus H shows a clear curvature; but one should not jump to the conclusion that the regression is not linear. Cook brings another example with constructed data, in which the regression function is clearly linear, without error term, and in which nevertheless the scatter plot of the response versus one of the predictors shows a similar curvature as in the mussel data.

PROBLEM 387. *Cook's constructed dataset is available as dataset* `reggra29` *in the* `ecmet` *package. Make a scatterplot matrix of the plot, then load it into* `XGobi` *and convince yourself that* $\mathtt{y}$ *depends linearly on* $\mathtt{x}_1$ *and* $\mathtt{x}_2$.

ANSWER. You need the commands `data(reggra29)` and then `pairs(reggra29)` to get the scatterplot matrix. Before you can access xgobi from R, you must give the command `library(xgobi)`. Then `xgobi(reggra29)`. The dependency is $\mathtt{y} = 3 + \mathtt{x}_1 + elem\mathtt{x}_2/2$. □

PROBLEM 388. *2 points Why can the scatter plot of the dependent variable against one of the independent variables be so misleading?*

ANSWER. Because the included independent variable becomes a proxy for the excluded variable. The effect of the excluded variable is mistaken to come from the included variable. Now if the included and the excluded variable are independent of each other, then the omission of the excluded variable increases the noise, but does not have a systematic effect. But if there is an empirical relationship between the included and the excluded variable, then this translates into a spurious relationship between included and dependent variables. The mathematics of this is discussed in Problem 328. □

PROBLEM 389. *Would it be possible in the scatter plot in* [**Coo98**, p. 217] *to reverse the signs of some of the variables in such a way that all correlations are positive?*

ANSWER. Yes, you have to reverse the signs of `6Below` and `AFDC`. Here are the instructions how to do the scatter plots: in arc, go to the load menu. (Ignore the close and the menu boxes, they don't seem to work.) Then type the path into the long box, `/usr/share/ecmet/xlispstat` and press return. This gives me only one option, `Minneapolis-schools.lsp`. I have to press 3 times on this until it jumps to the big box, then I can press enter on the big box to load the data. This gives me a bigger menu. Go to the MPLSchools menu, and to the add variable option. You have to type in `6BelNeg = (- 6Below)`, then enter, then `AFDCNeg = (- AFDC)`, and finally `BthPtsNeg = (- BthPts)`. Then go to the Graph&Fit menu, and select scatterplot matrix. Then you have to be careful about the order: first select AFDCNeg, in the left box and double click so that it jumps over to the right box. Then select HS, then BthPtsNeg, then 6BelNeg, then 6Above. Now the scatterplot matrix will be oriented all in 1 direction. □

33.2. Conditional Plots

In order to account for the effect of excluded variables in a scatter plot, the function `coplot` makes scatter plots in which the excluded variable is conditioned upon. The graphics demo has such a conditioning plot; here is the code (from the file `/usr/lib/R/demos/graphics/graphics.R`):


```
data(quakes)
coplot(long ~ lat | depth, data=quakes, pch=21)
```

33.3. Spinning

An obvious method to explore a more than two-dimensional structure graphically is to look at plots of $\mathtt{y}$ against various linear combinations of $\mathtt{x}$. Many statistical software packages have the ability to do so, but one of the most powerful ones is `XGobi`. Documentation about xgobi, which is more detailed than the `help(xgobi)` in R/`Splus` can be obtained by typing `man xgobi` while in unix. A nice brief documentation is [**Rip96**]. The official manual is is [**SCB91**] and [**BCS96**].

`XGobi` can be used as a stand-alone program or it can be invoked from inside R or `Splus`. In R, you must give the command `library(xgobi)` in order to make the function `xgobi` accessible.

The search for "interesting" projections of the data into one-, two-, or 3-dimensional spaces has been automated in *projection pursuit* regression programs. The basic reference is [**FS81**], but there is also the much older [**FT74**].

The most obvious graphical regression method consists in slicing or binning the data, and taking the mean of the data in each bin. But if you have too many explanatory variables, this local averaging becomes infeasible, because of the "curse of dimensionality." Consider a dataset with 1000 observations and 10 variables, all between 0 and 1. In order to see whether the data are uniformly distributed or whether they have some structure, you may consider splitting up the 10-dimensional unit cube into smaller cubes and counting the number of datapoints in each of these subcubes. The problem here is: if one makes those subcubes large enough that they contain more than 0 or 1 observations, then their coordinate lengths are not much smaller than the unit hypercube itself. Even with a side length of 1/2, which would be the largest reasonable side length, one needs 1024 subcubes to fill the hypercube, therefore the average number of data points is a little less than 1. By projecting instead of taking subspaces, projection pursuit regression does not have this problem of data scarcity.

Projection pursuit regression searches for an interesting and informative projection of the data by maximizing a criterion function. A logical candidate would for instance be the variance ratio as defined in (8.6.7), but there are many others.

About grand tours, projection pursuit guided tours, and manual tours see [**CBCH97**] and [**CB97**].

PROBLEM 390. *If you run* `XGobi` *from the menu in Debian GNU/Linux, it uses* `prim7`*, which is a 7-dimensional particle physics data set used as an example in* [**FT74**].

The following is from the help page for this dataset: There are 500 observations taken from a high energy particle physics scattering experiment which yields four particles. The reaction can be described completely by 7 independent measurements. The important features of the data are short-lived intermediate reaction stages which appear as protuberant "arms" in the point cloud.

The projection pursuit guided tour is the tool to use to understand this data set. Using all 7 variables turn on projection pursuit and optimize with the Holes index until a view is found that has a triangle and two arms crossing each other off one edge (this is very clear once you see it but the Holes index has a tendency to get stuck in another local maximum which doesn't have much structure). Brush the arms with separate colours and glyphs. Change to the Central Mass index and optimize. As new arms are revealed brush them and continue. When you have either run out of colours or time turn off projection pursuit and watch the data touring. Then it becomes clear that the underlying structure is a triangle with 5 or 6 arms (some appear to be 1-dimensional, some 2-dimensional) extending from the vertices.

33.4. Sufficient Plots

A different approach, which is less ad-hoc than projection pursuit, starts with the theory of conditional independence, see Section 2.10.3. A theoretical exposition of this approach is [**Coo98**] with web-site `www.stat.umn.edu/RegGraph`. This web site has the data and several short courses based on the book. Especially the file `www.stat.umn.edu/RegGraph/papers/interface.pdf` is a nice brief introduction. A more practically-oriented book, which teaches the software especially developed for this approach, is [**CW99**].

For any graphical procedure exploring linear combinations of the explanatory variables, the *structural dimension* d of a regression is relevant: it is the smallest number of distinct linear combinations of the predictors required to characterize the conditional distribution of $\mathsf{y}|\mathbf{x}$.

If the data follow a linear regression then their structural dimension is 1. But even if the regression is nonlinear but can be written in the form

$$\mathsf{y}|\mathbf{x} \sim g(\boldsymbol{\beta}^\top\mathbf{x}) + \sigma(\boldsymbol{\beta}^\top\mathbf{x})\varepsilon \tag{33.4.1}$$

with $\boldsymbol{\varepsilon}$ independent of $\mathbf{x}$, this is also a population with structural dimension of 1. If t is a monotonic transformation, then

$$t(\mathsf{y})|\mathbf{x} \sim g(\boldsymbol{\beta}^\top\mathbf{x}) + \sigma(\boldsymbol{\beta}^\top\mathbf{x})\varepsilon$$

is an even more general model with structural dimension 1.

If

$$\mathsf{y}|\mathbf{x} = \begin{cases} \boldsymbol{\beta}_1^\top\mathbf{x} + \varepsilon & \text{if } \mathsf{x}_1 \geq 0 \\ \boldsymbol{\beta}_2^\top\mathbf{x} + \varepsilon & \text{if } \mathsf{x}_1 < 0 \end{cases} \tag{33.4.2}$$

with $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ linearly independent, then the structural dimension is 2, since one needs 2 different linear combinations of $\mathbf{x}$ to characterize the distribution of y. If

$$\mathsf{y}|\mathbf{x} = \|\mathbf{x}\|^2 + \varepsilon \tag{33.4.3}$$

then this is a very simple relationship between $\mathbf{x}$ and y, but the structural dimension is k, the number of dimensions of $\mathbf{x}$, since the relationship is not intrinsically linear.

Problem 391. [**FS81**, p. 818] *Show that the regression function consisting of the interaction term between x_1 and x_2 only $\phi(\mathbf{x}) = \mathsf{x}_1\mathsf{x}_2$ has structural dimension 2, i.e., it can be written in the form $\phi(\mathbf{x}) = \sum_{m=1}^{2} s_m(\boldsymbol{\alpha}_m\mathbf{x})$ where s_m are smooth functions of one variable.*

Answer.

$$\boldsymbol{\alpha}_1 = \frac{1}{\sqrt{2}}\begin{bmatrix}1\\1\\\boldsymbol{o}\end{bmatrix} \quad \boldsymbol{\alpha}_2 = \frac{1}{\sqrt{2}}\begin{bmatrix}1\\-1\\\boldsymbol{o}\end{bmatrix} \quad s_1(z) = \frac{z^2}{2} \quad s_2(z) = -\frac{z^2}{2} \tag{33.4.4}$$

□

Problem 392. [**Coo98**, p. 62] *In the rubber data, `mnr` is the dependent variable y, and `temp` and `dp` form the two explanatory variables x_1 and x_2. Look at the data using `XGgobi` or some other spin program. What is the structural dimension of the data set?*

The rubber data are from [**Woo72**]*, and they are also discussed in* [**Ric**, p. 506]*. `mnr` is modulus of natural rubber, `temp` the temperature in degrees Celsius, and `dp` Dicumyl Peroxide in percent.*

Answer. The data are a plane that has been distorted by twisting and stretching. Since one needs a different view to get the best fit of the points in the upper-right corner than for the points in the lower-left corner, the structural dimension must be 2. □

If one looks at the scatter plots of y against all linear combinations of components of $\mathbf{x}$, and none of them show a relationship (either linear or nonlinear), then the structural dimension is zero.

Here are the instruction how to do graphical regression on the mussel data. Select the load menu (take the cursor down a little until it is black, then go back up), then press the `Check Data Dir` box, then double click `ARCG` so that it jumps into the big box. Then `Update/Open File` will give you a long list of selections, where you will find `mussels.lsp`. Double click on this so that it jumps into the big box, and then press on the `Update/Open File` box. Now for the Box-Cox transformation I first have to go to scatterplot matrices, then click on transformations, then to `find normalizing transformations`. It you just select the 4 predictors and then press the OK button, there will be an error message; apparently the starting values were not good enough. Try again, using `marginal Box-Cox Starting Values`. This will succeed, and the LR test for all transformations logs has a p-value of .14. Therefore

choose the log transform for all the predictors. (If we include all 5 variables, the LR test for all transformations to be log transformations has a p-value of 0.000.) Therefore transform the 4 predictor variables only to logs. There you see the very linear relationship between the predictors, and you see that all the scatter plots with the response are very similar. This is a sign that the structural dimension is 1 according to [**CW99**, pp. 435/6]. If that is the case, then a plot of the actual against the fitted values is a sufficient summary plot. For this, run the Fit Linear LS menu option, and then plot the dependent variable against the fitted value. Now the next question might be: what transformation will linearize this, and a log curve seems to fit well.

CHAPTER 34

Asymptotic Properties of the OLS Estimator

A much more detailed treatment of the contents of this chapter can be found in [**DM93**, Chapters 4 and 5].

Here we are concerned with the consistency of the OLS estimator for large samples. In other words, we assume that our regression model can be extended to encompass an arbitrary number of observations. First we assume that the regressors are nonstochastic, and we will make the following assumption:

$$\boldsymbol{Q} = \lim_{n\to\infty} \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X} \quad \text{exists and is nonsingular.} \tag{34.0.5}$$

Two examples where this is not the case. Look at the model $\mathsf{y}_t = \alpha+\beta t+\varepsilon_t$. Here $\boldsymbol{X} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ \vdots & \vdots \\ 1 & n \end{bmatrix}$. Therefore $\boldsymbol{X}^\top\boldsymbol{X} = \begin{bmatrix} 1+1+1+\cdots+1 & 1+2+3+\cdots+n \\ 1+2+3+\cdots+n & 1+4+9+\cdots+n^2 \end{bmatrix} = \begin{bmatrix} n & n(n+1)/2 \\ n(n+1)/2 & n(n+1)(2n+1)/6 \end{bmatrix}$, and $\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X} \to \begin{bmatrix} 1 & \infty \\ \infty & \infty \end{bmatrix}$. Here the assumption (34.0.5) does not hold, but one can still prove consistency and asymtotic normality, the estimators converge even faster than in the usual case.

The other example is the model $\mathsf{y}_t = \alpha + \beta\lambda^t + \varepsilon_t$ with a known λ with $-1 < \lambda < 1$. Here

$$\begin{aligned} \boldsymbol{X}^\top\boldsymbol{X} &= \begin{bmatrix} 1+1+\cdots+1 & \lambda+\lambda^2+\cdots+\lambda^n \\ \lambda+\lambda^2+\cdots+\lambda^n & \lambda^2+\lambda^4+\cdots+\lambda^{2n} \end{bmatrix} = \\ &= \begin{bmatrix} n & (\lambda-\lambda^{n+1})/(1-\lambda) \\ (\lambda-\lambda^{n+1})/(1-\lambda) & (\lambda^2-\lambda^{2n+2})/(1-\lambda^2) \end{bmatrix}. \end{aligned}$$

Therefore $\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X} \to \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$, which is singular. In this case, a consistent estimate of λ does not exist: future observations depend on λ so little that even with infinitely many observations there is not enough information to get the precise value of λ.

We will show that under assumption (34.0.5), $\hat{\boldsymbol{\beta}}$ and s^2 are *consistent*. However this assumption is really too strong for consistency. A weaker set of assumptions is the Grenander conditions, see [**Gre97**, p. 275]. To write down the Grenander conditions, remember that presently $\boldsymbol{X}$ depends on n (in that we only look at the first n elements of y and first n rows of $\boldsymbol{X}$), therefore also the column vectors $\boldsymbol{x}_j$ also depend of n (although we are not indicating this here). Therefore $\boldsymbol{x}_j^\top\boldsymbol{x}_j$ depends on n as well, and we will make this dependency explicit by writing $\boldsymbol{x}_j^\top\boldsymbol{x}_j = d_{nj}^2$. Then the first Grenander condition is $\lim_{n\to\infty} d_{nj}^2 = +\infty$ for all j. Second: for all i and k, $\lim_{n\to\infty}\max_{i=1\cdots n} x_{ij}/d_{nj}^2 = 0$ (here is a typo in Greene, he leaves the max out). Third: Sample correlation matrix of the columns of $\boldsymbol{X}$ minus the constant term converges to a nonsingular matrix.

Consistency means that the *probability limit* of the estimates converges towards the true value. For $\hat{\boldsymbol{\beta}}$ this can be written as $\operatorname{plim}_{n\to\infty}\hat{\boldsymbol{\beta}}_n = \boldsymbol{\beta}$. This means by definition that for all $\varepsilon > 0$ follows $\lim_{n\to\infty}\Pr[|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}| \le \varepsilon] = 1$.

The probability limit is one of several concepts of limits used in probability theory. We will need the following properties of the plim here:

(1) For nonrandom magnitudes, the probability limit is equal to the ordinary limit.

(2) It satisfies the Slutsky theorem, that for a continuous function g,

$$\operatorname{plim} g(\boldsymbol{z}) = g(\operatorname{plim}(\boldsymbol{z})). \tag{34.0.6}$$

(3) If the $\mathcal{MSE}$-matrix of an estimator converges towards the null matrix, then the estimator is consistent.

(4) Kinchine's theorem: the sample mean of an i.i.d. distribution is a consistent estimate of the population mean, even if the distribution does not have a population variance.

34.1. Consistency of the OLS estimator

For the proof of consistency of the OLS estimators $\hat{\boldsymbol{\beta}}$ and of s^2 we need the following result:

$$\operatorname{plim}\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{\varepsilon} = \boldsymbol{o}. \tag{34.1.1}$$

I.e., the true $\boldsymbol{\varepsilon}$ is asymptotically orthogonal to all columns of $\boldsymbol{X}$. This follows immediately from $\mathcal{MSE}[\boldsymbol{o};\boldsymbol{X}^\top\boldsymbol{\varepsilon}/n] = \mathcal{E}[\boldsymbol{X}^\top\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\boldsymbol{X}/n^2] = \sigma^2\boldsymbol{X}^\top\boldsymbol{X}/n^2$, which converges towards $\boldsymbol{O}$.

In order to prove consistency of $\hat{\boldsymbol{\beta}}$ and s^2, transform the formulas for $\hat{\boldsymbol{\beta}}$ and s^2 in such a way that they are written as continuous functions of terms each of which

converges for $n \to \infty$, and then apply Slutsky's theorem. Write $\hat{\boldsymbol{\beta}}$ as

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{\varepsilon} = \boldsymbol{\beta} + \Big(\frac{\boldsymbol{X}^\top \boldsymbol{X}}{n}\Big)^{-1}\frac{\boldsymbol{X}^\top \boldsymbol{\varepsilon}}{n} \tag{34.1.2}$$

$$\operatorname{plim}\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \lim\Big(\frac{\boldsymbol{X}^\top \boldsymbol{X}}{n}\Big)^{-1}\operatorname{plim}\frac{\boldsymbol{X}^\top \boldsymbol{\varepsilon}}{n} \tag{34.1.3}$$

$$= \boldsymbol{\beta} + \boldsymbol{Q}^{-1}\boldsymbol{o} = \boldsymbol{\beta}. \tag{34.1.4}$$

Let's look at the geometry of this when there is only one explanatory variable. The specification is therefore $\mathbf{y} = \boldsymbol{x}\beta + \boldsymbol{\varepsilon}$. The assumption is that $\boldsymbol{\varepsilon}$ is asymptotically orthogonal to $\boldsymbol{x}$. In small samples, it only happens by sheer accident with probability 0 that $\boldsymbol{\varepsilon}$ is orthogonal to $\boldsymbol{x}$. Only $\hat{\boldsymbol{\varepsilon}}$ is. But now let's assume the sample grows larger, i.e., the vectors $\mathbf{y}$ and $\boldsymbol{x}$ become very high-dimensional observation vectors, i.e. we are drawing here a two-dimensional subspace out of a very high-dimensional space. As more and more data are added, the observation vectors also become longer and longer. But if we divide each vector by $\sqrt{n}$, then the lengths of these normalized lenghts stabilize. The squared length of the vector $\boldsymbol{\varepsilon}/\sqrt{n}$ has the plim of σ^2. Furthermore, assumption (34.0.5) means in our case that $\operatorname{plim}_{n\to\infty}\frac{1}{n}\boldsymbol{x}^\top\boldsymbol{x}$ exists and is nonsingular. This is the squared length of $\frac{1}{\sqrt{n}}\boldsymbol{x}$. I.e., if we normalize the vectors by dividing them by $\sqrt{n}$, then they do not get longer but converge towards a finite length. And the result (34.1.1) $\operatorname{plim}\frac{1}{n}\boldsymbol{x}^\top\boldsymbol{\varepsilon} = 0$ means now that with this normalization, $\boldsymbol{\varepsilon}/\sqrt{n}$ becomes more and more orthogonal to $\boldsymbol{x}/\sqrt{n}$. I.e., if n is large enough, asymptotically, not only $\hat{\boldsymbol{\varepsilon}}$ but also the true $\boldsymbol{\varepsilon}$ is orthogonal to $\boldsymbol{x}$, and this means that asymptotically $\hat{\boldsymbol{\beta}}$ converges towards the true $\boldsymbol{\beta}$.

For the proof of consistency of s^2 we need, among others, that $\operatorname{plim}\frac{\boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon}}{n} = \sigma^2$, which is a consequence of Kinchine's theorem. Since $\hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}} = \boldsymbol{\varepsilon}^\top\boldsymbol{M}\boldsymbol{\varepsilon}$ it follows

$$\frac{\hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}}}{n-k} = \frac{n}{n-k}\boldsymbol{\varepsilon}^\top\Big(\frac{\boldsymbol{I}}{n} - \frac{\boldsymbol{X}}{n}\Big(\frac{\boldsymbol{X}^\top\boldsymbol{X}}{n}\Big)^{-1}\frac{\boldsymbol{X}^\top}{n}\Big)\boldsymbol{\varepsilon} =$$

$$= \frac{n}{n-k}\Big(\frac{\boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon}}{n} - \frac{\boldsymbol{\varepsilon}^\top\boldsymbol{X}}{n}\Big(\frac{\boldsymbol{X}^\top\boldsymbol{X}}{n}\Big)^{-1}\frac{\boldsymbol{X}^\top\boldsymbol{\varepsilon}}{n}\Big) \to 1\cdot\Big(\sigma^2 - \boldsymbol{o}^\top\boldsymbol{Q}^{-1}\boldsymbol{o}\Big).$$

34.2. Asymptotic Normality of the Least Squares Estimator

To show asymptotic normality of an estimator, multiply the sampling error by $\sqrt{n}$, so that the variance is stabilized.

We have seen $\operatorname{plim}\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{\varepsilon} = \boldsymbol{o}$. Now look at $\frac{1}{\sqrt{n}}\boldsymbol{X}^\top\boldsymbol{\varepsilon}_n$. Its mean is $\boldsymbol{o}$ and its covariance matrix $\sigma^2\frac{\boldsymbol{X}^\top\boldsymbol{X}}{n}$. Shape of distribution, due to a variant of the Central Limit Theorem, is asymptotically normal: $\frac{1}{\sqrt{n}}\boldsymbol{X}^\top\boldsymbol{\varepsilon}_n \to \mathsf{N}(\boldsymbol{o}, \sigma^2\boldsymbol{Q})$. (Here the convergence is convergence in distribution.)

We can write $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) = \Big(\frac{\boldsymbol{X}^\top\boldsymbol{X}}{n}\Big)^{-1}(\frac{1}{\sqrt{n}}\boldsymbol{X}^\top\boldsymbol{\varepsilon}_n)$. Therefore its limiting covariance matrix is $\boldsymbol{Q}^{-1}\sigma^2\boldsymbol{Q}\boldsymbol{Q}^{-1} = \sigma^2\boldsymbol{Q}^{-1}$, Therefore $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \to \mathsf{N}(\boldsymbol{o}, \sigma^2\boldsymbol{Q}^{-1})$ in distribution. One can also say: the asymptotic distribution of $\hat{\boldsymbol{\beta}}$ is $\mathsf{N}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1})$.

From this follows $\sqrt{n}(\boldsymbol{R}\hat{\boldsymbol{\beta}}_n - \boldsymbol{R}\boldsymbol{\beta}) \to \mathsf{N}(\boldsymbol{o}, \sigma^2\boldsymbol{R}\boldsymbol{Q}^{-1}\boldsymbol{R}^\top)$, and therefore

$$n(\boldsymbol{R}\hat{\boldsymbol{\beta}}_n - \boldsymbol{R}\boldsymbol{\beta})\big(\boldsymbol{R}\boldsymbol{Q}^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}_n - \boldsymbol{R}\boldsymbol{\beta}) \to \sigma^2\chi^2_i. \tag{34.2.1}$$

Divide by s^2 and replace in the limiting case $\boldsymbol{Q}$ by $\boldsymbol{X}^\top\boldsymbol{X}/n$ and s^2 by σ^2 to get

$$\frac{(\boldsymbol{R}\hat{\boldsymbol{\beta}}_n - \boldsymbol{R}\boldsymbol{\beta})\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}_n - \boldsymbol{R}\boldsymbol{\beta})}{\mathsf{s}^2} \to \chi^2_i \tag{34.2.2}$$

in distribution. All this is not a proof; the point is that in the denominator, the distribution is divided by the increasingly bigger number $n-k$, while in the numerator, it is divided by the constant i; therefore asymptotically the denominator can be considered 1.

The central limit theorems only say that for $n \to \infty$ these converge towards the χ^2, which is asymptotically equal to the F distribution. It is easily possible that before one gets to the limit, the F-distribution is better.

Problem 393. *Are the residuals $\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ asymptotically normally distributed?*

Answer. Only if the disturbances are normal, otherwise of course not! We can show that $\sqrt{n}(\boldsymbol{\varepsilon} - \hat{\boldsymbol{\varepsilon}}) = \sqrt{n}\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim \mathsf{N}(\boldsymbol{o}, \sigma^2\boldsymbol{X}\boldsymbol{Q}\boldsymbol{X}^\top)$. □

Now these results also go through if one has stochastic regressors. [**Gre97**, 6.7.7] shows that the above condition (34.0.5) with the lim replaced by plim holds if $\boldsymbol{x}_i$ and $\boldsymbol{\varepsilon}_i$ are an i.i.d. sequence of random variables.

Problem 394. *2 points In the regression model with random regressors $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, you only know that $\operatorname{plim}\frac{1}{n}\mathbf{X}^\top\mathbf{X} = \boldsymbol{Q}$ is a nonsingular matrix, and $\operatorname{plim}\frac{1}{n}\mathbf{X}^\top\boldsymbol{\varepsilon} = \boldsymbol{o}$. Using these two conditions, show that the OLS estimate is consistent.*

Answer. $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} = \boldsymbol{\beta} + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\varepsilon}$ due to (24.0.7), and

$$\operatorname{plim}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\varepsilon} = \operatorname{plim}(\frac{\mathbf{X}^\top\mathbf{X}}{n})^{-1}\frac{\mathbf{X}^\top\boldsymbol{\varepsilon}}{n} = \boldsymbol{Q}\boldsymbol{o} = \boldsymbol{o}.$$

□

CHAPTER 35

Least Squares as the Normal Maximum Likelihood Estimate

Now assume $\boldsymbol{\varepsilon}$ is multivariate normal. We will show that in this case the OLS estimator $\hat{\boldsymbol{\beta}}$ is at the same time the Maximum Likelihood Estimator. For this we need to write down the density function of $\mathbf{y}$. First look at one y_t which is $\mathsf{y}_t \sim \mathsf{N}(\boldsymbol{x}_t^\top \boldsymbol{\beta}, \sigma^2)$, where $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^\top \\ \vdots \\ \boldsymbol{x}_n^\top \end{bmatrix}$, i.e., $\boldsymbol{x}_t$ is the tth row of $\boldsymbol{X}$. It is written as a column vector, since we follow the "column vector convention." The (marginal) density function for this one observation is

$$f_{\mathsf{y}_t}(y_t) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y_t - \boldsymbol{x}_t^\top \boldsymbol{\beta})^2/2\sigma^2}. \tag{35.0.3}$$

Since the y_i are stochastically independent, their joint density function is the product, which can be written as

$$f_{\mathbf{y}}(\boldsymbol{y}) = (2\pi\sigma^2)^{-n/2} \exp\Big(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\Big). \tag{35.0.4}$$

To compute the maximum likelihood estimator, it is advantageous to start with the log likelihood function:

$$\log f_{\mathbf{y}}(\mathbf{y}; \boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}). \tag{35.0.5}$$

Assume for a moment that σ^2 is known. Then the MLE of $\boldsymbol{\beta}$ is clearly equal to the OLS $\hat{\boldsymbol{\beta}}$. Since $\hat{\boldsymbol{\beta}}$ does not depend on σ^2, it is also the maximum likelihood estimate when σ^2 is unknown. $\hat{\boldsymbol{\beta}}$ is a linear function of $\mathbf{y}$. Linear transformations of normal variables are normal. Normal distributions are characterized by their mean vector and covariance matrix. The distribution of the MLE of $\boldsymbol{\beta}$ is therefore $\hat{\boldsymbol{\beta}} \sim \mathsf{N}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1})$.

If we replace $\boldsymbol{\beta}$ in the log likelihood function (35.0.5) by $\hat{\boldsymbol{\beta}}$, we get what is called the log likelihood function with $\boldsymbol{\beta}$ "concentrated out."

$$\log f_{\mathbf{y}}(\mathbf{y}; \boldsymbol{\beta} = \hat{\boldsymbol{\beta}}, \sigma^2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top (\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}). \tag{35.0.6}$$

One gets the maximum likelihood estimate of σ^2 by maximizing this "concentrated" log likelihoodfunction. Taking the derivative with respect to σ^2 (consider σ^2 the name of a variable, not the square of another variable), one gets

$$\frac{\partial}{\partial\sigma^2}\log f_{\mathbf{y}}(\mathbf{y}; \hat{\boldsymbol{\beta}}) = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top (\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \tag{35.0.7}$$

Setting this zero gives

$$\tilde{\sigma}^2 = \frac{(\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top (\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})}{n} = \frac{\hat{\boldsymbol{\varepsilon}}^\top \hat{\boldsymbol{\varepsilon}}}{n}. \tag{35.0.8}$$

This is a scalar multiple of the unbiased estimate $\mathsf{s}^2 = \hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}}/(n-k)$ which we had earlier.

Let's look at the distribution of s^2 (from which that of its scalar multiples follows easily). It is a quadratic form in a normal variable. Such quadratic forms very often have χ^2 distributions.

Now recall equation 10.4.9 characterizing all the quadratic forms of multivariate normal variables that are χ^2's. Here it is again: Assume $\mathbf{y}$ is a multivariate normal vector random variable with mean vector $\boldsymbol{\mu}$ and covariance matrix $\sigma^2\boldsymbol{\Psi}$, and $\boldsymbol{\Omega}$ is a symmetric nonnegative definite matrix. Then $(\mathbf{y} - \boldsymbol{\mu})^\top \boldsymbol{\Omega}(\mathbf{y} - \boldsymbol{\mu}) \sim \sigma^2\chi^2_k$ iff

$$\boldsymbol{\Psi}\boldsymbol{\Omega}\boldsymbol{\Psi}\boldsymbol{\Omega}\boldsymbol{\Psi} = \boldsymbol{\Psi}\boldsymbol{\Omega}\boldsymbol{\Psi}, \tag{35.0.9}$$

and k is the rank of $\boldsymbol{\Psi}\boldsymbol{\Omega}$.

This condition is satisfied in particular if $\boldsymbol{\Psi} = \boldsymbol{I}$ (the identity matrix) and $\boldsymbol{\Omega}^2 = \boldsymbol{\Omega}$, and this is exactly our situation.

$$\hat{\sigma}^2 = \frac{(\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top (\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})}{n-k} = \frac{\boldsymbol{\varepsilon}^\top(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top)\boldsymbol{\varepsilon}}{n-k} = \frac{\boldsymbol{\varepsilon}^\top\boldsymbol{M}\boldsymbol{\varepsilon}}{n-k} \tag{35.0.10}$$

where $\boldsymbol{M}^2 = \boldsymbol{M}$ and $\operatorname{rank}\boldsymbol{M} = n-k$. (This last identity because for idempotent matrices, rank = tr, and we computed its tr above.) Therefore $\mathsf{s}^2 \sim \sigma^2\chi^2_{n-k}/(n-k)$, from which one obtains again unbiasedness, but also that $\operatorname{var}[\mathsf{s}^2] = 2\sigma^4/(n-k)$, a result that one cannot get from mean and variance alone.

Problem 395. *4 points Show that, if $\mathbf{y}$ is normally distributed, s^2 and $\hat{\boldsymbol{\beta}}$ are independent.*

Answer. We showed in question 300 that $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\varepsilon}}$ are uncorrelated, therefore in the normal case independent, therefore $\hat{\boldsymbol{\beta}}$ is also independent of any function of $\hat{\boldsymbol{\varepsilon}}$, such as $\hat{\sigma}^2$. □

PROBLEM 396. *Computer assignment: You run a regression with 3 explanatory variables, no constant term, the sample size is 20, the errors are normally distributed and you know that* $\sigma^2 = 2$. *Plot the density function of* s^2. *Hint: The command* `dchisq(x,df=25)` *returns the density of a Chi-square distribution with 25 degrees of freedom evaluated at* x. *But the number 25 was only taken as an example, this is not the number of degrees of freedom you need here.*

• a. *In the same plot, plot the density function of the Theil-Schweitzer estimate. Can one see from the comparison of these density functions why the Theil-Schweitzer estimator has a better* MSE*?*

ANSWER. Start with the Theil-Schweitzer plot, because it is higher. `> x <- seq(from = 0, to = 6, by = 0.01) > Density <- (19/2)*dchisq((19/2)*x, df=17) > plot(x, Density, type="l", lty=2) > lines(x,(17/2)*dchisq((17/2)*x, df=17)) > title(main = "Unbiased versus Theil-Schw Variance Estimate, 17 d.f.")` □

Now let us derive the maximum likelihood estimator in the case of nonspherical but positive definite covariance matrix. I.e., the model is $\mathsf{y} = \boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathsf{N}(\boldsymbol{o}, \sigma^2\boldsymbol{\Psi})$. The density function is

$$f_{\mathsf{y}}(\boldsymbol{y}) = (2\pi\sigma^2)^{-n/2}\,|\det\boldsymbol{\Psi}|^{-1/2}\exp\Big(-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^\top\boldsymbol{\Psi}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\Big). \tag{35.0.11}$$

PROBLEM 397. *Derive* (35.0.11) *as follows: Take a matrix* $\boldsymbol{P}$ *with the property that* $\boldsymbol{P}\boldsymbol{\varepsilon}$ *has covariance matrix* $\sigma^2\boldsymbol{I}$. *Write down the joint density function of* $\boldsymbol{P}\boldsymbol{\varepsilon}$. *Since* y *is a linear transformation of* $\boldsymbol{\varepsilon}$, *one can apply the rule for the density function of a transformed random variable.*

ANSWER. Write $\boldsymbol{\Psi} = \boldsymbol{Q}\boldsymbol{Q}^\top$ with $\boldsymbol{Q}$ nonsingular and define $\boldsymbol{P} = \boldsymbol{Q}^{-1}$ and $\boldsymbol{v} = \boldsymbol{P}\boldsymbol{\varepsilon}$. Then $\mathcal{V}[\boldsymbol{v}] = \sigma^2\boldsymbol{P}\boldsymbol{Q}\boldsymbol{Q}^\top\boldsymbol{P}^\top = \sigma^2\boldsymbol{I}$, therefore

$$f_{\boldsymbol{v}}(\boldsymbol{v}) = (2\pi\sigma^2)^{-n/2}\exp\Big(-\frac{1}{2\sigma^2}\boldsymbol{v}^\top\boldsymbol{v}\Big). \tag{35.0.12}$$

For the transformation rule, write $\boldsymbol{v}$, whose density function you know, as a function of y, whose density function you want to know. $\boldsymbol{v} = \boldsymbol{P}(\mathsf{y}-\boldsymbol{X}\boldsymbol{\beta})$; therefore the Jacobian matrix is $\partial\boldsymbol{v}/\partial\mathsf{y}^\top = \partial(\boldsymbol{P}\mathsf{y}-\boldsymbol{P}\boldsymbol{X}\boldsymbol{\beta})/\partial\mathsf{y}^\top = \boldsymbol{P}$, or one can see it also element by element

$$\begin{bmatrix} \frac{\partial \mathsf{v}_1}{\partial \mathsf{y}_1} & \cdots & \frac{\partial \mathsf{v}_1}{\partial \mathsf{y}_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial \mathsf{v}_n}{\partial \mathsf{y}_1} & \cdots & \frac{\partial \mathsf{v}_n}{\partial \mathsf{y}_n} \end{bmatrix} = \boldsymbol{P}, \tag{35.0.13}$$

therefore one has to do two things: first, substitute $\boldsymbol{P}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})$ for $\boldsymbol{v}$ in formula (35.0.12), and secondly multiply by the absolute value of the determinant of the Jacobian. Here is how to express the determinant of the Jacobian in terms of $\boldsymbol{\Psi}$: From $\boldsymbol{\Psi}^{-1} = (\boldsymbol{Q}\boldsymbol{Q}^\top)^{-1} = (\boldsymbol{Q}^\top)^{-1}\boldsymbol{Q}^{-1} = (\boldsymbol{Q}^{-1})^\top\boldsymbol{Q}^{-1} = \boldsymbol{P}^\top\boldsymbol{P}$ follows $(\det\boldsymbol{P})^2 = (\det\boldsymbol{\Psi})^{-1}$, hence $|\det\boldsymbol{P}| = \sqrt{\det\boldsymbol{\Psi}}$. □

From (35.0.11) one obtains the following log likelihood function:

$$\log f_{\mathsf{y}}(\mathsf{y}) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2}\ln\det[\boldsymbol{\Psi}] - \frac{1}{2\sigma^2}(\mathsf{y}-\boldsymbol{X}\boldsymbol{\beta})^\top\boldsymbol{\Psi}^{-1}(\mathsf{y}-\boldsymbol{X}\boldsymbol{\beta}). \tag{35.0.14}$$

Here, usually not only the elements of $\boldsymbol{\beta}$ are unknown, but also $\boldsymbol{\Psi}$ depends on unknown parameters. Instead of concentrating out $\boldsymbol{\beta}$, we will first concentrate out σ^2, i.e., we will compute the maximum of this likelihood function over σ^2 for any given set of values for the data and the other parameters:

$$\frac{\partial}{\partial\sigma^2}\log f_{\mathsf{y}}(\mathsf{y}) = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{(\mathsf{y}-\boldsymbol{X}\boldsymbol{\beta})^\top\boldsymbol{\Psi}^{-1}(\mathsf{y}-\boldsymbol{X}\boldsymbol{\beta})}{2\sigma^4} \tag{35.0.15}$$

$$\tilde{\sigma}^2 = \frac{(\mathsf{y}-\boldsymbol{X}\boldsymbol{\beta})^\top\boldsymbol{\Psi}^{-1}(\mathsf{y}-\boldsymbol{X}\boldsymbol{\beta})}{n}. \tag{35.0.16}$$

Whatever the value of $\boldsymbol{\beta}$ or the values of the unknown parameters in $\boldsymbol{\Psi}$, $\tilde{\sigma}^2$ is the value of σ^2 which, together with the given $\boldsymbol{\beta}$ and $\boldsymbol{\Psi}$, gives the highest value of the likelihood function. If one plugs this $\tilde{\sigma}^2$ into the likelihood function, one obtains the so-called "concentrated likelihood function" which then only has to be maximized over $\boldsymbol{\beta}$ and $\boldsymbol{\Psi}$:

$$\log f_{\mathsf{y}}(\mathsf{y};\tilde{\sigma}^2) = -\frac{n}{2}(1+\ln 2\pi - \ln n) - \frac{n}{2}\ln(\mathsf{y}-\boldsymbol{X}\boldsymbol{\beta})^\top\boldsymbol{\Psi}^{-1}(\mathsf{y}-\boldsymbol{X}\boldsymbol{\beta}) - \frac{1}{2}\ln\det[\boldsymbol{\Psi}] \tag{35.0.17}$$

This objective function has to be maximized with respect to $\boldsymbol{\beta}$ and the parameters entering $\boldsymbol{\Psi}$. If $\boldsymbol{\Psi}$ is known, then this is clearly maximized by the $\hat{\boldsymbol{\beta}}$ minimizing (26.0.9), therefore the GLS estimator is also the maximum likelihood estimator.

If $\boldsymbol{\Psi}$ depends on unknown parameters, it is interesting to compare the maximum likelihood estimator with the nonlinear least squares estimator. The objective function minimized by nonlinear least squares is $(\mathsf{y}-\boldsymbol{X}\boldsymbol{\beta})^\top\boldsymbol{\Psi}^{-1}(\mathsf{y}-\boldsymbol{X}\boldsymbol{\beta})$, which is the sum of squares of the innovation parts of the residuals. These two objective functions therefore differ by the factor $(\det[\boldsymbol{\Psi}])^{\frac{1}{n}}$, which only matters if there are unknown parameters in $\boldsymbol{\Psi}$. Asymptotically, the objective functions are identical.

Using the factorization theorem for sufficient statistics, one also sees easily that $\hat{\sigma}^2$ and $\hat{\boldsymbol{\beta}}$ together form sufficient statistics for σ^2 and $\boldsymbol{\beta}$. For this use the identity

$$\begin{aligned}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) &= (\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}) + (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^\top\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) \\ &= (n-k)\mathsf{s}^2 + (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^\top\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}).\end{aligned} \tag{35.0.18}$$

Therefore the observation $\boldsymbol{y}$ enters the likelihood function only through the two statistics $\hat{\boldsymbol{\beta}}$ and s^2. The factorization of the likelihood function is therefore the trivial factorization in which that part which does not depend on the unknown parameters but only on the data is unity.

PROBLEM 398. *12 points The log likelihood function in the linear model is given by* (35.0.5). *Show that the inverse of the information matrix is*

$$\begin{bmatrix} \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1} & \boldsymbol{o} \\ \boldsymbol{o}^\top & 2\sigma^4/n \end{bmatrix} \tag{35.0.19}$$

The information matrix can be obtained in two different ways. Its typical element has the following two forms:

$$\mathrm{E}[\frac{\partial \ln \ell}{\partial \theta_i}\frac{\partial \ln \ell}{\partial \theta_k}] = -\mathrm{E}[\frac{\partial^2 \ln \ell}{\partial \theta_i \partial \theta_k}], \tag{35.0.20}$$

or written as matrix derivatives

$$\mathcal{E}[\frac{\partial \ln \ell}{\partial \boldsymbol{\theta}}\frac{\partial \ln \ell}{\partial \boldsymbol{\theta}^\top}] = -\mathcal{E}[\frac{\partial^2 \ln \ell}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\top}]. \tag{35.0.21}$$

In our case $\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\beta} \\ \sigma^2 \end{bmatrix}$. *The expectation is taken under the assumption that the parameter values are the true values. Compute it both ways.*

ANSWER. The log likelihood function can be written as

$$\ln \ell = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln \sigma^2 - \frac{1}{2\sigma^2}(\mathsf{y}^\top\mathsf{y} - 2\mathsf{y}^\top\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta}). \tag{35.0.22}$$

The first derivatives were already computed for the maximum likelihood estimators:

$$\frac{\partial}{\partial \boldsymbol{\beta}^\top}\ln \ell = -\frac{1}{2\sigma^2}(2\mathsf{y}^\top\boldsymbol{X} + 2\boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{X}) = \frac{1}{\sigma^2}(\mathsf{y} - \boldsymbol{X}\boldsymbol{\beta})^\top\boldsymbol{X} = \frac{1}{\sigma^2}\boldsymbol{\varepsilon}^\top\boldsymbol{X} \tag{35.0.23}$$

$$\frac{\partial}{\partial \sigma^2}\ln \ell = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathsf{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\mathsf{y} - \boldsymbol{X}\boldsymbol{\beta}) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon} \tag{35.0.24}$$

By the way, one sees that each of these has expected value zero, which is a fact that is needed to prove consistency of the maximum likelihood estimator.

The formula with only one partial derivative will be given first, although it is more tedious: By doing $\frac{\partial}{\partial \boldsymbol{\beta}^\top}\left(\frac{\partial}{\partial \boldsymbol{\beta}^\top}\right)^\top$ we get a symmetric 2×2 partitioned matrix with the diagonal elements

$$\mathcal{E}[\frac{1}{\sigma^4}\boldsymbol{X}^\top\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\boldsymbol{X}] = \frac{1}{\sigma^2}\boldsymbol{X}^\top\boldsymbol{X} \tag{35.0.25}$$

and

$$\mathrm{E}[\Big(-\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon}\Big)^2] = \mathrm{var}[-\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon}] = \mathrm{var}[\frac{1}{2\sigma^4}\boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon}] = \frac{1}{4\sigma^8}2n\sigma^4 = \frac{n}{2\sigma^4} \tag{35.0.26}$$

One of the off-diagonal elements is $(\frac{n}{2\sigma^4} + \frac{1}{2\sigma^6}\boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon})\boldsymbol{\varepsilon}^\top\boldsymbol{X}$. Its expected value is zero: $\mathcal{E}[\boldsymbol{\varepsilon}] = \boldsymbol{o}$, and also $\mathcal{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon}] = \boldsymbol{o}$ since its ith component is $\mathrm{E}[\varepsilon_i\sum_j \varepsilon_j^2] = \sum_j \mathrm{E}[\varepsilon_i\varepsilon_j^2]$. If $i \neq j$, then ε_i is independent of ε_j^2, therefore $\mathrm{E}[\varepsilon_i\varepsilon_j^2] = 0 \cdot \sigma^2 = 0$. If $i = j$, we get $\mathrm{E}[\varepsilon_i^3] = 0$ since ε_i has a symmetric distribution.

It is easier if we differentiate once more:

$$\frac{\partial^2}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^\top}\ln \ell = -\frac{1}{\sigma^2}\boldsymbol{X}^\top\boldsymbol{X} \tag{35.0.27}$$

$$\frac{\partial^2}{\partial \boldsymbol{\beta}\partial \sigma^2}\ln \ell = -\frac{1}{\sigma^4}\boldsymbol{X}^\top(\mathsf{y} - \boldsymbol{X}\boldsymbol{\beta}) = -\frac{1}{\sigma^4}\boldsymbol{X}^\top\boldsymbol{\varepsilon} \tag{35.0.28}$$

$$\frac{\partial^2}{(\partial \sigma^2)^2}\ln \ell = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}(\mathsf{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\mathsf{y} - \boldsymbol{X}\boldsymbol{\beta}) = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon} \tag{35.0.29}$$

This gives the top matrix in [**JHG+88**, (6.1.24b)]:

$$\begin{bmatrix} -\frac{1}{\sigma^2}\boldsymbol{X}^\top\boldsymbol{X} & -\frac{1}{\sigma^4}(\boldsymbol{X}^\top\mathsf{y} - \boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta}) \\ -\frac{1}{\sigma^4}(\boldsymbol{X}^\top\mathsf{y} - \boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta})^\top & \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}(\mathsf{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\mathsf{y} - \boldsymbol{X}\boldsymbol{\beta}) \end{bmatrix} \tag{35.0.30}$$

Now assume that $\boldsymbol{\beta}$ and σ^2 are the true values, take expected values, and reverse the sign. This gives the information matrix

$$\begin{bmatrix} \sigma^{-2}\boldsymbol{X}^\top\boldsymbol{X} & \boldsymbol{o} \\ \boldsymbol{o}^\top & n/(2\sigma^4) \end{bmatrix} \tag{35.0.31}$$

For the lower righthand side corner we need that $\mathrm{E}[(\mathsf{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\mathsf{y} - \boldsymbol{X}\boldsymbol{\beta})] = \mathrm{E}[\boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon}] = n\sigma^2$.

Taking inverses gives (35.0.19), which is a lower bound for the covariance matrix; we see that s^2 with $\mathrm{var}[\mathsf{s}^2] = 2\sigma^4/(n-k)$ does not attain the bound. However one can show with other means that it is nevertheless efficient. □

CHAPTER 36

Bayesian Estimation in the Linear Model

The model is $\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim \mathsf{N}(\boldsymbol{o}, \sigma^2\boldsymbol{I})$. Both $\mathbf{y}$ and $\boldsymbol{\beta}$ are random. The distribution of $\boldsymbol{\beta}$, called the "prior information," is $\boldsymbol{\beta} \sim \mathsf{N}(\boldsymbol{\nu}, \tau^2\boldsymbol{A}^{-1})$. (Bayesians work with the precision matrix, which is the inverse of the covariance matrix). Furthermore $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$ are assumed independent. Define $\kappa^2 = \sigma^2/\tau^2$. To simplify matters, we assume that κ^2 is known.

Whether or not the probability is subjective, this specification implies that $\mathbf{y}$ and $\boldsymbol{\beta}$ are jointly Normal and

$$\begin{bmatrix}\mathbf{y}\\ \boldsymbol{\beta}\end{bmatrix} \sim \begin{bmatrix}\boldsymbol{X}\boldsymbol{\nu}\\ \boldsymbol{\nu}\end{bmatrix}, \tau^2\begin{bmatrix}\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{X}^\top + \kappa^2\boldsymbol{I} & \boldsymbol{X}\boldsymbol{A}^{-1}\\ \boldsymbol{A}^{-1}\boldsymbol{X}^\top & \boldsymbol{A}^{-1}\end{bmatrix}. \tag{36.0.32}$$

We can use theorem **??** to compute the best linear predictor $\hat{\hat{\boldsymbol{\beta}}}(\mathbf{y})$ of $\boldsymbol{\beta}$ on the basis of an observation of $\mathbf{y}$. Due to Normality, $\hat{\hat{\boldsymbol{\beta}}}$ is at the same time the conditional mean or "posterior mean" $\hat{\hat{\boldsymbol{\beta}}} = \mathcal{E}[\boldsymbol{\beta}|\mathbf{y}]$, and the $\mathcal{MSE}$-matrix is at the same time the variance of the posterior distribution of $\boldsymbol{\beta}$ given $\boldsymbol{y}$ $\mathcal{MSE}[\hat{\hat{\boldsymbol{\beta}}};\boldsymbol{\beta}] = \mathcal{V}[\hat{\hat{\boldsymbol{\beta}}}|\boldsymbol{y}]$. A proof is given as answer to Question **??**. Since one knows mean and variance of the posterior distribution, and since the posterior distribution is normal, the posterior distribution of $\boldsymbol{\beta}$ given $\boldsymbol{y}$ is known. This distribution is what the Bayesians are after. The posterior distribution combines *all* the information, prior information *and* sample information, about $\boldsymbol{\beta}$.

According to (**??**), this posterior mean can be written as

$$\hat{\hat{\boldsymbol{\beta}}} = \boldsymbol{\nu} + \boldsymbol{B}^*(\mathbf{y} - \boldsymbol{X}\boldsymbol{\nu}) \tag{36.0.33}$$

where $\boldsymbol{B}^*$ is the solution of the "normal equation" (**??**) which reads here

$$\boldsymbol{B}^*(\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{X}^\top + \kappa^2\boldsymbol{I}) = \boldsymbol{A}^{-1}\boldsymbol{X}^\top \tag{36.0.34}$$

The obvious solution of (36.0.34) is $\boldsymbol{B}^* = \boldsymbol{A}^{-1}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{X}^\top + \kappa^2\boldsymbol{I})^{-1}$, and according to (**??**), the $\mathcal{MSE}$-matrix of the predictor is

$$\tau^2(\boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{X}^\top + \kappa^2\boldsymbol{I})^{-1}\boldsymbol{X}\boldsymbol{A}^{-1})$$

These formulas are correct, but the Bayesians use mathematically equivalent formulas which have a simpler and more intuitive form. The solution of (36.0.34) can also be written as

$$\boldsymbol{B}^* = (\boldsymbol{X}^\top\boldsymbol{X} + \kappa^2\boldsymbol{A})^{-1}\boldsymbol{X}^\top, \tag{36.0.35}$$

and (36.0.33) becomes

$$\hat{\hat{\boldsymbol{\beta}}} = (\boldsymbol{X}^\top\boldsymbol{X} + \kappa^2\boldsymbol{A})^{-1}(\boldsymbol{X}^\top\mathbf{y} + \kappa^2\boldsymbol{A}\boldsymbol{\nu}) \tag{36.0.36}$$

$$= (\boldsymbol{X}^\top\boldsymbol{X} + \kappa^2\boldsymbol{A})^{-1}(\boldsymbol{X}^\top\boldsymbol{X}\hat{\boldsymbol{\beta}} + \kappa^2\boldsymbol{A}\boldsymbol{\nu}) \tag{36.0.37}$$

where $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\mathbf{y}$ is the OLS estimate. Bayesians are interested in $\hat{\hat{\boldsymbol{\beta}}}$ because this is the posterior mean. The $\mathcal{MSE}$-matrix, which is the posterior covariance matrix, can also be written as

$$\mathcal{MSE}[\hat{\hat{\boldsymbol{\beta}}};\boldsymbol{\beta}] = \sigma^2(\boldsymbol{X}^\top\boldsymbol{X} + \kappa^2\boldsymbol{A})^{-1} \tag{36.0.38}$$

Problem 399. *Show that* $\boldsymbol{B}^*$ *as defined in* (36.0.35) *satisfies* (36.0.34), *that* (36.0.33) *with this* $\boldsymbol{B}^*$ *becomes* (36.0.36), *and that* (36) *becomes* (36.0.38).

Answer. (36.0.35) in the normal equation (36.0.34) gives

$$\begin{aligned}(\boldsymbol{X}^\top\boldsymbol{X} + \kappa^2\boldsymbol{A})^{-1}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{X}^\top + \kappa^2\boldsymbol{I}) &= (\boldsymbol{X}^\top\boldsymbol{X} + \kappa^2\boldsymbol{A})^{-1}(\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{X}^\top + \kappa^2\boldsymbol{X}^\top) =\\ &= (\boldsymbol{X}^\top\boldsymbol{X} + \kappa^2\boldsymbol{A})^{-1}(\boldsymbol{X}^\top\boldsymbol{X} + \kappa^2\boldsymbol{A})\boldsymbol{A}^{-1}\boldsymbol{X}^\top = \boldsymbol{A}^{-1}\boldsymbol{X}^\top.\end{aligned} \tag{36.0.39}$$

Now the solution formula:

$$\hat{\hat{\boldsymbol{\beta}}} = \boldsymbol{\nu} + (\boldsymbol{X}^\top\boldsymbol{X} + \kappa^2\boldsymbol{A})^{-1}\boldsymbol{X}^\top(\mathbf{y} - \boldsymbol{X}\boldsymbol{\nu}) \tag{36.0.40}$$

$$= (\boldsymbol{X}^\top\boldsymbol{X} + \kappa^2\boldsymbol{A})^{-1}\Big((\boldsymbol{X}^\top\boldsymbol{X} + \kappa^2\boldsymbol{A})\boldsymbol{\nu} + \boldsymbol{X}^\top\mathbf{y} - \boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\nu}\Big) \tag{36.0.41}$$

$$= (\boldsymbol{X}^\top\boldsymbol{X} + \kappa^2\boldsymbol{A})^{-1}(\boldsymbol{X}^\top\mathbf{y} + \kappa^2\boldsymbol{A}\boldsymbol{\nu}). \tag{36.0.42}$$

For the formula of the $\mathcal{MSE}$ matrix one has to check that (36) times the inverse of (36.0.38) is the identity matrix, or that

$$\Big(\boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{X} + \kappa^2\boldsymbol{I})^{-1}\boldsymbol{X}\boldsymbol{A}^{-1}\Big)\Big(\boldsymbol{X}^\top\boldsymbol{X} + \kappa^2\boldsymbol{A}\Big) = \kappa^2\boldsymbol{I} \tag{36.0.43}$$

Multiplying out gives

$$\begin{aligned}\boldsymbol{A}^{-1}\boldsymbol{X}^\top\boldsymbol{X} + \kappa^2\boldsymbol{I} - \boldsymbol{A}^{-1}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{X} + \kappa^2\boldsymbol{I})^{-1}\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{X}^\top\boldsymbol{X} - \kappa^2\boldsymbol{A}^{-1}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{X} + \kappa^2\boldsymbol{I})^{-1}\boldsymbol{X} =\\ = \boldsymbol{A}^{-1}\boldsymbol{X}^\top\boldsymbol{X} + \kappa^2\boldsymbol{I} - \boldsymbol{A}^{-1}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{X} + \kappa^2\boldsymbol{I})^{-1}(\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{X} + \kappa^2\boldsymbol{I})\boldsymbol{X} = \kappa^2\boldsymbol{I}\end{aligned} \tag{36.0.44}$$

□

The formula (36.0.37) can be given two interpretations, neither of which is necessarily Bayesian. First interpretation: It is a matrix weighted average of the OLS estimate and $\boldsymbol{\nu}$, with the weights being the respective precision matrices. If $\boldsymbol{\nu} = \boldsymbol{o}$, then the matrix weighted average reduces to $\hat{\hat{\boldsymbol{\beta}}} = (\boldsymbol{X}^\top \boldsymbol{X} + \kappa^2 \boldsymbol{A})^{-1} \boldsymbol{X}^\top \mathbf{y}$, which has been called a "shrinkage estimator" (Ridge regression), since the "denominator" is bigger: instead of "dividing by" $\boldsymbol{X}^\top \boldsymbol{X}$ (strictly speaking, multiplying by $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$), one "divides" by $\boldsymbol{X}^\top \boldsymbol{X} + \kappa^2 \boldsymbol{A}$. If $\boldsymbol{\nu} \neq \boldsymbol{o}$ then the OLS estimate $\hat{\boldsymbol{\beta}}$ is "shrunk" not in direction of the origin but in direction of $\boldsymbol{\nu}$.

Second interpretation: It is as if, in addition to the data $\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, also an independent observation $\boldsymbol{\nu} = \boldsymbol{\beta} + \boldsymbol{\delta}$ with $\boldsymbol{\delta} \sim \mathsf{N}(\boldsymbol{o}, \tau^2 \boldsymbol{A}^{-1})$ was available, i.e., as if the model was

$$\begin{bmatrix} \mathbf{y} \\ \boldsymbol{\nu} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{I} \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \boldsymbol{\varepsilon} \\ \boldsymbol{\delta} \end{bmatrix} \quad \text{with} \quad \begin{bmatrix} \boldsymbol{\varepsilon} \\ \boldsymbol{\delta} \end{bmatrix} \sim \begin{bmatrix} \boldsymbol{o} \\ \boldsymbol{o} \end{bmatrix}, \tau^2 \begin{bmatrix} \sigma^2 \boldsymbol{I} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{A}^{-1} \end{bmatrix}. \tag{36.0.45}$$

The Least Squares objective function minimized by the GLS estimator $\boldsymbol{\beta} = \hat{\hat{\boldsymbol{\beta}}}$ in (36.0.45) is:

$$(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}) + \kappa^2 (\boldsymbol{\beta} - \boldsymbol{\nu})^\top \boldsymbol{A} (\boldsymbol{\beta} - \boldsymbol{\nu}). \tag{36.0.46}$$

In other words, $\hat{\hat{\boldsymbol{\beta}}}$ is chosen such that at the same time $\boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}}$ is close to $\mathbf{y}$ and $\hat{\hat{\boldsymbol{\beta}}}$ close to $\boldsymbol{\nu}$.

Problem 400. *Show that the objective function* (36.0.46) *is, up to a constant factor, the natural logarithm of the product of the prior density and the likelihood function. (Assume σ^2 and τ^2 known). Note: if $\boldsymbol{z} \sim \mathsf{N}(\boldsymbol{\theta}, \sigma^2 \boldsymbol{\Sigma})$ with nonsingular covariance matrix $\sigma^2 \boldsymbol{\Sigma}$, then its density function is*

$$f_{\boldsymbol{z}}(\boldsymbol{z}) = (2\pi\sigma^2)^{-n/2} \left|\det \boldsymbol{\Sigma}\right|^{-1/2} \exp\Big(-\frac{1}{2\sigma^2} (\boldsymbol{z} - \boldsymbol{\theta})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{z} - \boldsymbol{\theta})\Big). \tag{36.0.47}$$

Answer. Prior density $(2\pi\tau^2)^{-k/2} \left|\det \boldsymbol{A}\right|^{-1/2} \exp\Big(-\frac{(\boldsymbol{\beta}-\boldsymbol{\nu})^\top \boldsymbol{A} (\boldsymbol{\beta}-\boldsymbol{\nu})}{2\tau^2}\Big)$; likelihood function $(2\pi\sigma^2)^-$ the posterior density is then proportional to the product of the two:

$$\text{posterior} \propto \exp\Big(-\frac{(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}) + \kappa^2 (\boldsymbol{\beta} - \boldsymbol{\nu})^\top \boldsymbol{A} (\boldsymbol{\beta} - \boldsymbol{\nu})}{2\sigma^2}\Big). \tag{36.0.48}$$

□

Although frequentist and Bayesian approaches lead here to identical formulas, the interpretation is quite different. The BLUE/BLUP looks for best performance in repeated samples if $\mathbf{y}$, while the Bayesian posterior density function is the best update of the prior information about $\boldsymbol{\beta}$ by information coming from this one set of observations.

Here is a textbook example of how Bayesians construct the parameters of their prior distribution:

Problem 401. *As in Problem 274, we will work with the Cobb-Douglas production function, which relates output Q to the inputs of labor L and capital K as follows:*

$$Q_t = \mu K_t^\beta L_t^\gamma \exp(\varepsilon_t). \tag{36.0.49}$$

Setting $\mathsf{y}_t = \log Q_t$, $x_t = \log K_t$, $z_t = \log L_t$, and $\alpha = \log \mu$. one obtains the linear regression

$$\mathsf{y}_t = \alpha + \beta x_t + \gamma z_t + \varepsilon_t \tag{36.0.50}$$

Assume that the prior information about β, γ, and the returns to scale $\beta + \gamma$ is such that

$$\mathrm{E}[\beta] = \mathrm{E}[\gamma] = 0.5 \quad \mathrm{E}[\beta + \gamma] = 1.0 \tag{36.0.51}$$

$$\Pr[0.9 < \beta + \gamma < 1.1] = 0.9 \tag{36.0.52}$$

$$\Pr[0.2 < \beta < 0.8] = \Pr[0.2 < \gamma < 0.8] = 0.9 \tag{36.0.53}$$

About α assume that the prior information is such that

$$\mathrm{E}[\alpha] = 5.0, \quad \Pr[-10 < \alpha < 20] = 0.9 \tag{36.0.54}$$

and that our prior knowledge about α is not affected by (is independent of) our prior knowledge concerning β and γ. Assume that σ^2 is known and that it has the value $\sigma^2 = 0.09$. Furthermore, assume that our prior views about α, β, and γ can be adequately represented by a normal distribution. Compute from this the prior distribution of the vector $\boldsymbol{\beta} = \begin{bmatrix} \alpha & \beta & \gamma \end{bmatrix}^\top$.

Answer. This is [**JHG+88**, p. 288–290]. □

Here is my personal opinion what to think of this. I always get uneasy when I see graphs like [**JHG+88**, Figure 7.2 on p. 283]. The prior information was specified on pp. 277/8: the marginal propensity to consume is with high probability between 0.75 and 0.95, and there is a 50-50 chance that it lies above or below 0.85. The least squares estimate of the MPC is 0.9, with a reasonable confidence interval. There is no multicollinearity involved, since there is only one explanatory variable. I see no reason whatsoever to take issue with the least squares regression result, it matches my prior information perfectly. However the textbook tells me that as a Bayesian I have to modify what the data tell me and take the MPC to be 0.88. This is only because of the assumption that the prior information is normal.

I think the Bayesian procedure is inappropriate here because the situation is so simple. Bayesian procedures have the advantage that they are coherent, and therefore can serve as a guide in complex estimation situations, when the researcher is tempted

to employ ad-hoc procedures which easily become incoherent. The advantage of a Bayesian procedure is therefore that it prevents the researcher from stepping on his own toes too blatantly. In the present textbook situation, this advantage does not hold. On the contrary, the only situation where the researcher may be tempted to do something which he does not quite understand is in the above eliciation of prior information. It often happens that prior information gained in this way is self-contradictory, and the researcher is probably not aware what his naive assumptions about the variances of three linear combinations of two parameters imply for the correlation between them!

I can think of two justifications of Bayesian approaches. In certain situations the data are very insensitive, without this being a priori apparent. Widely different estimates give an almost as good fit to the data as the best one. In this case the researcher's prior information may make a big difference and it should be elicited.

Another justification of the Bayesian approach is the following: In many real-life situations the data manipulation and estimation which is called for is so complex that the researcher no longer knows what he is doing. In such a situation, a Bayesian procedure can serve as a guideline. The prior density may not be right, but at least everything is coherent.

CHAPTER 37

OLS With Random Constraint

A Bayesian considers the posterior density the full representation of the information provided by sample and prior information. Frequentists have discoveered that one can interpret the parameters of this density as estimators of the key unknown parameters, and that these estimators have good sampling properties. Therefore they have tried to re-derive the Bayesian formulas from frequentist principles.

If $\boldsymbol{\beta}$ satisfies the constraint $\boldsymbol{R\beta} = \boldsymbol{u}$ only approximately or with uncertainty, it has therefore become customary to specify

$$\boldsymbol{R\beta} = \boldsymbol{u} + \boldsymbol{\eta}, \quad \boldsymbol{\eta} \sim (\boldsymbol{o}, \tau^2\boldsymbol{\Phi}), \quad \boldsymbol{\eta} \text{ and } \boldsymbol{\varepsilon} \text{ uncorrelated.} \tag{37.0.55}$$

Here it is assumed $\tau^2 > 0$ and Φ positive definite.

Both interpretations are possible here: either $\boldsymbol{u}$ is a constant, which means necessarily that $\boldsymbol{\beta}$ is random, or $\boldsymbol{\beta}$ is as usual a constant and $\boldsymbol{u}$ is random, coming from whoever happened to do the research (this is why it is called "mixed estimation").

It is the correct procedure in this situation to do GLS on the model

$$\begin{bmatrix} \mathbf{y} \\ \boldsymbol{u} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{R} \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \boldsymbol{\varepsilon} \\ -\boldsymbol{\eta} \end{bmatrix} \text{ with } \begin{bmatrix} \boldsymbol{\varepsilon} \\ -\boldsymbol{\eta} \end{bmatrix} \sim \left(\begin{bmatrix} \boldsymbol{o} \\ \boldsymbol{o} \end{bmatrix}, \sigma^2 \begin{bmatrix} \boldsymbol{I} & \boldsymbol{O} \\ \boldsymbol{O} & \frac{1}{\kappa^2}\boldsymbol{I} \end{bmatrix} \right). \tag{37.0.56}$$

Therefore

$$\hat{\hat{\boldsymbol{\beta}}} = (\boldsymbol{X}^\top\boldsymbol{X} + \kappa^2\boldsymbol{R}^\top\boldsymbol{R})^{-1}(\boldsymbol{X}^\top\mathbf{y} + \kappa^2\boldsymbol{R}^\top\boldsymbol{u}). \tag{37.0.57}$$

where $\kappa^2 = \sigma^2/\tau^2$.

This $\hat{\hat{\boldsymbol{\beta}}}$ is the BLUE if in repeated samples $\boldsymbol{\beta}$ and $\boldsymbol{u}$ are drawn from such distributions that $\boldsymbol{R\beta} - \boldsymbol{u}$ has mean $\boldsymbol{o}$ and variance $\tau^2\boldsymbol{I}$, but $\mathcal{E}[\boldsymbol{\beta}]$ can be anything. If one considers both $\boldsymbol{\beta}$ and $\boldsymbol{u}$ fixed, then $\hat{\hat{\boldsymbol{\beta}}}$ is a biased estimator whose properties depend on how close the true value of $\boldsymbol{R\beta}$ is to $\boldsymbol{u}$.

Under the assumption of constant $\boldsymbol{\beta}$ and $\boldsymbol{u}$, the $\mathcal{MSE}$ matrix of $\hat{\hat{\boldsymbol{\beta}}}$ is smaller than that of the OLS $\hat{\boldsymbol{\beta}}$ if and only if the true parameter values $\boldsymbol{\beta}$, $\boldsymbol{u}$, and σ^2 satisfy

$$(\boldsymbol{R\beta} - \boldsymbol{u})^\top \left(\frac{2}{\kappa^2}\boldsymbol{I} + \boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top \right)^{-1} (\boldsymbol{R\beta} - \boldsymbol{u}) \leq \sigma^2. \tag{37.0.58}$$

This condition is a simple extension of (29.6.6).

An estimator of the form $\hat{\hat{\boldsymbol{\beta}}} = (\boldsymbol{X}^\top\boldsymbol{X} + \kappa^2\boldsymbol{I})^{-1}\boldsymbol{X}^\top\mathbf{y}$, where κ^2 is a constant, is called "ordinary ridge regression." Ridge regression can be considered the imposition of a random constraint, even though it does not hold—again in an effort to trade bias for variance. This is similar to the imposition of a constraint which does not hold. An explantation of the term "ridge" given by [**VU81**, p. 170] is that the ridge solutions are near a ridge in the likelihood surface (at a point where the ridge is close to the origin). This ridge is drawn in [**VU81**, Figures 1.4a and 1.4b].

Problem 402. *Derive from* (37.0.58) *the well-known formula that the $\mathcal{MSE}$ of ordinary ridge regression is smaller than that of the OLS estimator if and only if the true parameter vector satisfies*

$$\boldsymbol{\beta}^\top \left(\frac{2}{\kappa^2}\boldsymbol{I} + (\boldsymbol{X}^\top\boldsymbol{X})^{-1} \right)^{-1} \boldsymbol{\beta} \leq \sigma^2. \tag{37.0.59}$$

Answer. In (37.0.58) set $\boldsymbol{u} = \boldsymbol{o}$ and $\boldsymbol{R} = \boldsymbol{I}$. □

Whatever the true values of $\boldsymbol{\beta}$ and σ^2, there is always a $\kappa^2 > 0$ for which (37.0.59) or (37.0.58) holds. The corresponding statement for the trace of the $\mathcal{MSE}$-matrix has been one of the main justifications for ridge regression in [**HK70b**] and [**HK70a**], and much of the literature about ridge regression has been inspired by the hope that one can estimate κ^2 in such a way that the MSE is better everywhere. This is indeed done by the Stein-rule.

Ridge regression is reputed to be a good estimator when there is multicollinearity.

Problem 403. *(Not eligible for in-class exams) Assume* $\mathrm{E}[\mathrm{y}] = \mu$, $\mathrm{var}(\mathrm{y}) = \sigma^2$, *and you make n independent observations y_i. Then the best linear unbiased estimator of μ on the basis of these observations is the sample mean $\bar{\mathrm{y}}$. For which range of values of α is* $\mathrm{MSE}[\alpha\bar{\mathrm{y}}; \mu] < \mathrm{MSE}[\bar{\mathrm{y}}; \mu]$*? Unfortunately, this value depends on μ and can therefore not be used to improve the estimate.*

Answer.

$$\mathrm{MSE}[\alpha\bar{\mathrm{y}}; \mu] = \mathrm{E}\big[(\alpha\bar{\mathrm{y}} - \mu)^2\big] = \mathrm{E}\big[(\alpha\bar{\mathrm{y}} - \alpha\mu + \alpha\mu - \mu)^2\big] < \mathrm{MSE}[\bar{\mathrm{y}}; \mu] = \mathrm{var}[\bar{\mathrm{y}}] \tag{37.0.60}$$

$$\alpha^2\sigma^2/n + (1-\alpha)^2\mu^2 < \sigma^2/n \tag{37.0.61}$$

Now simplify it:

$$(1-\alpha)^2\mu^2 < (1-\alpha^2)\sigma^2/n = (1-\alpha)(1+\alpha)\sigma^2/n \tag{37.0.62}$$

This cannot be true for $\alpha \geq 1$, because for $\alpha = 1$ one has equality, and for $\alpha > 1$, the righthand side is negative. Therefore we are allowed to assume $\alpha < 1$, and can divide by $1 - \alpha$ without disturbing the inequality:

$$(1-\alpha)\mu^2 < (1+\alpha)\sigma^2/n \tag{37.0.63}$$

$$\mu^2 - \sigma^2/n < \alpha(\mu^2 + \sigma^2/n) \tag{37.0.64}$$

The answer is therefore

$$\frac{n\mu^2 - \sigma^2}{n\mu^2 + \sigma^2} < \alpha < 1. \tag{37.0.65}$$

□

PROBLEM 404. *(Not eligible for in-class exams) Assume* $\mathbf{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ *with* $\boldsymbol{\varepsilon} \sim (\boldsymbol{o}, \sigma^2\boldsymbol{I})$. *If prior knowledge is available that* $\boldsymbol{P\beta}$ *lies in an ellipsoid centered around* $\boldsymbol{p}$, *i.e.,* $(\boldsymbol{P\beta} - \boldsymbol{p})^\top \boldsymbol{\Phi}^{-1}(\boldsymbol{P\beta} - \boldsymbol{p}) \leq h$ *for some known positive definite symmetric matrix* $\boldsymbol{\Phi}$ *and scalar* h, *then one might argue that the* SSE *should be mimimized only for those* $\boldsymbol{\beta}$ *inside this ellipsoid. Show that this inequality constrained mimimization gives the same formula as OLS with a random constraint of the form* $\kappa^2(\boldsymbol{R\beta} - \boldsymbol{u}) \sim (\boldsymbol{o}, \sigma^2\boldsymbol{I})$ *(where* $\boldsymbol{R}$ *and* $\boldsymbol{u}$ *are appropriately chosen constants, while* κ^2 *depends on* $\mathbf{y}$. *You don't have to compute the precise values, simply indicate how* $\boldsymbol{R}$, $\boldsymbol{u}$, *and* κ^2 *should be determined.)*

ANSWER. Decompose $\boldsymbol{\Phi}^{-1} = \boldsymbol{C}^\top\boldsymbol{C}$ where $\boldsymbol{C}$ is square, and define $\boldsymbol{R} = \boldsymbol{CP}$ and $\boldsymbol{u} = \boldsymbol{Cp}$. The mixed estimator $\boldsymbol{\beta} = \boldsymbol{\beta}^*$ minimizes

$$(\mathbf{y} - \boldsymbol{X\beta})^\top(\mathbf{y} - \boldsymbol{X\beta}) + \kappa^4(\boldsymbol{R\beta} - \boldsymbol{u})^\top(\boldsymbol{R\beta} - \boldsymbol{u}) \tag{37.0.66}$$

$$= (\mathbf{y} - \boldsymbol{X\beta})^\top(\mathbf{y} - \boldsymbol{X\beta}) + \kappa^4(\boldsymbol{P\beta} - \boldsymbol{p})^\top\boldsymbol{\Phi}^{-1}(\boldsymbol{P\beta} - \boldsymbol{p}) \tag{37.0.67}$$

Choose κ^2 such that $\boldsymbol{\beta}^* = (\boldsymbol{X}^\top\boldsymbol{X} + \kappa^4\boldsymbol{P}^\top\boldsymbol{\Phi}^{-1}\boldsymbol{P})^{-1}(\boldsymbol{X}^\top\mathbf{y} + \kappa^4\boldsymbol{P}^\top\boldsymbol{\Phi}^{-1}\boldsymbol{p})$ satisfies the inequality constraint with equality, i.e., $(\boldsymbol{P\beta}^* - \boldsymbol{p})^\top\boldsymbol{\Phi}^{-1}(\boldsymbol{P\beta}^* - \boldsymbol{p}) = h$. □

ANSWER. Now take any $\boldsymbol{\beta}$ that satisfies $(\boldsymbol{P\beta} - \boldsymbol{p})^\top\boldsymbol{\Phi}^{-1}(\boldsymbol{P\beta} - \boldsymbol{p}) \leq h$. Then

$$(\mathbf{y} - \boldsymbol{X\beta}^*)^\top(\mathbf{y} - \boldsymbol{X\beta}^*) = (\mathbf{y} - \boldsymbol{X\beta}^*)^\top(\mathbf{y} - \boldsymbol{X\beta}^*) + \kappa^4(\boldsymbol{P\beta}^* - \boldsymbol{p})^\top\boldsymbol{\Phi}^{-1}(\boldsymbol{P\beta}^* - \boldsymbol{p}) - \kappa^4 h \tag{37.0.68}$$

(because $\boldsymbol{\beta}^*$ satisfies the inequality constraint with equality)

$$\leq (\mathbf{y} - \boldsymbol{X\beta})^\top(\mathbf{y} - \boldsymbol{X\beta}) + \kappa^4(\boldsymbol{P\beta} - \boldsymbol{p})^\top\boldsymbol{\Phi}^{-1}(\boldsymbol{P\beta} - \boldsymbol{p}) - \kappa^4 h \tag{37.0.69}$$

(because $\boldsymbol{\beta}^*$ minimizes (37.0.67))

$$\leq (\mathbf{y} - \boldsymbol{X\beta})^\top(\mathbf{y} - \boldsymbol{X\beta}) \tag{37.0.70}$$

(because $\boldsymbol{\beta}$ satisfies the inequality constraint). Therefore $\boldsymbol{\beta} = \boldsymbol{\beta}^*$ minimizes the inequality constrained problem. □

CHAPTER 38

Stein Rule Estimators

PROBLEM 405. *We will work with the regression model* $\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ *with* $\boldsymbol{\varepsilon} \sim \mathsf{N}(\boldsymbol{o}, \sigma^2\boldsymbol{I})$, *which in addition is "orthonormal," i.e., the* $\boldsymbol{X}$*-matrix satisfies* $\boldsymbol{X}^\top\boldsymbol{X} = \boldsymbol{I}$.

• a. *0 points Write down the simple formula for the OLS estimator* $\hat{\boldsymbol{\beta}}$ *in this model. Can you think of situations in which such an "orthonormal" model is appropriate?*

ANSWER. $\hat{\boldsymbol{\beta}} = \boldsymbol{X}^\top\mathbf{y}$. Sclove [**Scl68**] gives as examples: if one regresses on orthonormal polynomials, or on principal components. I guess also if one simply needs the means of a random vector. It seems the important fact here is that one can order the regressors; if this is the case then one can always make the Gram-Schmidt orthonormalization, which has the advantage that the jth orthonormalized regressor is a linear combination of the first j ordered regressors. □

• b. *0 points Assume one has Bayesian prior knowledge that* $\boldsymbol{\beta} \sim \mathsf{N}(\boldsymbol{o}, \tau^2\boldsymbol{I})$, *and* $\boldsymbol{\beta}$ *independent of* $\boldsymbol{\varepsilon}$. *In the general case, if prior information is* $\boldsymbol{\beta} \sim \mathsf{N}(\boldsymbol{\nu}, \tau^2\boldsymbol{A}^{-1})$, *the Bayesian posterior mean is* $\hat{\boldsymbol{\beta}}_M = (\boldsymbol{X}^\top\boldsymbol{X} + \kappa^2\boldsymbol{A})^{-1}(\boldsymbol{X}^\top\mathbf{y} + \kappa^2\boldsymbol{A}\boldsymbol{\nu})$ *where* $\kappa^2 = \sigma^2/\tau^2$. *Show that in the present case* $\hat{\boldsymbol{\beta}}_M$ *is proportional to the OLS estimate* $\hat{\boldsymbol{\beta}}$ *with proportionality factor* $(1 - \frac{\sigma^2}{\tau^2+\sigma^2})$, *i.e.,*

$$\hat{\boldsymbol{\beta}}_M = \hat{\boldsymbol{\beta}}(1 - \frac{\sigma^2}{\tau^2 + \sigma^2}). \tag{38.0.71}$$

ANSWER. The formula given is (36.0.36), and in the present case, $\boldsymbol{A}^{-1} = \boldsymbol{I}$. One can also view it as a regression with a random constraint $\boldsymbol{R}\boldsymbol{\beta} \sim (\boldsymbol{o}, \tau^2\boldsymbol{I})$ where $\boldsymbol{R} = \boldsymbol{I}$, which is mathematically the same as considering the know mean vector, i.e., the null vector, as additional observations. In either case one gets

(38.0.72)

$$\hat{\boldsymbol{\beta}}_M = (\boldsymbol{X}^\top\boldsymbol{X} + \kappa^2\boldsymbol{A})^{-1}\boldsymbol{X}^\top\mathbf{y} = (\boldsymbol{X}^\top\boldsymbol{X} + \kappa^2\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{X}^\top\mathbf{y} = (\boldsymbol{I} + \frac{\sigma^2}{\tau^2}\boldsymbol{I})^{-1}\boldsymbol{X}^\top\mathbf{y} = \hat{\boldsymbol{\beta}}(1 - \frac{\sigma^2}{\tau^2 + \sigma^2}),$$

i.e., it shrinks the OLS $\hat{\boldsymbol{\beta}} = \boldsymbol{X}^\top\mathbf{y}$. □

• c. *0 points Formula* (38.0.71) *can only be used for estimation if the ratio* $\sigma^2/(\tau^2 + \sigma^2)$ *is* known. *This is usually not the case, but it is possible to* estimate *both* σ^2 *and* $\tau^2 + \sigma^2$ *from the data. The use of such estimates instead the actual*

values of σ^2 *and* τ^2 *in the Bayesian formulas is sometimes called "empirical Bayes." Show that* $\mathrm{E}[\hat{\boldsymbol{\beta}}^\top\hat{\boldsymbol{\beta}}] = k(\tau^2 + \sigma^2)$, *and that* $\mathrm{E}[\mathbf{y}^\top\mathbf{y} - \hat{\boldsymbol{\beta}}^\top\hat{\boldsymbol{\beta}}] = (n-k)\sigma^2$, *where* n *is the number of observations and* k *is the number of regressors.*

ANSWER. Since $\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \sim \mathsf{N}(\boldsymbol{o}, \sigma^2\boldsymbol{X}\boldsymbol{X}^\top + \tau^2\boldsymbol{I})$, it follows $\hat{\boldsymbol{\beta}} = \boldsymbol{X}^\top\mathbf{y} \sim \mathsf{N}(\boldsymbol{o}, (\sigma^2 + \tau^2)\boldsymbol{I})$ (where we now have a k-dimensional identity matrix), therefore $\mathrm{E}[\hat{\boldsymbol{\beta}}^\top\hat{\boldsymbol{\beta}}] = k(\sigma^2 + \tau^2)$. Furthermore, since $\boldsymbol{M}\mathbf{y} = \boldsymbol{M}\boldsymbol{\varepsilon}$ regardless of whether $\boldsymbol{\beta}$ is random or not, σ^2 can be estimated in the usual manner from the SSE: $(n-k)\sigma^2 = \mathrm{E}[\hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}}] = \mathrm{E}[\hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}}] = \mathrm{E}[\mathbf{y}^\top\boldsymbol{M}\mathbf{y}] = \mathrm{E}[\mathbf{y}^\top\mathbf{y} - \hat{\boldsymbol{\beta}}^\top\hat{\boldsymbol{\beta}}]$ because $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{X}\boldsymbol{X}^\top$. □

• d. *0 points If one plugs the unbiased estimates of* σ^2 *and* $\tau^2 + \sigma^2$ *from part (c) into* (38.0.71), *one obtains a version of the so-called "James and Stein" estimator*

$$\hat{\boldsymbol{\beta}}_{JS} = \hat{\boldsymbol{\beta}}(1 - c\frac{\mathbf{y}^\top\mathbf{y} - \hat{\boldsymbol{\beta}}^\top\hat{\boldsymbol{\beta}}}{\hat{\boldsymbol{\beta}}^\top\hat{\boldsymbol{\beta}}}). \tag{38.0.73}$$

What is the value of the constant c *if one follows the above instructions? (This estimator has become famous because for* $k \geq 3$ *and* c *any number between 0 and* $2(n-k)/(n-k+2)$ *the estimator* (38.0.73) *has a uniformly lower* MSE *than the OLS* $\hat{\boldsymbol{\beta}}$, *where the* MSE *is measured as the trace of the* $\mathcal{MSE}$*-matrix.)*

ANSWER. $c = \frac{k}{n-k}$. I would need a proof that this is in the bounds. □

• e. *0 points The existence of the James and Stein estimator proves that the OLS estimator is "inadmissible." What does this mean? Can you explain why the OLS estimator turns out to be deficient exactly where it ostensibly tries to be strong? What are the practical implications of this?*

The properties of this estimator were first discussed in James and Stein [**JS61**], extending the work of Stein [**Ste56**].

Stein himself did not introduce the estimator as an "empirical Bayes" estimator, and it is not certain that this is indeed the right way to look at it. Especially this approach does not explain why the OLS cannot be uniformly improved upon if $k \leq 2$. But it is a possible and interesting way to look at it. If one pretends one has prior information, but does not really have it but "steals" it from the data, this "fraud" can still be successful.

Another interpretation is that these estimators are shrunk versions of unbiased estimators, and unbiased estimators always get better if one shrinks them a little. The only problem is that one cannot shrink them too much, and in the case of the normal distribution, the amount by which one has to shrink them depends on the unknown parameters. If one estimates the shrinkage factor, one usually does not know if the noise introduced by this estimated factor is greater or smaller than the savings. But in the case of the Stein rule, the noise is smaller than the savings.

PROBLEM 406. *0 points Return to the "orthonormal" model* $\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ *with* $\boldsymbol{\varepsilon} \sim \mathsf{N}(\boldsymbol{o}, \sigma^2\boldsymbol{I})$ *and* $\boldsymbol{X}^\top\boldsymbol{X} = \boldsymbol{I}$. *With the usual assumption of nonrandom* $\boldsymbol{\beta}$ *(and no prior information about* $\boldsymbol{\beta}$*), show that the* F*-statistic for the hypothesis* $\boldsymbol{\beta} = \boldsymbol{o}$ *is* $F = \frac{\hat{\boldsymbol{\beta}}^\top\hat{\boldsymbol{\beta}}/k}{(\mathbf{y}^\top\mathbf{y} - \hat{\boldsymbol{\beta}}^\top\hat{\boldsymbol{\beta}})/(n-k)}$.

ANSWER. $\mathsf{SSE}_r = \mathbf{y}^\top\mathbf{y}$, $\mathsf{SSE}_u = \mathbf{y}^\top\mathbf{y} - \hat{\boldsymbol{\beta}}^\top\hat{\boldsymbol{\beta}}$ as shown above, number of constraints is k. Use equation ... for the test statistic. □

• a. *0 points Now look at the following "pre-test estimator": Your estimate of* $\boldsymbol{\beta}$ *is the null vector* $\boldsymbol{o}$ *if the value of the* F*-statistic for the test* $\boldsymbol{\beta} = \boldsymbol{o}$ *is equal to or smaller than 1, and your estimate of* $\boldsymbol{\beta}$ *is the OLS estimate* $\hat{\boldsymbol{\beta}}$ *if the test statistic has a value bigger than 1. Mathematically, this estimator can be written in the form*

$$\hat{\boldsymbol{\beta}}_{PT} = I(F)\hat{\boldsymbol{\beta}}, \tag{38.0.74}$$

where F *is the* F *statistic derived in part (1) of this question, and* $I(F)$ *is the "indicator function" for* $F > 1$*, i.e.,* $I(F) = 0$ *if* $F \leq 1$ *and* $I(F) = 1$ *if* $F > 1$*. Now modify this pre-test estimator by using the following function* $I(F)$ *instead:* $I(F) = 0$ *if* $F \leq 1$ *and* $I(F) = 1 - 1/F$ *if* $F > 1$*. This is no longer an indicator function, but can be considered a continuous approximation to one. Since the discontinuity is removed, one can expect that it has, under certain circumstances, better properties than the indicator function itself. Write down the formula for this modified pre-test estimator. How does it differ from the Stein rule estimator* (38.0.73) *(with the value for c coming from the empirical Bayes approach)? Which estimator would you expect to be better, and why?*

ANSWER. This modified pre-test estimator has the form

$$\hat{\boldsymbol{\beta}}_{JS+} = \begin{cases} \boldsymbol{o} & \text{if } 1 - c\frac{\mathbf{y}^\top\mathbf{y} - \hat{\boldsymbol{\beta}}^\top\hat{\boldsymbol{\beta}}}{\hat{\boldsymbol{\beta}}^\top\hat{\boldsymbol{\beta}}} < 0 \\ \hat{\boldsymbol{\beta}}(1 - c\frac{\mathbf{y}^\top\mathbf{y} - \hat{\boldsymbol{\beta}}^\top\hat{\boldsymbol{\beta}}}{\hat{\boldsymbol{\beta}}^\top\hat{\boldsymbol{\beta}}}) & \text{otherwise} \end{cases} \tag{38.0.75}$$

It is equal to the Stein-rule estimator (38.0.73) when the estimated shrinkage factor $(1 - c\frac{\mathbf{y}^\top\mathbf{y} - \hat{\boldsymbol{\beta}}^\top\hat{\boldsymbol{\beta}}}{\hat{\boldsymbol{\beta}}^\top\hat{\boldsymbol{\beta}}})$ is positive, but the shrinkage factor is set 0 instead of turning negative. This is why it is commonly called the "positive part" Stein-rule estimator. Stein conjectured early on, and Baranchik [**Bar64**] showed that it is uniformly better than the Stein rule estimator: □

• b. *0 points Which lessons can one draw about pre-test estimators in general from this exercise?*

Stein rule estimators have not been used very much, they are not equivariant and the shrinkage seems arbitrary. Discussing them here brings out two things: the formulas for random constraints etc. are a pattern according to which one can build good operational estimators. And some widely used but seemingly ad-hoc procedures like pre-testing may have deeper foundations and better properties than the halfways sophisticated researcher may think.

PROBLEM 407. *6 points Why was it somewhat a sensation when Charles Stein came up with an estimator which is uniformly better than the OLS? Discuss the Stein rule estimator as empirical Bayes, shrinkage estimator, and discuss the "positive part" Stein rule estimator as a modified pretest estimator.*

CHAPTER 39

Random Regressors

Until now we always assumed that $\boldsymbol{X}$ was nonrandom, i.e., the hypothetical repetitions of the experiment used the same $\boldsymbol{X}$ matrix. In the nonexperimental sciences, such as economics, this assumption is clearly inappropriate. It is only justified because most results valid for nonrandom regressors can be generalized to the case of random regressors. To indicate that the regressors are random, we will write them as $\mathbf{X}$.

39.1. Strongest Assumption: Error Term Well Behaved Conditionally on Explanatory Variables

The assumption which we will discuss first is that $\mathbf{X}$ is random, but the classical assumptions hold *conditionally on* $\mathbf{X}$, i.e., the conditional expectation $\mathcal{E}[\boldsymbol{\varepsilon}|\mathbf{X}] = \boldsymbol{o}$, and the conditional variance-covariance matrix $\mathcal{V}[\boldsymbol{\varepsilon}|\mathbf{X}] = \sigma^2\boldsymbol{I}$. In this situation, the least squares estimator has all the classical properties *conditionally on* $\mathbf{X}$, for instance $\mathcal{E}[\hat{\boldsymbol{\beta}}|\mathbf{X}] = \beta$, $\mathcal{V}[\hat{\boldsymbol{\beta}}|\mathbf{X}] = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$, $\mathrm{E}[\mathrm{s}^2|\mathbf{X}] = \sigma^2$, etc.

Moreover, certain properties of the Least Squares estimator remain valid *unconditionally*. An application of the law of iterated expectations shows that the least squares estimator $\hat{\boldsymbol{\beta}}$ is still unbiased. Start with (24.0.7):

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\varepsilon} \tag{39.1.1}$$

$$\mathcal{E}[\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|\mathbf{X}] = \mathcal{E}[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\varepsilon}|\mathbf{X}] = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\,\mathcal{E}[\boldsymbol{\varepsilon}|\mathbf{X}] = \boldsymbol{o}. \tag{39.1.2}$$

$$\mathcal{E}[\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}] = \mathcal{E}\big[\mathcal{E}[\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|\mathbf{X}]\big] = \boldsymbol{o}. \tag{39.1.3}$$

Problem 408. *1 point In the model with random explanatory variables* $\mathbf{X}$ *you are considering an estimator* $\tilde{\boldsymbol{\beta}}$ *of* $\boldsymbol{\beta}$. *Which statement is stronger:* $\mathcal{E}[\tilde{\boldsymbol{\beta}}] = \boldsymbol{\beta}$, *or* $\mathcal{E}[\tilde{\boldsymbol{\beta}}|\mathbf{X}] = \boldsymbol{\beta}$. *Justify your answer.*

Answer. The second statement is stronger. The first statement follows from the second by the law of iterated expectations. □

Problem 409. *2 points Assume the regressors* $\mathbf{X}$ *are random, and the classical assumptions hold conditionally on* $\mathbf{X}$, *i.e.,* $\mathcal{E}[\boldsymbol{\varepsilon}|\mathbf{X}] = \boldsymbol{o}$ *and* $\mathcal{V}[\boldsymbol{\varepsilon}|\mathbf{X}] = \sigma^2\boldsymbol{I}$. *Show that* s^2 *is an unbiased estimate of* σ^2.

Answer. From the theory with nonrandom explanatory variables follows $\mathrm{E}[\mathrm{s}^2|\mathbf{X}] = \sigma^2$. Therefore $\mathrm{E}[\mathrm{s}^2] = \mathrm{E}\big[\mathrm{E}[\mathrm{s}^2|\mathbf{X}]\big] = \mathrm{E}[\sigma^2] = \sigma^2$. In words: if the expectation conditional on $\mathbf{X}$ does not depend on $\mathbf{X}$, then it is also the unconditional expectation. □

The law of iterated expectations can also be used to compute the unconditional $\mathcal{MSE}$ matrix of $\hat{\boldsymbol{\beta}}$:

$$\mathcal{MSE}[\hat{\boldsymbol{\beta}};\boldsymbol{\beta}] = \mathcal{E}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top] \tag{39.1.4}$$

$$= \mathcal{E}\big[\mathcal{E}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top|\mathbf{X}]\big] \tag{39.1.5}$$

$$= \mathcal{E}[\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}] \tag{39.1.6}$$

$$= \sigma^2\,\mathcal{E}[(\mathbf{X}^\top\mathbf{X})^{-1}]. \tag{39.1.7}$$

Problem 410. *2 points Show that* $\mathrm{s}^2(\mathbf{X}^\top\mathbf{X})^{-1}$ *is unbiased estimator of* $\mathcal{MSE}[\hat{\boldsymbol{\beta}};\boldsymbol{\beta}]$.

Answer.

$$\mathcal{E}[\mathrm{s}^2(\mathbf{X}^\top\mathbf{X})^{-1}] = \mathcal{E}\big[\mathcal{E}[\mathrm{s}^2(\mathbf{X}^\top\mathbf{X})^{-1}|\mathbf{X}]\big] \tag{39.1.8}$$

$$= \mathcal{E}[\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}] \tag{39.1.9}$$

$$= \sigma^2\,\mathcal{E}[(\mathbf{X}^\top\mathbf{X})^{-1}] \tag{39.1.10}$$

$$= \mathcal{MSE}[\hat{\boldsymbol{\beta}};\boldsymbol{\beta}] \quad \text{by (39.1.7).} \tag{39.1.11}$$

□

The Gauss-Markov theorem generalizes in the following way: Say $\tilde{\boldsymbol{\beta}}$ is an estimator, linear in $\mathbf{y}$, but not necessarily in $\mathbf{X}$, satisfying $\mathcal{E}[\tilde{\boldsymbol{\beta}}|\mathbf{X}] = \boldsymbol{\beta}$ (which is stronger than unbiasedness); then $\mathcal{MSE}[\tilde{\boldsymbol{\beta}};\boldsymbol{\beta}] \geq \mathcal{MSE}[\hat{\boldsymbol{\beta}};\boldsymbol{\beta}]$. Proof is immediate: we know by the usual Gauss-Markov theorem that $\mathcal{MSE}[\tilde{\boldsymbol{\beta}};\boldsymbol{\beta}|\mathbf{X}] \geq \mathcal{MSE}[\hat{\boldsymbol{\beta}};\boldsymbol{\beta}|\mathbf{X}]$, and taking expected values will preserve this inequality: $\mathcal{E}\big[\mathcal{MSE}[\tilde{\boldsymbol{\beta}};\boldsymbol{\beta}|\mathbf{X}]\big] \geq \mathcal{E}\big[\mathcal{MSE}[\hat{\boldsymbol{\beta}};\boldsymbol{\beta}|\mathbf{X}]\big]$, but this expected value is exactly the unconditional $\mathcal{MSE}$.

The assumption $\mathcal{E}[\boldsymbol{\varepsilon}|\mathbf{X}] = \boldsymbol{o}$ can also be written $\mathcal{E}[\mathbf{y}|\mathbf{X}] = \mathbf{X}\boldsymbol{\beta}$, and $\mathcal{V}[\boldsymbol{\varepsilon}|\mathbf{X}] = \sigma^2\boldsymbol{I}$ can also be written as $\mathcal{V}[\mathbf{y}|\mathbf{X}] = \sigma^2\boldsymbol{I}$. Both of these are assumptions about the conditional distribution $\mathbf{y}|\mathbf{X} = \boldsymbol{X}$ for all $\boldsymbol{X}$. This suggests the following broadening of the regression paradigm: $\mathbf{y}$ and $\mathbf{X}$ are jointly distributed random variables, and one is interested how $\mathbf{y}|\mathbf{X} = \boldsymbol{X}$ depends on $\boldsymbol{X}$. If the expected value of this distribution depends linearly, and the variance of this distribution is constant, then this is the linear regression model discussed above. But the expected value might also depend on $\mathbf{X}$ in a nonlinear fashion (nonlinear least squares), and the variance may not be constant—in which case the intuition that $\mathbf{y}$ is some function of $\mathbf{X}$ plus some error term may no longer be appropriate; $\mathbf{y}$ may for instance be the outcome of a binary choice, the probability of which depends on $\mathbf{X}$ (see chapter 69.2; the generalized linear model).

39.2. Contemporaneously Uncorrelated Disturbances

In many situations with random regressors, the condition $\mathcal{E}[\boldsymbol{\varepsilon}|\mathbf{X}] = \boldsymbol{o}$ is not satisfied. Instead, the columns of $\mathbf{X}$ are contemporaneously uncorrelated with $\boldsymbol{\varepsilon}$, but they may be correlated with past values of $\boldsymbol{\varepsilon}$. The main example here is regression with a lagged dependent variable. In this case, OLS is no longer unbiased, but asymptotically it still has all the good properties, it is asymptotically normal with the covariance matrix which one would expect. Asymptotically, the computer printout is still valid. This is a very important result, which is often used in econometrics, but most econometrics textbooks do not even start to prove it. There is a proof in [**Kme86**, pp. 749–757], and one in [**Mal80**, pp. 535–539].

Problem 411. *Since least squares with random regressors is appropriate whenever the disturbances are contemporaneously uncorrelated with the explanatory variables, a friend of yours proposes to test for random explanatory variables by checking whether the sample correlation coefficients between the residuals and the explanatory variables is significantly different from zero or not. Is this an appropriate statistic?*

Answer. No. The sample correlation coefficients are always zero! □

39.3. Disturbances Correlated with Regressors in Same Observation

But if $\boldsymbol{\varepsilon}$ is contemporaneously correlated with $\mathbf{X}$, then OLS is inconsistent. This can be the case in some dynamic processes (lagged dependent variable as regressor, and autocorrelated errors, see question 506), when there are, in addition to the relation which one wants to test with the regression, other relations making the righthand side variables dependent on the lefthand side variable, or when the righthand side variables are measured with errors. This is usually the case in economics, and econometrics has developed the technique of simultaneous equations estimation to deal with it.

Problem 412. *3 points What does one have to watch out for if some of the regressors are random?*

CHAPTER 40

The Mahalanobis Distance

Everything in this chapter is unpublished work, presently still in draft form. The aim is to give a motivation for the least squares objective function in terms of an initial measure of precision. The case of prediction is mathematically simpler than that of estimation, therefore this chapter will only discuss prediction. We assume that the joint distribution of $\mathbf{y}$ and $\boldsymbol{z}$ has the form

$$\begin{bmatrix}\mathbf{y}\\ \boldsymbol{z}\end{bmatrix} \sim \begin{bmatrix}\boldsymbol{X}\\ \boldsymbol{W}\end{bmatrix}\boldsymbol{\beta},\ \sigma^2\begin{bmatrix}\boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}} & \boldsymbol{\Omega}_{\mathbf{y}\boldsymbol{z}}\\ \boldsymbol{\Omega}_{\boldsymbol{z}\mathbf{y}} & \boldsymbol{\Omega}_{\boldsymbol{z}\boldsymbol{z}}\end{bmatrix}, \qquad (40.0.1)$$

$\sigma^2 > 0$, otherwise unk $\boldsymbol{\beta}$ unknown as well.

$\mathbf{y}$ is observed but $\boldsymbol{z}$ is not and has to be predicted. But assume we are not interested in the $\mathcal{MSE}$ since we do the experiment only once. We want to predict $\boldsymbol{z}$ in such a way that, whatever the true value of $\boldsymbol{\beta}$, the predicted value $\boldsymbol{z}^*$ "blends in" best with the given data $\boldsymbol{y}$.

There is an important conceptual difference between this criterion and the one based on the $\mathcal{MSE}$. The present criterion cannot be applied until after the data are known, therefore it is called a "final" criterion as opposed to the "initial" criterion of the $\mathcal{MSE}$. See Barnett [**Bar82**, pp. 157–159] for a good discussion of these issues.

How do we measure the degree to which a given data set "blend in," i.e., are not outliers for a given distribution? Hypothesis testing uses this criterion. The most often-used testing principle is: reject the null hypothesis if the observed value of a certain statistic is too much an outlier for the distribution which this statistic would have under the null hypothesis. If the statistic is a scalar, and if under the null hypothesis this statistic has expected value μ and standard deviation σ, then one often uses an estimate of $|x-\mu|/\sigma$, the number of standard deviations the observed value is away from the mean, to measure the "distance" of the observed value x from the distribution (μ, σ^2). The Mahalanobis distance generalizes this concept to the case that the test statistic is a vector random variable.

40.1. Definition of the Mahalanobis Distance

Since it is mathematically more convenient to work with the *squared* distance than with the distance itself, we will make the following thought experiment to

motivate the Mahalanobis distance. How could one generalize the squared scalar distance $(y-\mu)^2/\sigma^2$ for the distance of a vector value $\boldsymbol{y}$ from the distribution of the vector random variable $\mathbf{y} \sim (\boldsymbol{\mu}, \sigma^2\boldsymbol{\Omega})$? If all $\mathbf{y}_i$ have same variance σ^2, i.e., if $\boldsymbol{\Omega} = \boldsymbol{I}$, one might measure the squared distance of $\boldsymbol{y}$ from the distribution $(\boldsymbol{\mu}, \sigma^2\boldsymbol{\Omega})$ by $\frac{1}{\sigma^2}\max_i(y_i-\mu_i)^2$, but since the maximum from two trials is bigger than the value from one trial only, one should divide this perhaps by the expected value of such a maximum. If the variances are different, say σ_i^2, one might want to look a the number of standard deviations which the "worst" component of $\boldsymbol{y}$ is away from what would be its mean if $\boldsymbol{y}$ were an observation of $\mathbf{y}$, i.e., the squared distance of the obsrved vector from the distribution would be $\max_i \frac{(y_i-\mu_i)^2}{\sigma_i^2}$, again normalized by its expected value.

The principle actually used by the Mahalanobis distance goes only a small step further than the examples just cited. It is coordinate-free, i.e., any linear combinations of the elements of $\boldsymbol{y}$ are considered on equal footing with these elements themselves. In other words, it does not distinguish between variates and variables. The distance of a given vector value from a certain multivariate distribution is defined to be the distance of the "worst" *linear combination* of the elements of this vector from the univariate distribution of this linear combination, normalized in such a way that the expected value of this distance is 1.

Definition 40.1.1. Given a random n-vector $\mathbf{y}$ which has expected value and a nonsingular covariance matrix. The squared "Mahalanobis distance" or "statistical distance" of the observed value $\boldsymbol{y}$ from the distribution of $\mathbf{y}$ is defined to be

$$\mathrm{MHD}[\boldsymbol{y};\mathbf{y}] = \frac{1}{n}\max_{\boldsymbol{g}}\frac{\left(\boldsymbol{g}^\top\boldsymbol{y} - \mathrm{E}[\boldsymbol{g}^\top\mathbf{y}]\right)^2}{\mathrm{var}[\boldsymbol{g}^\top\mathbf{y}]}. \qquad (40.1.1)$$

If the denominator $\mathrm{var}[\boldsymbol{g}^\top\mathbf{y}]$ is zero, then $\boldsymbol{g} = \boldsymbol{o}$, therefore the numerator is zero as well. In this case the fraction is defined to be zero.

Theorem 40.1.2. *Let $\mathbf{y}$ be a vector random variable with $\mathcal{E}[\mathbf{y}] = \boldsymbol{\mu}$ and $\mathcal{V}[\mathbf{y}] = \sigma^2\boldsymbol{\Omega}$, $\sigma^2 > 0$ and $\boldsymbol{\Omega}$ positive definite. The squared Mahalanobis distance of the value $\boldsymbol{y}$ from the distribution of $\mathbf{y}$ is equal to*

$$\mathrm{MHD}[\boldsymbol{y};\mathbf{y}] = \frac{1}{n\sigma^2}(\boldsymbol{y}-\boldsymbol{\mu})^\top\boldsymbol{\Omega}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}) \qquad (40.1.2)$$

Proof. (40.1.2) is a simple consequence of (32.4.4). It is also somewhat intuitive since the righthand side of (40.1.2) can be considered a division of the square of $\boldsymbol{y}-\boldsymbol{\mu}$ by the covariance matrix of $\mathbf{y}$. □

The Mahalanobis distance is an asymmetric measure; a large value indicates a bad fit of the hypothetical population to the observation, while a value of, say, 0.1 does not necessarily indicate a better fit than a value of 1.

PROBLEM 413. *Let $\mathfrak{y}$ be a random n-vector with expected value $\boldsymbol{\mu}$ and nonsingular covariance matrix $\sigma^2\boldsymbol{\Omega}$. Show that the expected value of the Mahalobis distance of the observations of $\mathfrak{y}$ from the distribution of $\mathfrak{y}$ is 1, i.e.,*

$$\mathrm{E}\big[\mathrm{MHD}[\mathfrak{y};\mathfrak{y}]\big] = 1 \tag{40.1.3}$$

ANSWER.

(40.1.4)
$$\mathrm{E}[\frac{1}{n\sigma^2}(\mathfrak{y}-\boldsymbol{\mu})^\top\boldsymbol{\Omega}^{-1}(\mathfrak{y}-\boldsymbol{\mu})] = \mathrm{E}[\mathrm{tr}\,\frac{1}{n\sigma^2}\boldsymbol{\Omega}^{-1}(\mathfrak{y}-\boldsymbol{\mu})(\mathfrak{y}-\boldsymbol{\mu})^\top]\,\mathrm{tr}(\frac{1}{n\sigma^2}\boldsymbol{\Omega}^{-1}\sigma^2\boldsymbol{\Omega}) = \frac{1}{n}\,\mathrm{tr}(\boldsymbol{I}) = 1.$$

□

(40.1.2) is, up to a constant factor, the quadratic form in the exponent of the normal density function of $\mathfrak{y}$. For a normally distributed $\mathfrak{y}$, therefore, all observations located on the same density contour have equal distance from the distribution.

The Mahalanobis distance is also defined if the covariance matrix of $\mathfrak{y}$ is singular. In this case, certain nonzero linear combinations of the elements of $\mathfrak{y}$ are known with certainty. Certain vectors can therefore not possibly be realizations of $\mathfrak{y}$, i.e., the set of realizations of $\mathfrak{y}$ does not fill the whole $\mathbb{R}^n$.

PROBLEM 414. *2 points The random vector $\mathfrak{y} = \begin{bmatrix}\mathfrak{y}_1\\ \mathfrak{y}_2\\ \mathfrak{y}_3\end{bmatrix}$ has mean $\begin{bmatrix}1\\2\\-3\end{bmatrix}$ and covariance matrix $\frac{1}{3}\begin{bmatrix}2&-1&-1\\-1&2&-1\\-1&-1&2\end{bmatrix}$. Is this covariance matrix singular? If so, give a linear combination of the elements of $\mathfrak{y}$ which is known with certainty. And give a value which can never be a realization of $\mathfrak{y}$. Prove everything you state.*

ANSWER. Yes, it is singular;

$$\begin{bmatrix}2&-1&-1\\-1&2&-1\\-1&-1&2\end{bmatrix}\begin{bmatrix}1\\1\\1\end{bmatrix} = \begin{bmatrix}0\\0\\0\end{bmatrix} \tag{40.1.5}$$

I.e., $\mathfrak{y}_1+\mathfrak{y}_2+\mathfrak{y}_3 = 0$ because its variance is 0 and its mean is zero as well since $\begin{bmatrix}1&1&1\end{bmatrix}\begin{bmatrix}1\\2\\-3\end{bmatrix} = 0$. □

DEFINITION 40.1.3. Given a vector random variable $\mathfrak{y}$ which has a mean and a covariance matrix. A value $\boldsymbol{y}$ has infinite statistical distance from this random variable, i.e., it cannot possibly be a realization of this random variable, if a vector of coefficients $\boldsymbol{g}$ exists such that $\mathrm{var}[\boldsymbol{g}^\top\mathfrak{y}] = 0$ but $\boldsymbol{g}^\top\boldsymbol{y} \neq \boldsymbol{g}^\top\mathcal{E}[\mathfrak{y}]$. If such a $\boldsymbol{g}$ does not exist, then the squared Mahalanobis distance of $\boldsymbol{y}$ from $\mathfrak{y}$ is defined as in (40.1.1), with n replaced by $\mathrm{rank}[\boldsymbol{\Omega}]$. If the denominator in (40.1.1) is zero, then it no longer necessarily follows that $\boldsymbol{g} = \boldsymbol{o}$ but it nevertheless follows that the numerator is zero, and the fraction should in this case again be considered zero.

If $\boldsymbol{\Omega}$ is singular, then the inverse $\boldsymbol{\Omega}^{-1}$ in formula (40.1.2) must be replaced by a "g-inverse." A g-inverse of a matrix $\boldsymbol{A}$ is any matrix $\boldsymbol{A}^-$ which satisfies $\boldsymbol{A}\boldsymbol{A}^-\boldsymbol{A} = \boldsymbol{A}$. G-inverses always exist, but they are usually not unique.

PROBLEM 415. *a is a scalar. What is its g-inverse a^-?*

THEOREM 40.1.4. *Let $\mathfrak{y}$ be a random variable with $\mathcal{E}[\mathfrak{y}] = \boldsymbol{\mu}$ and $\mathcal{V}[\mathfrak{y}] = \sigma^2\boldsymbol{\Omega}$, $\sigma^2 > 0$. If it is not possible to express the vector $\boldsymbol{y}$ in the form $\boldsymbol{y} = \boldsymbol{\mu}+\boldsymbol{\Omega}\boldsymbol{a}$ for some $\boldsymbol{a}$, then the squared Mahalanobis distance of $\boldsymbol{y}$ from the distribution of $\mathfrak{y}$ is infinite, i.e., $\mathrm{MHD}[\boldsymbol{y};\mathfrak{y}] = \infty$; otherwise*

$$\mathrm{MHD}[\boldsymbol{y};\mathfrak{y}] = \frac{1}{\sigma^2\,\mathrm{rank}[\boldsymbol{\Omega}]}(\boldsymbol{y}-\boldsymbol{\mu})^\top\boldsymbol{\Omega}^-(\boldsymbol{y}-\boldsymbol{\mu}) \tag{40.1.6}$$

Now we will dicuss how a given observation vector can be extended by additional observations in such a way that the Mahalanobis distance of the whole vector from its distribution is minimized.

40.2. The Conditional Mahalanobis Distance

Now let us assume that after the observation of $\mathfrak{y}$ additional observations become available. I.e., the scenario now is

$$\begin{bmatrix}\mathfrak{y}\\ \boldsymbol{z}\end{bmatrix} \sim \begin{bmatrix}\boldsymbol{\mu}\\ \boldsymbol{\nu}\end{bmatrix},\ \sigma^2\begin{bmatrix}\boldsymbol{\Omega}_{\mathfrak{y}\mathfrak{y}} & \boldsymbol{\Omega}_{\mathfrak{y}z}\\ \boldsymbol{\Omega}_{z\mathfrak{y}} & \boldsymbol{\Omega}_{zz}\end{bmatrix}, \qquad \sigma^2 > 0. \tag{40.2.1}$$

Assume and

$$\mathrm{rank}[\boldsymbol{\Omega}_{\mathfrak{y}\mathfrak{y}}] = p \qquad \text{and} \qquad \mathrm{rank}\begin{bmatrix}\boldsymbol{\Omega}_{\mathfrak{y}\mathfrak{y}} & \boldsymbol{\Omega}_{\mathfrak{y}z}\\ \boldsymbol{\Omega}_{z\mathfrak{y}} & \boldsymbol{\Omega}_{zz}\end{bmatrix} = r. \tag{40.2.2}$$

In this case we define the conditional Mahalanobis distance of an observation $\boldsymbol{z}$ given the prior observation $\boldsymbol{y}$ to be

(40.2.3)
$$\mathrm{MHD}[\boldsymbol{z};\mathfrak{y},\boldsymbol{z}|\boldsymbol{y}] = \frac{1}{(r-p)\sigma^2}\Big(\begin{bmatrix}\boldsymbol{y}-\boldsymbol{\mu}\\ \boldsymbol{z}-\boldsymbol{\nu}\end{bmatrix}^\top\begin{bmatrix}\boldsymbol{\Omega}_{\mathfrak{y}\mathfrak{y}} & \boldsymbol{\Omega}_{\mathfrak{y}z}\\ \boldsymbol{\Omega}_{z\mathfrak{y}} & \boldsymbol{\Omega}_{zz}\end{bmatrix}^-\begin{bmatrix}\boldsymbol{y}-\boldsymbol{\mu}\\ \boldsymbol{z}-\boldsymbol{\nu}\end{bmatrix} - (\boldsymbol{y}-\boldsymbol{\mu})^\top\boldsymbol{\Omega}_{\mathfrak{y}\mathfrak{y}}^-(\boldsymbol{y}-\boldsymbol{\mu})\Big)$$

This is again a nonnegative measure whose expected value is 1, and if the underlying distribution is Normal, this is the same as the Mahalanobis distance of $\boldsymbol{z}$ in its distribution conditionally on $\mathfrak{y} = \boldsymbol{y}$.

40.3. First Scenario: Minimizing relative increase in Mahalanobis distance if distribution is known

We start with a situation where the expected values of the random vectors $\mathfrak{y}$ and $\boldsymbol{z}$ are known, and their joint covariance matrix is known up to an unknown scalar factor $\sigma^2 > 0$. We will write this as

$$\begin{bmatrix}\mathfrak{y}\\ \boldsymbol{z}\end{bmatrix} \sim \begin{bmatrix}\boldsymbol{\mu}\\ \boldsymbol{\nu}\end{bmatrix},\ \sigma^2\begin{bmatrix}\boldsymbol{\Omega}_{\mathfrak{y}\mathfrak{y}} & \boldsymbol{\Omega}_{\mathfrak{y}z}\\ \boldsymbol{\Omega}_{z\mathfrak{y}} & \boldsymbol{\Omega}_{zz}\end{bmatrix}, \qquad \sigma^2 > 0. \tag{40.3.1}$$

$\boldsymbol{\Omega}_{\mathsf{y}\mathsf{y}}$ has rank p and $\begin{bmatrix} \boldsymbol{\Omega}_{\mathsf{y}\mathsf{y}} & \boldsymbol{\Omega}_{\mathsf{y}z} \\ \boldsymbol{\Omega}_{z\mathsf{y}} & \boldsymbol{\Omega}_{zz} \end{bmatrix}$ has rank r. Since σ^2 is not known, one cannot compute the Mahalanobis distance of the observed and/or conjectured values $\boldsymbol{y}$ and $\boldsymbol{z}$ from their distribution. But if one works with the relative increase in the Mahalanobis distance if $\boldsymbol{z}$ is added to $\boldsymbol{y}$, then σ^2 cancels out. In order to measure how well the conjectured value $\boldsymbol{z}$ fits together with the observed $\boldsymbol{y}$ we will therefore divide the Mahalanobis distance of the vector composed of $\boldsymbol{y}$ and $\boldsymbol{z}$ from its distribution by the Mahalanobis distance of $\boldsymbol{y}$ alone from *its* distribution:

$$\frac{\frac{1}{r\sigma^2}\begin{bmatrix} \boldsymbol{y}-\boldsymbol{\mu} \\ \boldsymbol{z}-\boldsymbol{\nu} \end{bmatrix}^\top \begin{bmatrix} \boldsymbol{\Omega}_{\mathsf{y}\mathsf{y}} & \boldsymbol{\Omega}_{\mathsf{y}z} \\ \boldsymbol{\Omega}_{z\mathsf{y}} & \boldsymbol{\Omega}_{zz} \end{bmatrix}^- \begin{bmatrix} \boldsymbol{y}-\boldsymbol{\mu} \\ \boldsymbol{z}-\boldsymbol{\nu} \end{bmatrix}}{\frac{1}{p\sigma^2}(\boldsymbol{y}-\boldsymbol{\mu})^\top \boldsymbol{\Omega}_{\mathsf{y}\mathsf{y}}^-(\boldsymbol{y}-\boldsymbol{\mu})}. \tag{40.3.2}$$

This relative measure is independent of σ^2, and if y is observed but $\boldsymbol{z}$ is not, one can predict $\boldsymbol{z}$ by that value $\boldsymbol{z}^*$ which minimizes this relative contribution.

An equivalent criterion which leads to mathematically simpler formulas is to divide the *conditional* Mahalanobis distance of $\boldsymbol{z}$ given $\boldsymbol{y}$ by the Mahalanobis distance of $\boldsymbol{y}$ from y:

$$\frac{\frac{1}{(r-p)\sigma^2}\left(\begin{bmatrix} \boldsymbol{y}-\boldsymbol{\mu} \\ \boldsymbol{z}-\boldsymbol{\nu} \end{bmatrix}^\top \begin{bmatrix} \boldsymbol{\Omega}_{\mathsf{y}\mathsf{y}} & \boldsymbol{\Omega}_{\mathsf{y}z} \\ \boldsymbol{\Omega}_{z\mathsf{y}} & \boldsymbol{\Omega}_{zz} \end{bmatrix}^- \begin{bmatrix} \boldsymbol{y}-\boldsymbol{\mu} \\ \boldsymbol{z}-\boldsymbol{\nu} \end{bmatrix} - (\boldsymbol{y}-\boldsymbol{\mu})^\top \boldsymbol{\Omega}_{\mathsf{y}\mathsf{y}}^-(\boldsymbol{y}-\boldsymbol{\mu})\right)}{\frac{1}{p\sigma^2}(\boldsymbol{y}-\boldsymbol{\mu})^\top \boldsymbol{\Omega}_{\mathsf{y}\mathsf{y}}^-(\boldsymbol{y}-\boldsymbol{\mu})}. \tag{40.3.3}$$

We already solved this minimization problem in chapter **??**. By (**??**), the minimum value of this relative contribution is zero, and the value of $\boldsymbol{z}$ which minimizes this relative contribution is the same as the value of the best linear predictor of $\boldsymbol{z}$, i.e., the value assumed by the linear predictor which minimizes the $\mathcal{MSE}$ among all linear predictors.

40.4. Second Scenario: One Additional IID Observation

In the above situation, we could *minimize* the relative increase in the Mahalanobis distance (instead of selecting its *minimax* value) because all parameters of the underlying distribution were known. The simplest situation in which they are not known, and therefore we must resort to minimizing the relative increase in the Mahalanobis distance for the most unfavorable value of this unknown parameter, is the following: A vector y of n i.i.d. observations is given with unknown mean μ and variance $\sigma^2 > 0$. The squared Mahalanobis distance of these data from their population is $\frac{1}{n\sigma^2}(\mathsf{y}-\boldsymbol{\iota}\mu)^\top(\mathsf{y}-\boldsymbol{\iota}\mu)$. it depends on the unknown μ and σ^2. How can we predict a $n+1$st observation in such a way as to minimize the worst possible *relative increase in this Mahalanobis distance?*

Minimizing the maximum possible *relative* increase in the Mahalanobis distance due to y_{n+1} is the same as minimizing

$$\mathsf{q} = \frac{(\mu-\mathsf{y}_{n+1})^2}{(\boldsymbol{\iota}\mu-\mathsf{y})^\top(\boldsymbol{\iota}\mu-\mathsf{y})/n} \tag{40.4.1}$$

We will show that the prediction $\mathsf{y}_{n+1} = \bar{\mathsf{y}}$ is the solution to this minimax problem, and that the minimax value of q is $\mathsf{q} = 1$.

We will show that (1) for $\mathsf{y}_{n+1} = \bar{\mathsf{y}}$, $\mathsf{q} \leq 1$ for all values of μ, but one can find μ for which q is arbitrarily close to 1, and (2) if $\mathsf{y}_{n+1} \neq \bar{\mathsf{y}}$, then $\mathsf{q} > 1$ for certain values of μ.

In the proof of step (1), the case $\mathsf{y}_1 = \mathsf{y}_2 = \cdots = \mathsf{y}_n$ must be treated separately. If this condition holds (which is always the case if $n = 1$, but otherwise it is a special case occurring with zero probability), and someone predicts y_{n+1} by a value different from the value taken by all previous realizations of y, i.e., if $\mathsf{y}_1 = \mathsf{y}_2 = \cdots = \mathsf{y}_n \neq \mathsf{y}_{n+1}$, then q is unbounded and becomes infinite if μ takes the same value as the observed y_i. If, on the other hand, $\mathsf{y}_1 = \mathsf{y}_2 = \cdots = \mathsf{y}_{n+1}$, then $\mathsf{q} = 1$ if μ does not take the same value as the y_i, and q is a 1+0/0 undefined value otherwise, but by continuity we can say $\mathsf{q} = 1$ for all μ. Therefore $\mathsf{y}_{n+1} = \bar{\mathsf{y}}$ is the best predictor in this special case.

Now turn to the regular case in which not all observed y_i are equal. Re-write q as

$$\mathsf{q} = \frac{(\mu-\mathsf{y}_{n+1})^2}{(\boldsymbol{\iota}\bar{\mathsf{y}}-\mathsf{y})^\top(\boldsymbol{\iota}\bar{\mathsf{y}}-\mathsf{y})/n + (\mu-\bar{\mathsf{y}})^2} \tag{40.4.2}$$

If $\mathsf{y}_{n+1} = \bar{\mathsf{y}}$, then

$$\mathsf{q} = \frac{(\mu-\bar{\mathsf{y}})^2}{(\boldsymbol{\iota}\bar{\mathsf{y}}-\mathsf{y})^\top(\boldsymbol{\iota}\bar{\mathsf{y}}-\mathsf{y})/n + (\mu-\bar{\mathsf{y}})^2} \leq 1, \tag{40.4.3}$$

and it gets arbitrarily close to 1 for large absolute values of μ. Therefore the supremum associated with $\mathsf{y}_{n+1} = \bar{\mathsf{y}}$, which we claim to be the predictor which gives the lowest supremum, is 1.

For step (2) of the proof we have to show: if y_{n+1} is not equal to $\bar{\mathsf{y}}$, then q can assume values larger than 1. To show this, we will find for a given y and y_{n+1} that parameter value μ for which this relative increase is highest, and the value of the highest relative increase.

Look at the μ defined by either one of the following equations

$$\mu - \bar{\mathsf{y}} = -\frac{(\boldsymbol{\iota}\bar{\mathsf{y}}-\mathsf{y})^\top(\boldsymbol{\iota}\bar{\mathsf{y}}-\mathsf{y})/n}{\mathsf{y}_{n+1}-\bar{\mathsf{y}}} \tag{40.4.4}$$

or equivalently

$$\mu - \mathsf{y}_{n+1} = -\frac{(\mathsf{y}_{n+1} - \bar{\mathsf{y}})^2 + (\boldsymbol{\iota}\bar{\mathsf{y}} - \mathbf{y})^\top(\boldsymbol{\iota}\bar{\mathsf{y}} - \mathbf{y})/n}{\mathsf{y}_{n+1} - \bar{\mathsf{y}}} \tag{40.4.5}$$

With this μ, equation (40.4.2) for q becomes

$$\mathsf{q} = \frac{\frac{\left((\mathsf{y}_{n+1}-\bar{\mathsf{y}})^2+(\boldsymbol{\iota}\bar{\mathsf{y}}-\mathbf{y})^\top(\boldsymbol{\iota}\bar{\mathsf{y}}-\mathbf{y})/n\right)^2}{(\mathsf{y}_{n+1}-\bar{\mathsf{y}})^2}}{(\boldsymbol{\iota}\bar{\mathsf{y}} - \mathbf{y})^\top(\boldsymbol{\iota}\bar{\mathsf{y}} - \mathbf{y})/n + \frac{((\boldsymbol{\iota}\bar{\mathsf{y}}-\mathbf{y})^\top(\boldsymbol{\iota}\bar{\mathsf{y}}-\mathbf{y})/n)^2}{(\mathsf{y}_{n+1}-\bar{\mathsf{y}})^2}} \tag{40.4.6}$$

$$= \frac{\left((\mathsf{y}_{n+1} - \bar{\mathsf{y}})^2 + (\boldsymbol{\iota}\bar{\mathsf{y}} - \mathbf{y})^\top(\boldsymbol{\iota}\bar{\mathsf{y}} - \mathbf{y})/n\right)^2}{(\mathsf{y}_{n+1} - \bar{\mathsf{y}})^2(\boldsymbol{\iota}\bar{\mathsf{y}} - \mathbf{y})^\top(\boldsymbol{\iota}\bar{\mathsf{y}} - \mathbf{y})/n + \left((\boldsymbol{\iota}\bar{\mathsf{y}} - \mathbf{y})^\top(\boldsymbol{\iota}\bar{\mathsf{y}} - \mathbf{y})/n\right)^2} \tag{40.4.7}$$

$$= \frac{1}{(\boldsymbol{\iota}\bar{\mathsf{y}} - \mathbf{y})^\top(\boldsymbol{\iota}\bar{\mathsf{y}} - \mathbf{y})/n}\frac{\left((\mathsf{y}_{n+1} - \bar{\mathsf{y}})^2 + (\boldsymbol{\iota}\bar{\mathsf{y}} - \mathbf{y})^\top(\boldsymbol{\iota}\bar{\mathsf{y}} - \mathbf{y})/n\right)^2}{(\mathsf{y}_{n+1} - \bar{\mathsf{y}})^2 + (\boldsymbol{\iota}\bar{\mathsf{y}} - \mathbf{y})^\top(\boldsymbol{\iota}\bar{\mathsf{y}} - \mathbf{y})/n} \tag{40.4.8}$$

$$= \frac{(\mathsf{y}_{n+1} - \bar{\mathsf{y}})^2 + (\boldsymbol{\iota}\bar{\mathsf{y}} - \mathbf{y})^\top(\boldsymbol{\iota}\bar{\mathsf{y}} - \mathbf{y})/n}{(\boldsymbol{\iota}\bar{\mathsf{y}} - \mathbf{y})^\top(\boldsymbol{\iota}\bar{\mathsf{y}} - \mathbf{y})/n} = \frac{(\mathsf{y}_{n+1} - \bar{\mathsf{y}})^2}{(\boldsymbol{\iota}\bar{\mathsf{y}} - \mathbf{y})^\top(\boldsymbol{\iota}\bar{\mathsf{y}} - \mathbf{y})/n} + 1 \tag{40.4.9}$$

which is clearly greater than 1. This concludes the proof that $\bar{\mathsf{y}}$ minimaxes the relative increase in the Mahalanobis distance over all values of μ and σ^2.

40.5. Third Scenario: one additonal observation in a Regression Model

Our third scenario starts with an observation $\boldsymbol{y}$ of the random n-vector $\mathbf{y} \sim (\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$, where the nonrandom $\boldsymbol{X}$ has full column rank, and the parameters $\boldsymbol{\beta}$ and $\sigma^2 > 0$ are unknown. The squared Mahalanobis distance of this observation from its population is

$$\text{MHD}[\boldsymbol{y}; \mathbf{y}] = \text{MHD}\big[\boldsymbol{y}; (\boldsymbol{X\beta}, \sigma^2\boldsymbol{I}_n)\big] = \frac{1}{n\sigma^2}(\boldsymbol{y} - \boldsymbol{X\beta})^\top(\boldsymbol{y} - \boldsymbol{X\beta}) \tag{40.5.1}$$

If an $n+1$st observation y_{n+1} becomes available, associated with the row of regressor values $\boldsymbol{x}_{n+1}^\top$, then the Mahalanobis distance of the total vector is

$$\text{MHD}\Big[\begin{bmatrix}\boldsymbol{y}\\ y_{n+1}\end{bmatrix}; \begin{bmatrix}\mathbf{y}\\ \mathsf{y}_{n+1}\end{bmatrix}\Big] = \text{MHD}\Big[\begin{bmatrix}\boldsymbol{y}\\ y_{n+1}\end{bmatrix}; \Big(\begin{bmatrix}\boldsymbol{X}\\ \boldsymbol{x}_{n+1}^\top\end{bmatrix}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_{n+1}\Big)\Big] = \tag{40.5.2}$$

$$= \frac{1}{(n+1)\sigma^2}\big((\boldsymbol{y} - \boldsymbol{X\beta})^\top(\boldsymbol{y} - \boldsymbol{X\beta}) + (y_{n+1} - \boldsymbol{x}_{n+1}^\top\boldsymbol{\beta})^2\big) \tag{40.5.3}$$

and the conditional Mahalanobis distance is

$$\text{MHD}[y_{n+1}; \mathsf{y}_{n+1}|\boldsymbol{y}] = (y_{n+1} - \boldsymbol{x}_{n+1}^\top\boldsymbol{\beta})^2 \tag{40.5.4}$$

Both Mahalanobis distances (40.5.1) and (40.5.4) are unknown since they depend on the unknown parameters. However we will show here that whatever the true value of the parameters, the *ratio* of the conditional divided by the original Mahalanobis distance is never greater than

$$\frac{\text{MHD}[y_{n+1}; \mathsf{y}_{n+1}|\boldsymbol{y}]}{\text{MHD}[\boldsymbol{y}; \mathbf{y}]} \leq \frac{n}{k}\Big(\boldsymbol{x}_{n+1}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_{n+1} + \frac{(\boldsymbol{x}_{n+1}^\top\hat{\boldsymbol{\beta}} - y_{n+1})^2}{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top\boldsymbol{X}^\top\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}\Big). \tag{40.5.5}$$

Equation (40.5.5) is the ratio of these distances in the worst case. From it one sees immediately that that value of y_{n+1} which minimizes the maximum possible *ratio of the conditional Mahalanobis distance divided by the old*, this maximum taken over all possible values of $\boldsymbol{\beta}$ and σ^2, is exactly the OLS predicted value of y_{n+1} on the basis of the given data.

Leaving out the degrees of freedom n and k we have to find the minimax value of

$$\mathsf{q} = \frac{(y_{n+1} - \boldsymbol{x}_{n+1}^\top\boldsymbol{\beta})^2}{(\boldsymbol{y} - \boldsymbol{X\beta})^\top(\boldsymbol{y} - \boldsymbol{X\beta})} \tag{40.5.6}$$

We will show that the OLS prediction $y_{n+1} = \boldsymbol{x}_{n+1}^\top\hat{\boldsymbol{\beta}}$ is the solution to this minimax problem, and that the minimax value of q is $\mathsf{q} = \boldsymbol{x}_{n+1}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_{n+1}$.

This proof will proceed in two steps. (1) For $y_{n+1} = \boldsymbol{x}_{n+1}^\top\hat{\boldsymbol{\beta}}$, $\mathsf{q} \leq \boldsymbol{x}_{n+1}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_{n+1}$ for all values of $\boldsymbol{\beta}$, and whatever g-inverse was used in (40.5.6), but one can find $\boldsymbol{\beta}$ for which q is arbitrarily close to $\boldsymbol{x}_{n+1}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_{n+1}$. (2) If $y_{n+1} \neq \boldsymbol{x}_{n+1}^\top\hat{\boldsymbol{\beta}}$, then $\mathsf{q} > \boldsymbol{x}_{n+1}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_{n+1}$ for certain values of $\boldsymbol{\beta}$, and again independent of the choice of g-inverse in (40.5.6).

In the proof of step (1), the case $\boldsymbol{y} = \boldsymbol{X}\tilde{\boldsymbol{\beta}}$ for some $\tilde{\boldsymbol{\beta}}$ must be treated separately. If this condition holds (which is always the case if rank $\boldsymbol{X} = n$, but otherwise it only occurs with zero probability), and someone predicts y_{n+1} by a value different than $\boldsymbol{x}_{n+1}^\top\tilde{\boldsymbol{\beta}}$, then q is unbounded as the true $\boldsymbol{\beta}$ approaches $\boldsymbol{\beta} \to \tilde{\boldsymbol{\beta}}$.

If, on the other hand, $\boldsymbol{y}_{n+1}$ is predicted by $\boldsymbol{y}_{n+1} = \boldsymbol{x}_{n+1}^\top\tilde{\boldsymbol{\beta}}$, then

$$\mathsf{q} = \frac{\big(\boldsymbol{x}_{n+1}^\top(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})\big)^2}{(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top\boldsymbol{X}^\top\boldsymbol{X}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})} \leq \boldsymbol{x}_{n+1}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_{n+1} \tag{40.5.7}$$

if $\boldsymbol{\beta} \neq \tilde{\boldsymbol{\beta}}$ (with equality holding if $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \lambda(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_{n+1}$ for some $\lambda \neq 0$), and $\mathsf{q} = 0$ if $\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}$.

Now turn to the regular case in which the vector $\mathbf{y}$ cannot be written in the form $\mathbf{y} = \boldsymbol{X}\tilde{\boldsymbol{\beta}}$ for any $\tilde{\boldsymbol{\beta}}$. Re-write (40.5.6) as

$$\mathsf{q} = \frac{(\mathsf{y}_{n+1} - \boldsymbol{x}_{n+1}^\top\boldsymbol{\beta})^2}{(\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top\boldsymbol{X}^\top\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})} \tag{40.5.8}$$

If $\mathsf{y}_{n+1} = \boldsymbol{x}_{n+1}^\top \hat{\boldsymbol{\beta}}$, then

(40.5.9)
$$\mathsf{q} = \frac{\big(\boldsymbol{x}_{n+1}^\top(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\big)^2}{(\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \boldsymbol{X}^\top \boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})} \leq \frac{\big(\boldsymbol{x}_{n+1}^\top(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\big)^2}{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \boldsymbol{X}^\top \boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})} \leq \boldsymbol{x}_{n+1}^\top$$

One gets arbitrarily close to $\boldsymbol{x}_{n+1}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_{n+1}$ for $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \lambda(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_{n+1}$ and λ sufficiently large.

For step (2) of the proof we have to show: if y_{n+1} is not equal to $\boldsymbol{x}_{n+1}^\top\hat{\boldsymbol{\beta}}$, then q can assume values larger than $\boldsymbol{x}_{n+1}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_{n+1}$. To show this, we will find for a given $\mathbf{y}$ and y_{n+1} that parameter value $\tilde{\boldsymbol{\beta}}$ for which this relative increase is highest, and the value of the highest relative increase.

Here is the rough draft for the continuation of this proof. Here I am solving the first order condition and I am not 100 percent sure whether it is a global maximum. For the derivation which follows this won't matter, but I am on the lookout for a proof that it is. After I am done with this, this need not even be in the proof, all that needs to be in the proof is that this highest value (wherever I get it from) gives a value of q that is greater than $n\boldsymbol{x}_{n+1}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_{n+1}$.

Since there is no $\tilde{\boldsymbol{\beta}}$ with $\mathbf{y} = \boldsymbol{X}\tilde{\boldsymbol{\beta}}$, the quadratic form $b^2 = (\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$ is positive. The relative increase can therefore be written in the form

$$\mathsf{q} = n\frac{(\zeta - a)^2}{b^2 + \xi^2} = n\frac{u}{v} \tag{40.5.10}$$

where $\zeta = \boldsymbol{x}_{n+1}^\top\boldsymbol{\beta}$, $a = \mathsf{y}_{n+1}$, and $\xi^2 = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top\boldsymbol{X}^\top\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$. Hence $\frac{\partial\xi^2}{\partial\boldsymbol{\beta}^\top} = -2(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top\boldsymbol{X}^\top\boldsymbol{X} =: -2\boldsymbol{\alpha}^\top$. To get the highest relative increase, we need the roots of

$$\left(\frac{u}{v}\right)' = \frac{u'v - v'u}{v^2} \tag{40.5.11}$$

Here, in a provisional notation which will be replaced by matrix differentiation eventually, the prime represents the derivative with respect to β_i. I will call $-2\alpha_i = \frac{\partial\xi^2}{\partial\beta_i}$. Since the numerator is always positive, we only have to look for the roots of $u'v - v'u = 2x_{n+1,i}(\zeta - a)(b^2 + \xi^2) + 2\alpha_i(\zeta - a)^2$. Here it is in matrix differentiation:

$$\frac{\partial u}{\partial\boldsymbol{\beta}}v - \frac{\partial v}{\partial\boldsymbol{\beta}}u = \frac{\partial\zeta}{\partial\boldsymbol{\beta}}2(\zeta - a)(b^2 + \xi^2) - \frac{\partial\xi^2}{\partial\boldsymbol{\beta}}(\zeta - a)^2 = 0 \tag{40.5.12}$$

One root is $\zeta = a$, which is clearly a minimum, not a maximum. Division by $2\zeta - a$ gives

$$\boldsymbol{x}_{n+1}(b^2 + \xi^2) + (\boldsymbol{X}^\top\boldsymbol{X})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\zeta - a) = 0 \tag{40.5.13}$$

$$(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_{n+1}(b^2 + \xi^2) = -(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\zeta - a) \tag{40.5.14}$$

In other words,

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = -\gamma(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_{n+1} \qquad \text{where} \quad \gamma = \frac{b^2 + \xi^2}{\zeta - a}. \tag{40.5.15}$$

One can express ξ^2 and ζ in terms of γ:

$$\xi^2 = \gamma^2\boldsymbol{x}_{n+1}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_{n+1} \tag{40.5.16}$$

and

$$\zeta = \boldsymbol{x}_{n+1}^\top\hat{\boldsymbol{\beta}} + \gamma\boldsymbol{x}_{n+1}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_{n+1} \tag{40.5.17}$$

Therefore the defining identity for γ becomes

$$\gamma = \frac{b^2 + \xi^2}{\zeta - a} = \frac{b^2 + \gamma^2\boldsymbol{x}_{n+1}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_{n+1}}{\boldsymbol{x}_{n+1}^\top\hat{\boldsymbol{\beta}} - \mathsf{y}_{n+1} + \gamma\boldsymbol{x}_{n+1}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_{n+1}} \tag{40.5.18}$$

or equivalently

$$\gamma(\boldsymbol{x}_{n+1}^\top\hat{\boldsymbol{\beta}} - \mathsf{y}_{n+1}) + \gamma^2\boldsymbol{x}_{n+1}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_{n+1} = b^2 + \gamma^2\boldsymbol{x}_{n+1}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_{n+1} \tag{40.5.19}$$

The terms with γ^2 fall out and one obtains

$$\gamma = \frac{b^2}{\boldsymbol{x}_{n+1}^\top\hat{\boldsymbol{\beta}} - \mathsf{y}_{n+1}}. \tag{40.5.20}$$

The worst $\boldsymbol{\beta}$ is

$$\boldsymbol{\beta} = \hat{\boldsymbol{\beta}} - (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_{n+1}\frac{(\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})}{\mathsf{y}_{n+1} - \boldsymbol{x}_{n+1}^\top\hat{\boldsymbol{\beta}}}. \tag{40.5.21}$$

Now I should start from here and prove it new, and also do the proof that all other values of β give better q.

(40.5.16) and (40.5.17) into (40.5.10) gives

$$\mathsf{q} = \frac{\big(\boldsymbol{x}_{n+1}^\top\hat{\boldsymbol{\beta}} - \mathsf{y}_{n+1} + \gamma\boldsymbol{x}_{n+1}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_{n+1}\big)^2}{b^2 + \gamma^2\boldsymbol{x}_{n+1}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_{n+1}}, \tag{40.5.22}$$

and using $b^2 = \gamma(\boldsymbol{x}_{n+1}^\top\hat{\boldsymbol{\beta}} - \mathsf{y}_{n+1})$ one gets

(40.5.23)
$$\begin{aligned}\mathsf{q} = \frac{1}{\gamma}\,\frac{\big(\boldsymbol{x}_{n+1}^\top\hat{\boldsymbol{\beta}} - \mathsf{y}_{n+1} + \gamma\boldsymbol{x}_{n+1}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_{n+1}\big)^2}{\boldsymbol{x}_{n+1}^\top\hat{\boldsymbol{\beta}} - \mathsf{y}_{n+1} + \gamma\boldsymbol{x}_{n+1}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_{n+1}} &= \frac{\boldsymbol{x}_{n+1}^\top\hat{\boldsymbol{\beta}} - \mathsf{y}_{n+1}}{\gamma} + \boldsymbol{x}_{n+1}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_{n+1} \\ &= \frac{(\boldsymbol{x}_{n+1}^\top\hat{\boldsymbol{\beta}} - \mathsf{y}_{n+1})^2}{b^2} + \boldsymbol{x}_{n+1}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_{n+1}\end{aligned}$$

This is clearly bigger than $\boldsymbol{x}_{n+1}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_{n+1}$. This is what he had to show.

By the way, the excess of this q over the minimum value is proportional to the F-statistic for the test whether the $n+1$st observation comes from the same value $\boldsymbol{\beta}$. (The only difference is that numerator and denominator are not divided by their degrees of freedom). This gives a new interpretation of the F-test, and also of the F-confidence regions (which will be more striking if we predict more than one observation). F-confidence regions are conjectured observations for which the minimax value of the Mahalanobis ratio stays below a certain bound.

Now the proof that it is the worst $\boldsymbol{\beta}$. Without loss of generality we can write and $\boldsymbol{\beta}$ as follows in terms of a $\boldsymbol{\delta}$:

$$\boldsymbol{\beta} = \hat{\boldsymbol{\beta}} - \gamma(\boldsymbol{X}^\top\boldsymbol{X})^{-1}(\boldsymbol{x}_{n+1}+\boldsymbol{\delta}) \qquad \text{where} \quad \gamma = \frac{(\mathsf{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\mathsf{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})}{\mathsf{y}_{n+1}-\boldsymbol{x}_{n+1}^\top\hat{\boldsymbol{\beta}}} \tag{40.5.24}$$

(this is a sign change from above). Then

$$(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top\boldsymbol{X}^\top\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) = \gamma^2(\boldsymbol{x}_{n+1}+\boldsymbol{\delta})^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}(\boldsymbol{x}_{n+1}+\boldsymbol{\delta}) \tag{40.5.25}$$

and

$$\mathsf{y}_{n+1}-\boldsymbol{x}_{n+1}^\top\boldsymbol{\beta} = \mathsf{y}_{n+1}-\boldsymbol{x}_{n+1}^\top\hat{\boldsymbol{\beta}} + \gamma\boldsymbol{x}_{n+1}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}(\boldsymbol{x}_{n+1}+\boldsymbol{\delta}) \tag{40.5.26}$$

therefore we get

$$\mathsf{q} = \frac{\big(\mathsf{y}_{n+1}-\boldsymbol{x}_{n+1}^\top\hat{\boldsymbol{\beta}}+\gamma\boldsymbol{x}_{n+1}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}(\boldsymbol{x}_{n+1}+\boldsymbol{\delta})\big)^2}{(\mathsf{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\mathsf{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})+\gamma^2(\boldsymbol{x}_{n+1}+\boldsymbol{\delta})^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}(\boldsymbol{x}_{n+1}+\boldsymbol{\delta})}. \tag{40.5.27}$$

We have to show that the maximum value

$$\frac{(\boldsymbol{x}_{n+1}^\top\hat{\boldsymbol{\beta}}-\mathsf{y}_{n+1})^2}{(\mathsf{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\mathsf{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})} + \boldsymbol{x}_{n+1}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_{n+1} \tag{40.5.28}$$

is bigger than this, i.e., we have to show

$$\frac{(\boldsymbol{x}_{n+1}^\top\hat{\boldsymbol{\beta}}-\mathsf{y}_{n+1})^2}{(\mathsf{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\mathsf{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})}+\boldsymbol{x}_{n+1}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_{n+1} \geq \frac{\big(\mathsf{y}_{n+1}-\boldsymbol{x}_{n+1}^\top\hat{\boldsymbol{\beta}}+\gamma\boldsymbol{x}_{n+1}^\top(\boldsymbol{X}^\top\boldsymbol{X})^-}{(\mathsf{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\mathsf{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})+\gamma^2(\boldsymbol{x}_{n+1}+\boldsymbol{\delta})^\top(} \tag{40.5.29}$$

Since the denominator on the right hand side is positive, we can multiply both sides with it. Using the notation $b^2 = (\mathsf{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\mathsf{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})$ and the identities

$\gamma^2\frac{(\boldsymbol{x}_{n+1}^\top\hat{\boldsymbol{\beta}}-\mathsf{y}_{n+1})^2}{b^2} = b^2$ and $\gamma(\boldsymbol{x}_{n+1}^\top\hat{\boldsymbol{\beta}}-\mathsf{y}_{n+1}) = b^2$ we get

$$\begin{aligned}(\boldsymbol{x}_{n+1}^\top\hat{\boldsymbol{\beta}}-\mathsf{y}_{n+1})^2 + b^2\boldsymbol{x}_{n+1}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_{n+1} + b^2(\boldsymbol{x}_{n+1}+\boldsymbol{\delta})^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}(\boldsymbol{x}_{n+1}+\boldsymbol{\delta}) + \\ + \gamma^2(\boldsymbol{x}_{n+1}+\boldsymbol{\delta})^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}(\boldsymbol{x}_{n+1}+\boldsymbol{\delta})\boldsymbol{x}_{n+1}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_{n+1} \geq \\ \geq (\mathsf{y}_{n+1}-\boldsymbol{x}_{n+1}^\top\hat{\boldsymbol{\beta}})^2 + 2b^2\boldsymbol{x}_{n+1}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}(\boldsymbol{x}_{n+1}+\boldsymbol{\delta}) + \gamma^2\big(\boldsymbol{x}_{n+1}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}(\boldsymbol{x}_{n+1}+\boldsymbol{\delta})\big)^2\end{aligned} \tag{40.5.30}$$

or

$$\begin{aligned}b^2\boldsymbol{x}_{n+1}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_{n+1} + b^2(\boldsymbol{x}_{n+1}+\boldsymbol{\delta})^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}(\boldsymbol{x}_{n+1}+\boldsymbol{\delta}) + \\ + \gamma^2(\boldsymbol{x}_{n+1}+\boldsymbol{\delta})^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}(\boldsymbol{x}_{n+1}+\boldsymbol{\delta})\boldsymbol{x}_{n+1}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_{n+1} - \\ - 2b^2\boldsymbol{x}_{n+1}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}(\boldsymbol{x}_{n+1}+\boldsymbol{\delta}) - \gamma^2\big(\boldsymbol{x}_{n+1}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}(\boldsymbol{x}_{n+1}+\boldsymbol{\delta})\big)^2 \geq 0\end{aligned} \tag{40.5.31}$$

Collecting terms we get

$$b^2\boldsymbol{\delta}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{\delta} + \gamma^2\big(\boldsymbol{\delta}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{\delta}\boldsymbol{x}_{n+1}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_{n+1} - (\boldsymbol{x}_{n+1}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{\delta})^2\big) \geq 0 \tag{40.5.32}$$

which certainly holds. These steps can be reversed, which concludes the proof.

CHAPTER 41

Interval Estimation

We will first show how the least squares principle can be used to construct confidence regions, and then we will derive the properties of these confidence regions.

41.1. A Basic Construction Principle for Confidence Regions

The least squares objective function, whose minimum argument gave us the BLUE, naturally allows us to generate *confidence intervals* or higher-dimensional *confidence regions*. A confidence region for $\boldsymbol{\beta}$ based on $\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$ can be constructed as follows:

- Draw the OLS estimate $\hat{\boldsymbol{\beta}}$ into k-dimensional space; it is the vector which minimizes $\mathsf{SSE} = (\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$.
- For every other vector $\tilde{\boldsymbol{\beta}}$ one can define the sum of squared errors associated with that vector as $\mathsf{SSE}_{\tilde{\boldsymbol{\beta}}} = (\mathbf{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}})^\top(\mathbf{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}})$. Draw the level hypersurfaces (if $k = 2$: level lines) of this function. These are ellipsoids centered on $\hat{\boldsymbol{\beta}}$.
- Each of these ellipsoids is a confidence region for $\boldsymbol{\beta}$. Different confidence regions differ by their coverage probabilities.
- If one is only interested in certain coordinates of $\boldsymbol{\beta}$ and not in the others, or in some other linear transformation $\boldsymbol{\beta}$, then the corresponding confidence regions are the corresponding transformations of this ellipse. Geometrically this can best be seen if this transformation is an orthogonal projection; then the confidence ellipse of the transformed vector $\boldsymbol{R}\boldsymbol{\beta}$ is also a projection or "shadow" of the confidence region for the whole vector. Projections of the same confidence region have the same confidence level, independent of the direction in which this projection goes.

The confidence regions for $\boldsymbol{\beta}$ with coverage probability π will be written here as $\mathsf{B}_{\boldsymbol{\beta};\pi}$ or, if we want to make its dependence on the observation vector $\mathbf{y}$ explicit, $B_{\boldsymbol{\beta};\pi}(\mathbf{y})$. These confidence regions are level lines of the SSE, and mathematically, it is advantageous to define these level lines by their level relative to the minimum level, i.e., as as the set of all $\tilde{\boldsymbol{\beta}}$ for which the quotient of the attained $\mathsf{SSE}_{\tilde{\boldsymbol{\beta}}} =$

$(\mathbf{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}})^\top(\mathbf{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}})$ divided by the smallest possible $\mathsf{SSE} = (\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$ is smaller or equal a given number. In formulas,

$$\tilde{\boldsymbol{\beta}} \in B_{\boldsymbol{\beta};\pi}(\mathbf{y}) \iff \frac{(\mathbf{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}})^\top(\mathbf{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}})}{(\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})} \leq c_{\pi;n-k,k} \tag{41.1.1}$$

It will be shown below, in the discussion following (41.2.1), that $c_{\pi;n-k,k}$ only depends on π (the confidence level), $n - k$ (the degrees of freedom in the regression), and k (the dimension of the confidence region).

To get a geometric intuition of this principle, look at the case $k = 2$, in which the parameter vector $\boldsymbol{\beta}$ has only two components. For each possible value $\tilde{\boldsymbol{\beta}}$ of the parameter vector, the associated sum of squared errors is $\mathsf{SSE}_{\tilde{\boldsymbol{\beta}}} = (\mathbf{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}})^\top(\mathbf{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}})$. This a quadratic function of $\tilde{\boldsymbol{\beta}}$, whose level lines form concentric ellipses as shown in Figure 1. The center of these ellipses is the unconstrained least squares estimate. Each of the ellipses is a confidence region for $\boldsymbol{\beta}$ for a different confidence level.

If one needs a confidence region not for the whole vector $\boldsymbol{\beta}$ but, say, for i linearly independent linear combinations $\boldsymbol{R}\boldsymbol{\beta}$ (here $\boldsymbol{R}$ is a $i \times k$ matrix with full *row* rank), then the above principle applies in the following way: the vector $\tilde{\boldsymbol{u}}$ lies in the confidence region for $\boldsymbol{R}\boldsymbol{\beta}$ generated by $\mathbf{y}$ for confidence level π, notation $\mathsf{B}_{\boldsymbol{R}\boldsymbol{\beta};\pi}$, if and only if there is a $\tilde{\boldsymbol{\beta}}$ in the confidence region (41.1.1) (with the parameters adjusted to reflect the dimensionality of $\tilde{\boldsymbol{u}}$) which satisfies $\boldsymbol{R}\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{u}}$:

$$\tilde{\boldsymbol{u}} \in B_{\boldsymbol{R}\boldsymbol{\beta};\pi}(\mathbf{y}) \iff \text{exist } \tilde{\boldsymbol{\beta}} \text{ with } \tilde{\boldsymbol{u}} = \boldsymbol{R}\tilde{\boldsymbol{\beta}} \text{ and } \frac{(\mathbf{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}})^\top(\mathbf{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}})}{(\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})} \leq c_{\pi;n-k,i} \tag{41.1.2}$$

Problem 416. *Why does one have to change the value of c when one goes over to the projections of the confidence regions?*

Answer. Because the projection is a many-to-one mapping, and vectors which are not in the original ellipsoid may still end up in the projection. □

Again let us illustrate this with the 2-dimensional case in which the confidence region for $\boldsymbol{\beta}$ is an ellipse, as drawn in Figure 1, called $B_{\boldsymbol{\beta};\pi}(\mathbf{y})$. Starting with this ellipse, the above criterion defines individual confidence intervals for linear combinations $u = \boldsymbol{r}^\top\boldsymbol{\beta}$ by the rule: $\tilde{u} \in B_{\boldsymbol{r}^\top\boldsymbol{\beta};\pi}(\mathbf{y})$ iff a $\tilde{\boldsymbol{\beta}} \in B_{\boldsymbol{\beta}}(\mathbf{y})$ exists with $\boldsymbol{r}^\top\tilde{\boldsymbol{\beta}} = \tilde{u}$. For $\boldsymbol{r} = \left[\begin{smallmatrix}1\\0\end{smallmatrix}\right]$, this interval is simply the projection of the ellipse on the horizontal axis, and for $\boldsymbol{r} = \left[\begin{smallmatrix}0\\1\end{smallmatrix}\right]$ it is the projection on the vertical axis.

The same argument applies for all vectors $\boldsymbol{r}$ with $\boldsymbol{r}^\top\boldsymbol{r} = 1$. The inner product of two vectors is the length of the first vector times the length of the projection of the second vector on the first. If $\boldsymbol{r}^\top\boldsymbol{r} = 1$, therefore, $\boldsymbol{r}^\top\tilde{\boldsymbol{\beta}}$ is simply the length of the orthogonal projection of $\tilde{\boldsymbol{\beta}}$ on the line generated by the vector $\boldsymbol{r}$. Therefore

the confidence interval for $\boldsymbol{r}^\top\boldsymbol{\beta}$ is simply the projection of the ellipse on the line generated by $\boldsymbol{r}$. (This projection is sometimes called the "shadow" of the ellipse.)

The confidence region for $\boldsymbol{R\beta}$ can also be defined as follows: $\tilde{\boldsymbol{u}}$ lies in this confidence region if and only if the "best" $\hat{\hat{\boldsymbol{\beta}}}$ which satisfies $\boldsymbol{R}\hat{\hat{\boldsymbol{\beta}}} = \tilde{\boldsymbol{u}}$ lies in the confidence region (41.1.1), this best $\hat{\hat{\boldsymbol{\beta}}}$ being, of course, the constrained least squares estimate subject to the constraint $\boldsymbol{R\beta} = \tilde{\boldsymbol{u}}$, whose formula is given by (29.3.13). The confidence region for $\boldsymbol{R\beta}$ consists therefore of all $\tilde{\boldsymbol{u}}$ for which the constrained least squares estimate $\hat{\hat{\boldsymbol{\beta}}} = \hat{\boldsymbol{\beta}} - (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{u}})$ satisfies condition (41.1.1):

$$\tilde{\boldsymbol{u}} \in B_{\boldsymbol{R\beta}}(\mathsf{y}) \iff \frac{(\mathsf{y} - \boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}})^\top(\mathsf{y} - \boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}})}{(\mathsf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\mathsf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})} \leq c_{\pi;n-k,i} \tag{41.1.3}$$

One can also write it as

$$\tilde{\boldsymbol{u}} \in B_{\boldsymbol{R\beta}}(\mathsf{y}) \iff \frac{\mathsf{SSE}_{\text{constrained}}}{\mathsf{SSE}_{\text{unconstrained}}} \leq c_{\pi;n-k,i} \tag{41.1.4}$$

i.e., those $\tilde{\boldsymbol{u}}$ are in the confidence region which, if imposed as a constraint on the regression, will not make the SSE too much bigger.

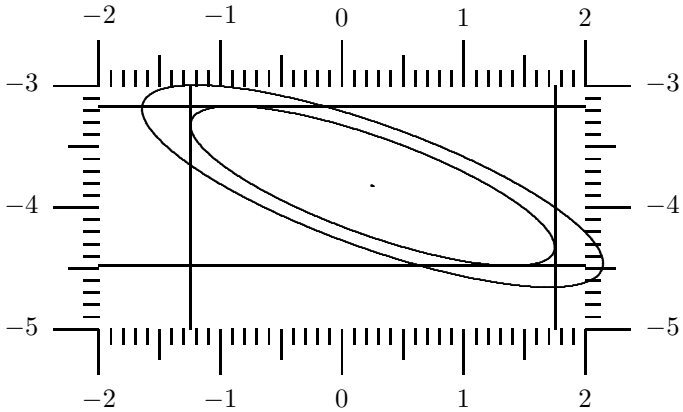


FIGURE 1. Confidence Ellipse with "Shadows"

In order to transform (41.1.3) into a mathematically more convenient form, write it as

$$\tilde{\boldsymbol{u}} \in B_{\boldsymbol{R\beta};\pi}(\mathsf{y}) \iff \frac{(\mathsf{y} - \boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}})^\top(\mathsf{y} - \boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}}) - (\mathsf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\mathsf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})}{(\mathsf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\mathsf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})} \leq c_{\pi;n-k,i} - 1$$

and then use (29.7.2) to get

$$\tilde{\boldsymbol{u}} \in B_{\boldsymbol{R\beta};\pi}(\mathsf{y}) \iff \frac{(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{u}})^\top\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{u}})}{(\mathsf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\mathsf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})} \leq c_{\pi;n-k,i} - 1 \tag{41.1.5}$$

This formula has the great advantage that $\hat{\hat{\boldsymbol{\beta}}}$ no longer appears in it. The condition whether $\tilde{\boldsymbol{u}}$ belongs to the confidence region is here formulated in terms of $\hat{\boldsymbol{\beta}}$ alone.

PROBLEM 417. *Using* (18.2.12), *show that* (41.1.1) *can be rewritten as*

$$\tilde{\boldsymbol{\beta}} \in B_{\boldsymbol{\beta};\pi}(\mathsf{y}) \iff \frac{(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})^\top\boldsymbol{X}^\top\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})}{(\mathsf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\mathsf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})} \leq c_{\pi;n-k,k} - 1 \tag{41.1.6}$$

Verify that this is the same as (41.1.5) *in the special case* $\boldsymbol{R} = \boldsymbol{I}$.

PROBLEM 418. *You have run a regression with intercept, but you are not interested in the intercept per se but need a joint confidence region for all slope parameters. Using the notation of Problem 361, show that this confidence region has the form*

$$\tilde{\boldsymbol{\beta}} \in B_{\boldsymbol{\beta};\pi}(\mathsf{y}) \iff \frac{(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})^\top\underline{\boldsymbol{X}}^\top\underline{\boldsymbol{X}}(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})}{(\underline{\mathsf{y}} - \underline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})^\top(\underline{\mathsf{y}} - \underline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})} \leq c_{\pi;n-k,k-1} - 1 \tag{41.1.7}$$

I.e., we are sweeping the means out of both regressors and dependent variables, and then we act as if the regression never had an intercept and use the formula for the full parameter vector (41.1.6) *for these transformed data (except that the number of degrees of freedom* $n-k$ *still reflects the intercept as one of the explanatory variables).*

ANSWER. Write the full parameter vector as $\begin{bmatrix}\alpha\\ \boldsymbol{\beta}\end{bmatrix}$ and $\boldsymbol{R} = \begin{bmatrix}\boldsymbol{o} & \boldsymbol{I}\end{bmatrix}$. Use (41.1.5) but instead of $\tilde{\boldsymbol{u}}$ write $\tilde{\boldsymbol{\beta}}$. The only tricky part is the following which uses (30.0.37):

$$\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top = \begin{bmatrix}\boldsymbol{o} & \boldsymbol{I}\end{bmatrix}\begin{bmatrix}1/n + \bar{\boldsymbol{x}}^\top(\underline{\boldsymbol{X}}^\top\underline{\boldsymbol{X}})^{-1}\bar{\boldsymbol{x}} & -\bar{\boldsymbol{x}}^\top(\underline{\boldsymbol{X}}^\top\underline{\boldsymbol{X}})^{-1}\\ -(\underline{\boldsymbol{X}}^\top\underline{\boldsymbol{X}})^{-1}\bar{\boldsymbol{x}} & (\underline{\boldsymbol{X}}^\top\underline{\boldsymbol{X}})^{-1}\end{bmatrix}\begin{bmatrix}\boldsymbol{o}^\top\\ \boldsymbol{I}\end{bmatrix} = (\underline{\boldsymbol{X}}^\top\underline{\boldsymbol{X}})^{-1} \tag{41.1.8}$$

The denominator is $(\mathsf{y} - \boldsymbol{\iota}\hat{\alpha} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\mathsf{y} - \boldsymbol{\iota}\hat{\alpha} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$, but since $\hat{\alpha} = \bar{\mathsf{y}} - \bar{\boldsymbol{x}}^\top\hat{\boldsymbol{\beta}}$, see problem 242, this denominator can be rewritten as $(\underline{\mathsf{y}} - \underline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})^\top(\underline{\mathsf{y}} - \underline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})$. □

PROBLEM 419. *3 points We are in the simple regression* $\mathsf{y}_t = \alpha + \beta x_t + \varepsilon_t$. *If one draws, for every value of* x, *a 95% confidence interval for* $\alpha + \beta x$, *one gets a "confidence band" around the fitted line, as shown in Figure 2. Is the probability that this confidence band covers the true regression line* over its whole length *equal to 95%, greater than 95%, or smaller than 95%? Give a good verbal reasoning for your answer. You should make sure that your explanation is consistent with the fact that the confidence interval is random and the true regression line is fixed.*

FIGURE 2. Confidence Band for Regression Line

41.2. Coverage Probability of the Confidence Regions

The probability that any *given known value* $\tilde{\boldsymbol{u}}$ lies in the confidence region (41.1.3) depends on the unknown $\boldsymbol{\beta}$. But we will show now that the "coverage probability" of the region, i.e., the probability with which the confidence region contains the *unknown true value* $\boldsymbol{u} = \boldsymbol{R\beta}$, does *not* depend on any unknown parameters.

To get the coverage probability, we must substitute $\tilde{\boldsymbol{u}} = \boldsymbol{R\beta}$ (where $\boldsymbol{\beta}$ is the true parameter value) in (41.1.5). This gives

(41.2.1)
$$\boldsymbol{R\beta} \in B_{\boldsymbol{R\beta};\pi}(\mathsf{y}) \iff \frac{(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{R\beta})^\top \big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{R\beta})}{(\mathsf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\mathsf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})} \leq c_{\pi;n-k,i} - 1$$

Let us look at numerator and denominator separately. Under the Normality assumption, $\boldsymbol{R}\hat{\boldsymbol{\beta}} \sim \mathsf{N}(\boldsymbol{R\beta}, \sigma^2\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top)$. Therefore, by (10.4.9), the distribution of the numerator of (41.2.1) is

$$(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{R\beta})^\top \big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{R\beta}) \sim \sigma^2\chi^2_i. \tag{41.2.2}$$

This probability distribution only depends on one unknown parameter, namely, σ^2. Regarding the denominator, remember that, by (24.4.2), $(\mathsf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\mathsf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \boldsymbol{\varepsilon}^\top\boldsymbol{M}\boldsymbol{\varepsilon}$, and if we apply (10.4.9) to this we can see that

$$(\mathsf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\mathsf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \sim \sigma^2\chi^2_{n-k} \tag{41.2.3}$$

Furthermore, numerator and denominator are independent. To see this, look first at $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\varepsilon}}$. By Problem 300 they are uncorrelated, and since they are also jointly Normal, it follows that they are independent. If $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\varepsilon}}$ are independent, any functions of $\hat{\boldsymbol{\beta}}$ are independent of any functions of $\hat{\boldsymbol{\varepsilon}}$. The numerator in the test

statistic (41.2.1) is a function of $\hat{\boldsymbol{\beta}}$ and the denominator is a function of $\hat{\boldsymbol{\varepsilon}}$; therefore they are independent, as claimed. Lastly, if we divide numerator by denominator, the unknown "nuisance parameter" σ^2 in their probability distributions cancels out, i.e., the distribution of the quotient is fully known.

To sum up: if $\tilde{\boldsymbol{u}}$ is the true value $\tilde{\boldsymbol{u}} = \boldsymbol{R\beta}$, then the test statistic in (41.2.1) can no longer be observed, but its *distribution* is is known; it is a χ^2_i divided by an independent χ^2_{n-k}. Therefore, for every value c, the probability that the confidence region (41.1.5) contains the true $\boldsymbol{R\beta}$ can be computed, and conversely, for any desired coverage probability, the appropriate critical value c can be computed. As claimed, this critical value only depends on the confidence level π and $n-k$ and i.

41.3. Conventional Formulas for the Test Statistics

In order to get this test statistic into the form in which it is conventionally tabulated, we must divide both numerator and denominator of (41.1.5) by their degrees of freedom, to get a χ^2_i/i divided by an independent $\chi^2_{n-k}/(n-k)$. This quotient is called a F-distribution with i and $n-k$ degrees of freedom.

The F-distribution is defined as $\mathsf{F}_{i,j} = \frac{\chi^2_i/i}{\chi^2_j/j}$ instead of the seemingly simpler formula $\frac{\chi^2_i}{\chi^2_j}$, because the division by the degrees of freedom makes all F-distributions and the associated critical values similar; an observed value below 4 is insignificant, but greater values may be signficant depending on the number of parameters.

Therefore, instead of , the condition deciding whether a given vector $\tilde{\boldsymbol{u}}$ lies in the confidence region for $\boldsymbol{R\beta}$ with confidence level $\pi = 1-\alpha$ is formulated as follows:

(41.3.1)
$$\frac{(\mathsf{SSE}_{\text{constrained}} - \mathsf{SSE}_{\text{unconstrained}})/\text{number of constraints}}{\mathsf{SSE}_{\text{unconstr.}}/(\text{numb. of obs.} - \text{numb. of coeff. in unconstr. model})} \leq F_{(i,n-k;\alpha)}$$

Here the constrained SSE is the SSE in the model estimated with the constraint $\boldsymbol{R\beta} = \tilde{\boldsymbol{u}}$ imposed, and $F_{(i,n-k;\alpha)}$ is the upper α quantile of the F distribution with i and $n-k$ degrees of freedom, i.e., it is that scalar c for which a random variable F which has a F distribution with i and $n-k$ degrees of freedom satisfies $\Pr[\mathsf{F} \geq c] = \alpha$.

41.4. Interpretation in terms of Studentized Mahalanobis Distance

The division of numerator and denominator by their degrees of freedom also gives us a second intuitive interpretation of the test statistic in terms of the Mahalanobis distance, see chapter 40. If one divides the denominator by its degrees of freedom, one gets an unbiased estimate of σ^2

$$\mathsf{s}^2 = \frac{1}{n-k}(\mathsf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\mathsf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}). \tag{41.4.1}$$

Therefore from (41.1.5) one gets the following alternative formula for the joint confidence region $B(\mathbf{y})$ for the vector parameter $\boldsymbol{u} = \boldsymbol{R\beta}$ for confidence level $\pi = 1-\alpha$:

(41.4.2)
$$\tilde{\boldsymbol{u}} \in B_{\boldsymbol{R\beta};1-\alpha}(\mathbf{y}) \iff \frac{1}{\mathrm{s}^2}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{u}})^\top \big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{u}}) \le iF_{(i,n-k;\alpha)}$$

Here $\hat{\boldsymbol{\beta}}$ is the least squares estimator of $\boldsymbol{\beta}$, and $\mathrm{s}^2 = (\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})/(n-k)$ the unbiased estimator of σ^2. Therefore $\hat{\boldsymbol{\Sigma}} = \mathrm{s}^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$ is the estimated covariance matrix as available in the regression printout. Therefore $\hat{\mathbf{V}} = \mathrm{s}^2\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top$ is the estimate of the covariance matrix of $\boldsymbol{R}\hat{\boldsymbol{\beta}}$. Another way to write (41.4.2) is therefore

$$B(\mathbf{y}) = \{\tilde{\boldsymbol{u}} \in \mathbb{R}^i\colon\ (\boldsymbol{R}\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{u}})^\top\hat{\mathbf{V}}^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{u}}) \le iF_{(i,n-k;\alpha)}\}. \tag{41.4.3}$$

This formula allows a suggestive interpretation. whether $\tilde{\boldsymbol{u}}$ lies in the confidence region or not depends on the Mahalanobis distance of the actual value of $\boldsymbol{R}\hat{\boldsymbol{\beta}}$ would have from the distribution which $\boldsymbol{R}\hat{\boldsymbol{\beta}}$ would have if the true parameter vector were to satisfy the constraint $\boldsymbol{R\beta} = \tilde{\boldsymbol{u}}$. It is not the Mahalanobis distance itself but only an estimate of it because σ^2 is replaced by its unbiased estimate s^2.

These formulas are also useful for drawing the confidence ellipses. The r which you need in equation (10.3.22) in order to draw the confidence ellipse is $r = \sqrt{iF_{(i,n-k;\alpha)}}$. This is the same as the local variable `mult` in the following S-function to draw this ellipse: its arguments are the center point (a 2-vector `d`), the estimated covariance matrix (a 2×2 matrix `C`), the degrees of freedom in the denominator of the F-distribution (the scalar `df`), and the confidence level (the scalar `level` between 0 and 1 which defaults to 0.95 if not specified).

```
confelli <-
function(b, C, df, level = 0.95, xlab = "", ylab = "", add=T, prec=51)

# Plot an ellipse with "covariance matrix" C, center b, and P-content
# level according the F(2,df) distribution.
# Sent to S-NEWS on May 19, 1999 by Roger Koenker
# Department of Economics
# University of Illinois
# Champaign, IL 61820
# url: http://www.econ.uiuc.edu
# email roger@ysidro.econ.uiuc.edu
# vox: 217-333-4558
# fax: 217-244-6678.
# Included in the ecmet package with his permission.
```



```
{
d <- sqrt(diag(C))
dfvec <- c(2, df)
phase <- acos(C[1, 2]/(d[1] * d[2]))
angles <- seq( - (PI), PI, len = prec)
mult <- sqrt(dfvec[1] * qf(level, dfvec[1], dfvec[2]))
xpts <- b[1] + d[1] * mult * cos(angles)
ypts <- b[2] + d[2] * mult * cos(angles + phase)
if(add) lines(xpts, ypts)
else plot(xpts, ypts, type = "l", xlab = xlab, ylab = ylab)
}
```

The mathematics why this works is in Problem 166.

Problem 420. *3 points In the regression model $\mathbf{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ you observe $\mathbf{y}$ and the (nonstochastic) $\boldsymbol{X}$ and you construct the following confidence region $\mathsf{B}(\mathbf{y})$ for $\boldsymbol{R\beta}$, where $\boldsymbol{R}$ is a $i \times k$ matrix with full row rank:*

(41.4.4)
$$\mathsf{B}(\mathbf{y}) = \{\boldsymbol{u} \in \mathbb{R}^i\colon\ (\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{u})^\top(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{u}) \le i\mathrm{s}^2F_{(i,n-k;\alpha)}\}.$$

Compute the probability that B contains the true $\boldsymbol{R\beta}$.

Answer.

$$\Pr[B(\mathbf{y}) \ni \boldsymbol{R\beta}] = \Pr[(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{R\beta})^\top(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{R\beta}) \le iF_{(i,n-k;\alpha)}\mathrm{s}^2] = \tag{41.4.5}$$

$$= \Pr[\frac{(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{R\beta})^\top(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{R\beta})/i}{\mathrm{s}^2} \le F_{(i,n-k;\alpha)}] = 1-\alpha \tag{41.4.6}$$

□

This interpretation with the Mahalanobis distance is commonly used for the construction of t-Intervals. A t-interval is a special case of the above confidence region for the case $i = 1$. The confidence interval with confidence level $1-\alpha$ for the scalar parameter $u = \boldsymbol{r}^\top\boldsymbol{\beta}$, where $\boldsymbol{r} \neq \boldsymbol{o}$ is a vector of constant coefficients, can be written as

$$\mathsf{B}(\mathbf{y}) = \{u \in \mathbb{R}\colon\ |u - \boldsymbol{r}^\top\hat{\boldsymbol{\beta}}| \le t_{(n-k;\alpha/2)}\mathrm{s}_{\boldsymbol{r}^\top\hat{\boldsymbol{\beta}}}\}. \tag{41.4.7}$$

What do those symbols mean? $\hat{\boldsymbol{\beta}}$ is the least squares estimator of $\boldsymbol{\beta}$. $t_{(n-k;\alpha/2)}$ is the upper $\alpha/2$-quantile of the t distribution with $n-k$ degrees of freedom, i.e., it is that scalar c for which a random variable t which has a t distribution with $n-k$ degrees of freedom satisfies $\Pr[\mathsf{t} \ge c] = \alpha/2$. Since by symmetry $\Pr[\mathsf{t} \le -c] = \alpha/2$ as well, one obtains the inequality relevant for a two-sided test:

$$\Pr[|\mathsf{t}| \ge t_{(n-k;\alpha/2)}] = \alpha. \tag{41.4.8}$$

Finally, $\mathrm{s}_{\boldsymbol{r}^\top\hat{\boldsymbol{\beta}}}$ is the estimated standard deviation of $\boldsymbol{r}^\top\hat{\boldsymbol{\beta}}$.

It is computed by the following three steps: First write down the variance of $\boldsymbol{r}^\top\hat{\boldsymbol{\beta}}$:

$$\operatorname{var}[\boldsymbol{r}^\top\hat{\boldsymbol{\beta}}] = \sigma^2\boldsymbol{r}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{r}. \tag{41.4.9}$$

Secondly, replace σ^2 by its unbiased estimator $\mathsf{s}^2 = (\mathsf{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\mathsf{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})/(n-k)$, and thirdly take the square root. This gives $\mathsf{s}_{\boldsymbol{r}^\top\hat{\boldsymbol{\beta}}} = \mathsf{s}\sqrt{\boldsymbol{r}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{r}}$.

Problem 421. *Which element(s) on the right hand side of* (41.4.7) *depend(s) on* y*?*

Answer. $\hat{\boldsymbol{\beta}}$ depends on y, and also $\mathsf{s}_{\boldsymbol{r}^\top\hat{\boldsymbol{\beta}}}$ depends on y through s^2. □

Let us verify that the coverage probability, i.e., the probability that the confidence interval constructed using formula (41.4.7) contains the true value $\boldsymbol{r}^\top\boldsymbol{\beta}$, is, as claimed, $1-\alpha$:

(41.4.10)

$$\Pr[\mathsf{B}(\mathsf{y}) \ni \boldsymbol{r}^\top\boldsymbol{\beta}] = \Pr[\left|\boldsymbol{r}^\top\boldsymbol{\beta} - \boldsymbol{r}^\top\hat{\boldsymbol{\beta}}\right| \le t_{(n-k;\alpha/2)}\mathsf{s}_{\boldsymbol{r}^\top\hat{\boldsymbol{\beta}}}]$$

$$= \Pr\left[\left|\boldsymbol{r}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon}\right| \le t_{(n-k;\alpha/2)}\mathsf{s}\sqrt{\boldsymbol{r}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{r}}\right] \tag{41.4.11}$$

$$= \Pr[\left|\frac{\boldsymbol{r}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon}}{\mathsf{s}\sqrt{\boldsymbol{r}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{r}}}\right| \le t_{(n-k;\alpha/2)}] \tag{41.4.12}$$

$$= \Pr[\left|\frac{\boldsymbol{r}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon}}{\sigma\sqrt{\boldsymbol{r}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{r}}}\Big/\frac{\mathsf{s}}{\sigma}\right| \le t_{(n-k;\alpha/2)}] = 1-\alpha, \tag{41.4.13}$$

This last equality holds because the expression left of the big slash is a standard normal, and the expression on the right of the big slash is the square root of an independent χ^2_{n-k} divided by $n-k$. The random variable between the absolute signs has therefore a t-distribution, and (41.4.13) follows from (41.4.8).

In R, one obtains $t_{(n-k;\alpha/2)}$ by giving the command `qt(1-alpha/2,n-p)`. Here `qt` stands for t-quantile [**BCW96**, p. 48]. One needs `1-alpha/2` instead of `alpha/2` because it is the usual convention for quantiles (or cumulative distribution functions) to be defined as lower quantiles, i.e., as the probabilities of a random variable being $\le$ a given number, while test statistics are usually designed in such a way that the significant values are the high values, i.e., for testing one needs the upper quantiles.

There is a basic duality between confidence intervals and hypothesis tests. Chapter 42 is therefore a discussion of the same subject under a slightly different angle:

CHAPTER 42

Three Principles for Testing a Linear Constraint

We work in the model $\mathbf{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ with normally distributed errors $\boldsymbol{\varepsilon} \sim \mathsf{N}(\boldsymbol{o}, \sigma^2\boldsymbol{I})$. There are three basic approaches to test the null hypothesis $\boldsymbol{R\beta} = \boldsymbol{u}$. In the linear model, these three approaches are mathematically equivalent, but if one goes over to nonlinear least squares or maximum likelihood estimators, they lead to different (although asymptotically equivalent) tests.

(1) ("Wald Criterion") Compute the vector of OLS estimates $\hat{\boldsymbol{\beta}}$, and reject the null hypothesis if $\boldsymbol{R}\hat{\boldsymbol{\beta}}$ is "too far away" from $\boldsymbol{u}$. For this criterion one only needs the unconstrained estimator, not the constrained one.

(2) ("Likelihood Ratio Criterion") Estimate the model twice: once with the constraint $\boldsymbol{R\beta} = \boldsymbol{u}$, and once without the constraint. Reject the null hypothesis if the model with the constraint imposed has a much worse fit than the model without the constraint.

(3) ("Lagrange Multiplier Criterion") This third criterion is based on the constrained estimator only. It has two variants. In its "score test" variant, one rejects the null hypothesis if the vector of derivatives of the unconstrained least squares objective function, evaluated at the constrained estimate $\hat{\hat{\boldsymbol{\beta}}}$, is too far away from $\boldsymbol{o}$. In the variant which has given this Criterion its name, one rejects if the vector of Lagrange multipliers needed for imposing the constraint is too far away from $\boldsymbol{o}$.

Many textbooks inadvertently and implicitly distinguish between (1) and (2) as follows: they introduce the t-test for one parameter by principle (1), and the F-test for several parameters by principle (2). Later, the student is surprised to find out that the t-test and the F-test in one dimension are equivalent, i.e., that the difference between t-test and F-test has nothing to do with the dimension of the parameter vector to be tested. Some textbooks make the distinction between (1) and (2) *explicit*. For instance [**Chr87**, p. 29ff] distinguishes between "testing linear parametric functions" and "testing models." However the distinction between all 3 principles has been introduced into the linear model only after the discovery that these three principles give different but asymptotically equivalent tests in the Maximum Likelihood estimation. Compare [**DM93**, Chapter 3.6] about this.

42.1. Mathematical Detail of the Three Approaches

(1) For the "Wald criterion" we must specify what it means that $\boldsymbol{R}\hat{\boldsymbol{\beta}}$ is "too far away" from $\boldsymbol{u}$. The Mahalanobis distance gives such a criterion: If the true $\boldsymbol{\beta}$ satisfies $\boldsymbol{R\beta} = \boldsymbol{u}$, then $\boldsymbol{R}\hat{\boldsymbol{\beta}} \sim (\boldsymbol{u}, \sigma^2\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top)$, and the Mahalanobis distance of the observed value of $\boldsymbol{R}\hat{\boldsymbol{\beta}}$ from this distribution is a logical candidate for the Wald criterion. The only problem is that σ^2 is not known, therefore we have to use the "studentized" Mahalanobis distance in which σ^2 is replaced by s^2. Conventionally, in the conterxt of linear regression, the Mahalanobis distance is also divided by the number of degrees of freedom; this normalizes its expected value to 1. Replacing σ^2 by s^2 and dividing by i gives the test statistic

$$\frac{1}{i}\frac{(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{u})^\top\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{u})}{\mathsf{s}^2}. \tag{42.1.1}$$

(2) Here are the details for the second approach, the "goodness-of-fit criterion." In order to compare the fit of the models, we look at the attained SSE's. Of course, the constrained SSE_r is always larger than the unconstrained SSE_u, even if the true parameter vector satisfies the constraint. But if we divide SSE_r by its degrees of freedom $n + i - k$, it is an unbiased estimator of σ^2 if the constraint holds and it is biased upwards if the constraint does not hold. The unconstrained SSE_u, divided by its degrees of freedom, on the other hand, is always an unbiased estimator of σ^2. If the constraint holds, the SSE's divided by their respective degrees of freedom should give roughly equal numbers. According to this, a feasible test statistic would be

$$\frac{\mathsf{SSE}_r/(n+i-k)}{\mathsf{SSE}_u/(n-k)} \tag{42.1.2}$$

and one would reject if this is too much > 1. The following variation of this is more convenient, since its distribution does not depend on n, k and i separately, but only through $n - k$ and i.

$$\frac{(\mathsf{SSE}_r - \mathsf{SSE}_u)/i}{\mathsf{SSE}_u/(n-k)} \tag{42.1.3}$$

It still has the property that the numerator is an unbiased estimator of σ^2 if the constraint holds and biased upwards if the constraint does not hold, and the denominator is always an unbiased estimator. Furthermore, in this variation, the numerator and denominator are independent random variables. If this test statistic is much larger than 1, then the constraints are incompatible with the data and the null hypothesis must be rejected. The statistic (42.1.3) can also be written as

$$\frac{(\mathsf{SSE}_{\text{constrained}} - \mathsf{SSE}_{\text{unconstrained}})/\text{number of constraints}}{\mathsf{SSE}_{\text{unconstrained}}/(\text{numb. of observations} - \text{numb. of coefficients in unconstr. model})} \tag{42.1.4}$$

The equivalence of formulas (42.1.1) and (42.1.4) is a simple consequence of (29.7.2).

(3) And here are the details about the score test variant of the Lagrange multiplier criterion: The Jacobian of the least squares objective function is

$$\frac{\partial}{\partial \boldsymbol{\beta}^\top}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) = -2(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^\top\boldsymbol{X}. \tag{42.1.5}$$

This is a row vector consisting of all the partial derivatives. Taking its transpose, in order to get a column vector, and plugging the constrained least squares estimate $\hat{\hat{\boldsymbol{\beta}}}$ into it gives $-2\boldsymbol{X}^\top(\boldsymbol{y}-\boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}})$. Again we need the Mahalanobis distance of this observed value from the distribution which the random variable

$$-2\boldsymbol{X}^\top(\mathsf{y}-\boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}}) \tag{42.1.6}$$

has if the true $\boldsymbol{\beta}$ satisfies $\boldsymbol{R}\boldsymbol{\beta}=\boldsymbol{u}$. If this constraint is satisfied, $\hat{\hat{\boldsymbol{\beta}}}$ is unbiased, therefore (42.1.6) has expected value zero. Furthermore, if one premultiplies (29.7.1) by $\boldsymbol{X}^\top$ one gets $\boldsymbol{X}^\top(\mathsf{y}-\boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}}) = \boldsymbol{R}^\top\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{u})$, therefore $\mathcal{V}[\boldsymbol{X}^\top(\mathsf{y}-\boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}})] = \boldsymbol{R}^\top\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}\boldsymbol{R}$; and now one can see that $\frac{1}{\sigma^2}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$ is a g-inverse of this covariance matrix. Therefore the Malahalanobis distance of the observed value from the distribution is

$$\frac{1}{\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}})^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{y}-\boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}}) \tag{42.1.7}$$

The Lagrange multiplier statistic is based on the restricted estimator alone. If one wanted to take this principle seriously one would have to to replace σ^2 by the unbiased estimate from the restricted model to get the "score form" of the Lagrange Multiplier Test statistic. But in the linear model this leads to it that the denominator in the test statistic is no longer independent of the numerator, and since the test statistic as a function of the ratio of the constrained and unconstrained estimates of σ^2 anyway, one will only get yet another monotonic transformation of the same test statistic. If one were to use the unbiased estimate from the unrestricted model, one would exactly get the Wald statistic back, as one can verify using (29.3.13).

This same statistic can also be motivated in terms of the Lagrange multipliers, and this is where this testing principle has its name from, although the applications usually use the score form. According to (29.3.12), the Lagrange multiplier is $\boldsymbol{\lambda} = 2\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{u})$. If the constraint holds, then $\mathcal{E}[\boldsymbol{\lambda}]=\boldsymbol{o}$, and $\mathcal{V}[\boldsymbol{\lambda}] = 4\sigma^2\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}$. The Mahalanobis distance of the observed value from this distribution is

$$\boldsymbol{\lambda}^\top(\mathcal{V}[\boldsymbol{\lambda}])^{-1}\boldsymbol{\lambda} = \frac{1}{4\sigma^2}\boldsymbol{\lambda}^\top\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\boldsymbol{\lambda} \tag{42.1.8}$$

Using (29.7.1) one can verify that this is the same as (42.1.7).

Problem 422. *Show that* (42.1.7) *is equal to the righthand side of* (42.1.8).

Problem 423. *10 points Prove that* $\hat{\hat{\boldsymbol{\varepsilon}}}^\top\hat{\hat{\boldsymbol{\varepsilon}}}-\hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}}$ *can be written alternatively in the following five ways:*

$$\hat{\hat{\boldsymbol{\varepsilon}}}^\top\hat{\hat{\boldsymbol{\varepsilon}}}-\hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}} = (\hat{\boldsymbol{\beta}}-\hat{\hat{\boldsymbol{\beta}}})^\top\boldsymbol{X}^\top\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\hat{\hat{\boldsymbol{\beta}}}) \tag{42.1.9}$$

$$= (\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{u})^\top(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{u}) \tag{42.1.10}$$

$$= \frac{1}{4}\boldsymbol{\lambda}^\top\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\boldsymbol{\lambda} \tag{42.1.11}$$

$$= \hat{\hat{\boldsymbol{\varepsilon}}}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\hat{\hat{\boldsymbol{\varepsilon}}} \tag{42.1.12}$$

$$= (\hat{\hat{\boldsymbol{\varepsilon}}}-\hat{\boldsymbol{\varepsilon}})^\top(\hat{\hat{\boldsymbol{\varepsilon}}}-\hat{\boldsymbol{\varepsilon}}) \tag{42.1.13}$$

Furthermore show that

$$\boldsymbol{X}^\top\boldsymbol{X} \quad \textit{is } \sigma^2 \textit{ times a g-inverse of} \quad \mathcal{V}[\hat{\boldsymbol{\beta}}-\hat{\hat{\boldsymbol{\beta}}}] \tag{42.1.14}$$

$$(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top)^{-1} \quad \textit{is } \sigma^2 \textit{ times the inverse of} \quad \mathcal{V}[\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{u}] \tag{42.1.15}$$

$$\frac{1}{4}\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top \quad \textit{is } \sigma^2 \textit{ times the inverse of} \quad \mathcal{V}[\boldsymbol{\lambda}] \tag{42.1.16}$$

$$(\boldsymbol{X}^\top\boldsymbol{X})^{-1} \quad \textit{is } \sigma^2 \textit{ times a g-inverse of} \quad \mathcal{V}[\boldsymbol{X}^\top(\mathsf{y}-\boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}})] \tag{42.1.17}$$

$$\boldsymbol{I} \quad \textit{is } \sigma^2 \textit{ times a g-inverse of} \quad \mathcal{V}[\hat{\hat{\boldsymbol{\varepsilon}}}-\hat{\boldsymbol{\varepsilon}}] \tag{42.1.18}$$

and show that $-2\boldsymbol{X}^\top(\mathsf{y}-\boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}})$ *is the gradient of the* SSE *objective function evaluated at* $\hat{\hat{\boldsymbol{\beta}}}$. *By the way, one should be a little careful in interpreting* (42.1.12) *because* $\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$ *is not* σ^2 *times the g-inverse of* $\mathcal{V}[\hat{\hat{\boldsymbol{\varepsilon}}}]$.

Answer.

$$\hat{\hat{\boldsymbol{\varepsilon}}} = \boldsymbol{y}-\boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}+\hat{\boldsymbol{\varepsilon}}-\boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}} = \boldsymbol{X}(\hat{\boldsymbol{\beta}}-\hat{\hat{\boldsymbol{\beta}}})+\hat{\boldsymbol{\varepsilon}}, \tag{42.1.19}$$

and since $\boldsymbol{X}^\top\hat{\boldsymbol{\varepsilon}}=\boldsymbol{o}$, the righthand decomposition is an orthogonal decomposition. This gives (42.1.9) above:

$$\hat{\hat{\boldsymbol{\varepsilon}}}^\top\hat{\hat{\boldsymbol{\varepsilon}}} = (\hat{\boldsymbol{\beta}}-\hat{\hat{\boldsymbol{\beta}}})^\top\boldsymbol{X}^\top\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\hat{\hat{\boldsymbol{\beta}}})+\hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}}, \tag{42.1.20}$$

Using (29.3.13) one obtains $\mathcal{V}[\hat{\boldsymbol{\beta}}-\hat{\hat{\boldsymbol{\beta}}}] = \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$. This is a singular matrix, and one verifies immediately that $\frac{1}{\sigma^2}\boldsymbol{X}^\top\boldsymbol{X}$ is a g-inverse of it.

To obtain (42.1.10), which is (29.7.2), one has to plug (29.3.13) into (42.1.20). Clearly, $\mathcal{V}[\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{u}] = \sigma^2\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top$.

For (42.1.11) one needs the formula for the Lagrange multiplier (29.3.12).

□

The test statistic defined alternatively either by (42.1.1) or (42.1.4) or (42.1.7) or (42.1.8) has the following nice properties:

- $\mathrm{E}(\mathsf{SSE}_u) = \mathrm{E}(\hat{\varepsilon}^\top \hat{\varepsilon}) = \sigma^2(n-k)$, which holds whether or not the constraint is true. Furthermore it was shown earlier that

$$\mathrm{E}(\mathsf{SSE}_r - \mathsf{SSE}_u) = \sigma^2 i + (\boldsymbol{R\beta} - \boldsymbol{u})^\top (\boldsymbol{R}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{R}^\top)^{-1}(\boldsymbol{R\beta} - \boldsymbol{u}), \tag{42.1.21}$$

  i.e., this expected value is equal to $\sigma^2 i$ if the constraint is true, and larger otherwise. If one divides SSE_u and $\mathsf{SSE}_r - \mathsf{SSE}_u$ by their respective degrees of freedom, as is done in (42.1.4), one obtains therefore: the denominator is always an unbiased estimator of σ^2, regardless of whether the null hypothesis is true or not. The numerator is an unbiased estimator of σ^2 when the null hypothesis is correct, and has a *positive* bias otherwise.
- If the distribution of ε is normal, then numerator and denominator are independent. The numerator is a function of $\hat{\boldsymbol{\beta}}$ and the denominator one of $\hat{\varepsilon}$, and $\hat{\boldsymbol{\beta}}$ and $\hat{\varepsilon}$ are independent.
- Again under assumption of normality, numerator and denominator are distributed as $\sigma^2\chi^2$ with i and $n-k$ degrees of freedom, divided by their respective degrees of freedom. If one divides them, the common factor σ^2 cancels out, and the ratio has a F distribution. Since both numerator and denominator have the same expected value σ^2, the value of this F distribution should be in the order of magnitude of 1. If it is much larger than that, the null hypothesis is to be rejected. (Precise values in the F-tables).

42.2. Examples of Tests of Linear Hypotheses

Some tests can be read off directly from the computer printouts. One example is the t-tests for an individual component of $\boldsymbol{\beta}$. The situation is $\mathsf{y} = \boldsymbol{X\beta} + \varepsilon$, where $\boldsymbol{\beta} = \begin{bmatrix}\beta_1 & \cdots & \beta_k\end{bmatrix}^\top$, and we want to test $\beta_j = u$. Here $\boldsymbol{R} = \boldsymbol{e}_j = \begin{bmatrix}0 & \cdots & 0 & 1 & 0 & \cdots & 0\end{bmatrix}$, with the 1 on the jth place, and $\boldsymbol{u}$ is the 1-vector u, and $i = 1$. Therefore $\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top = d_{jj}$, the jth diagonal element of $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$, and (42.1.1) becomes

$$\frac{(\hat{\beta}_j - u)^2}{\mathsf{s}^2 d_{jj}} \sim F_{1,n-k} \quad \text{when } H \text{ is true.} \tag{42.2.1}$$

This is the square of a random variable which has a t-distribution:

$$\frac{\hat{\beta}_j - u}{\mathsf{s}\sqrt{d_{jj}}} \sim t_{n-k} \quad \text{when } H \text{ is true.} \tag{42.2.2}$$

This latter test statistic is simply $\hat{\beta}_j - u$ divided by the estimated standard deviation of $\hat{\beta}_j$.

If one wants to test that a certain linear combination of the parameter values is equal to (or bigger than or smaller than) a given value, say $\boldsymbol{r}^\top\boldsymbol{\beta} = u$, one can use a

t-test as well. The test statistic is, again, simply $\boldsymbol{r}^\top\hat{\boldsymbol{\beta}} - u$ divided by the estimated standard deviation of $\boldsymbol{r}^\top\hat{\boldsymbol{\beta}}$:

$$\frac{\boldsymbol{r}^\top\hat{\boldsymbol{\beta}} - u}{\mathsf{s}\sqrt{\boldsymbol{r}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{r}}} \sim t_{n-k} \quad \text{when } H \text{ is true.} \tag{42.2.3}$$

By this one can for instance also test whether the sum of certain regression coefficients is equal to 1, or whether two regression coefficients are equal to each other (but not the hypothesis that three coefficients are equal to each other).

Many textbooks use the Wald criterion to derive the t-test, and the Likelihood-Ratio criterion to derive the F-test. Our approach showed that the Wald criterion can be used for simultaneous testing of several hypotheses as well. The t-test is equivalent to an F-test if only one hypothesis is tested, i.e., if $\boldsymbol{R}$ is a row vector. The only difference is that with the t-test one can test one-sided hypotheses, with the F-test one cannot.

Next let us discuss the test for the existence of a relationship, "the" F-test which every statistics package performs automatically whenever the regression has a constant term: it is the test whether all the slope parameters are zero, such that only the intercept may take a nonzero value.

Problem 424. *4 points In the model $\mathsf{y} = \boldsymbol{X\beta} + \varepsilon$ with intercept, show that the test statistic for testing whether all the slope parameters are zero is*

$$\frac{(\mathsf{y}^\top\boldsymbol{X}\hat{\boldsymbol{\beta}} - n\bar{\mathsf{y}}^2)/(k-1)}{(\mathsf{y}^\top\mathsf{y} - \mathsf{y}^\top\boldsymbol{X}\hat{\boldsymbol{\beta}})/(n-k)} \tag{42.2.4}$$

This is [**Seb77**, equation (4.26) on p. 110]. *What is the distribution of this test statistic if the null hypothesis is true (i.e., if all the slope parameters are zero)?*

Answer. The distribution is $\sim \mathsf{F}_{k-1,n-k}$. (42.2.4) is most conveniently derived from (42.1.4). In the constrained model, which has only a constant term and no other explanatory variables, i.e., $\mathsf{y} = \iota\mu + \varepsilon$, the BLUE is $\hat{\mu} = \bar{\mathsf{y}}$. Therefore the constrained residual sum of squares $\mathsf{SSE}_{\text{const.}}$ is what is commonly called SST ("total" or, more precisely, "corrected total" sum of squares):

$$\mathsf{SSE}_{\text{const.}} = \mathsf{SST} = (\mathsf{y} - \iota\bar{\mathsf{y}})^\top(\mathsf{y} - \iota\bar{\mathsf{y}}) = \mathsf{y}^\top(\mathsf{y} - \iota\bar{\mathsf{y}}) = \mathsf{y}^\top\mathsf{y} - n\bar{\mathsf{y}}^2 \tag{42.2.5}$$

while the unconstrained residual sum of squares is what is usually called SSE:

$$\mathsf{SSE}_{\text{unconst.}} = \mathsf{SSE} = (\mathsf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\mathsf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \mathsf{y}^\top(\mathsf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \mathsf{y}^\top\mathsf{y} - \mathsf{y}^\top\boldsymbol{X}\hat{\boldsymbol{\beta}}. \tag{42.2.6}$$

This last equation because $\boldsymbol{X}^\top(\mathsf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \boldsymbol{X}^\top\hat{\varepsilon} = \boldsymbol{o}$. A more elegant way is perhaps

$$\mathsf{SSE}_{\text{unconst.}} = \mathsf{SSE} = \hat{\varepsilon}^\top\hat{\varepsilon} = \mathsf{y}^\top\boldsymbol{M}^\top\boldsymbol{M}\mathsf{y} = \mathsf{y}^\top\boldsymbol{M}\mathsf{y} = \mathsf{y}^\top\mathsf{y} - \mathsf{y}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\mathsf{y} = \mathsf{y}^\top\mathsf{y} - \mathsf{y}^\top\boldsymbol{X}\hat{\boldsymbol{\beta}} \tag{42.2.7}$$

According to (18.3.12) we can write $\mathsf{SSR} = \mathsf{SST} - \mathsf{SSE}$, therefore the F-statistic is

$$\frac{\mathsf{SSR}/(k-1)}{\mathsf{SSE}/(n-k)} = \frac{(\mathbf{y}^\top \boldsymbol{X}\hat{\boldsymbol{\beta}} - n\bar{\mathbf{y}}^2)/(k-1)}{(\mathbf{y}^\top \mathbf{y} - \mathbf{y}^\top \boldsymbol{X}\hat{\boldsymbol{\beta}})/(n-k)} \sim \mathsf{F}_{k-1,n-k} \quad \text{if } H_0 \text{ is true.} \tag{42.2.8}$$

□

PROBLEM 425. *2 points Can one compute the value of the* F*-statistic testing for the existence of a relationship if one only knows the coefficient of determination* $R^2 = \mathsf{SSR}/\mathsf{SST}$*, the number of observations* n*, and the number of regressors (counting the constant term as one of the regressors)* k*?*

ANSWER.

$$F = \frac{\mathsf{SSR}/(k-1)}{\mathsf{SSE}/(n-k)} = \frac{n-k}{k-1}\frac{\mathsf{SSR}}{\mathsf{SST}-\mathsf{SSR}} = \frac{n-k}{k-1}\frac{R^2}{1-R^2}. \tag{42.2.9}$$

□

Other, similar F-tests are: the F-test that all among a number of additional variables have the coefficient zero, the F-test that three or more coefficients are equal. One can use the t-test for testing whether two coefficients are equal, but not for three. It may be possible that the t-test for $\beta_1 = \beta_2$ does not reject and the t-test for $\beta_2 = \beta_3$ does not reject either, but the t-test for $\beta_1 = \beta_3$ does reject!

PROBLEM 426. *4 points* [**Seb77**, exercise 4b.5 on p. 109/10] *In the model* $\mathbf{y} = \boldsymbol{\beta} + \boldsymbol{\varepsilon}$ *with* $\boldsymbol{\varepsilon} \sim \mathsf{N}(\boldsymbol{o}, \sigma^2\boldsymbol{I})$ *and subject to the constraint* $\boldsymbol{\iota}^\top\boldsymbol{\beta} = 0$*, which we had in Problem 348, compute the test statistic for the hypothesis* $\beta_1 = \beta_3$*.*

ANSWER. In this problem, the "unconstrained" model for the purposes of testing is already constrained, it is subject to the constraint $\boldsymbol{\iota}^\top\boldsymbol{\beta} = 0$. The "constrained" model has the additional constraint $\boldsymbol{R}\boldsymbol{\beta} = \begin{bmatrix} 1 & 0 & -1 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} = 0$. In Problem 348 we computed the "unconstrained" estimates $\hat{\boldsymbol{\beta}} = \mathbf{y} - \boldsymbol{\iota}\bar{\mathbf{y}}$ and $\mathsf{s}^2 = n\bar{\mathbf{y}}^2 = (\mathsf{y}_1 + \cdots + \mathsf{y}_n)^2/n$. You are allowed to use this without proving it again. Therefore $\boldsymbol{R}\hat{\boldsymbol{\beta}} = \mathsf{y}_1 - \mathsf{y}_3$; its variance is $2\sigma^2$, and the F test statistic is $\frac{n(\mathsf{y}_1-\mathsf{y}_3)^2}{2(\mathsf{y}_1+\cdots+\mathsf{y}_n)^2} \sim F_{1,1}$. The "unconstrained" model had 4 parameters subject to one constraint, therefore it had 3 free parameters, i.e.,$k = 3$, $n = 4$, and $j = 1$. □

Another important F-test is the "Chow test" named by its popularizer Chow [**Cho60**]: it tests whether two regressions have equal coefficients (assuming that the disturbance variances are equal). For this one has to run three regressions. If the first regression has n_1 observations and sum of squared error SSE_1, and the second regression n_2 observations and SSE_2, and the combined regression (i.e., the restricted model) has SSE_r, then the test statistic is

$$\frac{(\mathsf{SSE}_r - \mathsf{SSE}_1 - \mathsf{SSE}_2)/k}{(\mathsf{SSE}_1 + \mathsf{SSE}_2)/(n_1 + n_2 - 2k)}. \tag{42.2.10}$$

If $n_2 < k$, the second regression cannot be run by itself. In this case, the unconstrained model has "too many" parameters: they give an exact fit for the second group of observations $\mathsf{SSE}_2 = 0$, and in addition not all parameters are identified. In effect this second regression has only n_2 parameters. These parameters can be considered dummy variables for every observation, i.e., this test can be interpreted to be a test whether the n_2 additional observations come from the same population as the n_1 first ones. The test statistic becomes

$$\frac{(\mathsf{SSE}_r - \mathsf{SSE}_1)/n_2}{\mathsf{SSE}_1/(n_1 - k)}. \tag{42.2.11}$$

This latter is called the "predictive Chow test," because in its Wald version it looks at the prediction errors involving observations in the second regression.

The following is a special case of the Chow test, in which one can give a simple formula for the test statistic.

PROBLEM 427. *Assume you have* n_1 *observations* $\mathsf{u}_j \sim \mathsf{N}(\mu_1, \sigma^2)$ *and* n_2 *observations* $\mathsf{v}_j \sim \mathsf{N}(\mu_2, \sigma^2)$*, all independent of each other, and you want to test whether* $\mu_1 = \mu_2$*. (Note that the variances are known to be equal).*

- a. *2 points Write the model in the form* $\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$*.*

ANSWER.

$$\begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\iota}_1\mu_1 + \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\iota}_2\mu_2 + \boldsymbol{\varepsilon}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{\iota}_1 & \boldsymbol{o} \\ \boldsymbol{o} & \boldsymbol{\iota}_2 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \end{bmatrix}. \tag{42.2.12}$$

here $\boldsymbol{\iota}_1$ and $\boldsymbol{\iota}_2$ are vectors of ones of appropriate lengths. □

- b. *2 points Compute* $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$ *in this case.*

ANSWER.

$$\boldsymbol{X}^\top\boldsymbol{X} = \begin{bmatrix} \boldsymbol{\iota}_1^\top & \boldsymbol{o}^\top \\ \boldsymbol{o}^\top & \boldsymbol{\iota}_2^\top \end{bmatrix} \begin{bmatrix} \boldsymbol{\iota}_1 & \boldsymbol{o} \\ \boldsymbol{o} & \boldsymbol{\iota}_2 \end{bmatrix} = \begin{bmatrix} n_1 & 0 \\ 0 & n_2 \end{bmatrix} \tag{42.2.13}$$

$$(\boldsymbol{X}^\top\boldsymbol{X})^{-1} = \begin{bmatrix} \frac{1}{n_1} & 0 \\ 0 & \frac{1}{n_2} \end{bmatrix} \tag{42.2.14}$$

□

- c. *2 points Compute* $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\mathbf{y}$ *in this case.*

ANSWER.

$$\boldsymbol{X}^\top\mathbf{y} = \begin{bmatrix} \boldsymbol{\iota}_1^\top & \boldsymbol{o}^\top \\ \boldsymbol{o}^\top & \boldsymbol{\iota}_2^\top \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n_1} \mathsf{u}_i \\ \sum_{j=1}^{n_2} \mathsf{v}_j \end{bmatrix} \tag{42.2.15}$$

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\mathbf{y} = \begin{bmatrix} \frac{1}{n_1} & 0 \\ 0 & \frac{1}{n_2} \end{bmatrix} \begin{bmatrix} \sum_{i=1}^{n_1} \mathsf{u}_i \\ \sum_{j=1}^{n_2} \mathsf{v}_j \end{bmatrix} = \begin{bmatrix} \bar{\mathsf{u}} \\ \bar{\mathsf{v}} \end{bmatrix} \tag{42.2.16}$$

□

• d. *3 points Compute* $\mathsf{SSE} = (\mathbf{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\mathbf{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})$ *and* s^2*, the unbiased estimator of* σ^2*, in this case.*

Answer.

$$\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = \begin{bmatrix}\mathbf{u}\\ \boldsymbol{v}\end{bmatrix} - \begin{bmatrix}\boldsymbol{\iota}_1 & \boldsymbol{o}\\ \boldsymbol{o} & \boldsymbol{\iota}_2\end{bmatrix}\begin{bmatrix}\bar{\mathsf{u}}\\ \bar{\mathsf{v}}\end{bmatrix} = \begin{bmatrix}\mathbf{u} - \boldsymbol{\iota}_1\bar{\mathsf{u}}\\ \boldsymbol{v} - \boldsymbol{\iota}_2\bar{\mathsf{u}}\end{bmatrix} \tag{42.2.17}$$

$$\mathsf{SSE} = \sum_{i=1}^{n_1}(\mathsf{u}_i - \bar{\mathsf{u}})^2 + \sum_{j=1}^{n_2}(\mathsf{v}_j - \bar{\mathsf{v}})^2 \tag{42.2.18}$$

$$\mathsf{s}^2 = \frac{\sum_{i=1}^{n_1}(\mathsf{u}_i - \bar{\mathsf{u}})^2 + \sum_{j=1}^{n_2}(\mathsf{v}_j - \bar{\mathsf{v}})^2}{n_1 + n_2 - 2} \tag{42.2.19}$$

□

• e. *1 point Next, the hypothesis* $\mu_1 = \mu_2$ *must be written in the form* $\boldsymbol{R}\boldsymbol{\beta} = u$*. Since in the present case* $\boldsymbol{R}$ *has just has one row, it should be written as a row-vector* $\boldsymbol{R} = \boldsymbol{r}^\top$*, and since the vector* $\boldsymbol{u}$ *has only one component, it should be written as a scalar* u*, i.e., the hypothesis should be written in the form* $\boldsymbol{r}^\top\boldsymbol{\beta} = u$*. What are* $\boldsymbol{r}$ *and* u *in our case?*

Answer. Since $\boldsymbol{\beta} = \begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}$, the constraint can be written as

$$\begin{bmatrix}1 & -1\end{bmatrix}\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix} = 0 \qquad \text{i.e.,} \qquad \boldsymbol{r} = \begin{bmatrix}1\\ -1\end{bmatrix} \qquad \text{and} \qquad u = 0 \tag{42.2.20}$$

□

• f. *2 points Compute the standard deviation of* $\boldsymbol{r}^\top\hat{\boldsymbol{\beta}}$*.*

Answer. First compute the variance and then take the square root.

$$\operatorname{var}[\boldsymbol{r}^\top\hat{\boldsymbol{\beta}}] = \sigma^2\boldsymbol{r}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{r} = \sigma^2\begin{bmatrix}1 & -1\end{bmatrix}\begin{bmatrix}\frac{1}{n_1} & 0\\ 0 & \frac{1}{n_2}\end{bmatrix}\begin{bmatrix}1\\ -1\end{bmatrix} = \sigma^2\Big(\frac{1}{n_1} + \frac{1}{n_2}\Big) \tag{42.2.21}$$

One can also see this without matrix algebra. $\operatorname{var}[\bar{\mathsf{u}} = \sigma^2\frac{1}{n_1}$, $\operatorname{var}[\bar{\mathsf{v}} = \sigma^2\frac{1}{n_2}$, and since $\bar{\mathsf{u}}$ and $\bar{\mathsf{v}}$ are independent, the variance of the difference is the sum of the variances. □

• g. *2 points Use* (42.2.3) *to derive the formula for the* t*-test.*

Answer. The test statistic is $\bar{\mathsf{u}} - \bar{\mathsf{v}}$ divided by its estimated standard deviation, i.e.,

$$\frac{\bar{\mathsf{u}} - \bar{\mathsf{v}}}{\mathsf{s}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2} \quad \text{when } H \text{ is true.} \tag{42.2.22}$$

□

Problem 428. [**Seb77**, exercise 4d-3] *Given* $n+1$ *observations* y_j *from a* $\mathsf{N}(\mu, \sigma^2)$*. After the first* n *observations, it is suspected that a sudden change in the mean of the distribution occurred, i.e., that* $\mathsf{y}_{n+1} \sim \mathsf{N}(\nu, \sigma^2)$ *with* $\nu \neq \mu$*. We will*

use here three different approaches to derive the same test statistic for testing the hypothesis that the $n+1$*st observation has the same population mean as the previous observations, i.e., that* $\nu = \mu$*, against the two-sided alternative. The formulas for this statistic should be given in terms of the observations* y_i*. It is recommended to use the notation* $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ *and* $\bar{\bar{y}} = \frac{1}{n+1}\sum_{j=1}^{n+1} y_j$*.*

• a. *3 points First you should derive this statistic by testing whether* $\nu - \mu = 0$ *(the "Wald principle"). For this you must compute the BLUE of* $\nu - \mu$ *and its standard deviation and construct the* t *statistic from this.*

Answer. BLUE of μ is $\bar{\mathsf{y}} = \frac{1}{n}\sum_{i=1}^{n}\mathsf{y}_i$, and that of ν is y_{n+1}. BLUE of $\nu - \mu$ is $\bar{\mathsf{y}} - \mathsf{y}_{n+1}$. Because of independence $\operatorname{var}[\bar{\mathsf{y}} - \mathsf{y}_{n+1}] = \operatorname{var}[\bar{\mathsf{y}}] + \operatorname{var}[\mathsf{y}_{n+1}] = \sigma^2((1/n) + 1) = \sigma^2(n+1)/n$. Standard deviation is $\sigma\sqrt{(n+1)/n}$.

For the denominator in the t-statistic you need the s^2 from the unconstrained regression, which is

$$\mathsf{s}^2 = \frac{1}{n-1}\sum_{j=1}^{n}(\mathsf{y}_j - \bar{\mathsf{y}})^2 \tag{42.2.23}$$

What happened to the $(n+1)$st observation here? It always has a zero residual. And the factor $1/(n-1)$ should really be written $1/(n+1-2)$: there are $n+1$ observations and 2 parameters.

Divide $\bar{\mathsf{y}} - \mathsf{y}_{n+1}$ by its standard deviation and replace σ by s (the square root of s^2) to get the t statistic

$$\frac{\bar{\mathsf{y}} - \mathsf{y}_{n+1}}{\mathsf{s}\sqrt{1 + \frac{1}{n}}} \tag{42.2.24}$$

□

• b. *2 points One can interpret this same formula also differently (and this is why this test is sometimes called the "predictive" Chow test). Compute the Best Linear Unbiased Predictor of* y_{n+1} *on the basis of the first* n *observations, call it* $\hat{\hat{\mathsf{y}}}(n+1)_{n+1}$*. Show that the predictive residual* $\mathsf{y}_{n+1} - \hat{\hat{\mathsf{y}}}(n+1)_{n+1}$*, divided by the square root of* $\operatorname{MSE}[\hat{\hat{\mathsf{y}}}(n+1)_{n+1}; \mathsf{y}_{n+1}]$*, with* σ *replaced by* s *(based on the first* n *observations only), is equal to the above* t *statistic.*

Answer. BLUP of y_{n+1} based on first n observations is $\bar{\mathsf{y}}$ again. Since it is unbiased, $\operatorname{MSE}[\bar{\mathsf{y}}; \mathsf{y}_{n+1}] = \operatorname{var}[\bar{\mathsf{y}} - \mathsf{y}_{n+1}] = \sigma^2(n+1)/n$. From now on everything is as in part a. □

• c. *6 points Next you should show that the above two formulas are identical to the statistic based on comparing the* SSE*s of the constrained and unconstrained models (the likelihood ratio principle). Give a formula for the constrained* SSE_r*, the unconstrained* SSE_u*, and the* F*-statistic.*

Answer. According to the Likelihood Ratio principle, one has to compare the residual sums of squares in the regressions under the assumption that the mean did not change with that under the assumption that the mean changed. If the mean did not change (constrained model), then $\bar{\bar{y}}$ is the

OLS of μ. In order to make it easier to derive the difference between constrained and unconstrained SSE, we will write the constrained SSE as follows:

$$\mathsf{SSE}_r = \sum_{j=1}^{n+1}(y_j - \bar{y})^2 = \sum_{j=1}^{n+1} y_j^2 - (n+1)\bar{y}^2 = \sum_{j=1}^{n+1} y_j^2 - \frac{1}{n+1}(n\bar{y} + y_{n+1})^2$$

If one allows the mean to change (unconstrained model), then $\bar{y}$ is the BLUE of μ, and y_{n+1} is the BLUE of ν.

$$\mathsf{SSE}_u = \sum_{j=1}^{n}(y_j - \bar{y})^2 + (y_{n+1} - y_{n+1})^2 = \sum_{j=1}^{n} y_j^2 - n\bar{y}^2.$$

Now subtract:

$$\begin{aligned}\mathsf{SSE}_r - \mathsf{SSE}_u &= y_{n+1}^2 + n\bar{y}^2 - \frac{1}{n+1}(n\bar{y} + y_{n+1})^2 \\ &= y_{n+1}^2 + n\bar{y}^2 - \frac{1}{n+1}(n^2\bar{y}^2 + 2n\bar{y}y_{n+1} + y_{n+1}^2) \\ &= (1 - \frac{1}{n+1})y_{n+1}^2 + (n - \frac{n^2}{n+1})\bar{y}^2 - \frac{n}{n+1}2\bar{y}y_{n+1} \\ &= \frac{n}{n+1}(y_{n+1} - \bar{y})^2.\end{aligned}$$

Interestingly, this depends on the first n observations only through $\bar{y}$.

Since the unconstrained model has $n+1$ observations and 2 parameters, the test statistic is

$$\frac{\mathsf{SSE}_r - \mathsf{SSE}_u}{\mathsf{SSE}_u/(n+1-2)} = \frac{\frac{n}{n+1}(y_{n+1} - \bar{y})^2}{\sum_1^n (y_j - \bar{y})^2/(n-1)} = \frac{(y_{n+1} - \bar{y})^2\, n(n-1)}{\sum_1^n (y_j - \bar{y})^2\,(n+1)} \sim F_{1,n-1} \tag{42.2.25}$$

This is the square of the t statistic (42.2.24). □

42.2.1. Goodness of Fit Test.

PROBLEM 429. [**Seb77**, pp. 117–119] *Given a regression model with k independent variables. There are n observations of the vector of independent variables, and for each of these n values there is not one but $r > 1$ different replicated observations of the dependent variable. This model can be written*

$$\mathsf{y}_{mq} = \sum_{j=1}^{k} x_{mj}\beta_j + \varepsilon_{mq} \qquad \text{or} \quad \mathsf{y}_{mq} = \boldsymbol{x}_m^\top \boldsymbol{\beta} + \varepsilon_{mq}, \tag{42.2.26}$$

where $m = 1, \ldots, n$, $j = 1, \ldots, k$, $q = 1, \ldots, r$, and $\boldsymbol{x}_m^\top$ is the mth row *of the $\boldsymbol{X}$-matrix. For simplicity we assume that r does not depend on m, each observation of the independent variables has the same number of repetitions. We also assume that the $n \times k$ matrix $\boldsymbol{X}$ has full column rank.*

• a. *2 points In this model it is possible to test whether the regression line is in fact a straight line. If it is not a straight line, then each observation of the dependent*

variables $\boldsymbol{x}_m$ has a different coefficient vector $\boldsymbol{\beta}_m$ associated with it, i.e., the model is

$$\mathsf{y}_{mq} = \sum_{j=1}^{k} x_{mj}\beta_{mj} + \varepsilon_{mq} \qquad \text{or} \quad \mathsf{y}_{mq} = \boldsymbol{x}_m^\top \boldsymbol{\beta}_m + \varepsilon_{mq}. \tag{42.2.27}$$

This unconstrained model does not have enough information to estimate any of the individual coefficients β_{mj}. Explain how it is nevertheless still possible to compute SSE_u.

ANSWER. Even though the individual coefficients β_{mj} are not identified, their linear combination $\eta_m = \boldsymbol{x}_m^\top \boldsymbol{\beta}_m = \sum_{j=1}^k x_{mj}\beta_{mj}$ is identified; one unbiased estimator, although by far not the best one, is any individual observation y_{mq}. This linear combination is all one needs to compute SSE_u, the sum of squared errors in the unconstrained model. □

• b. *2 points Writing your estimate of $\eta_m = \boldsymbol{x}_m^\top \boldsymbol{\beta}_m$ as $\tilde{\eta}_m$, give the formula of the sum of squared errors of this estimate, and by taking the first order conditions, show that the unconstrained least squares estimate of η_m is $\hat{\eta}_m = \bar{\mathsf{y}}_{m\cdot}$ for $m = 1, \ldots, n$, where $\bar{\mathsf{y}}_{m\cdot} = \frac{1}{r}\sum_{q=1}^r \mathsf{y}_{mq}$ (i.e., the dot in the subscript indicates taking the mean).*

ANSWER. If we know the $\tilde{\eta}_m$ the sum of squared errors no longer depents on the independent observations $\boldsymbol{x}_m$ but is simply

$$\mathsf{SSE}_u = \sum_{m,q}(\mathsf{y}_{mq} - \tilde{\eta}_m)^2 \tag{42.2.28}$$

First order conditions are

$$\frac{\partial}{\partial \tilde{\eta}_h}\sum_{m,q}(\mathsf{y}_{mq} - \tilde{\eta}_m)^2 = \frac{\partial}{\partial \tilde{\eta}_h}\sum_{q}(\mathsf{y}_{hq} - \tilde{\eta}_h)^2 = -2\sum_{q}(\mathsf{y}_{hq} - \tilde{\eta}_h) = 0 \tag{42.2.29}$$

□

• c. *1 point The sum of squared errors associated with this least squares estimate is the unconstrained sum of squared errors SSE_u. How would you set up a regression with dummy variables which would give you this SSE_u?*

ANSWER. The unconstrained model should be regressed in the form $\mathsf{y}_{mq} = \eta_m + \varepsilon_{mq}$. I.e., string out the matrix $\mathbf{Y}$ as a vector and for each column of $\mathbf{Y}$ introduce a dummy variable which is $= 1$ if the given observation was originally in this colum. □

• d. *2 points Next turn to the constrained model* (42.2.26). *If $\boldsymbol{X}$ has full column rank, then it is fully identified. Writing $\tilde{\beta}_j$ for your estimates of β_j, give a formula for the sum of squared errors of this estimate. By taking the first order conditions, show that the estimate $\hat{\boldsymbol{\beta}}$ is the same as the estimate in the model without replicated observations*

$$z_m = \sum_{j=1}^{k} x_{mj}\beta_j + \varepsilon_m, \tag{42.2.30}$$

where $\mathsf{z}_m = \bar{\mathsf{y}}_{m\cdot}$ *as defined above.*

• e. *2 points If* SSE_c *is the* SSE *in the constrained model* (42.2.26) *and* SSE_b *the* SSE *in* (42.2.30)*, show that* $\mathsf{SSE}_c = r \cdot \mathsf{SSE}_b + \mathsf{SSE}_u$.

Answer. For every m we have $\sum_q(\mathsf{y}_{mq} - \boldsymbol{x}_m^\top\hat{\boldsymbol{\beta}})^2 = \sum_q(\mathsf{y}_{mq} - \bar{\mathsf{y}}_{m\cdot})^2 + r(\mathsf{y}_{m\cdot} - \boldsymbol{x}_m^\top\hat{\boldsymbol{\beta}})^2$; therefore $\mathsf{SSE}_c = \sum_{m,q}(\mathsf{y}_{mq} - \bar{\mathsf{y}}_{m\cdot})^2 + r\sum_m(\mathsf{y}_{m\cdot} - \boldsymbol{x}_m^\top\hat{\boldsymbol{\beta}})^2$; □

• f. *3 points Write down the formula of the* F*-test in terms of* SSE_u *and* SSE_c *with a correct accounting of the degrees of freedom, and give this formula also in terms of* SSE_u *and* SSE_b.

Answer. Unconstrained model has n parameters, and constrained model has k parameters; the number of additional "constraints" is therefore $n-k$. This gives the F-statistic

$$\frac{(\mathsf{SSE}_c - \mathsf{SSE}_u)/(n-k)}{\mathsf{SSE}_u/n(r-1)} = \frac{r\mathsf{SSE}_b/(n-k)}{\mathsf{SSE}_u/n(r-1)} \tag{42.2.31}$$

□

42.3. The F-Test Statistic is a Function of the Likelihood Ratio

Problem 430. *The critical region of the generalized likelihood ratio test can be written as*

$$C = \{y_1, \dots, y_n\colon \frac{\sup_{\boldsymbol{\theta}\in\Omega}\ell(y_1,\dots,y_n;\theta_1,\dots,\theta_k)}{\sup_{\boldsymbol{\theta}\in\omega}\ell(y_1,\dots,y_n;\theta_1,\dots,\theta_k)} \geq k\}, \tag{42.3.1}$$

where ω *refers to the null and* Ω *to the alternative hypothesis (it is assumed that the hypotheses are nested, i.e.,* $\omega \subset \Omega$*). In other words, one rejects the hypothesis if the maximal achievable likelihood level with the restriction imposed is much lower than that without the restriction. If* $\hat{\boldsymbol{\theta}}$ *is the unrestricted and* $\hat{\hat{\boldsymbol{\theta}}}$ *the restricted maximum likelihood estimator, then the test statistic is*

$$LR = 2(\log\ell(\mathbf{y},\hat{\boldsymbol{\theta}}) - \log\ell(\mathbf{y},\hat{\hat{\boldsymbol{\theta}}})) \to \chi^2_i \tag{42.3.2}$$

where i *is the number of restrictions. In this exercise we are proving that the* F*-test in the linear model is equivalent to the generalized likelihood ratio test. (You should assume here that both* $\boldsymbol{\beta}$ *and* σ^2 *are unknown.) All this is in* [**Gre97**, p. 304].

• a. *1 point Since we only have constraints on* $\boldsymbol{\beta}$ *and not on* σ^2*, it makes sense to first compute the concentrated likelihood function with* σ^2 *concentrated out. Derive the formula for this concentrated likelihood function which is given in* [**Gre97**, just above (6.88)].

Answer.

$$\text{Concentrated } \log\ell(\mathbf{y};\boldsymbol{\beta}) = -\frac{n}{2}\Big(1 + \log 2\pi + \log\frac{1}{n}(\mathbf{y}-\boldsymbol{X}\boldsymbol{\beta})^\top(\mathbf{y}-\boldsymbol{X}\boldsymbol{\beta})\Big) \tag{42.3.3}$$

□

• b. *2 points In the case of a linear restriction, show that* LR *is connected with the* F*-statistic* F *as follows:*

$$LR = n\log\Big(1 + \frac{i}{n-k}F\Big) \tag{42.3.4}$$

Answer. $LR = -n\Big(\log\frac{1}{n}\hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}} - \log\frac{1}{n}\hat{\hat{\boldsymbol{\varepsilon}}}^\top\hat{\hat{\boldsymbol{\varepsilon}}}\Big) = n\log\frac{\hat{\hat{\boldsymbol{\varepsilon}}}^\top\hat{\hat{\boldsymbol{\varepsilon}}}}{\hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}}}$ In order to connect this with the F statistic note that

$$F = \frac{n-k}{i}\Big(\frac{\hat{\hat{\boldsymbol{\varepsilon}}}^\top\hat{\hat{\boldsymbol{\varepsilon}}}}{\hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}}} - 1\Big) \tag{42.3.5}$$

□

42.4. Tests of Nonlinear Hypotheses

Make linear approximation, need Jacobian for this. Here is an example where a nonlinear hypothesis arises naturally:

Problem 431. [**Gre97**, Example 7.14 on p. 361]*: The model*

$$C_t = \alpha + \beta Y_t + \gamma C_{t-1} + \varepsilon_t \tag{42.4.1}$$

has different long run and short run propensities to consume. Give formulas for both.

Answer. Short-run is β; to compute the long run propensity, which would prevail in the stationary state when $C_t = C_{t-1}$, write $C_\infty = \alpha + \beta Y_\infty + \gamma C_\infty + \varepsilon_\infty$ or $C_\infty(1-\gamma) = \alpha + \beta Y_\infty + \varepsilon_\infty$ or $C_\infty = \alpha/(1-\gamma) + \beta/(1-\gamma)Y_\infty + \varepsilon_t/(1-\gamma)$. Therefore long run propensity is $\delta = \beta/(1-\gamma)$. □

42.5. Choosing Between Nonnested Models

Throwing all regressors into the same regression is a straightforward way out but not very good. J-test (the J comes from "joint") is better: throw the predicted values of one of the two models as a regressor into the other model and test whether this predicted value has a nonzero coefficient. Here is more detail: if the null hypothesis is that model 1 is right, then throw the predicted value of model 2 into model 1 and test the null hypothesis that the coefficient of this predicted value is zero. If Model 1 is right, then this additional regressor leaves all other estimators unbiased, and the true coefficient of the additional regressor is 0. If Model 2 is right, then asymptotically, this additional regressor should be the only regressor in the combined model with a nonzero coefficient (its coefficient is = 1 asymptotically, and all the other regressors should have coefficient zero.) Whenever nonnested hypotheses are tested, is is possible that both hypotheses are rejected, or that neither hypothesis is rejected by this criterion.

CHAPTER 43

Multiple Comparisons in the Linear Model

Due to the isomorphism of tests and confidence intervals, we will keep this whole discussion in terms of confidence intervals.

43.1. Rectangular Confidence Regions

Assume you are interested in two linear combinations of $\boldsymbol{\beta}$ at the same time, i.e., you want separate confidence intervals for them. If you use the Cartesian product (or the intersection, depending on how you look at it) of the individual confidence intervals, the confidence level of this rectangular confidence region will of necessity be different that of the individual intervals used to form this region. If you want the joint confidence region to have confidence level 95%, then the individual confidence intervals must have a confidence level higher than 95%, i.e., they must be be wider.

There are two main approaches for compute the confidence levels of the individual intervals, one very simple one which is widely applicable but which is only approximate, and one more specialized one which is precise in some situations and can be taken as an approximation in others.

43.1.1. Bonferroni Intervals. To derive the first method, the Bonferroni intervals, assume you have individual confidence intervals R_i for parameter ϕ_i. In order to make simultaneous inferences about the whole parameter vector $\boldsymbol{\phi} = \begin{bmatrix} \phi_1 \\ \vdots \\ \phi_i \end{bmatrix}$ you take the Cartesian product $R_1 \times R_2 \times \cdots \times R_i$; it is defined by $\begin{bmatrix} \phi_1 \\ \vdots \\ \phi_i \end{bmatrix} \in R_1 \times R_2 \times \cdots \times R_i$ if and only if $\phi_i \in R_i$ for all i.

Usually it is difficult to compute the precise confidence level of such a rectangular set. If one cannot be precise, it is safer to understate the confidence level. The following inequality from elementary probability theory, called the Bonferroni inequality, gives a lower bound for the confidence level of this Cartesian product: Given i events E_i with $\Pr[E_i] = 1 - \alpha_i$; then $\Pr[\bigcap E_i] \geq 1 - \sum \alpha_i$. Proof:

$\Pr[\bigcap E_i] = 1 - \Pr[\bigcup E_i'] \geq 1 - \sum \Pr[E_i']$. The so-called Bonferroni bounds therefore have the individual levels $1 - \alpha/i$. Instead of $\gamma_j = \alpha/i$ one can also take any other $\gamma_i \geq 0$ with $\sum \gamma_i = \alpha$. For small α and small i this is an amazingly precise method.

43.1.2. The Multivariate t Distribution. Let $\boldsymbol{z} \sim \mathsf{N}(\boldsymbol{o}, \sigma^2 \boldsymbol{\Psi})$ where $\boldsymbol{\Psi}$ is positive definite and has ones in the diagonal:

$$\boldsymbol{\Psi} = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \cdots & \rho_{1i} \\ \rho_{12} & 1 & \rho_{23} & \cdots & \rho_{2i} \\ \rho_{13} & \rho_{23} & 1 & \cdots & \rho_{3i} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{1i} & \rho_{2i} & \rho_{3i} & \cdots & 1 \end{bmatrix} \tag{43.1.1}$$

Let $\mathsf{s}^2 \sim \frac{\sigma^2}{\nu}\chi^2_\nu$ be independent of $\boldsymbol{z}$. Then $\mathbf{t} = \boldsymbol{z}/\mathsf{s}$ has a multivariate t distribution with ν degrees of freedom. This is clearly what one needs for simultaneous t intervals, since this is the joint distribution of the statistics used to construct t intervals. Each t_i has a t distribution. For certain special cases of $\boldsymbol{\Psi}$, certain quantiles of this joint distribution have been calculated and tabulated. This allows to compute the precise confidence levels of multiple t intervals in certain situations.

Problem 432. *Show that the correlation coefficient between t_i and t_j is ρ_{ij} But give a verbal argument that the t_i are not independent, even if the $\rho_{ij} = 0$, i.e. z_i are independent. (This means, one cannot get the quantiles of their maxima from individual quantiles.)*

Answer. First we have $\mathrm{E}[\mathsf{t}_j] = \mathrm{E}[\mathsf{z}_j]\,\mathrm{E}[\frac{1}{\mathsf{s}}] = 0$, since z_j and s are independent. Therefore (43.1.2)

$$\mathrm{cov}[\mathsf{t}_i, \mathsf{t}_j] = \mathrm{E}[\mathsf{t}_i\mathsf{t}_j] = \mathrm{E}[\mathrm{E}[\mathsf{t}_i\mathsf{t}_j]|\mathsf{s}] = \mathrm{E}[\mathrm{E}[\frac{1}{\mathsf{s}^2}\mathsf{z}_i\mathsf{z}_j|\mathsf{s}]] = \mathrm{E}[\frac{1}{\mathsf{s}^2}\mathrm{E}[\mathsf{z}_i\mathsf{z}_j|\mathsf{s}]] = \mathrm{E}[\frac{1}{\mathsf{s}^2}\,\mathrm{E}[\mathsf{z}_i\mathsf{z}_j]] = \mathrm{E}[\frac{\sigma^2}{\mathsf{s}^2}]\rho_{ij}.$$

In particular, $\mathrm{var}[\mathsf{t}_i] = \mathrm{E}[\frac{\sigma^2}{\mathsf{s}^2}]$, and the statement follows. □

43.1.3. Studentized Maximum Modulus and Related Intervals. Look at the special case where all ρ_{ij} are equal, call them ρ. Then the following quantiles have been tabulated by [**HH71**], and reprinted in [**Seb77**, pp. 404–410], where they are called $u^\alpha_{i,\nu,\rho}$:

$$\Pr[\big(\max_{j=1,\ldots,i} |\mathsf{t}_i|\big) \leq u^\alpha_{i,\nu,\rho}] = 1 - \alpha, \tag{43.1.3}$$

where $\mathbf{t} = \begin{bmatrix} \mathsf{t}_1 \\ \vdots \\ \mathsf{t}_i \end{bmatrix}$ is a multivariate equicorrelated t with ν degrees of freedom and correlation coefficient ρ.

If one needs only *two* joint confidence intervals, i.e., if $i = 2$, then there are only two off-diagonal elements in the dispersion matrix, which must be equal by symmetry. A 2×2 dispersion matrix is therefore always "equicorrelated." The values of the $u^{\alpha}_{2,n-k,\rho}$ can therefore be used to compute simultaneous confidence intervals for any *two* parameters in the regression model. For ρ one must use the actual correlation coefficient between the OLS estimates of the respective parameters, which is known precisely.

PROBLEM 433. *In the model* $\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, *with* $\boldsymbol{\varepsilon} \sim (\boldsymbol{o}, \sigma^2\boldsymbol{I})$, *give a formula for the correlation coefficient between* $\boldsymbol{g}^\top\hat{\boldsymbol{\beta}}$ *and* $\boldsymbol{h}^\top\hat{\boldsymbol{\beta}}$, *where* $\boldsymbol{g}$ *and* $\boldsymbol{h}$ *are arbitrary constant vectors.*

ANSWER. This is in Seber, [**Seb77**, equation (5.7) on p. 128].

$$\rho = \boldsymbol{g}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{h}/\sqrt{(\boldsymbol{g}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{g})(\boldsymbol{h}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{h})} \tag{43.1.4}$$

□

But in certain situations, those equicorrelated quantiles can also be applied for testing more than two parameters. The most basic situation in which this is the case is the following: you have $n \times m$ observations $\mathsf{y}_{ij} = \mu_i + \varepsilon_{ij}$, and the $\boldsymbol{\varepsilon}_{ij} \sim \mathsf{NID}(0, \sigma^2)$. Then the equicorrelated t quantiles allow you to compute precise joint confidence intervals for all μ_i. Define $\mathsf{s}^2 = \sum_{i,j}(\mathsf{y}_{ij} - \bar{\mathsf{y}}_{i\cdot})^2/(n(m-1))$, and define $\boldsymbol{z}$ by $z_i = (\bar{\mathsf{y}}_{i\cdot} - \mu_i)\sqrt{m}$. These z_i are normal with mean zero and dispersion matrix $\sigma^2\boldsymbol{I}$, and they are independent of s^2. Therefore one gets confidence intervals

$$\mu_i \in \bar{\mathsf{y}}_{i\cdot} \pm u^{\alpha}_{n,n(m-1),0}\mathsf{s}/\sqrt{m}. \tag{43.1.5}$$

This simplest example is a special case of "orthogonal regression," in which $\boldsymbol{X}^\top\boldsymbol{X}$ is a diagonal matrix. One can do the same procedure also in other cases of orthogonal regression, such as a regression with orthogonal polynomials as explanatory variables.

Now return to the situation of the basic example, but assume that the first row of the matrix $\mathbf{Y}$ of observations is the reference group, and one wants to know whether the means of the other groups are significantly different than that first group. Give the first row the subscript $i = 0$. Then use $\mathsf{z}_i = (\bar{\mathsf{y}}_{i\cdot} - \bar{\mathsf{y}}_{0\cdot})\sqrt{m}/\sqrt{2}$, $i = 1, \ldots, n$. One obtains again the multivariate t, this time $\rho = 1/2$. Miller calls these intervals "many-one intervals."

PROBLEM 434. *Assume again we are in the situation of our basic example, revert to counting i from 1 to n. Construct simultaneous confidence intervals for the difference between the individual means and the grand mean.*

ANSWER. One uses

$$z_i = \Big(\bar{\mathsf{y}}_{i\cdot} - \bar{\mathsf{y}}_{\cdot\cdot} - (\mu_i - \mu)\Big)\Big/\sqrt{\frac{n-1}{mn}}, \tag{43.1.6}$$

where $\bar{\mathsf{y}}_{\cdot\cdot}$ is the grand sample mean and μ its population counterpart. Since $\bar{\mathsf{y}}_{\cdot\cdot} = \frac{1}{n}\sum\bar{\mathsf{y}}_{i\cdot\cdot}$ one obtains $\operatorname{cov}[\bar{\mathsf{y}}_{i\cdot}, \bar{\mathsf{y}}_{\cdot\cdot}] = \frac{1}{n}\operatorname{var}[\bar{\mathsf{y}}_{i\cdot}] = \frac{\sigma^2}{mn}$. Therefore $\operatorname{var}[\bar{\mathsf{y}}_{i\cdot} - \bar{\mathsf{y}}_{\cdot\cdot}] = \sigma^2\left(\frac{1}{m} - \frac{2}{mn} + \frac{1}{mn}\right) = \sigma^2\left(\frac{n-1}{mn}\right)$. And the correlation coefficient is $1/(n-1)$. □

43.1.4. Studentized Range. This is a famous example, it is not in Seber [**Seb77**], but we should at least know what it is. Just as the projected F intervals are connected with the name of Scheffé, these intervals are connected with the name of Tukey. Again in the situation of our basic example one uses $\bar{\mathsf{y}}_{i\cdot} - \bar{\mathsf{y}}_{k\cdot}$ to build confidence intervals for $\mu_i - \mu_k$ for all pairs i, k: $i \neq k$. This is no longer the equicorrelated case. (Such simultaneous confidence intervals are useful if one knows that one will compare means, but one does not know a priori which means.)

PROBLEM 435. *Again in our basic example, define*

$$\boldsymbol{z} = \frac{1}{\sqrt{2}}\begin{bmatrix} \bar{\mathsf{y}}_{1\cdot} - \bar{\mathsf{y}}_{2\cdot} \\ \bar{\mathsf{y}}_{1\cdot} - \bar{\mathsf{y}}_{3\cdot} \\ \bar{\mathsf{y}}_{1\cdot} - \bar{\mathsf{y}}_{4\cdot} \\ \bar{\mathsf{y}}_{2\cdot} - \bar{\mathsf{y}}_{3\cdot} \\ \bar{\mathsf{y}}_{2\cdot} - \bar{\mathsf{y}}_{4\cdot} \\ \bar{\mathsf{y}}_{3\cdot} - \bar{\mathsf{y}}_{4\cdot} \end{bmatrix}.$$

Compute the correlation matrix of $\boldsymbol{z}$.

ANSWER. Write $\boldsymbol{z} = \frac{1}{\sqrt{2}}\boldsymbol{A}\bar{\mathbf{y}}$, therefore $\mathcal{V}[\boldsymbol{z}] = \frac{\sigma^2}{2m}\boldsymbol{A}\boldsymbol{A}^\top$ where

$$\boldsymbol{A} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{bmatrix} \qquad \mathcal{V}[\boldsymbol{z}] = \frac{1}{2m}\boldsymbol{A}\boldsymbol{A}^\top = \frac{1}{2m}\begin{bmatrix} 2 & 1 & 1 & -1 & -1 & 0 \\ 1 & 2 & 1 & 1 & 0 & -1 \\ 1 & 1 & 2 & 0 & 1 & 1 \\ -1 & 1 & 0 & 2 & 1 & -1 \\ -1 & 0 & 1 & 1 & 2 & 1 \\ 0 & -1 & 1 & -1 & 1 & 2 \end{bmatrix} \tag{43.1.7}$$

□

43.2. Relation between F-test and t-tests.

Assume you have constructed the t-intervals for several different linear combinations of the two parameters β_1 and β_2. In the (β_1, β_2)-plane, each of these intervals can be represented by a band delimited by parallel straight lines. If one draws many of these bands, their intersection becomes an ellipse, which has the same shape as the joint F-confidence region for β_1 and β_2, but it is smaller, i.e., it comes from an F-test for a lower significance level α

The F-test, say for $\beta_1 = \beta_2 = 0$, is therefore equivalent not to two but to infinitely many t-tests, one for each linear combination of β_1 and β_2, but each of these t tests

has a higher confidence level than that of the F test. This is the right way how to look at the F test.

What are situations in which one would want to obtain a F-confidence region in order to get information about many different linear combinations of the parameters at the same time?

For instance, one examines a regression output and looks at all parameters and computes linear combinations of parameters of interest, and believes they are significant if their t-tests reject. This whole procedure is sometimes considered as a misuse of statistics, "data-snooping," but Scheffé argued it was justified if one raises the significance level to that of the F test implied by the infinitely many t tests of all linear combinations of $\boldsymbol{\beta}$.

Or one looks at only certain kinds of linear combinations, for instance, at all contrasts, i.e., linear combinations whose coefficients sum to zero. This is a very thorough way to ascertain that all parameters are equal.

Or if one wants to draw a confidence band around the whole regression line.

PROBLEM 436. *Someone fits a regression with 18 observations, one explanatory variable and a constant term, and then draws around each point of the regression line a standard 95%* t *interval. What is the probability that the band created in this way covers the true regression line over its entire length? Note: the* `Splus` *commands* `qf(1-alpha,df1,df2)` *and* `qt(1-alpha/2,df)` *give quantiles, and the commands* `pf(critval,df1,df2)` *and* `pt(critval,df)` *give the cumulative distribution function of* F *and* t *distributions.*

ANSWER. Instead of $n = 18$ and $k = 2$ we do it for arbitrary n and k. We need a α such that

$$t_{(n-k;0.025)} = \sqrt{2F_{(k,n-k;\alpha)}} \tag{43.2.1}$$

$$\frac{1}{2}(t_{(n-k;0.025)})^2 = F_{(k,n-k;\alpha)} \tag{43.2.2}$$

$$1-\alpha = \Pr[\mathsf{F}_{k,n-k} \leq \frac{1}{2}(t_{(n-k;0.025)})^2] \tag{43.2.3}$$

The `Splus` command is `obsno<-18; conflev<-pf((qt(0.975,obsno-2)^2/2,2,obsno-2)`. The value is 0.8620989. □

PROBLEM 437. *6 points Which options do you have if you want to test more than one hypothesis at the same time? Describe situations in which one* F*-test is better than two* t*-tests (i.e., in which an elliptical confidence region is better than a rectangular one). Are there also situations in which you might want two* t*-tests instead of one* F*-test?*

In the one-dimensional case this confidence region is identical to the t-interval. But if one draws for $i = 2$ the confidence ellipse generated by the F-test and the two intervals generated by the t-tests into the same diagram, one obtains the picture as in figure 5.1 of Seber [**Seb77**], p. 131. In terms of hypothesis testing this means:

there are values for which the F test does not reject but one or both t tests reject, and there are values for which one or both t-tests fail to reject but the F-test rejects. The reason for this confusing situation is that one should not compare t tests and F tests at the same confidence level. The relationship between those testing procedures becomes clear if one compares the F test at a given confidence level to t tests at a certain higher confidence level.

We need the following math for this. For a positive definite $\boldsymbol{\Psi}$ and arbitrary $\boldsymbol{x}$ it follows from (A.5.6) that

$$\boldsymbol{x}^\top \boldsymbol{\Psi}^{-1}\boldsymbol{x} = \max_{\boldsymbol{g}:\ \boldsymbol{g}\neq\boldsymbol{o}} \frac{(\boldsymbol{g}^\top\boldsymbol{x})^2}{\boldsymbol{g}^\top\boldsymbol{\Psi}\boldsymbol{g}}. \tag{43.2.4}$$

Applying this to the situation here we have

$$(\boldsymbol{u}-\boldsymbol{R}\hat{\boldsymbol{\beta}})^\top(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top)^{-1}(\boldsymbol{u}-\boldsymbol{R}\hat{\boldsymbol{\beta}}) = \max_{\boldsymbol{g}:\ \boldsymbol{g}\neq\boldsymbol{o}} \frac{(\boldsymbol{g}^\top\boldsymbol{u}-\boldsymbol{g}^\top\boldsymbol{R}\hat{\boldsymbol{\beta}})^2}{\boldsymbol{g}^\top\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\boldsymbol{g}}. \tag{43.2.5}$$

Now the maximum of a set is smaller or equal to $iF_{(i,n-q;\alpha)}\mathsf{s}^2$ if and only if each element of this set is smaller or equal. Therefore the F-confidence region (41.4.3) can also be written as

$$R(\mathsf{y}) = \{\boldsymbol{u}\in\mathbb{R}^i:\ \frac{(\boldsymbol{g}^\top\boldsymbol{u}-\boldsymbol{g}^\top\boldsymbol{R}\hat{\boldsymbol{\beta}})^2}{\boldsymbol{g}^\top\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\boldsymbol{g}} \leq iF_{(i,n-q;\alpha)}\mathsf{s}^2 \text{ for all } \boldsymbol{g}\neq\boldsymbol{o}\} \tag{43.2.6}$$

$$= \{\boldsymbol{u}\in\mathbb{R}^i:\ (\boldsymbol{g}^\top\boldsymbol{u}-\boldsymbol{g}^\top\boldsymbol{R}\hat{\boldsymbol{\beta}})^2 \leq iF_{(i,n-q;\alpha)}\mathsf{s}^2\boldsymbol{g}^\top\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\boldsymbol{g} \text{ for all } \boldsymbol{g}\} \tag{43.2.7}$$

$$= \{\boldsymbol{u}\in\mathbb{R}^i:\ |\boldsymbol{g}^\top\boldsymbol{u}-\boldsymbol{g}^\top\boldsymbol{R}\hat{\boldsymbol{\beta}}| \leq \sqrt{iF_{(i,n-q;\alpha)}}\mathsf{s}_{\boldsymbol{g}^\top\boldsymbol{R}\hat{\boldsymbol{\beta}}} \text{ for all } \boldsymbol{g}\} \tag{43.2.8}$$

$$= \bigcap_{\boldsymbol{g}}\{\boldsymbol{u}\in\mathbb{R}^i:\ |\boldsymbol{g}^\top\boldsymbol{u}-\boldsymbol{g}^\top\boldsymbol{R}\hat{\boldsymbol{\beta}}| \leq \sqrt{iF_{(i,n-q;\alpha)}}\mathsf{s}_{\boldsymbol{g}^\top\boldsymbol{R}\hat{\boldsymbol{\beta}}}\}. \tag{43.2.9}$$

It is sufficient to take the intersection over all $\boldsymbol{g}$ with unit length. What does each of these regions intersected look like? First note that the $i\times 1$ vector $\boldsymbol{u}$ lies in that region if and only if $\boldsymbol{g}^\top\boldsymbol{u}$ lies in a t-interval for $\boldsymbol{g}^\top\boldsymbol{R}\boldsymbol{\beta}$, whose confidence level is no longer α but is $\gamma = \Pr[|\mathsf{t}| \leq \sqrt{iF_{(i,n-q;\alpha)}}]$, where t is distributed as a t with $n-q$ degrees of freedom. Geometrically, in Seber [**Seb77**]'s figure 5.1, these confidence regions can be represented by all the bands *tangent* to the ellipse.

Taking only the vertical and the horizontal band tangent to the ellipse, one has now the following picture: if one of the t-tests rejects, then the F-test rejects too. But it may be possible that the F-test rejects but neither of the two t-tests rejects.

In this case, there must be some other linear combination of the two variables for which the $\mathfrak{t}$ test rejects.

Another example for simultaneous $\mathfrak{t}$-tests, this time derived from Hotelling's T^2, is given in Johnson and Wichern [**JW88**, chapter 5]. It is very similar to the above; we will do here only the large-sample development:

43.3. Large-Sample Simultaneous Confidence Regions

Assume every row $\mathfrak{y}_i$ of the $n \times p$ matrix $\mathbf{Y}$ is an independent drawing from a population with mean $\boldsymbol{\mu}$ and dispersion matrix $\boldsymbol{\Sigma}$. If n is much larger than p, then one can often do all tests regarding the unknown $\boldsymbol{\mu}$ in terms of the sample mean $\bar{\mathfrak{y}}$, which one may assume to be normally distributed, and whose true dispersion matrix may be assumed to be know and to be equal to the sample dispersion matrix of the $\mathfrak{y}_i$, divided by n.

Therefore it makes sense to look at the following model (the $\mathfrak{y}$ in this model is equal to the $\bar{\mathfrak{y}}$ in the above model, and the $\boldsymbol{\Sigma}$ in this model is equal to $\mathbf{S}/n$, or any other consistent estimate, for that matter, in the above model):

Assume $\mathfrak{y} \sim \mathsf{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with unknown $\boldsymbol{\mu}$ and known $\boldsymbol{\Sigma}$. We allow $\boldsymbol{\Sigma}$ to be singular, i.e., there may be some nonzero linear combinations $\boldsymbol{g}^\top \mathfrak{y}$ which have zero variance. Let q be the rank of $\boldsymbol{\Sigma}$. Then a simultaneous $1-\alpha$ confidence region for all linear combinations of $\boldsymbol{\mu}$ is

$$\boldsymbol{g}^\top \boldsymbol{\mu} \in \boldsymbol{g}^\top \mathfrak{y} \pm \sqrt{\chi_q^{2(\alpha)}} \sqrt{\boldsymbol{g}^\top \boldsymbol{\Sigma} \boldsymbol{g}} \tag{43.3.1}$$

where $\chi_q^{2(\alpha)}$ is the upper α-quantile of the χ^2 distribution with q degrees of freedom, i.e., $\Pr[\chi_q^2 \geq \chi_q^{2(\alpha)}] = \alpha$.

Proof: For those $\boldsymbol{g}$ with $\operatorname{var}[\boldsymbol{g}^\top \mathfrak{y}] = 0$, i.e., $\boldsymbol{g}^\top \boldsymbol{\Sigma} \boldsymbol{g} = 0$, the confidence interval has 100 percent coverage probability (despite its zero length); therefore we only have to worry about those $\boldsymbol{g}$ with $\boldsymbol{g}^\top \boldsymbol{\Sigma} \boldsymbol{g} \neq 0$:

$$\Pr[\boldsymbol{g}^\top \boldsymbol{\mu} \in \boldsymbol{g}^\top \mathfrak{y} \pm \sqrt{\chi_q^{2(\alpha)}} \sqrt{\boldsymbol{g}^\top \boldsymbol{\Sigma} \boldsymbol{g}} \text{ for all } \boldsymbol{g}\ ] = \tag{43.3.2}$$

$$= \Pr[\boldsymbol{g}^\top \boldsymbol{\mu} \in \boldsymbol{g}^\top \mathfrak{y} \pm \sqrt{\chi_q^{2(\alpha)}} \sqrt{\boldsymbol{g}^\top \boldsymbol{\Sigma} \boldsymbol{g}} \text{ for all } \boldsymbol{g} \text{ with } \boldsymbol{g}^\top \boldsymbol{\Sigma} \boldsymbol{g} \neq 0\ ] = \tag{43.3.3}$$

$$= \Pr[\frac{(\boldsymbol{g}^\top(\boldsymbol{\mu} - \mathfrak{y}))^2}{\boldsymbol{g}^\top \boldsymbol{\Sigma} \boldsymbol{g}} \leq \chi_q^{2(\alpha)} \text{ for all } \boldsymbol{g} \text{ with } \boldsymbol{g}^\top \boldsymbol{\Sigma} \boldsymbol{g} \neq 0\ ] = \tag{43.3.4}$$

$$= \Pr[\max_{\boldsymbol{g}:\ \boldsymbol{g}^\top \boldsymbol{\Sigma} \boldsymbol{g} \neq 0} \frac{(\boldsymbol{g}^\top(\boldsymbol{\mu} - \mathfrak{y}))^2}{\boldsymbol{g}^\top \boldsymbol{\Sigma} \boldsymbol{g}} \leq \chi_q^{2(\alpha)}] = \tag{43.3.5}$$

$$= \Pr[(\mathfrak{y} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^- (\mathfrak{y} - \boldsymbol{\mu}) \leq \chi_q^{2(\alpha)}] = 1 - \alpha \tag{43.3.6}$$

One can apply the maximization theorem here because $\mathfrak{y} - \boldsymbol{\mu}$ can be written in the form $\boldsymbol{\Sigma} \boldsymbol{u}$ for some $\boldsymbol{u}$.

Now as an example let's do Johnson and Wichern, example 5.7 on pp. 194–197. In a survey, people in a city are asked which bank is their primary savings bank. The answers are collected as rows in the $\mathbf{Y}$ matrix. The columns correspond to banks A, B, C, D, to some other bank, and to having no savings. Each row has exactly one 1 in the column corresponding to the respondent's primary savings bank, zeros otherwise. The people with no savings will be ignored, i.e., their rows will be trimmed from the matrix together with the last column. After this trimming, $\mathbf{Y}$ has 5 columns, and there are 355 respondents in these five categories. It is assumed that the rows of $\mathbf{Y}$ are independent, which presupposes sampling with replacement, i.e., the sampling is done in such a way that theoretically the same people might be asked twice (or the sample is small compared with the population). The probability distribution of each of these rows, say here the ith row, is the multinomial distribution whose parameters form the p-vector $\boldsymbol{p}$ (of nonnegative elements adding up to 1). Its means, variances, and covariances can be computed according to the rules for discrete distributions:

$$\mathrm{E}[\mathfrak{y}_{ij}] = 1(p_j) + 0(1 - p_j) = p_j \tag{43.3.7}$$

$$\operatorname{var}[\mathfrak{y}_{ij}] = \mathrm{E}[\mathfrak{y}_{ij}^2] - (\mathrm{E}[\mathfrak{y}_{ij}])^2 = p_j - p_j^2 = p_j(1 - p_j) \quad \text{because } \mathfrak{y}_{ij}^2 = \mathfrak{y}_{ij} \tag{43.3.8}$$

$$\operatorname{cov}[\mathfrak{y}_{ij}, \mathfrak{y}_{ik}] = \mathrm{E}[\mathfrak{y}_{ij}\mathfrak{y}_{ik}] - \mathrm{E}[\mathfrak{y}_{ij}]\,\mathrm{E}[\mathfrak{y}_{ik}] = -p_j p_k \quad \text{because } \mathfrak{y}_{ij}\mathfrak{y}_{ik} = 0 \tag{43.3.9}$$

The p_i can be estimated by the ith sample means. From these sample means one also obtains an estimate $\mathbf{S}$ of the dispersion matrix of the rows of $\mathbf{Y}$. This estimate is singular (as is the true dispersion matrix), it has rank $r - 1$, since every row of the $\mathbf{Y}$-matrix adds up to 1. Provided $n - r$ is large, which means here that $n\hat{p}_k \geq 20$ for each category k, one can use the normal asymptotics, and gets as simultaneous confidence interval for all linear combinations

$$\boldsymbol{g}^\top \boldsymbol{p} \in \boldsymbol{g}^\top \hat{\boldsymbol{p}} \pm \sqrt{\chi_{r-1}^{2\ (\alpha)}} \sqrt{\frac{\boldsymbol{g}^\top \mathbf{S} \boldsymbol{g}}{n}} \tag{43.3.10}$$

A numerical example illustrating the width of these confidence intervals is given in [**JW88**, p. 196].

CHAPTER 44

Sample SAS Regression Output

```
dep variable:  wagerate
analysis of variance

                          sum of          mean
source          df       squares        square      F value    prob>F

model            6    1553.90611     258.98435       20.931    0.0001
error          547    6768.00436   12.37295129
c total        553    8321.91046

root mse  3.517521      R-square       0.1867
dep mean  4.817097      adj R-sq       0.1778
c.v.      73.02159

parameter estimates
                       parameter      standard    t for H0:
variable        df      estimate         error  parameter=0  prob>|t|

intercep         1    0.36435820    1.21195551        0.301    0.7638
age              1   -0.01661574    0.02937263       -0.566    0.5718
educatn          1    0.41699451    0.05107535        8.164    0.0001
xperienc         1    0.02981372    0.02958059        1.008    0.3140
gender           1   -1.73266849    0.41844140       -4.141    0.0001
white            1    0.33807525    0.84295802        0.401    0.6885
black            1   -0.19974753    0.87184661       -0.229    0.8189
```

TABLE 1. Sample SAS regression output

Table 1 is the output of a SAS run. The dependent variable is the $\mathbf{y}$ variable, here it has the name `wagerate`. "Analysis" is the same as "decomposition," and

"variance" is here the sample variance or, say better, the sum of squares. "Analysis of variance" is a decomposition of the "corrected total" sum of squares $\sum_{j=1}^{n}(y_j-\bar{y})^2 = 8321.91046$ into its "explained" part $\sum_{j=1}^{n}(\hat{y}_j-\bar{y})^2 = 1553.90611$, the sum of squares whose "source" is the "model," and its "unexplained" part, the sum of squared "errors" $\sum_{j=1}^{n}(y_j-\hat{y}_j)^2$, which add up to 6768.00436 here. The "degrees of freedom" are a dimensionality constant; the d.f. of the corrected total sum of squares (SST) is the number of observations minus 1, while the d.f. of the SSE is the number of observations minus the number of parameters (intercept and slope parameters) in the regression. The d.f. of the sum of squares due to the model consists in the number of slope parameters (not counting the intercept) in the model.

The "mean squares" are the corresponding sum of squares divided by their degrees of freedom. This "mean sum of squares due to error" should not be confused with the "mean squared error" of an estimator $\hat{\theta}$ of θ, defined as $\text{MSE}[\hat{\theta};\theta] = \text{E}[(\hat{\theta}-\theta)^2]$. One can think of the mean sum of squares due to error as the sample analog of the $\text{MSE}[\hat{\mathbf{y}};\mathbf{y}]$; it is as the same time an unbiased estimate of the disturbance variance σ^2. The mean sum of squares explained by the model is an unbiased estimate of σ^2 if all slope coefficients are zero, and is larger otherwise. The F value is the mean sum of squares explained by the model divided by the mean sume of squares due to error, this is the value of the test statistic for the F-test that all slope parameters are zero. The p-value `prob` > `F` gives the probability of getting an even larger F-value when the null hypothesis is true, i.e., when all parameters are indeed zero. To reject the null hypothesis at significance level α, this p-value must be smaller than α.

The `root mse` is the square root of the mean sum of squares due to error, it is an estimate of σ. The `dependent mean` is simply $\bar{y}$. The "coefficient of variation" (`c.v.`) is 100 times `root mse` divided by `dependent mean`. The `R-square` is `ss(model)` divided by `ss(c. total)`, and the `adjusted R-square` is $\bar{R}^2 = 1 - \frac{\text{SSE}/(n-k)}{\text{SST}/(n-1)}$.

For every parameter, including the intercept, the estimated value is printed, and next to it the estimate of its standard deviation. The next column has the t-value, which is the estimated value divided by its estimated standard deviation. This is the test statistic for the null hypothesis that this parameter is zero. The `prob>|t|` value indicates the significance for the two-sided test.

PROBLEM 438. *What does the* `c` *stand for in* `c total` *in Table 1?*

PROBLEM 439. *4 points Using the sample SAS regression output in Table 1, test at the 5% significance level that the coefficient of* `gender` *is* -1.0*, against the alternative that it is* < -1.0.

ANSWER. `-1.73266849-(-1)=-.73266849` must be divided by `0.41844140`, which gives `-1.7509465`, and then we must look it up in the t-table or, since there are so many observations, the normal table. It is a one-sided test, therefore the critical value is `-1.645`, therefore reject the null hypothesis. □

PROBLEM 440. *Here is part of the analysis of variance table printed out by the SAS regression procedure: If you don't have a calculator, simply give the answers in*

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE
MODEL	4	1529.68955	382.42239
ERROR	549	6792.22092	12.37198710
C TOTAL	553	8321.91046	

expressions like $\sqrt{1529.68955}/7$ *etc.*

• a. *1 point The SSE of this regression is* [ ]

• b. *2 points The estimated standard deviation of the disturbance is* [ ]

• c. *1 point Besides a constant term, the regression has* [ ] *explanatory variables.*

• d. *1 point The dataset has* [ ] *observations.*

• e. *3 points Make an* F *test of the null hypothesis that the slope parameters of all the explanatory variables are zero, at the 5% significance level. The observed value of the* F *statistic is is* [ ]*, and the critical value is* [ ] *in this case.*

ANSWER. $\Pr[F_{4,\infty} > 3.32] = 0.01$; $\Pr[F_{4,\infty} > 2.37] = 0.05$. The observed F value is 382.42/12.372=3 which is significant up to the level 0.0001, therefore reject H_0. □

• f. *Here is the printout of the analysis of variance table after additional explanatory variables were included in the above regression (i.e., the dependent variable is the same, and the set of explanatory variables contains all variables used in the above regression, plus some additional ones).*

• g. *1 point There were* [ ] *additional variables.*

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE
MODEL	6	1553.90611	258.98435
ERROR	547	6768.00436	12.37295129
C TOTAL	553	8321.91046	

• h. *3 points Make an F test at the 5% significance level of the null hypothesis that all* additional *variables (i.e., those variables* not *included in the first regression) had slope coefficients zero. The value of the test statistic is* [ ] *(show how you derived it), and the critical value is* [ ]

ANSWER. Therefore you *cannot* reject. □

ANSWER. $SSE_u = 6768.0$, $SSE_r = 6792.2$, and the number of restrictions is 2. The F-statistic is therefore

$$\frac{(6792.2 - 6768.0)/2}{6768.0/547} = 0.9779(0.97861?) < 3.00 = F_{.005;2,547} \tag{44.0.11}$$

Therefore you cannot reject the null hypothesis. □

CHAPTER 45

Flexible Functional Form

So far we have assumed that the mean of the dependent variable is a linear function of the explanatory variables. In this chaper, this assumption will be relaxed. We first discuss the case where the explanatory variables are categorical variables. For categorical variables (gender, nationality, occupations, etc.), the concept of linearity does not make sense, and indeed, it is customary to fit arbitrary numerical functions of these categorical variables. One can do this also if one has numerical variables which assume only a limited number of values (such as the number of people in a household). As long as there are repeated observations for each level of these variables, it is possible to introduce a different dummy variables for every level, and in this way also allow arbitrary functions. Linear restrictions between the coefficients of these dummies can be interpreted as the selection of more restricted functional spaces.

45.1. Categorical Variables: Regression with Dummies and Factors

If the explanatory variables are categorical then it is customary to fit arbitrary functions of these variables. This can be done with the use of dummy variables, or by the use of variables coded as "factors." If there are more than two categories, you need several several regressors taking only the values 0 and 1, which is why they are called "dummy variables." One regressor with several levels 0,1,2, etc. is too restrictive. The "factor" data type in R allows to code several levels in one variable, which will automatically be expanded into a set of dummy variables. Therefore let us first discuss dummy variables.

If one has a categorical variable which has j possible outcomes, the simplest and most obvious thing to do would be to generate j regressors into the equation, each taking the value 1 if the observation has this level, and the value 0 otherwise. But if one does this, one has to leave the intercept out of the regression, otherwise one gets perfect multicollinearity. Usually in practice one keeps the intercept and omits one of the dummy variables. This makes it a little more difficult to interpret the dummy variables.

PROBLEM 441. *In the intermediate econometrics textbook* [**WW79**], *the following regression line is estimated:*

$$\mathsf{b}_t = 0.13 + .068y_t + 0.23w_t + \hat{\varepsilon}_t, \tag{45.1.1}$$

where b_t *is the public purchase of Canadian government bonds (in billion \$),* y_t *is the national income, and* w_t *is a dummy variable with the value* $w_t = 1$ *for the war years 1940–45, and zero otherwise.*

• a. *1 point This equation represents two regression lines, one for peace and one for war, both of which have the same slope, but which have different intercepts. What is the intercept of the peace time regression, and what is that of the war time regression line?*

ANSWER. In peace, $w_t = 0$, therefore the regression reads $\mathsf{b}_t = 0.13 + .068y_t + \hat{\varepsilon}_t$, therefore the intercept is .13. In war, $w_t = 1$, therefore $\mathsf{b}_t = 0.13 + .068y_t + 0.23 + \hat{\varepsilon}_t$, therefore the intercept is $.13 + .23 = .36$. □

• b. *1 point What would the estimated equation have been if, instead of* w_t, *they had used a variable* p_t *with the values* $p_t = 0$ *during the war years, and* $p_t = 1$ *otherwise? (Hint: the coefficient for* p_t *will be negative, because the intercept in peace times is below the intercept in war times).*

ANSWER. Now the intercept of the whole equation is the intercept of the war regression line, which is .36, and the coefficient of p_t is the difference between peace and war intercepts, which is -.23.

$$\mathsf{b}_t = .36 + .068y_t - .23p_t + \hat{\varepsilon}_t. \tag{45.1.2}$$

□

• c. *1 point What would the estimated equation have been if they had thrown in both* w_t *and* p_t, *but left out the intercept term?*

ANSWER. Now the coefficient of w_t is the intercept in the war years, which is .36, and the coefficient of p_t is the intercept in the peace years, which is .13.

$$\mathsf{b}_t = .36w_t + .13p_t + .068y_t + \hat{\varepsilon}_t? \tag{45.1.3}$$

□

• d. *2 points What would the estimated equation have been, if bond sales and income had been measured in millions of dollars instead of billions of dollars? (1 billion = 1000 million.)*

ANSWER. From $\mathsf{b}_t = 0.13 + .068y_t + 0.23w_t + \hat{\varepsilon}_t$ follows $1000b_t = 130 + .068 \cdot 1000y_t + 230w_t + 1000\hat{\varepsilon}_t$, or

$$\mathsf{b}_t^{(m)} = 130 + .068y_t^{(m)} + 230w_t + \hat{\varepsilon}_t^{(m)}, \tag{45.1.4}$$

where $\mathsf{b}_t^{(m)}$ is bond sales in millions (i.e., $\mathsf{b}_t^{(m)} = 1000b_t$), and $y_t^{(m)}$ is national income in millions (i.e., $y_t^{(m)} = 1000y_t$). □

PROBLEM 442. *5 points Assume you run a time series regression* $\mathbf{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$, *but you have reason to believe that the values of the parameter* $\boldsymbol{\beta}$ *are not equal in all time periods* t. *What would you do?*

ANSWER. Include dummies, run separate regressions for subperiods, use a varying parameter model. □

There are various ways to set it up. Threshold effects might be represented by the following dummies:

$$\begin{bmatrix} \boldsymbol{\iota} & \boldsymbol{o} & \boldsymbol{o} & \boldsymbol{o} \\ \boldsymbol{\iota} & \boldsymbol{\iota} & \boldsymbol{o} & \boldsymbol{o} \\ \boldsymbol{\iota} & \boldsymbol{\iota} & \boldsymbol{\iota} & \boldsymbol{o} \\ \boldsymbol{\iota} & \boldsymbol{\iota} & \boldsymbol{\iota} & \boldsymbol{\iota} \end{bmatrix} \tag{45.1.5}$$

In the example in Problem 441, the slope of the numerical variables does not change with the levels of the categorical variables, in other words, there is no interaction between those variables, but each variable makes a separate contribution to the response variable. The presence of interaction can be modeled by including products of the dummy variables with the response variable with whom interaction exists.

How do you know the interpretation of the coefficients of a given set of dummies? Write the equation for every category separately. E.g. [**Gre97**, p. 383]: Winter $\mathbf{y} = \beta_1 + \beta_5\boldsymbol{x}$, Spring $\mathbf{y} = \beta_1 + \beta_2 + \beta_5\boldsymbol{x}$ Summer $\mathbf{y} = \beta_1 + \beta_3 + \beta_5\boldsymbol{x}$, Autumn $\mathbf{y} = \beta_1 + \beta_4 + \beta_5\boldsymbol{x}$. I.e. the overall intercept β_1 is the intercept in Winter, the coefficient for the first seasonal dummy β_2 is the difference between Spring and Winter, that for the second dummy β_3 difference between Summer and Winter, and β_4 the difference between Autumn and Winter.

If the slope differs too, do

$$\begin{bmatrix} \boldsymbol{\iota} & \boldsymbol{o} & \boldsymbol{x} & \boldsymbol{o} \\ \boldsymbol{\iota} & \boldsymbol{\iota} & \boldsymbol{x} & \boldsymbol{x} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\iota} & \boldsymbol{d} & \boldsymbol{x} & \boldsymbol{d} * \boldsymbol{x} \end{bmatrix} \tag{45.1.6}$$

where $*$ denotes the Hadamard product of two matrices (their element-wise multiplication). This last term is called an interaction term.

An alternative to using dummy variables is to use factor variables. If one includes a factor variable into a regression formula, the statistical package converts it into a set of dummies. Look at Section 22.5 for an example how to use factor variables instead of dummies in R.

45.2. Flexible Functional Form for Numerical Variables

Here the issue is: how to find the right transformation of the explanatory variables before running the regression? Each of the methods to be discussed has a smoothing parameter.

To fix notation, assume for now that only one explanatory variable $\boldsymbol{x}$ is given and you want to estimate the model $\mathbf{y} = f(\boldsymbol{x}) + \boldsymbol{\varepsilon}$ with the usual assumption $\boldsymbol{\varepsilon} \sim \boldsymbol{o}, \sigma^2\boldsymbol{I}$. But whereas the regression model specified that f is an affine function, we allow f to be an element of an appropriate larger function space. The size of this space is characterized by a so-called *smoothing parameter*.

45.2.1. Polynomial Regression.

The most frequently used method is *polynomial regression*, i.e., one chooses f to be a polynomial of *order* m (i.e. it has m terms, including the constant term) or *degree* $m-1$ (i.e. the highest power is x^{m-1}). $f(x) = \theta_0 + \theta_1 x + \cdots + \theta_{m-1}x^{m-1}$.

Motivation: This is a seamless generalization of ordinary least squares, since affine functions are exactly polynomials of degree 1 (order 2). Taylor's theorem says that any $f \in W^m[a,b]$ can be approximated by a polynomial of order m (degree $m-1$) plus a remainder term which can be written as an integral involving the mth derivative, see [**Eub88**, (3.5) on p. 90]. The Weierstrass Approximation Theorem say that any continuous function over a closed and bounded interval can be uniformly approximated by polynomials of sufficiently high degree.

Here one has to decide what degree to use, the degree of the polynomial plays here the role of the smoothing parameter.

Some practical hints:

For higher degree polynomials don't use the "power basis" $1, x, x^2, \ldots, x^{m-1}$, but there are two reasonable choices. Either one can use Legendre polynomials [**Eub88**, (3.10) and (3.11) on p. 54], which are obtained from the power basis by Gram-Schmidt orthonormalization over the interval $[a,b]$. This does not make the design matrix orthogonal, but at least one should expect it not to be too ill-conditioned, and the roots and the general shape of Legendre polynomials is well-understood. As the second main choice one may also select polynomials that make the design-matrix itself exactly orthonormal. The `Splus`-function `poly` does that.

The jth Legendre polynomial has exactly j real roots in the interval [**Dav75**, Chapter X], [**Sze59**, Chapter III]. The orthogonal polynomials probably have a similar property. This gives another justification for using polynomial regession, which is similar to the justification one sometimes reads for using Fourier-series: The data have high-frequency and low-frequency components, and one wants to filter out the low-frequency components.

In practice, polynomials do not always give a good fit. There are better alternatives available, which will be discussed in turn.

45.2.2. The Box-Cox Transformation.

An early attempt used in Econometrics was to use a family of functions which is not as complete as the polynomials but which ecompasses many functional forms encountered in Economics. These functions

are only defined for $x > 0$ and have the form

$$B(x,\lambda) = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \ln(x) & \text{if } \lambda = 0 \end{cases} \tag{45.2.1}$$

[**DM93**, p. 484] have a plot with the curves for $\lambda = 1.5$, 1, 0.5, 0, -0.5, and -1. They point out some serious disadvantage of this transformation: if $\lambda \neq 0$, $B(x,\lambda)$ is bounded eihter from below or above. For $\lambda < 0$, $B(x,\lambda)$ cannot be greater than $-1/\lambda$, and for $\lambda > 0$, it cannot be less than $-1/\lambda$.

About the Box-Cox transformation read [**Gre97**, 10.4]

45.2.3. Kernel Estimates. For nonparametric estimates look at [**Loa99**], it has the R-package `locfit`.

Figure 1.1 is a good example: it is actuarial data, which are roughly fitted by a straight line, but a better idea of the accelerations and decelerations can be very useful for a life insurance company.

Chapter 1 gives a historical overview: Spencer's rule from 1904 was designed for computational convenience (for hand-calculations), and it reproduces polynomials up to the 3rd degree. Figure 2.1 illustrates how local regression is done. Pp. 18/19: emphasis on fitted values, not on the parameter estimates. There are two important parameters: the bandwidth and the degree of the polynomial. To see the effects of bandwidth, see the plots on p. 21: using our data we can do plots of the sort `plot(locfit(r~year,data=uslt,alpha=0.1,deg=3),get.data=T)` and then vary `alpha` and `deg`.

PROBLEM 443. *What kind of smoothing would be best for the time series of the variable* `r` *(profit rate) in dataset* `uslt`*?*

PROBLEM 444. *Locally constant smooths are not good at the edges, and also not at the maxima and minima of the data. Why not?*

The kernel estimator can be considered a local fit of a constant. Straight lines are better, and cubic parabolas even better. Quadratic ones not as good.

The birth rate data which require smoothing with a varying bandwidth are interesting, see Simonoff p. 157, description in the text on p. 158.

45.2.4. Regression Splines. About the word "spline," [**Wah90**, p. vii] writes: "The mechanical spline is a thin reedlike strip that was used to draw curves needed in the fabrication of cross sections of ships' hulls. Ducks or weights were placed on the strip to force it to go through given points, and the free portion of the strip would assume a position in space that minimized the bending energy."

One of the drawbacks of polynomial regression is that its fit is global. One method to provide for local fits is to fit a piecewise polynomial. A spline is a piecewise

polynomial of order m (degree $m-1$) spliced together at given "knots" so that all derivatives coincide up to and including the $m-2$nd one. Polynomial splines are generalizations of polynomials: whereas one can characterize polynomials of order m (degree $m-1$) as functions whose $m-1$st derivative is constant, polynomial splines are functions whose $m-1$st derivative is *piecewise* constant. This is the smoothest way to put different polynomials together. Compare the `Splus`-function `bs`.

If one starts with a cubic spline, i.e., a spline of order 4, and postulates in addition that the 2nd derivative is zero outside the boundary points, one obtains what is called a "natural cubic spline"; compare the `Splus`-function `ns`. There is exactly one natural spline going through n datapoints.

One has to choose the order and the location of the knots. The most popular are cubic splines, and higher orders do not seem to add much, therefore it is more important to concentrate on a good selection of the knots. here are some guidelines how to choose knots, taken from [**Eub88**, p. 357]:

For $m = 2$, linear splines, place knots at points where the data exhibit a change in slope.

For $m = 3$, quadratic splines, locate knots near local maxima, minima or inflection points in the data.

For $m = 4$, cubic splines, arrange the knots so that they are close to inflexion points in the data and not more than one extreme point (maximum or minimum) and one inflection point occurs between any two knots.

It is also possible to determine the number of knots and select their location so as to optimize the fit. But this is a hairy minimization problem; [**Eub88**, p. 362] gives some shortcuts.

Extensions: Sometime one wants knots which are not so smooth, this can be obtained by letting several knots coincide. Or one wants polynomials of different degrees in the different segments.

[**Gre97**, pp. 389/90] has a nice example for a linear spline. Each of 3 different age groups has a different slope and a different intercept: $t < t_*$, $t_* \leq t < t_{**}$, and $t_{**} \leq t$. These age groups are coded by the matrix $\boldsymbol{D}$ consisting of two dummy variables, one for $t \geq t_*$ and one for $t \geq t_{**}$. I.e, $\boldsymbol{D} = \begin{bmatrix} \boldsymbol{d}^{(1)} & \boldsymbol{d}^{(2)} \end{bmatrix}$ where $\boldsymbol{d}_j^{(1)} = 1$ if age $t_j \geq t_*$ and $\boldsymbol{d}_j^{(2)} = 1$ if $t_j \geq t_{**}$. Throwing $\boldsymbol{D}$ into the regression allows for different intercepts in these different age groups.

In order to allow for the slopes with respect to t to vary too, we need a matrix $\boldsymbol{E}$, again consisting of 2 columns, so that $e_{j1} = t_j$ if $t_j \geq t^*$ and 0 otherwise; and $e_{j2} = t_j$ if if $t_j \geq t_{**}$, and 0 otherwise. Each column of $\boldsymbol{E}$ is the corresponding column of $\boldsymbol{D}$ element-wise multiplied with $\boldsymbol{t}$, i.e., $\boldsymbol{E} = \begin{bmatrix} \boldsymbol{d}^{(1)} * \boldsymbol{t} & \boldsymbol{d}^{(2)} * \boldsymbol{t} \end{bmatrix}$.

If one then writes the model as $\mathbf{y} = \boldsymbol{D}\boldsymbol{\gamma} + \boldsymbol{E}\boldsymbol{\delta} + \boldsymbol{X}\boldsymbol{\beta}$ one gets an unconstrained model with 3 different slopes and 3 different intercepts. Assume for example there

are 3 observations in the first age group, 2 in the second, and 4 in the third, then

(45.2.2)
$$\boldsymbol{D} = \begin{bmatrix} \boldsymbol{d}^{(1)} & \boldsymbol{d}^{(2)} \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \qquad \boldsymbol{E} = \begin{bmatrix} \boldsymbol{d}^{(1)} * \boldsymbol{t} & \boldsymbol{d}^{(2)} * \boldsymbol{t} \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ t_4 & 0 \\ t_5 & 0 \\ t_6 & t_6 \\ t_7 & t_7 \\ t_8 & t_8 \\ t_9 & t_9 \end{bmatrix} \qquad \boldsymbol{X} = \begin{bmatrix} \boldsymbol{\iota} & t & \boldsymbol{x} \end{bmatrix}$$

Written column by column:

$$\mathfrak{y} = \beta_1 + \beta_2 \boldsymbol{t} + \beta_3 \boldsymbol{x} + \gamma_1 \boldsymbol{d}^{(1)} + \delta_1 \boldsymbol{d}^{(1)} * \boldsymbol{t} + \gamma_2 \boldsymbol{d}^{(2)} + \delta_2 \boldsymbol{d}^{(2)} * \boldsymbol{t} + \boldsymbol{\varepsilon} \tag{45.2.3}$$

This is how [**Gre97**, equation (8.3) on p. 389] should be understood. The jth observation has the form

$$\mathfrak{y}_j = \beta_1 + \beta_2 t_j + \beta_3 x_j + \gamma_1 d_j^{(1)} + \delta_1 d_j^{(1)} t_j + \gamma_2 d_j^{(2)} + \delta_2 d_j^{(2)} t_j + \varepsilon_j \tag{45.2.4}$$

An observation at the year t_* has, according to the formula for $\geq t_*$, the form

$$\mathfrak{y}_* = \beta_1 + \beta_2 t_* + \beta_3 x_* + \gamma_1 + \delta_1 t_* + \varepsilon_* \tag{45.2.5}$$

but had the formula for $< t_*$ still applied, the equation would have been

$$\mathfrak{y}_* = \beta_1 + \beta_2 t_* + \beta_3 x_* + \varepsilon_* \tag{45.2.6}$$

For these two equations to be equal, which means that the two regression lines intersect at x_*, we have to impose the constraint $\gamma_1 + \delta_1 t_* = 0$

Similarly, an observation at the year t_{**} has, according to the formula for $\geq t_{**}$, the form

$$\mathfrak{y}_{**} = \beta_1 + \beta_2 t_{**} + \beta_3 x_{**} + \gamma_1 + \gamma_2 + \delta_1 t_{**} + \delta_2 t_{**} + \varepsilon_{**} \tag{45.2.7}$$

but had the formula for $< t_{**}$ still applied, it would have been

$$\mathfrak{y}_{**} = \beta_1 + \beta_2 t_{**} + \beta_3 x_{**} + \gamma_1 + \delta_1 t_{**} + \varepsilon_{**} \tag{45.2.8}$$

Again equality of these two representations requires $\gamma_2 + \delta_2 t_{**} = 0$.

These two constraints can be written as

$$\begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix} = - \begin{bmatrix} t_* & 0 \\ 0 & t_{**} \end{bmatrix} \begin{bmatrix} \delta_1 \\ \delta_2 \end{bmatrix} \qquad \text{or} \quad \boldsymbol{\gamma} = -\boldsymbol{W}\boldsymbol{\delta}, \quad \text{where} \quad \boldsymbol{W} = \begin{bmatrix} t_* & 0 \\ 0 & t_{**} \end{bmatrix}. \tag{45.2.9}$$

Plugging this into $\mathfrak{y} = \boldsymbol{D}\boldsymbol{\gamma} + \boldsymbol{E}\boldsymbol{\delta} + \boldsymbol{X}\boldsymbol{\beta}$ gives $\mathfrak{y} = -\boldsymbol{D}\boldsymbol{W}\boldsymbol{\delta} + \boldsymbol{E}\boldsymbol{\delta} + \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{F}\boldsymbol{\delta} + \boldsymbol{X}\boldsymbol{\beta}$ where $\boldsymbol{F}$ is a dummy matrix of the form

$$\boldsymbol{F} = \boldsymbol{E} - \boldsymbol{D}\boldsymbol{W} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ t_4 - t_* & 0 \\ t_5 - t_* & 0 \\ t_6 - t_* & t_6 - t_{**} \\ t_7 - t_* & t_7 - t_{**} \\ t_8 - t_* & t_8 - t_{**} \\ t_9 - t_* & t_9 - t_{**} \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 2 & 0 \\ 3 & 1 \\ 4 & 2 \\ 5 & 3 \\ 6 & 4 \end{bmatrix} \tag{45.2.10}$$

Since each column is a continuous function of time without jumps, it is clear that the fitted values are also continuous functions of time.

Here are a few remarks about intrinsically linear functions, following [**Gre97**, (8.4) on p. 390]; Say $\mathfrak{y}$ is the vector of observations of the *dependent* variable, and $\boldsymbol{Z}$ is the data matrix of the *explanatory variables*, for instance in the above example $\boldsymbol{Z}$ would contain only $\boldsymbol{t}$, $\boldsymbol{x}$, and perhaps a categorical variable denoting the different age groups. In the above example this categorical variable can be formed as a function of $\boldsymbol{t}$; but in other dummy variable settings such a categorical variable is necessary. Now the *regressand* is not necessarily $\mathfrak{y}$ but may be a transformation $g(\mathfrak{y})$, and the k *regressors* have the form $f_i(\boldsymbol{Z})$ where the functions f_i are linearly independent. For instance $f_1(\boldsymbol{Z}) = \begin{bmatrix} z_{11} & z_{21} & \dots \end{bmatrix}^\top$ may pick out the first column of $\boldsymbol{Z}$, and $f_2(\boldsymbol{Z}) = \begin{bmatrix} z_{11}^2 & z_{21}^2 & \dots \end{bmatrix}^\top$ the square of the first column. The functions g and f_i define the relationship between the given economic variables and the variables in the regression. [**Gre97**, Definition 81 on p. 396] says something about the relationship between the parameters of interest and the regression coefficients: if the k regression coefficients $\beta_1, \dots, \beta_k$ can be written as k one-to-one possibly nonlinear functions of the k underlying parameters $\theta_1, \dots, \theta_k$, then the model is intrinsically linear in $\boldsymbol{\theta}$.

[**Gre97**, p. 391/2] brings the example of a regression with an interaction term:

$$\mathfrak{y} = \boldsymbol{\iota}\beta_1 + \boldsymbol{s}\beta_2 + \boldsymbol{w}\beta_3 + \boldsymbol{s} * \boldsymbol{w}\beta_4 + \boldsymbol{\varepsilon} \tag{45.2.11}$$

Say the underlying parameters of interest are $\frac{\partial \operatorname{E}[\mathfrak{y}_t]}{\partial s_t} = \beta_2 + \beta_4 w_t$, $\frac{\partial \operatorname{E}[\mathfrak{y}_t]}{\partial w_t} = \beta_3 + \beta_4 s_t$, and the second derivative $\frac{\partial^2 \operatorname{E}[\mathfrak{y}_t]}{\partial s_t \partial w_t} = \beta_4$. Here the first parameter of interest depends on the value of the explanatory variables, and one has to select a value; usually one takes the mean or some other central value, but for braking distance some extreme value may be more interesting.

[**Gre97**, example 8.4 on p. 396] is a maximum likelihood model that can also be estimated as an intrinsically linear regression model. I did not find the reference where he discussed this earlier, perhaps I have to look in the earlier edition. Here

maximum likelihood is far better. Greene asks why and answers: least squares does not use one of the sufficient statistics.

[**Gre97**, example 8.5 on p. 397/8] starts with a CES production function, then makes a Taylor development, and this Taylor development is an intrinsically linear regression of the 4 parameters involved. Greene computes the Jacobian matrix necessary to get the variances. He compares that with doing nonlinear least squares on the production function directly, and gets widely divergent parameter estimates.

45.2.5. Smoothing Splines. This seems the most promising approach. If one estimates a function by a polynomial of order m or degree $m-1$, then this means that one sets the mth derivative zero. An approximation to a polynomial would be a function whose mth derivative is small. We will no longer assume that the fitting functions are themselves polynomials, but we will assume that $f \in W^m[a,b]$ which means f itself and its derivatives up to and including the $m-1$st derivative are absolutely continuous over a closed and bounded interval $[a,b]$, and the mth derivative is square integrable over $[a,b]$.

If we allow such a general f, then the estimation criterion can no longer be the minimization of the sum of squared errors, because in this case one could simply choose an interpolant of the data, i.e., a f which satisfies $f(x_i) = \mathsf{y}_i$ for all i. Instead, the estimation criterion must be a constrained or penalized least squares criterion (analogous to OLS with an exact or random linear constraint) which has a penalty for the mth order derivative. The idea of smoothing splines is to minimize the objective function

$$\big(\mathsf{y} - f(\boldsymbol{x})\big)^\top \big(\mathsf{y} - f(\boldsymbol{x})\big) + \lambda \int_a^b \big(f^{(m)}(x)\big)^2\, dx \tag{45.2.12}$$

Of course, only the values which f takes on the observed x_i are relevant; but for each sequence of observations there is one polynomial which minimizes this objective function, and this is *a natural spline with the observed values as breakpoints.*

45.2.6. Local regression, Kernel Operators. A different approach is to run locally weighted regressions. Here the response surface at a given value of the independent variable is estimated by a linear regressions which only encompasses the points in the neighborhood of the independent variable. `Splus` command `loess`.

If this local regression only has an intercept, it is also known as a "kernel smoother." But locally linear smoothers perform better at the borders of the sample than locally constant ones.

45.3. More than One Explanatory Variable: Backfitting

Until now we restricted the discussion to the case of one explanatory variable. The above discussion can be extended to smooth functions of several explanatory variables, but this numerically very complex, and the results are hard to interpret.

But in many real-life situations one has to do with several additive effects without interaction. This is much easier to estimate and to interpret.

One procedure here is "projection pursuit regression" [**FS81**]. Denoting the ith row of $\boldsymbol{X}$ with $\boldsymbol{x}_i$, the model is

$$\mathsf{y}_i = \sum_{j=1}^k f_j(\boldsymbol{\alpha}_j^\top \boldsymbol{x}_i) + \varepsilon_i, \tag{45.3.1}$$

which can also be written as

$$\mathsf{y} = \sum_{j=1}^k f_j(\boldsymbol{X}\boldsymbol{\alpha}_j) + \boldsymbol{\varepsilon} \tag{45.3.2}$$

Here one estimates k arbitrary functions of certain linear combinations of the explanatory variables $\boldsymbol{\alpha}_j^\top \boldsymbol{x}_i$ along with the linear combinations themselves. This is implemented in `Splus` in the function `ppreg`.

The matter will be easier if one already knows that the columns of the $\boldsymbol{X}$-matrix are the relevant variables and only their transformation has to be estimated. This gives the additive model

$$\mathsf{y} = \sum_{j=1}^k f_j(\boldsymbol{x}_j) + \boldsymbol{\varepsilon}, \tag{45.3.3}$$

where $\boldsymbol{x}_j$ is the jth *column* of $\boldsymbol{X}$. The beauty is that one can specify here different univariate smoothing techniques for the individual variables and then combine it all into a joint fit by the method of *back-substitution.* Back-substitution is an iterative procedure by which one obtains the joint fit by an iteration only involving fits on one explanatory variable each. One starts with some initial set of functions $f_i^{(0)}$ and then, cycling through $j = 1, \ldots, k, 1, \ldots, k, \ldots$ one fits the residual $\mathsf{y} - \sum_{k \neq j} f_k(\boldsymbol{x}_k)$ as a function of $\boldsymbol{x}_j$. This looks like a crude heuristic device, but it has a deep theoretical justification.

If the fitting procedure, with respect to the jth explanatory variable, can be written as $\hat{\mathsf{y}}_j = \boldsymbol{S}_j \boldsymbol{x}_j$ (but the more common notation is to write it $\boldsymbol{f}_j = \boldsymbol{S}_j \boldsymbol{x}_j$), then this backfitting works because the joint fit $\mathsf{y} = \boldsymbol{f}_1 + \cdots + \boldsymbol{f}_k + \hat{\boldsymbol{\varepsilon}}$ is a solution

of the equation

$$\begin{bmatrix} \boldsymbol{I} & \boldsymbol{S}_1 & \boldsymbol{S}_1 & \cdots & \boldsymbol{S}_1 \\ \boldsymbol{S}_2 & \boldsymbol{I} & \boldsymbol{S}_2 & \cdots & \boldsymbol{S}_2 \\ \boldsymbol{S}_3 & \boldsymbol{S}_3 & \boldsymbol{I} & \cdots & \boldsymbol{S}_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{S}_k & \boldsymbol{S}_k & \boldsymbol{S}_k & \cdots & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{f}_1 \\ \boldsymbol{f}_2 \\ \boldsymbol{f}_3 \\ \vdots \\ \boldsymbol{f}_k \end{bmatrix} = \begin{bmatrix} \boldsymbol{S}_1 \mathfrak{y} \\ \boldsymbol{S}_2 \mathfrak{y} \\ \boldsymbol{S}_3 \mathfrak{y} \\ \vdots \\ \boldsymbol{S}_k \mathfrak{y} \end{bmatrix}, \tag{45.3.4}$$

and the iteration is known as a numerical iteration procedure to solve this system of equations, called the Gauss-Seidel algorithm.

In the case of an OLS fit, in which one estimates k parameters $\hat{\beta}_j$, $j = 1, \ldots, k$, so that $\boldsymbol{f}_j = \boldsymbol{x}_j \hat{\beta}_j$, the univariate projection functions are $\boldsymbol{S}_j = \boldsymbol{x}_j (\boldsymbol{x}_j^\top \boldsymbol{x}_j)^{-1} \boldsymbol{x}_j^\top$, therefore the estimation equation reads

(45.3.5)

$$\begin{bmatrix} \boldsymbol{I} & \boldsymbol{x}_1(\boldsymbol{x}_1^\top \boldsymbol{x}_1)^{-1}\boldsymbol{x}_1^\top & \boldsymbol{x}_1(\boldsymbol{x}_1^\top \boldsymbol{x}_1)^{-1}\boldsymbol{x}_1^\top & \cdots & \boldsymbol{x}_1(\boldsymbol{x}_1^\top \boldsymbol{x}_1)^{-1}\boldsymbol{x}_1^\top \\ \boldsymbol{x}_2(\boldsymbol{x}_2^\top \boldsymbol{x}_2)^{-1}\boldsymbol{x}_2^\top & \boldsymbol{I} & \boldsymbol{x}_2(\boldsymbol{x}_2^\top \boldsymbol{x}_2)^{-1}\boldsymbol{x}_2^\top & \cdots & \boldsymbol{x}_2(\boldsymbol{x}_2^\top \boldsymbol{x}_2)^{-1}\boldsymbol{x}_2^\top \\ \boldsymbol{x}_3(\boldsymbol{x}_3^\top \boldsymbol{x}_3)^{-1}\boldsymbol{x}_3^\top & \boldsymbol{x}_3(\boldsymbol{x}_3^\top \boldsymbol{x}_3)^{-1}\boldsymbol{x}_3^\top & \boldsymbol{I} & \cdots & \boldsymbol{x}_3(\boldsymbol{x}_3^\top \boldsymbol{x}_3)^{-1}\boldsymbol{x}_3^\top \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{x}_k(\boldsymbol{x}_k^\top \boldsymbol{x}_k)^{-1}\boldsymbol{x}_k^\top & \boldsymbol{x}_k(\boldsymbol{x}_k^\top \boldsymbol{x}_k)^{-1}\boldsymbol{x}_k^\top & \boldsymbol{x}_k(\boldsymbol{x}_k^\top \boldsymbol{x}_k)^{-1}\boldsymbol{x}_k^\top & \cdots & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_1 \hat{\beta}_1 \\ \boldsymbol{x}_2 \hat{\beta}_2 \\ \boldsymbol{x}_3 \hat{\beta}_3 \\ \vdots \\ \boldsymbol{x}_k \hat{\beta}_k \end{bmatrix} =$$

It can be shown that this equation is equivalent to the OLS normal equations.

CHAPTER 46

Transformation of the Response Variable

So far we have concentrated on transformations of the predictor variables. If one transforms the response variable then one will also get different variances and covariances of the error terms. We will first discuss a procedure which treats the predictor variables and the response variables symmetrically, and then we will look at transformations of the response variable which give justice to the special position the response variable has in the regression model.

46.1. Alternating Least Squares and Alternating Conditional Expectations

In this section we will discuss a technique realized in the `ace` procedure of `Splus`. The classic reference is [**BF85**] (although they emphasize too much a certain implementation, the use of fast smoothers, instead of giving the general theory). a good survey is [**HT90**, Chapter 7], and [**Buj90**] has some interesting meta-analysis: he points out that this methodology is related to canonical correlation, optimal scoring, dual scaling, reciprocal averaging, simultaneous linear regression, alternating least squares, correspondence analysis, nonlinear multivariate analysis, and homogeneity analysis. This is why some of the pathologies of ACE arise. He cites [**VdBDL83**] as a central basic article.

46.1.1. ACE with one response and just one predictor variable. Assume x and y are two random variables (but they may also be categorical variables or random vectors). Their *maximal correlation* $\operatorname{corr}^*[\mathsf{x},\mathsf{y}]$ is the maximal value of $\operatorname{corr}[\phi(\mathsf{x}),\theta(\mathsf{y})]$, where ϕ and θ are two real-valued mappings of the space in which x and y are defined, with $0<\operatorname{var}[\phi(\mathsf{x})]<\infty$ and $0<\operatorname{var}[\theta(\mathsf{y})]<\infty$. The maximal correlation has the following three properties:

- $0\leq\operatorname{corr}^*[\mathsf{x},\mathsf{y}]\leq 1$. (Note that the usual correlation coefficient is between -1 and 1.)
- $\operatorname{corr}^*[\mathsf{x},\mathsf{y}]=0$ if and only if x and y are independent.
- $\operatorname{corr}^*[\mathsf{x},\mathsf{y}]=1$ if and only if two functions u and v exist with $u(\mathsf{x})=v(\mathsf{y})$ and $\operatorname{var}[u(\mathsf{x})]=\operatorname{var}[v(\mathsf{y})]>0$.

The functions ϕ and θ can be understood to describe the functional, as opposed to the stochastic, relationship between the variables.

If the two variables are jointly normal, then their correlation coefficient is at the same time their maximal correlation. (But the jointly normal is not the only distribution with this property, see [**Buj90**].) If you start with two jointly normal variables, it is therefore not possible to find transformations of each variable separately which increase their correlation. This can also be formulated as follows: If $\begin{bmatrix}\mathsf{x} & \mathsf{y}\end{bmatrix}^\top$ has a bivariate normal distribution, and ϕ and θ are two functions of one variable each, with $0<\operatorname{var}[\phi(\mathsf{x})]<\infty$ and $0<\operatorname{var}[\theta(\mathsf{y})]<\infty$, then $\operatorname{corr}[\phi(\mathsf{x}),\theta(\mathsf{y})]\leq\operatorname{corr}[\mathsf{x},\mathsf{y}]$. Proof in [**KS79**, section 33.44]. Often therefore the ϕ and θ which give the highest correlation between two variables are very similar to those univariate transformations which make each variable separately as normal as possible.

Note that such transformations can be applied to categorical variables too. Say you have two categorical random variables. Knowing their joint distribution amounts to knowing for every cell, i.e., for every pair of possible outcomes of these categories, the probability that this cell is reached. The sample equivalent would be a contingency table. Then one can ask: which "scores" does one have to assign to the levels of each of the two categories so that the resulting real-valued random variables have maximal correlation? This can be solved by an eigenvalue problem. This is discussed in [**KS79**, sections 33.47–49].

How can one find the optimal θ and ψ in the continuous case? Since correlation coefficients are invariant under affine transformations, such optimal transformations are unique only up to a constant coefficient and an intercept. Here without proof the following procedure, called "alternative conditional expectations:" let ϕ_1 and θ_1 be the identical functions $\phi_1(\mathsf{x})=\mathsf{x}$ and $\theta_1(\mathsf{y})=\mathsf{y}$. Then do recursively for $i=2,\dots$ the following: $\phi_i(\mathsf{x})=\operatorname{E}[\theta_{i-1}(\mathsf{y})|\mathsf{x}]$ and $\theta_i(\mathsf{y})=\operatorname{E}[\phi_i(\mathsf{x})|\mathsf{y}]$. Remember that $\operatorname{E}[\mathbf{y}|\mathbf{x}]$ is a function of $\mathbf{x}$, and this function will be $\phi_2(\mathbf{x})$. In order to prevent this recursion to become an increasingly steep or flat line, one does not exactly use this recursion but rescales one of the variables, say θ, after each step so that it has zero mean and unit variance.

46.1.2. ACE with more than 2 Variables. How can that be generalized to a multivariate situation? Let us look at the case where y remains a scalar but $\mathbf{x}$ is a k-vector. One can immediately speak of their maximal correlation again if one maximizes over functions ϕ of one variable and θ of k variables. In the case of joint normality, the above result generalizes to the following: the optimal ϕ can be chosen to be the identity, and the optimal θ is linear; it can be cvhosen to be the best linear predictor.

In the case of several variables, one can also ask for second-best and third-best etc. solutions, which are required to be uncorrelated with the better solutions and

maximize the correlation subject to this constraint. They can in principle already be defined if both variables are univariate, but in this case they are usually just simple polynomials in the best solutions. In the multivariate case, these next-best solutions may be of interest of their own. Not only the optimal but also these next-best transformations give rise to linear regressions (Buja and Kass, Comment to [**BF85**], p. 602).

46.1.3. Restrictions of the Functions over which to Maximize. If one looks at several variables, this procedure of maximizing the correlation is also interesting if one restricts the classes of functions to maximize.

The *linear* (or, to be more precise, *affine*) counterpart of maximal correlation is already known to us. The best linear predictor can be characterized as that linear combination of the components of $\mathbf{x}$ which has maximal correlation with y. The maximum value of this correlation coefficient is called the multiple correlation coefficient.

An in-between step between the set of all functions and the set of all linear functions is the one realized in the `ace` procedure in `Splus`. It uses all those functions of k variables with can be written as *linear combinations* or, without loss of generality, as *sums* of functions of one variable each. Therefore one wants functions $\phi_1, \dots, \phi_k$ and θ which maximize the correlation of $\phi_1(\mathsf{x}_1) + \cdots + \phi_k(\mathsf{x}_k)$ with $\theta(\mathsf{y})$. This can again be done by "backfitting," which is a simple recursive algorithm using only bivariate conditional expectations at every step. Each step does the following: for the given best estimates of $\phi_1, \dots, \phi_{i-1}, \phi_{i+1}, \cdots, \phi_k$ and θ one gets the best estimate of ϕ_i as $\phi_i(\mathsf{x}_i) = \mathrm{E}[\theta(\mathsf{y}) - \sum_{j:\ j\neq i} \phi_j(\mathsf{x}_j)|\mathsf{x}_i]$.

If one does not know the joint distributions but has samples then one can replace the conditonal expectations by a "Smoother" using the datapoints. One such procedure is the function `supsmu` in `Splus`, described in [**HT90**, p. 70]. Functions will not be given in closed form but one gets their graph by plotting the untransformed against the transformed variables.

46.1.4. Cautions About the ace Procedure. There are certain features which one should be aware of before using this procedure.

First, this is a procedure which treats both variables symmetrically. The regression model between the variables is not a fixed point. If the variables satisfy the regression specification $\mathsf{y} = \boldsymbol{\beta}^\top \mathbf{x} + \varepsilon$ with $\boldsymbol{\varepsilon}$ independent of the vector $\mathbf{x}$, then the optimal transformations will not be the simple multiples of the components of $\boldsymbol{\beta}$, although they will usually be close to them. This symmetry makes `ace` more appropriate for general multivariate models, like correlation analysis, than for regression. The `avas` procedure, which will be discussed next, is a modification of `ace` which seems to work better for regression.

Secondly, there are situations in which the functions of x and y which have highest correlation are not very interesting functions.

Here is an example in which the function with the highest correlation may be uninteresting. If y and one of the x_i change sign together, one gets correlation of 1 by predicting the sign of y by the sign of x_i and ignoring all other components of $\mathbf{x}$.

Here is another example in which the function with the highest correlation may be uninteresting. Let $\begin{bmatrix}\mathsf{x} & \mathsf{y}\end{bmatrix}^\top$ be a mixture consisting of

$$\begin{bmatrix}\mathsf{x} & \mathsf{y}\end{bmatrix}^\top = \begin{cases} \begin{bmatrix}\mathsf{x}' & \mathsf{y}'\end{bmatrix}^\top & \text{with probability } 1-\alpha \\ \begin{bmatrix}\mathsf{x}'' & \mathsf{y}''\end{bmatrix}^\top & \text{with probability } \alpha \end{cases} \tag{46.1.1}$$

where x' and y' are *independent* random variables which have density functions, while x'' and y'' are discrete, and let $D_{\mathsf{x}''}$ and $D_{\mathsf{y}''}$ be the supports of them, i.e., the finite sets of values which these variables are able to assume. One would expect the maximal correlation to converge toward zero if $\alpha \to 0$, but in reality, as long as $\alpha > 0$ the maximum correlation is always equal to one, even if x'' and y'' are independent of each other. The functions which achieve this are the indicator functions $\phi = I[\mathsf{x} \in D_{\mathsf{x}''}]$ and $\theta = I[\mathsf{y} \in D_{\mathsf{y}''}]$. In other words, the functions which have the highest correlations may be uninteresting. But in this case it is clear that one should also look for the second and third choices. This is one of the remedies proposed in [**BF85**].

Another potential source of trouble is that the optimal functions are not always uniquely determined. Or sometimes, the eigenvalues of optimal and next-best solutions cross each other, i.e., in a continuous modification of the data one will get abrupt changes in the optimal transformations. All this is alleviated if one not only looks at the optimal functions, but also the second-best solutions.

An automated procedure such as `ace` may lead to strange results due to errors in setting up the problem, errors which one would easily catch if one had to do it by hand. This is not a particularity of `ace` itself but a danger of any automated procedure. (E.g., people run regressions without looking at the data.) The example Prebibon and Vardi in their comment to [**BF85**] on p. 600 is interesting: If the plot consists of two parallel regression lines, one would, if one did it by hand, never dream of applying a transformation, but one would look for the additional variable distinguishing the two regimes. An automatic application of `ace` gives a zig-zag line, see figure 2 on p. 600.

Of course, `ace` makes all significance levels in the ensuing regression invalid. Tradeoff between parametric and nonparametric methods.

46.2. Additivity and Variance Stabilizing Transformations (avas)

This is a modification of `ace` which seems to be more appropriate for regression, although it does not have the nice theoretical foundation which `ace` has in the theory of maximal correlation. [**Tib88**] is main source about the `avas` method.

Again let's discuss the simplest case with two variables first. Assume x and y are such that there exist two functions ϕ and θ so that one can write

$$\theta(\mathsf{y}) = \phi(\mathsf{x}) + \varepsilon \tag{46.2.1}$$

where $\mathrm{E}[\varepsilon] = 0$ and ε independent of x. Then our purpose is to find these functions. `ace` does not really do that because the correlation is not necessarily maximized by these transformations. But from (46.2.1) follow the following two necessary, but not sufficient conditions:

$$\mathrm{E}[\theta(\mathsf{y})|\mathsf{x}] = \phi(\mathsf{x}) \tag{46.2.2}$$

$$\mathrm{var}[\theta(\mathsf{y})|\phi(\mathsf{x})] = \text{constant}. \tag{46.2.3}$$

There is nothing natural about these conditions other than that they can be implemented numerically and that their iteration usually finds the stationary values.

The following iterative procedure corresponds to this: let ϕ_1 and θ_1 be the identical functions $\phi_1(\mathsf{x}) = \mathsf{x}$ and $\theta_1(\mathsf{y}) = \mathsf{y}$. Then do recursively for $i = 2, \ldots$ the following: The transformation involving x is the same as in the ACE procedure: $\phi_i(\mathsf{x}) = \mathrm{E}[\theta_{i-1}(\mathsf{y})|\mathsf{x}]$. But the transformation of y is different: $\theta_i(\mathsf{y})$ is a transformation of $\theta_{i-1}(\mathsf{y})$ which attempts to make $\mathrm{var}[\theta_i(\mathsf{y})|\phi_i(\mathsf{x}) = u]$ independent of u. You need three steps to construct it:

- Compute $v_i(u) = \mathrm{var}[\theta_{i-1}(\mathsf{y})|\phi_i(\mathsf{x}) = u]$.
- Compute $h_i(t) = \int_0^t \frac{du}{\sqrt{v_i(u)}}$. In other words, it is a function whose derivative is $h_i'(t) = \frac{1}{\sqrt{v_i(t)}}$.
- Set $\theta_i(\mathsf{y}) = h_i(\theta_{i-1}(\mathsf{y}))$.

We want to show that this transformation indeed stabilizes the variance. Let us first see how one can (asymptotically) obtain the variance of $h(\mathsf{z})$: let $u = \mathrm{E}[\mathsf{z}]$, make Taylor development of h around u: $h(z) = h(u) + h'(u)(z - u)$, therefore asymptotically $\mathrm{var}[h(\mathsf{z})] = (h'(u))^2 \,\mathrm{var}[\mathsf{z}]$.

To apply this for our purposes, pick a certain value of u. Make a Taylor development of $\theta_i(y) = h_i(\theta_{i-1}(y))$ around $\mathrm{E}[\theta_{i-1}(\mathsf{y})|\phi_i(\mathsf{x}) = u] = u$, which reads $\theta_i(y) = h_i(u) + h_i'(u)(\theta_{i-1}(y) - u)$. Therefore $\mathrm{var}[\theta_i(\mathsf{y})|\phi_i(\mathsf{x}) = u] = (h_i'(u))^2 \mathrm{var}[\theta_{i-1}(\mathsf{y})|\phi_i(\mathsf{x}) = u] = \frac{1}{v_i(u)} v_i(u) = 1$. This asymptotic expression for the variance is independent of the u chosen.

In this procedure, therefore, only the transformations of the independent variables are designed to achieve linearity. Those of the dependent variables are designed

to equalize the variance. This is a rule of thumb one should always consider in selecting transformations: It makes sense to use transformations of the y axis to get homoskedasticity, and then transformations of the x axis for straightening out the regression line.

In the case of several predictor variables, the same "backfitting" procedure is used which `ace` uses.

Again this is not exactly the iterative procedure chosen; in order to avoid ambiguity in the result, the `avas` procedure normalizes at each step the function θ so that it has zero mean and unit variance.

46.3. Comparing ace and avas

Note the following differences between `ace` and `avas`: (1) in `avas`, the tranformation of the dependent variable θ is by construction a strictly increasing function, while the `ace` θ is not necessarily monotone. (2) If the joint density function of x and y is concentrated in the first and third quadrant only, say, then `ace` just uses the sign of x to predict the sign of y with correlation one. This "collapsing" does not happen withe `avas`, which fits the two areas separately. (3) In the two regression lines case where `ace` gave a Z-shaped function, `avas` gives a linear function, i.e., it takes the mean over both regression lines. One might wonder which is better.

Both methods find certain transformations which satisfy certain mathematical criteria. These criteria may not always be the ones one is most interested in, but in many cases they are. These transformations are not given in a closed form but the transformed values of the given data are computed by a numerical procedure. Their suggested use is to plot the transformed against the untransformed data in order to see which transformations the data ask for, and let this be a guide for choosing simple analytical transformations (log, square root, polynomials, etc.) If these transformations give funny results, this may be a diagnostic tool regarding the model.

Problem 445. *8 points What does one have to pay attention to if one transforms data in a regression equation? Discuss methods in which the data decide about the functional form of the transformation.*

CHAPTER 47

Density Estimation

[**Sim96**] is a very encompassing text. A more elementary introduction with good explanations is [**WJ95**]. This also has some plots with datasets relevant to economics, see pp. 1, 11, and there is a R and `Splus`-package called `KernSmooth` associated with it (but this package does not contain the datasets). A more applied book is [**BA97**], which goes together with the R and `Splus`-package `sm`.

47.1. How to Measure the Precision of a Density Estimator

Let $\hat{\mathsf{f}}$ be the estimated density and f the true density. Then for every fixed value u, the estimation error at u is $\hat{\mathsf{f}}(u) - f(u)$. This is a random variable which depends on u as a nonrandom parameter. Its expected value is the bias at u $\mathrm{E}[\hat{\mathsf{f}}(u) - f(u)]$, and the expected value of its square is the MSE at u $\mathrm{E}\big[\big(\hat{\mathsf{f}}(u) - f(u)\big)^2\big]$. This is a measure of the precision of the density estimate at point u only.

The *overall* deviation of the estimated density from the true density can be measured by the integrated squared error (ISE) $\int_{u=-\infty}^{+\infty}\big(\hat{\mathsf{f}}(u) - f(u)\big)^2\,du$, This is a random variable; for each observation vector, it gives a different number. The mean integrated squared error (MISE) is the expected value of the ISE, and at the same time (as long as integration and formation of the expected value can be interchanged) it is the integral of the MSE at u over all us: MISE= $\int_{u=-\infty}^{+\infty}\big(\hat{\mathsf{f}}(u) - f(u)\big)^2\,du = \int_{u=-\infty}^{+\infty}\mathrm{E}\big[\big(\hat{\mathsf{f}}(u) - f(u)\big)^2\big]\,du$. The asymptotic value of this is called AMISE.

The MSE is the variance plus the squared bias. In density estimation, bias arises if one overmoothes, and variance increases if one undersmoothes.

47.2. The Histogram

Histograms are density estimates. They are easy to understand, easy to construct, and do not require advanced graphical tools.

Here the number of bins is important. Too few bins lead to oversmoothing, too many to undersmoothing. [**Sim96**, p. 16] has some math how to compute the MISE of a histogram, and which bin size is optimal. If the underlying distribution

is Normal, the optimal bin width is

$$h = 3.491\sigma n^{-1/3}$$

This is often used also for non-Normal distributions, but if these distributions are bimodal, then one needs narrower bins. The R/S-function `dpih` (which stands for Direct Plug In Histogram) in the library `KernSmooth` uses more sophisticated methods to select an optimal bin width.

Also the anchor positions can have a big impact on the appearance of a histogram. To demonstrate this, `cd /usr/share/ecmet/xlispstat/anchor-position` then do xlispstat, then `(load "fde")`, then `(fde-demo)`, and pick `animate anchor-moving`.

Regarding the labeling on the vertical axis of a histogram there is a naive and a more sophisticated approach. The naive approach gives the number of data points in each bin. The preferred, more sophisticated approach is to divide the total number of points in each bin by the overall size of the dataset and by the bin width. In this way one gets the *relative frequency density*. With this normalization, the total area under the histogram is 1 and the histogram is directly comparable with other estimates of the probability density function.

47.3. The Frequency Polygon

Derived from histogram by connecting the mid-points of each bin. Gives a better approximation to the actual density. Now the optimal bin width for a Normal is

$$h = 2.15\sigma n^{-1/5}$$

Dominates the histogram, and is not really more difficult to construct. Simonoff argues that one should never draw histograms, only frequency polygons.

47.4. Kernel Densities

For every observation draw a standard Normal with that point as the mode, and then add them up. An illustration is `sm.script(sp_build)`. It can also be a Normal with variance different than 1; the greater the variance, the smoother the density estimate. Instead of the Normal density one can also take other smoothing kernels, i.e., functions k with $\int_{u=-\infty}^{+\infty} k(u)\,du = 1$ and $\int_{u=-\infty}^{+\infty} uk(u)\,du = 0$. An often-used kernel is the Triweight kernel $\frac{35}{32}(1-x^2)^3$ for $|x| \leq 1$ and 0 otherwise, but these kernel functions may also assume negative values (in which case they are no longer densities). The choice of the functional form of the kernel is much less important than the bandwidth, i.e., the variance of the kernel (if interpreted as a density function) $\mu_2 = \int_{u=-\infty}^{+\infty} u^2 k(u)\,du = 1$,

PROBLEM 446. *If $u \mapsto k(u)$ is the kernel, and $\mathsf{x} = \begin{bmatrix}\mathsf{x}_1 & \cdots & \mathsf{x}_n\end{bmatrix}^\top$ the data vector, then $\hat{\mathsf{f}}(u) = \frac{1}{n}\sum_{i=1}^n k(u-\mathsf{x}_i)$ is the kernel estimate of the density at u.*

- a. *3 points Compute the mean of the kernel estimator at u.*

ANSWER. $\mathrm{E}[\hat{\mathsf{f}}(u)] = \frac{1}{n}\sum_{i=1}^n \mathrm{E}[k(u-\mathsf{x}_i)]$ but since all x_i are assumed to come from the same distribution, it follows $\mathrm{E}[\hat{\mathsf{f}}(u)] = \mathrm{E}[k(u-\mathsf{x})] = \int_{x=-\infty}^{+\infty} k(u-x)f(x)\,dx$. □

- b. *4 points Assuming the x_i are independent, show that*

$$\mathrm{var}[\hat{\mathsf{f}}(u)] = \frac{1}{n}\Big(\int_{x=-\infty}^{+\infty} k^2(u-x)f(x)\,dx - \big(\int_{x=-\infty}^{+\infty} k(u-x)f(x)\,dx\big)^2\Big). \tag{47.4.1}$$

ANSWER.

$$\mathrm{var}[\hat{\mathsf{f}}(u)] = \frac{1}{n^2}\sum_{i=1}^n \mathrm{var}[k(u-\mathsf{x}_i)] \tag{47.4.2}$$

$$= \frac{1}{n}\,\mathrm{var}[k(u-\mathsf{x})] \tag{47.4.3}$$

$$= \frac{1}{n}\Big(\mathrm{E}\big[\big(k(u-\mathsf{x})\big)^2\big] - \big(\mathrm{E}[k(u-\mathsf{x})]\big)^2\Big) \tag{47.4.4}$$

$$= \frac{1}{n}\Big(\int_{x=-\infty}^{+\infty} k^2(u-x)f(x)\,dx - \big(\int_{x=-\infty}^{+\infty} k(u-x)f(x)\,dx\big)^2\Big). \tag{47.4.5}$$

□

47.5. Transformational Kernel Density Estimators

This approach transforms the data first, then estimates a density of the transformed data, and then re-transforms this density to the original scale. For instance the income distribution can use this, see [**WJ95**, p. 11].

47.6. Confidence Bands

The variance of a density estimate is usually easier to compute than the bias. One method to get a confidence band is to draw additional curves with 2 estimated pointwise standard deviations above and below the plot. This makes the assumption that the bias is 0. Therefore it is not really usable for inference, but it may give some idea whether certain features of the plot should be taken seriously. `sm.script(air_band)` Another approach is bootstrapping. `sm.script(air_boot)`. The expected value of the bootstrapped density functions is $\hat{f}$ (and not f; therefore bootstrapping will not reveal the bias but it does reveal the variance of the density estimate.

47.7. Other Approaches to Density Estimation

Variable bandwidth methods

Nearest Neighbor methods

Orthogonal Series Methods: project the data on an orthogonal base and only use the first few terms. Advantage: here one actually knows the functional form of the estimated density. See [**BA97**, pp. 19–21].

47.8. Two-and Three-Dimensional Densities

`sm.script(air_dens)` and `sm.script(air_imag)` give different representations of a two-dimensional density; `sm.script(air_cont)` gives the evolution over time (dotted line is first, dashed second, and solid line third).

`sm.script(mag_scat)` is the plot of a dataset containing 3-dimensional directions (longitude and latitude). Here is a kernel function and a smoothed representation of this dataset: `sm.script(mag_dens)`.

PROBLEM 447. *Write a function that translates the latitude and longitude data of the* `magrem` *dataset into a 3-dimensional dataset which can be loaded into xgobi.*

Here is a 3-dimensional rendering of the geyser data: `provide.data(geys3d)` and then `xgobi(geys3d)`. The script which draws a 3-dimensional density contour does not work right now: `sm.script(geys_td)`.

47.9. Other Characterizations of Distributions

Instead of the density function one can also give smoothed versions of the empirical cumulative distribution function, or of the hazard function $\frac{f(u)}{1-F(u)}$.

47.10. Quantile-Quantile Plots

The QQ-plot is a plot of the quantile functions, as defined in (3.4.14), of two different distributions against each other.

The graph of a cumulative distribution function is given in Figure 1, and the corresponding quantile function is given in Figure 2. The bullets on the beginning of the lines in the cumulative distribution function indicate that the line includes its infimum but not its supremum. The quantile function has the bullets at the end of the lines.

The "theoretical QQ plot" of two distributions which have distribution functions F_1 and F_2 and quantile functions F_1^{-1} and F_2^{-1} is the set of all $(x_1, x_2) \in \mathbb{R}^2$ for which there exists a p such that $x_1 = F_1^{-1}(p)$ and $x_2 = F_2^{-1}(p)$.

If both distributions are continuous and strictly increasing, then the theoretical QQ-plot is continuous as well. If the cumulative distribution functions have horizontal straight line segments, then the theoretical QQ-plot has gaps. If one of the

two distribution functions is a step function and the other is continuous, then the theoretical QQ-plot is a step function; and if both distribution functions are step functions, then the theoretical QQ-plot consists of isolated points.

Here is a practical instruction how to construct a QQ plot from the given cumulative distribution functions: Draw the cumulative distribution functions of the two distributions which you want to compare into the same diagram. Then, for every value p between 0 and 1 plot the abscisse of the intersection of the horizontal line with height p with the first cumulative distribution function against the abscisse of its intersection with the second. If there are horizontal line segments in these distribution functions, then the suprema of these line segments should be used. If the cumulative distribution functions is a step function stepping over p, then the value at which the step occurs should be used.

If the QQ-plot is a straight line, then the two distributions are either identical, or the underlying random variables differ only by a scale factor. The plots have special sensitivity regarding differences in the tail areas of the two distributions.

PROBLEM 448. *Let F_1 be the cumulative distribution function of random variable x_1, and F_2 that of the variable x_2 whose distribution is the same as that of $\alpha\mathsf{x}_1$, where α is a positive constant. Show that the theoretical QQ plot of these two distributions is contained in the straight line $q_2 = \alpha q_1$.*

ANSWER. $(x_1, x_2) \in$ QQ-plot $\iff$ a p exists with $x_1 = F_1^{-1}(p) = \inf\{u\colon \Pr[x_1 \le u] \ge p\}$ and $x_2 = F_2^{-1}(p) = \inf\{u\colon \Pr[x_2 \le u] \ge p\} = \inf\{u\colon \Pr[\alpha x_1 \le u] \ge p\}$. Write $v = u/\alpha$, i.e., $u = \alpha v$; then $x_2 = \inf\{\alpha v\colon \Pr[\alpha x_1 \le \alpha v] \ge p\} = \inf\{\alpha v\colon \Pr[x_1 \le v] \ge p\} = \alpha \inf\{v\colon \Pr[x_1 \le v] \ge p\} = \alpha x_1$.

□

In other words, if one makes a QQ plot of a normal with mean zero and variance 2 on the vertical axis against a normal with mean zero and variance 1 on the horizontal axis, one gets a straight line with slope 2. This makes such plots so valuable, since visual inspection can easily discriminate whether a curve is a straight line or not. To repeat, QQ plots have the great advantage that one only needs to know the correct distribution up to a scale factor!

QQ-plots can not only be used to compare two probability measures, but an important application is to decide whether a given *sample* comes from a given distribution by plotting the quantile function of the empirical distribution of the sample, compare (3.4.17). against the quantile function of the given cumulative distribution function. Since empirical cumulative distribution functions and quantile functions are step functions, the resulting theoretical QQ plot is also a step function.

In order to make it easier to compare this QQ plot with a straight line, one usually does not draw the full step function but one chooses one point on the face of each step, so that the plot contains one point per observation. This is like plotting

the given sample against a very regular sample from the given distribution. Where on the face of each step should one choose these points? One wants to choose that ordinate where the first step in an empirical cumulative distribution function should usually be.

It is a mathematically complicated problem to compute for instance the "usual location" (say, the "expected value") of the smallest of 50 normally distributed variables. But there is one simple method which gives roughly the right locations independently of the distribution used. Draw the cumulative distribution function (cdf) which you want to test against, and then draw between the zero line and the line $p = 1$ n parallel lines which divide the unit strip into $n + 1$ equidistant strips. The intersection points of these n lines with the cdf will roughly give the locations where the smallest, second smallest, etc., of a sample of n normally distributed observations should be found.

For a mathematical justification of this, make the following thought experiment. Assume you have n observations from a uniform distribution on the unit interval. Where should you expect the smallest observation to be? The answer is given by the simple result that the expected value of the smallest observation is $1/(n + 1)$, the expected value of the second-smallest observation is $2/(n + 1)$, etc. In other words, in the average, the n observations, cut the unit interval into $n + 1$ intervals of equal distance.

Therefore we do know where the first step of an empirical cumulative distribution function of a uniform random variable should be, and it is a very simple formula. But this can be transferred to the general case by the following fact: if one plugs any random variable into its cumulative distribution, one obtains a uniform distribution! These locations will therefore give, strictly speaking, the usual values of the smallest, second smallest etc. observation of $F_{\mathsf{x}}(\mathsf{x})$, but the usual values for x itself cannot be far from this.

If one plots the data on the vertical axis versus the standard normal on the horizontal axis (the default for the R-function `qqnorm`), then an S-shaped plot indicates a light-tailed distribution, an inverse S says that the distribution is heavy-tailed (try `qqnorm(rt(25,df=1))` as an example), a C is left-skewed, and an inverse C, a J, is right-skewed. A right-skewed, or positively skewed, distribution is one which has a long right tail, like the lognormal `qqnorm(rlnorm(25))` or chisquare `qqnorm(rchisq(25,df=3))`.

The classic reference which everyone has read and which explains it all is [**Gum58**, pp. 28–34 and 46/47]. Also [**WG68**] is useful, many examples.

47.11. Testing for Normality

[**Vas76**] is a test for Normality based on entropy.

CHAPTER 48

Measuring Economic Inequality

48.1. Web Resources about Income Inequality

- UNU/Wider-UNDP World Income Inequality Database (4500 Gini-coefficients) `www.wider.unu.edu/wiid/wiid.htm`
- Luxembourg Income Study (detailed household surveys) http://lissy.ceps.lu
- World Bank site on Inequality, Poverty, and Socio-economic Performance http://www.worldbank.org/poverty/inequal/index.htm
- World Bank Data on Poverty and Inequality http://www.worldbank.org/poverty
- World Bank Living Standard Measurement Surveys http://www.worldbank.org/
- University of Texas Inequality Project (Theil indexes on manufacturing and industrial wages for 71 countries) http://utip.gov.utexas.edu/
- MacArthur Network on the Effects of Inequality on Economic Performance, Institute of International Studies, University of California, Berkeley. http://glob
- Inter-American Development Bank site on Poverty and Inequality http://www.ia

48.2. Graphical Representations of Inequality

See [**Cow77**] (there is now a second edition out, but our library only has this one):

- Pen's Parade: Suppose that everyone's height is proportional to his or her income. Line everybody in the population up in order of height. Then the curve which their heights draws out is "Pen's Parade." It is the empirical quantile function of the sample. It highlights the presence of any extremely large income and to a certain extent abnormally small incomes. But the information on middle income receivers is not so obvious.
- Histogram or other estimates of the underlying density function. Suppose you are looking down on a field. On one side, there is a long straight fence marked off with income categories: the physical distance between any two points directly corresponds to the income differences they represent. Get the whole population to come into the field and line up in the strip of land marked off by the piece of fence corresponding to their income bracket. The

shape you will see is a histogram of the income distribution. This shows more clearly what is happening in the middle income ranges. But perhaps it is not so readily apparent what is happening in the upper tail. This can be remedied by taking the distribution of the logarithm of income, i.e., arrange the income markers on the fence in such a way that equal physical distances mean equal income ratios.
- Lorenz curve: Line up everybody in ascending order and let them parade by. You have a big "cake" representing the overall sum of incomes. As each person passes, hand him or her his or her share of the cake, i.e., a piece of cake representing the proportion of income that person receives. Make a diagram indicating how much of the cake has been handed out, versus the number of people that have passed by. This gives the Lorenz curve. The derivative of the Lorenz curve is Pen's parade. The mean income is that point at which the slope is parallel to the diagonal. A straight line means total equality.

48.3. Quantitative Measures of Income Inequality

Some ot the above graphical representations are suggestive of quantitative measures, other quantitative measures arose from different considerations:

- Relative mean deviation: draw into Pen's parade a horizontal line at the average income, and use as measure of inequality the area between the Parade curve and this horizontal line, divided by the total area under the Parade curve.
- Gini coefficient: the area between the Lorenz curve and the diagonal line, times 2 (so that a Gini coefficient of 100% would mean: one person owns everything, and a Gini of 0 means total equality.
- Theil's entropy measure: Say x_i is person i's income, $\bar{\mathsf{x}}$ is the average income, and n the population count. Then the person's income share is $\mathsf{s}_i = \frac{\mathsf{x}_i}{n\bar{\mathsf{x}}}$. The entropy of this income distribution, according to (3.11.2), but with natural logarithms instead of base 2, is

$$\sum_{i=1}^{n} \mathsf{s}_i \ln \frac{1}{\mathsf{s}_i} \tag{48.3.1}$$

and the maximum possible entropy, obtained if income distribution is equal, is

$$\sum_{i=1}^{n} \frac{1}{n} \ln n = \ln n \tag{48.3.2}$$

Subtract the actual entropy of the income distribution from this maximal entropy to get Theil's measure

$$\frac{1}{n}\sum_{i=1}^{n}\tfrac{\mathsf{x}_i}{\bar{\mathsf{x}}}\ln\big(\tfrac{\mathsf{x}_i}{\bar{\mathsf{x}}}\big) \tag{48.3.3}$$

- Coefficient of variation (standard deviation divided by mean).
- Herfindahl's index

$$\sum_{i=1}^{n}\mathsf{s}_i^2 = \frac{1}{n^2}\sum_{i=1}^{n}\big(\tfrac{\mathsf{x}_i}{\bar{\mathsf{x}}}\big)^2 \tag{48.3.4}$$

- The following distance-function related inequality measure generalizes both Theil's and (up to an affine transformation) Herfindahl's indices: choose $\beta \geq 0$ and set $h(s) = \ln s$ if $\beta = 0$ and $h(s) = -\frac{1}{\beta}s^\beta$ otherwise and define

$$\frac{1}{1+\beta}\Big(\sum_{i=1}^{n}\tfrac{1}{n}h(\tfrac{1}{n}) - \sum_{i=1}^{n}\mathsf{s}_i h(\mathsf{s}_i)\Big) \tag{48.3.5}$$

which can also be written as

$$\frac{1}{1+\beta}\sum_{i=1}^{n}\mathsf{s}_i\big(h(\tfrac{1}{n}) - h(\mathsf{s}_i)\big) \tag{48.3.6}$$

i.e., the difference of the overall measure from the smallest possible measure is at the same time the weighted average of the differences of $h(\mathsf{s}_i)$ from $h(\frac{1}{n})$.

Problem 449. *Show that* (48.3.3) *is the difference between* (48.3.1) *and* (48.3.2)

Answer.

$$\begin{aligned}\ln n - \sum_{i=1}^{n}\frac{\mathsf{x}_i}{n\bar{\mathsf{x}}}\ln\frac{n\bar{\mathsf{x}}}{\mathsf{x}_i} &= \ln n - \frac{1}{n}\sum_{i=1}^{n}\frac{\mathsf{x}_i}{\bar{\mathsf{x}}}\big(\ln n + \ln\frac{\bar{\mathsf{x}}}{\mathsf{x}_i}\big)\\ &= \ln n - \frac{1}{n}\sum_{i=1}^{n}\frac{\mathsf{x}_i}{\bar{\mathsf{x}}}\ln n - \frac{1}{n}\sum_{i=1}^{n}\frac{\mathsf{x}_i}{\bar{\mathsf{x}}}\ln\frac{\bar{\mathsf{x}}}{\mathsf{x}_i}\\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{\mathsf{x}_i}{\bar{\mathsf{x}}}\ln\frac{\mathsf{x}_i}{\bar{\mathsf{x}}}\end{aligned} \tag{48.3.7}$$

□

Problem 450. *7 points Show that if one takes a small amount of income share* $d\mathsf{s}$ *from person 2 and adds it to person 1, then the inequality measure defined in*

(48.3.5) *changes by* $\big(h(\mathsf{s}_2) - h(\mathsf{s}_1)\big)d\mathsf{s}$. *Hint: if* $\beta \neq 0$,

$$\frac{\partial I}{\partial \mathsf{s}_i} = \frac{\partial}{\partial \mathsf{s}_i}\mathsf{s}_i h(\mathsf{s}_i) = \frac{1}{\beta}\mathsf{s}_i^\beta = -h(\mathsf{s}_i). \tag{48.3.8}$$

If one therefore takes $d\mathsf{s}$ *away from 2 and gives it to 1, I changes by*

$$dI = \Big(-\frac{\partial I}{\partial \mathsf{s}_2} + \frac{\partial I}{\partial \mathsf{s}_1}\Big)d\mathsf{s} = \Big(h(\mathsf{s}_2) - h(\mathsf{s}_1)\Big)d\mathsf{s} \tag{48.3.9}$$

If $\beta = 0$, *only small modifications apply.*

Answer. If $\beta = 0$, then

$$I = \Big(\ln(\tfrac{1}{n}) - \sum_{i=1}^{n}\mathsf{s}_i\ln(\mathsf{s}_i)\Big) \tag{48.3.10}$$

therefore one has in this case

$$\frac{\partial I}{\partial \mathsf{s}_i} = \frac{\partial}{\partial \mathsf{s}_i}\mathsf{s}_i h(\mathsf{s}_i) = -\ln(\mathsf{s}_i) - 1 = -h(\mathsf{s}_i) - 1 \tag{48.3.11}$$

But the extra -1 cancels in the difference. □

Interpretation: if $h(\mathsf{s}_2) - h(\mathsf{s}_1) = h(\mathsf{s}_4) - h(\mathsf{s}_3)$ *then for the purposes of this inequality measure, the distance between 2 and 1 is the same as the distance between 4 and 3. These inequality measures are therefore based on very specific notions of what constitutes inequality.*

48.4. Properties of Inequality Measures

e invariance: In order to make economic sense, the measures must be invariant under a change to different monetary units. As long as the inequality measures are functions of the nominal incomes only, without a real anchor (such as: make global inequality measures higher if more people live beyond a subsistence level), this invariance under changes of the monetary unit also makes them invariant under proportional changes of everyone's incomes.

f Population It one doubles the population, with the newcomers having exactly the same income distribution as the original population, then the income distribution measure should not change.

of Transfers A hypothetical transfer of income from a richer to a poorer person should decrease inequality.

of Transfers If the effect only depends on the distances of donor and recipient expressed by the$h$ function.

Variances and Theil's entropy measure can be decomposed into "within" and "between" measures, whereas the Gini coefficient cannot.

CHAPTER 49

Distributed Lags

In the simplest case of one explanatory variable only, the model is

$$\mathsf{y}_t = \alpha + \beta_0 x_t + \beta_1 x_{t-1} + \cdots + \beta_N x_{t-N} + \varepsilon_t \tag{49.0.1}$$

This can be written in the form

$$\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \text{where} \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_1 & x_0 & \cdots & x_{1-N} \\ 1 & x_2 & x_1 & \cdots & x_{2-N} \\ \vdots & \vdots & & \vdots & \\ 1 & x_n & x_{n-1} & \cdots & x_{n-N} \end{bmatrix}. \tag{49.0.2}$$

Note that $\boldsymbol{X}$ contains presample values.

Two problems: lag length often not known, and $\boldsymbol{X}$ matrix often highly multicollinear.

How to determine lag length? Sometimes it is done by the adjusted $\bar{R}^2$. [**Mad88**, p. 357] says this will lead to too long lags and proposes remedies.

Assume we know for sure that lag length is not greater than M. [**JHG$^+$88**, pp. 723–727] recommends the following "general-to-specific" specification procedure for finding the lag length: First run the regression with M lags; if the t-test for the parameter of the Mth lag is significant, we say the lag length is M. If it is insignificant, run the regression with $M-1$ lags and test again for the last coefficient: If the t-test for the parameter of the $M-1$st coefficient is significant, we say the lag length is $M-1$, etc.

The significance level of this test depends on M and on the true lag length. Since we never know the true lag length for sure, we will never know the true significance level for sure. The calculation which follows now allows us to compute this significance level under the assumption that the N given by the test is the correct N. Furthermore this calculation only gives us the one-sided significance level: the null hypothesis is not that the true lag length is $= N$, but that the true lag length is $\leq N$.

Assume the null hypothesis is true, i.e., that the true lag length is $\leq N$. Since we assume we know for sure that the true lag length is $\leq M$, the null hypothesis is equivalent to: $\beta_{N+1} = \beta_{N+2} = \cdots = \beta_M = 0$. Now assume that we apply the

above procedure and the null hypothesis holds. The significance level of our test is the probability that our procedure rejects the null although the null is true. In other words, it is the probability that either the first t-test rejects, or the first t-test accepts and the second t-test rejects, or the first two t-tests accept and the third t-test rejects, etc, all under the assumption that the true β_i are zero. In all, the lag length is overstated if at least one of the $M-N$ t-tests rejects. Therefore if we define the event C_i to be the rejection of the ith t-test, and define $Q_j = C_1 \cup \cdots \cup C_j$, then $\Pr[Q_j] = \Pr[Q_{j-1} \cup C_j] = \Pr[Q_{j-1}] + \Pr[C_j] - \Pr[Q_{j-1} \cap C_j]$. [**JHG$^+$88**, p. 724] says, and a proof can be found in [**And66**] or [**And71**, pp. 34–43], that the test statistics of the different t-tests are independent of each other. Therefore one can write $\Pr[Q_j] = \Pr[Q_{j-1}] + \Pr[C_j] - \Pr[Q_{j-1}]\Pr[C_j]$.

Examples: Assuming all t-tests are carried out at the 5% significance level, and two tests are insiginficant before the first rejection occurs. I.e., the test indicates that the true lag length is $\leq M-2$. Assuming that the true lag length is indeed $\leq M-2$, the probability of falsely rejecting the hypothesis that the Mth and $M-1$st lags are zero is $0.05 + 0.05 - 0.05^2 = 0.1 - 0.0025 = 0.0975$. For three and four tests the levels are 0.1426 and 0.1855. For 1% significance level and two tests it would be $0.01 + 0.01 - 0.01^2 = 0.0200 - 0.0001 = 0.0199$. For 1% significance level and three tests it would be $0.0199 + 0.01 - 0.000199 = 0.029701$.

Problem 451. *Here are excerpts from SAS outputs, estimating a consumption function. The dependent variable is always the same, GCN72, the quarterly personal consumption expenditure for nondurable goods, in 1972 constant dollars, 1948–1985. The explanatory variable is GYD72, personal income in 1972 constant dollars (deflated by the price deflator for nondurable goods), lagged 0–8 quarters.*

VARIABLE	DF	PARAMETER ESTIMATE	STANDARD ERROR	T FOR H0: PARAMETER=0	PROB > |T|
INTERCEP	1	65.61238269	0.88771664	73.911	0.0001
GYD72	1	0.13058204	0.000550592	237.167	0.0001

VARIABLE	DF	PARAMETER ESTIMATE	STANDARD ERROR	T FOR H0: PARAMETER=0	PROB > |T|
INTERCEP	1	65.80966177	0.85890869	76.620	0.0001
GYD72	1	0.07778248	0.01551323	5.014	0.0001
GYD72L1	1	0.05312929	0.01560094	3.406	0.0009

• a. *3 points Make a sequential test how long you would like to have the lag length.*

VARIABLE	DF	PARAMETER ESTIMATE	STANDARD ERROR	T FOR H0: PARAMETER=0	PROB > |T|
INTERCEP	1	65.87672382	0.84399982	78.053	0.0001
GYD72	1	0.08289905	0.01537243	5.393	0.0001
GYD72L1	1	0.008943833	0.02335691	0.383	0.7023
GYD72L2	1	0.03932710	0.01569029	2.506	0.0133

VARIABLE	DF	PARAMETER ESTIMATE	STANDARD ERROR	T FOR H0: PARAMETER=0	PROB > |T|
INTERCEP	1	65.99593829	0.82873058	79.635	0.0001
GYD72	1	0.08397167	0.01507688	5.570	0.0001
GYD72L1	1	0.01413009	0.02298584	0.615	0.5397
GYD72L2	1	-0.007354543	0.02363040	-0.311	0.7561
GYD72L3	1	0.04063255	0.01561334	2.602	0.0102

VARIABLE	DF	PARAMETER ESTIMATE	STANDARD ERROR	T FOR H0: PARAMETER=0	PROB > |T|
INTERCEP	1	66.07544717	0.80366736	82.217	0.0001
GYD72	1	0.09710692	0.01518481	6.395	0.0001
GYD72L1	1	-0.000042518	0.02272008	-0.002	0.9985
GYD72L2	1	0.001564270	0.02307528	0.068	0.9460
GYD72L3	1	-0.01713777	0.02362498	-0.725	0.4694
GYD72L4	1	0.05010149	0.01573309	3.184	0.0018

VARIABLE	DF	PARAMETER ESTIMATE	STANDARD ERROR	T FOR H0: PARAMETER=0	PROB > |T|
INTERCEP	1	66.15803761	0.78586731	84.185	0.0001
GYD72	1	0.09189381	0.01495668	6.144	0.0001
GYD72L1	1	0.01675422	0.02301415	0.728	0.4678
GYD72L2	1	-0.01061389	0.02297260	-0.462	0.6448
GYD72L3	1	-0.008377491	0.02330072	-0.360	0.7197
GYD72L4	1	-0.000826189	0.02396660	-0.034	0.9725
GYD72L5	1	0.04296552	0.01551164	2.770	0.0064

Answer. If all tests are made at 5% significance level, reject that there are 8 or 7 lags, and go with 6 lags. □

• b. *5 points What is the probability of type I error of the test you just described?*

Answer. For this use the fact that the t-statistics are independent. There is a 5% probability of incorrectly rejecting the first t-test and also a 5% probability of incorrectly rejecting the second

VARIABLE	DF	PARAMETER ESTIMATE	STANDARD ERROR	T FOR H0: PARAMETER=0	PROB > |T|
INTERCEP	1	66.22787177	0.77701222	85.234	0.0001
GYD72	1	0.08495948	0.01513456	5.614	0.0001
GYD72L1	1	0.02081719	0.02281536	0.912	0.3631
GYD72L2	1	0.001067395	0.02335633	0.046	0.9636
GYD72L3	1	-0.01567316	0.02327465	-0.673	0.5018
GYD72L4	1	0.003008501	0.02374452	0.127	0.8994
GYD72L5	1	0.004766535	0.02369258	0.201	0.8408
GYD72L6	1	0.03304355	0.01563169	2.114	0.0363

VARIABLE	DF	PARAMETER ESTIMATE	STANDARD ERROR	T FOR H0: PARAMETER=0	PROB > |T|
INTERCEP	1	66.29560292	0.77062598	86.028	0.0001
GYD72	1	0.08686483	0.01502757	5.780	0.0001
GYD72L1	1	0.01258635	0.02301437	0.547	0.5853
GYD72L2	1	0.004612589	0.02321459	0.199	0.8428
GYD72L3	1	-0.005511693	0.02366979	-0.233	0.8162
GYD72L4	1	-0.002789862	0.02372100	-0.118	0.9065
GYD72L5	1	0.008280160	0.02354535	0.352	0.7256
GYD72L6	1	-0.001408690	0.02383478	-0.059	0.9530
GYD72L7	1	0.02951031	0.01551907	1.902	0.0593

VARIABLE	DF	PARAMETER ESTIMATE	STANDARD ERROR	T FOR H0: PARAMETER=0	PROB > |T|
INTERCEP	1	66.36142439	0.77075066	86.100	0.0001
GYD72	1	0.08619326	0.01500496	5.744	0.0001
GYD72L1	1	0.01541463	0.02307449	0.668	0.5052
GYD72L2	1	-0.002721499	0.02388376	-0.114	0.9094
GYD72L3	1	-0.001837498	0.02379826	-0.077	0.9386
GYD72L4	1	0.003802403	0.02424060	0.157	0.8756
GYD72L5	1	0.004328457	0.02370310	0.183	0.8554
GYD72L6	1	0.000718960	0.02384368	0.030	0.9760
GYD72L7	1	0.006305240	0.02404827	0.262	0.7936
GYD72L8	1	0.02002826	0.01587971	1.261	0.2094

t-test. The probability of incorrectly rejecting at least one of the two tests is therefore $0.05 + 0.05 - 0.05 \cdot 0.05 = 0.1 - 0.0025 = 0.0975$. For 1% it is (for two tests) $0.01 + 0.01 - 0.01 \cdot 0.01 = 0.0199$, but three tests will be necessary! □

• c. *3 points Which common problem of an estimation with lagged explanatory variables is apparent from this printout? What would be possible remedies for this problem?*

Answer. The explanatory variables are highly multicollinear, therefore use Almon lags or something similar. Another type of problem is: increase of type I errors with increasing number of steps, start with small significance levels! □

Secondly: what to do about multicollinearity? Prior information tells you that the true lag coefficients probably do not go in zigzag, but follow a smooth curve. This information can be incorporated into the model by pre-selecting a family of possible lag contours from which that should be chosen that fits best, i.e. by doing constrained least squares. The simplest such assumption is that the lag coefficients lie on a polynomial or degree d (polynomial distributed lags, often called Almon lags). Since linear combinations of polynomials are again polynomials, this restricts the $\boldsymbol{\beta}$ vectors one has to choose from to a subspace of k-dimensional space.

Usually this is done by the imposition of linear constraints. One might explicitly write it as linear constraints of the form $\boldsymbol{R\beta} = \boldsymbol{o}$, since polynomials of dth order are characterized by the fact that the dth differences of the coefficients are constant, or their $d+1$st differences zero. (This gives one linear constraint for every position in $\boldsymbol{\beta}$ for which the dth difference can be computed.)

But here it is more convenient to incorporate these restrictions into the regression equation and in this way end up with a regression with fewer explanatory variables. Any $\boldsymbol{\beta}$ with a polynomial lag structure has the form $\boldsymbol{\beta} = \boldsymbol{H\alpha}$ for the $(d+1)\times 1$ vector $\boldsymbol{\alpha}$, where the columns of $\boldsymbol{H}$ simply are polynomials:

$$\begin{bmatrix}\beta_0\\ \beta_1\\ \beta_2\\ \beta_3\\ \beta_4\end{bmatrix} = \begin{bmatrix}1 & 0 & 0 & 0\\ 1 & 1 & 1 & 1\\ 1 & 2 & 4 & 8\\ 1 & 3 & 9 & 27\\ 1 & 4 & 16 & 64\end{bmatrix}\begin{bmatrix}\alpha_0\\ \alpha_1\\ \alpha_2\\ \alpha_3\end{bmatrix} \tag{49.0.3}$$

More examples for such $\boldsymbol{H}$-matrices are in [**JHG+88**, p. 730]. Then the specification $\mathsf{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ becomes $\mathsf{y} = \boldsymbol{XH\alpha} + \boldsymbol{\varepsilon}$. I.e., one estimates the coefficients of $\boldsymbol{\alpha}$ by an ordinary regression again, and even in the presence of polynomial distributed lags one can use the ordinary F-test, impose other linear constraints, do "GLS" in the usual way, etc. (SAS allows for an autoregressive error structure in addition to the lags). The `pdlreg` procedure in SAS also uses a $\boldsymbol{H}$ whose first column contains a zero order polynomial, the second a first order polynomial, etc. But it does not use these exact polynomials shown above but chooses the polynomials in such a way that they are orthogonal to each other. The elements of $\boldsymbol{\alpha}$ are called `X**0` (coefficient of the zero order polynomial), `X**1`, etc.

In order to determine the degree of the polynomial one might use the same procedure on this reparametrized regression which one used before to determine the lag length.

About endpoint restrictions: The polynomial determines the coefficients β_0 through β_M, with the other β_j being zero. Endpoint restrictions (the SAS options `last`, `first`, or `both`) determine that either the polynomial is such that its formula also gives $\beta_{M+1} = 0$ or $\beta_{-1} = 0$ or both. This may prevent, for instance, the last lagged coefficient from becomeing negative if all the others are positive. But experience shows that in many cases such endpoint restrictions are not a good idea.

Alternative specifications of the lag coefficients: Shiller lag: In 1973, long before smoothing splines became popular, Shiller in [**Shi73**] proposed a joint minimization of SSE and k times the squared sum of $d+1$st differences on lag coefficients. He used a Bayesian approach; Maddala classical method. This is the BLUE if one replaces the exact linear constraint by a random linear constraint.

Problem 452. *Which problems does one face if one estimates a regression with lags in the explanatory variables? How can these problems be overcome?*

49.1. Geometric lag

Even more popular than polynomial lags are geometric lags. Here the model is

$$\mathsf{y}_t = \alpha + \gamma x_t + \gamma\lambda x_{t-1} + \gamma\lambda^2 x_{t-2} + \cdots + \varepsilon_t \tag{49.1.1}$$

$$= \alpha + \beta(1-\lambda)x_t + \beta(1-\lambda)\lambda x_{t-1} + \beta(1-\lambda)\lambda^2 x_{t-2} + \cdots + \varepsilon_t. \tag{49.1.2}$$

Here the second line is written in a somewhat funny way in order to make the $w_t = (1-\lambda)\lambda^t$, the weights with which β is distributed over the lags, sum to one. Here it is tempting to do the following *Koyck-transformation:* lag this equation by one and premultipy by λ to get

$$\lambda\mathsf{y}_{t-1} = \lambda\alpha + \beta(1-\lambda)\lambda x_{t-1} + \beta(1-\lambda)\lambda^2 x_{t-2} + \beta(1-\lambda)\lambda^3 x_{t-3} + \cdots + \lambda\varepsilon_{t-1}. \tag{49.1.3}$$

Now subtract:

$$\mathsf{y}_t = \alpha(1-\lambda) + \lambda\mathsf{y}_{t-1} + \beta(1-\lambda)x_t + \varepsilon_t - \lambda\varepsilon_{t-1}. \tag{49.1.4}$$

This has a lagged dependent variable. This is not an accident, as the follwing discussion suggests.

49.2. Autoregressive Distributed Lag Models

[**DM93**, p. 679] say that (49.0.1) is not a good model because it is not a dynamic model, i.e., y_t depends on lagged values of x_t but not on lagged values of itself. As a consequence, only the current values of the error term ε_t affect y_t. But if the error term is thought of as reflecting the combined influence of many variables that are unavoidably omitted from the regression, one might want to have the possibility that these omitted variables have a lagged effect on y_t just as x_t does. Therefore it is natural to allow lagged values of y_t to enter the regression along with lagged values of x_t:

(49.2.1)
$$\mathsf{y}_t = \alpha+\beta_1\mathsf{y}_{t-1}+\beta_2\mathsf{y}_{t-2}+\cdots+\beta_p\mathsf{y}_{t-p}+\gamma_0 x_t+\gamma_1 x_{t-1}+\cdots+\gamma_N x_{t-q}+\varepsilon_t \quad \varepsilon_t \sim \mathrm{IID}(0,\sigma^2)$$

This is called an $ADL(p,q)$ model. A widely encountered special case is the $ADL(1,1)$ model

$$\mathsf{y}_t = \alpha + \beta_1\mathsf{y}_{t-1} + \gamma_0 x_t + \gamma_1 x_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \mathrm{IID}(0,\sigma^2) \tag{49.2.2}$$

This has the following special cases: distributed lag model with geometric lags ($\gamma_1 = 0$), static model with AR(1) errors ($\gamma_1 = -\beta_1\gamma_0$), partial adjustment model ($\gamma_1 = 0$), model in the first differences ($\beta_1 = 1$), ($\gamma_1 = -\gamma_0$).

This lagged dependent variable is not an obstacle to running OLS, in light of the results discussed under "random regressors."

We will discuss two models which give rise to such a lag structure: either with the *desired level achieved incompletely* as the *dependent* variable (Partial Adjustment models), or with an *adaptively formed expected level* as the *explanatory* variable. In the first case, OLS on the Koyck transformation is consistent, in the other case it is not, but alternative methods are available.

Partial Adjustment. Here the model is

$$\mathsf{y}^*_t = \alpha + \beta x_t + \varepsilon_t, \tag{49.2.3}$$

where y^*_t is not the actual but the desired level of y_t. These y^*_t are not observed, but the assumption is made that the actual values of y_t adjust to the desired levels as follows:

$$\mathsf{y}_t - \mathsf{y}_{t-1} = (1-\lambda)(\mathsf{y}^*_t - \mathsf{y}_{t-1}). \tag{49.2.4}$$

Solving (49.2.4) for y_t gives $\mathsf{y}_t = \lambda\mathsf{y}_{t-1} + (1-\lambda)\mathsf{y}^*_t$. If one substitutes (49.2.3) into this, one gets

$$\mathsf{y}_t = \alpha(1-\lambda) + \beta(1-\lambda)x_t + \lambda\mathsf{y}_{t-1} + (1-\lambda)\varepsilon_t \tag{49.2.5}$$

If one were to repeatedly lag this equation, premultipy by λ, and reinsert, one would get

$$\begin{aligned}\mathsf{y}_t = \alpha + \beta(1-\lambda)x_t + \beta(1-\lambda)\lambda x_{t-1} + \beta(1-\lambda)\lambda^2 x_{t-2} + \cdots \\ \cdots + +(1-\lambda)\varepsilon_t + \lambda(1-\lambda)\varepsilon_{t-1} + \cdots.\end{aligned} \tag{49.2.6}$$

These are geometrically declining lags, and (49.2.5) is their Koyck transform. It should be estimated in the form (49.2.5). It has a lagged dependent variable, but contemporaneously uncorrelated, therefore OLS is consistent and has all desired asymptotic properties.

The next question is about *Adaptive Expectations*, an example where regression on the Koyck-transformation leads to inconsistent results.

Problem 453. *Suppose the simple regression model is modified so that y_t is, up to a disturbance term, a linear function not of x_t but of what the economic agents at time t consider to be the "permanent" level of x, call it x^*_t. One example would be a demand relationship in which the quantity demanded is a function of permanent price. The demand for oil furnaces, for instance, depends on what people expect the price of heating oil to be in the long run. Another example is a consumption function with permanent income as explanatory variable. Then*

$$\mathsf{y}_t = \alpha + \beta x^*_t + \varepsilon_t, \qquad \varepsilon_t \sim \mathrm{IID}(0,\sigma^2) \tag{49.2.7}$$

Here x^*_t is the economic agents' perceptions of the permanent level of x_t. Usually the x^*_t are not directly observed. In order to link x^*_t to the observed actual (as opposed to permanent) values x_t, assume that in every time period t the agents modify their perception of the permanent level based on their current experience x_t as follows:

$$x^*_t - x^*_{t-1} = (1-\lambda)(x_t - x^*_{t-1}). \tag{49.2.8}$$

I.e., the adjustment which they apply to their perception of the permanent level in period t, $x^*_t - x^*_{t-1}$, depends on by how much last period's permanent level differs from the present period's actual level; more precisely, it is $1-\lambda$ times this difference. Here $1-\lambda$ represents some number between zero and one, which does not change over time. We are using $1-\lambda$ instead of λ in order to make the formulas below a little simpler and to have the notation consistent with the partial adjustment model.

- a. *1 point Show that* (49.2.8) *is equivalent to*

$$x^*_t = \lambda x^*_{t-1} + (1-\lambda)x_t \tag{49.2.9}$$

- b. *2 points Derive the following regression equation from* (49.2.7) *and* (49.2.9)*:*

$$\mathsf{y}_t = \alpha(1-\lambda) + \beta(1-\lambda)x_t + \lambda\mathsf{y}_{t-1} + \eta_t \tag{49.2.10}$$

and write the new disturbances η_t in terms of the old ones ε_t. What is the name of the mathematical transformation you need to derive (49.2.10)*?*

Answer. Lag (49.2.7) by 1 and premultiply by λ (the Koyck-transformation) to get

$$\lambda \mathfrak{y}_{t-1} = \alpha\lambda + \lambda\beta x^*_{t-1} + \lambda\varepsilon_{t-1} \tag{49.2.11}$$

Subtract this from (49.2.7) to get

$$\mathfrak{y}_t - \lambda\mathfrak{y}_{t-1} = \alpha(1+\lambda) + \beta x^*_t - \beta\lambda x^*_{t-1} + \varepsilon_t - \lambda\varepsilon_{t-1} \tag{49.2.12}$$

Now use (49.2.9) in the form $x^*_t - \lambda x^*_{t-1} = (1-\lambda)x_t$ to get (49.2.10). The new disturbances are $\eta_t = \varepsilon_t - \lambda\varepsilon_{t-1}$.

□

• c. *2 points Argue whether or not it is possible to get consistent estimates $\hat\alpha$ and $\hat\beta$ by applying OLS to the equation*

$$\mathfrak{y}_t = \alpha_0 + \beta_0 x_t + \lambda\mathfrak{y}_{t-1} + \eta_t \tag{49.2.13}$$

and then setting

$$\hat\alpha = \frac{\hat\alpha_0}{1-\hat\lambda} \quad \textit{and} \quad \hat\beta = \frac{\hat\beta_0}{1-\hat\lambda}. \tag{49.2.14}$$

Answer. OLS is inconsistent because $\mathfrak{y}_{t-1}$ and ε_{t-1}, therefore also $\mathfrak{y}_{t-1}$ and η_t are correlated. (It is also true that η_{t-1} and η_t are correlated, but this is *not* the reason of the inconsistency). □

• d. *1 point In order to get an alternative estimator, show that repeated application of* (49.2.8) *gives*

$$x^*_t = (1-\lambda)\big(x_t + \lambda x_{t-1} + \lambda^2 x_{t-2} + \cdots + \lambda^{t-1}x_1\big) + \lambda^t x^*_0 \tag{49.2.15}$$

Answer. Rearranging (49.2.9) one obtains

$$x^*_t = (1-\lambda)x_t + \lambda x^*_{t-1} \tag{49.2.16}$$

$$= (1-\lambda)x_t + \lambda\big(1-\lambda)x_{t-1} + \lambda x^*_{t-1}\big) \tag{49.2.17}$$

$$= (1-\lambda)(x_t + \lambda x_{t-1}) + \lambda x^*_{t-1} \tag{49.2.18}$$

$$= (1-\lambda)\big(x_t + \lambda x_{t-1} + \lambda^2 x_{t-2} + \cdots + \lambda^{t-1}x_1\big) + \lambda^t x^*_0 \tag{49.2.19}$$

□

• e. *2 points If λ is known, show how α and β can be estimated consistently by OLS from the following equation, which is gained from inserting* (49.2.15) *into* (49.2.7)*:*

$$\mathfrak{y}_t = \alpha + \beta(1-\lambda)\big(x_t + \lambda x_{t-1} + \lambda^2 x_{t-2} + \cdots + \lambda^{t-1}x_1\big) + \beta x^*_0\lambda^t + \varepsilon_t. \tag{49.2.20}$$

How many regressors are in this equation? Which are the unknown parameters? Describe exactly how you get these parameters from the coefficients of these regressors.

Answer. Three regressors: intercept, $(1-\lambda)\big(x_t + \lambda x_{t-1} + \lambda^2 x_{t-2} + \cdots + \lambda^{t-1}x_1\big)$, and λ^t. In the last term, λ^t is the explanatory variable. A regression gives estimates of α, β, and a "prediction" of x^*_0. Note that the sum whose coefficient is β has many elements for high t, and few elements for low t. Also note that the λ^t-term becomes very small, i.e., only the first few observations of this "variable" count. This is why the estimate of x^*_0 is not consistent, i.e., increasing the sample size will not get an arbitrarily precise estimate of this value. Will the estimate of σ^2 be consistent? □

• f. *1 point What do you do if λ is not known?*

Answer. Since λ is usually not known, one can do the above procedure for all values along a grid from 0 to 1 and then pick the value of λ which gives the best SSE. Zellner and Geisel did this in [**ZG70**], and their regression can be reproduced in R with the commands `data(geizel)` and then `plot((1:99)/100, geizel.regression(geizel$c, geizel$y, 99), xlab="lambda", ylab="sse")`. They got two local minima for λ, and that local minimum which was smaller corresponded to a $\beta > 1$ which had to be ruled out for economic reasons. Their results are described in [**Kme86**, p. 534]. They were re-estimated with more recent data in [**Gre93**, pp. 531–533], where this paradox did not arise. □

Here is R-code to compute the regressors in (49.2.19), and to search for the best λ.

```
"geizel.regressors" <- function(x, lambda)
  { lngth <- length(x)
    lampow <- z <- vector(mode="numeric",length=lngth)
    lampow[[1]] <- lambda
    z[[1]] <- x[[1]]
    for (ii in 2:lngth)
      { lampow[[ii]] <- lambda*lampow[[ii-1]]
        z[[ii]] <- x[[ii]] + lambda*z[[ii-1]] }
    data.frame(z=(1-lambda)*z, lampow=lampow) }

"geizel.regression" <- function(y, x, nn, lambdamin=1/(nn+1), lambdamax=nn/(n
  #The original Zellner Geisel article has an intercept.
  { sigmavec <- vector(mode="numeric",length=nn)
    for (ii in 1:nn)
      { sigmavec[[ii]] <-
          summary(lm(y ~ z + lampow - 1,
                     data=geizel.regressors(x,
                       lambdamin+(ii-1)*(lambdamax-lambdamin)/(nn-1))))$sigma
    sigmavec }
```

If one does this with the DHSY data (see chapter **??**, one first has to seasonally adjust the data. `ecmet-script(geizel)` de-seasonalizes and de-trends the data as it was done implicitly in model (**??**). Then the optimal λ is 0.28, and the marginal propensity to consume for this optimal λ is 0.323, which is much too low.

PROBLEM 454. *4 points There are at least two different situations how a geometric lag can come about. In one of those two situations, regression after performing the "Koyck transformation" leads to consistent estimates, in the other it does not. Explain.*

PROBLEM 455. *7 points Zvi Griliches in* [**Gri67**] *considers the problem of distinguishing between the following two models: A partial adjustment model*

$$\mathfrak{y}_t = \alpha + \beta x_t + \rho \mathfrak{y}_{t-1} + \nu_t \tag{49.2.21}$$

where ν_t *obeys all classical assumptions, and a simple regression model with an autoregressive disturbance*

$$\mathfrak{y}_t = \alpha + \beta x_t + \varepsilon_t \qquad \text{where} \quad \varepsilon_t = \rho\varepsilon_{t-1} + \nu_t. \tag{49.2.22}$$

• a. *Suppose you have data on* x_t *and* $\mathfrak{y}_t$ *for* n *consecutive periods. Describe in detail a test procedure that might enable you (if there is enough information in the data) to decide which of the two models is appropriate. (*Note*: Spell out the null and the alternative hypothesis and propose a test statistic and state its distribution.)*

Hint: *Models* (49.2.21) *and* (49.2.22) *are not nested, but they can both be written as special cases of the same umbrella specification.*

ANSWER. This umbrella specification is

$$\mathfrak{y}_t = \gamma + \beta x_t + \rho \mathfrak{y}_{t-1} + \delta x_{t-1} + \nu_t \tag{49.2.23}$$

with well-behaved disturbances. Clearly, (49.2.21) follows from (49.2.23) by setting $\gamma = \alpha$ and $\delta = 0$. To see that (49.2.22) follows as well, write it as

$$\mathfrak{y}_t = \alpha + \beta x_t + \rho\varepsilon_{t-1} + \nu_t \tag{49.2.24}$$

and insert $\varepsilon_{t-1} = \mathfrak{y}_{t-1} - \alpha - \beta x_{t-1}$. You get (49.2.23) with $\gamma = \alpha(1-\rho)$ and $\delta = -\beta\rho$. The two models make therefore two different statements about δ, the coefficient of x_{t-1}. (49.2.21) has the constraint $\delta = 0$, while (49.2.22) has $\delta = -\beta\rho$.

To test whether the data reject the first hypothesis, run a simple t-test for $\delta = 0$. To test whether the data reject the second hypothesis, test the nonlinear constraint $\beta\rho + \delta = 0$. The likelihood-ratio test is the neatest way. If the data reject one test and accept the other, then one is lucky. If the data accept both, then one can argue that there is not enough information to discriminate between the two models. If the data reject both, then one has exceptionally bad data (assuming the umbrella hypothesis is correct). □

Other alternatives:

Schmidt's polynomial geometric lag: not necessary to decide over maximum length of the lag.

What is desired is usually a hump, and this can be modeled according to density functions: Pascal lag: the weights are $w_i = \binom{1+r-1}{1}(1-\lambda)^r\lambda^i$; estimate by MLE. Gamma-lag $w_i = i^{s-1}e^{-i}$, $s > 0$, integer; does *not* add up to one! Not recommonded because: $w_0 = 0$ for $s > 1$, and w_1 is always the same. Modified Gamma-lag: $w_i = (i+1)^{\alpha/(1-\alpha)}\lambda^i$; $0 \le \alpha < 1$.

CHAPTER 50

Investment Models

50.1. Accelerator Models

The assumption is that for output Q, a capital stock of aQ is necessary. Therefore accelerator

$$\Delta K = a\Delta Q. \tag{50.1.1}$$

But it does not fit, the estimated a is much too small for a reasonable capital-output ratio.

PROBLEM 456. *Plot the capital stock of your industry against value added, and also plot the first differences against each other. Interpret your results.*

Now the flexible accelerator has the following two basic equations:

$$K_t^* = aQ_t \tag{50.1.2}$$

$$K_t - K_{t-1} = (1-\gamma)(K_t^* - K_{t-1}) \tag{50.1.3}$$

This can either be used to generate a relation between capital stock and output, or a relation between investment and output. For the relation between capital stock and output write (50.1.3) as

$$K_t = (1-\gamma)K_t^* + \gamma K_{t-1} \tag{50.1.4}$$

and plug in (50.1.2) to get

$$K_t = a(1-\gamma)Q_t + \gamma K_{t-1} \tag{50.1.5}$$

This is a convenient form for estimation, i.e., one has to regress K_t on Q_t and K_{t-1}.

But one may also eliminate K_{t-1} on the righthand side of (50.1.5) by using the lagged version of (50.1.5):

$$K_t = a(1-\gamma)Q_t + a(1-\gamma)\gamma Q_{t-1} + \gamma^2 K_{t-2}. \tag{50.1.6}$$

Since $\gamma < 1$ and K_{t-j} is bounded, one obtains in the limit

$$K_t = a(1-\gamma)\sum_{j=0}^{\infty}\gamma^j Q_{t-j}. \tag{50.1.7}$$

The other alternative is to get a relation between output and investment. This is convenient when no capital stock data are available. The figure for investment usually refers to both replacement investment and net investment, i.e.,

$$I_t = K_t - K_{t-1} + D_t, \tag{50.1.8}$$

where D_t is depreciation.

The usual treatment of depreciation is to set $D_t = \delta K_{t-1}$, with δ either estimated or obtained from additional information.

Therefore

$$I_t = K_t - (1-\delta)K_{t-1} \tag{50.1.9}$$

Now substitute equation (50.1.4) for K_t to get:

$$I_t = (1-\gamma)K_t^* - (1-\gamma-\delta)K_{t-1}. \tag{50.1.10}$$

In order to eliminate the term with K_{t-1} on the righthand side, use the following trick:

$$I_t - (1-\delta)I_{t-1} = (1-\gamma)\big(K_t^* - (1-\delta)K_{t-1}^*\big) \tag{50.1.11}$$

$$- (1-\gamma-\delta)\big(K_{t-1} - (1-\delta)K_{t-2}\big). \tag{50.1.12}$$

The last term on the righthand side is equal to $(1-\gamma-\delta)I_{t-1}$, and by setting $K_t^* = aQ_t$ one obtains

$$I_t = a(1-\gamma)Q_t - a(1-\gamma)(1-\delta)Q_{t-1} + \gamma I_{t-1}. \tag{50.1.13}$$

Therefore one can estimate these parameters by regressing I_t on Q_t, Q_{t-1}, and I_{t-1}.

PROBLEM 457. *Estimate the parameters γ from the relationship* (50.1.5) *and from the relationship* (50.1.13). *Which result do you consider better?*

50.2. Jorgenson's Model

Jorgenson assumes a Cobb-Douglas production function

$$Q = F(K,L) = K^\alpha L^\beta, \tag{50.2.1}$$

where $\alpha + \beta < 1$ (but β will not be estimated). The marginal product of capital is $F_K = \alpha K^{\alpha-1}L^\beta = \alpha Q/K$. Although Jorgenson's theoretical derivation starts with te assumption that firms seek to maximize the present discounted value of their cash flow, he makes such strong supplementary assumptions that this is equivalent to the firms maximizing their net revenue at every instant. In other words, their desired capital input and labor input are such that $F_L = w(t)/p(t)$ and $F_K = c(t)/p(t)$, where $c(t) = (\delta + r)q(t) - \dot{q}(t)$ is the user cost of capital ($q(t)$ is the price index for capital goods). However firms are not at the desired path, and in order to

reach this path they pursue the following strategy: they hire enough labor to satisfy the marginal product condition for labor given their actual capital stock, i.e., they produce the profit maximizing amount given this capital stock. The capital stock which they consider their desired capital stock at time t is that amount of capital which would be optimal for producing the output they are actually producing at time t, i.e., $K_t^* = \frac{\alpha \cdot Q_t \cdot p_t}{c_t}$.

As they approach this capital stock, they also hire more labor to fulfill the marginal product condition for labor, therefore their output rises, and therefore their desired capital stock will rise also. What the firms therefore consider their desired capital stock is not yet the optimal path, but as long as they have not reached this optimal path, they see a discrepancy between their actual and desired capital stock.

In other words, although Jorgenson claims to be modelling a very neoclassical forward-looking optimizing behavior, he ands up estimating an equation in which firms start with the situation they are in and go from there.

The investment orders placed at time t are assumed to be $K_t^* - K_{t-1}^* = \Delta K_t^*$. Now Jorgenson makes the following assumptions about the actual deliveries of these investment goods. The portion $\mu_0 \Delta K_t^*$ will be delivered in the same period it is ordered, the portion $\mu_1 \Delta K_t^*$ will be delivered one period later, the portion $\mu_2 \Delta K_t^*$ will be delivered two periods later, etc. The coefficients $\mu_0, \mu_1, \mu_2, \ldots$ do not change over time and are also independent of the absolute size of ΔK_t^*. There is also replacement investment, which is assumed ordered early enough that it will be delivered on time. Therefore one obtains for investment:

$$I_t = \sum_{i=0}^{\infty} \mu_i \Delta K_{t-i}^* + \delta K_{t-1}. \tag{50.2.2}$$

Substituting $\Delta K_{t-i}^* = \alpha \cdot \Delta\Big(\frac{Qp}{c}\Big)_{t-i}$ one gets the following regression equation:

$$I_t = \alpha \sum_{i=0}^{\infty} \mu_i \Delta\Big(\frac{Qp}{c}\Big)_{t-i} + \delta K_{t-1}. \tag{50.2.3}$$

This can be considered a modification of the accelerator model in which the output variables are modified by price variables, in order to capture the effects of prices on investment.

If one has annual data, one might use this for estimation, assuming there are at most 3 or 4 lags. The coefficient α is identified because $\sum_{i=0}^{\infty} \mu_i = 1$.

Problem 458. *Run the Jorgonson equation* (50.2.3) *for some finite number of lags, and then run the same equation leaving out the cost-of-capital adjustment terms, i.e., just looking at it as a simple accelerator model. Do you get better results?*

Jorgenson, who works with quarterly data, makes the following assumption about the lag structure $\mu_0, \mu_1, \ldots$. Using the lag operator, the regression to be estimated is

$$I_t - \delta K_{t-1} = \alpha \sum_{i=0}^{\infty} \mu_i \boldsymbol{L}^i \Delta\Big(\frac{Qp}{c}\Big). \tag{50.2.4}$$

Jorgenson assumes a rational lag, i.e., the delivery lags are $\sum_i \mu_i \boldsymbol{L}^i = \frac{\sum \gamma_i \boldsymbol{L}^i}{\sum \omega_i \boldsymbol{L}^i}$, where both sums are rather short, i.e., the only nonzero coefficients γ_i and ω_i may be $\omega_0 = 1$, ω_1, ω_2, and γ_3, γ_4, and γ_5. Multiplying the regression equation through by $\sum \omega_i \boldsymbol{L}^i$ gives

$$I_t - \delta K_{t-1} + \omega_1(I_{t-1} - \delta K_{t-2}) + \omega_2(I_{t-2} - \delta K_{t-3}) = \alpha \sum_{i=3}^{5} \gamma_i \boldsymbol{L}^i \Delta\Big(\frac{Qp}{c}\Big). \tag{50.2.5}$$

Therefore $I_t - \delta K_{t-1}$ is the dependent variable, and the other variables the independent variables. α is identified here because $\omega_0 + \omega_1 + \omega_2 = \gamma_3 + \gamma_4 + \gamma_5$.

Jorgenson's results are that $\alpha = 0.01$, which is very small, it would indicate that only 1% of the sales revenues goes to the owners of capital, the rest goes to the laborers.

Problem 459. *Use the given data for an estimation of the investment function along the lines suggested by Jorgenson.*

50.3. Investment Function Project

We will work with annual data for the 2-digit SIC manufacturing industries, which are the following:

(50.3.1)

20 :	Food and Kindred Products		30 :	Rubber Products
21 :	Tobacco Manufactures*		31 :	Leather and Leather Products*
22 :	Textile Mill Products		32 :	Stone, Clay, and Glass Products
23 :	Apparel and Related Products*		33 :	Primary Metal Industries
24 :	Lumber and Products*		34 :	Fabricated Metal Products
25 :	Furniture and Fixtures*		35 :	Machinery (Except Electrical)
26 :	Paper and Allied Products		36 :	Electrical Equipment
27 :	Printing and Publishing Industries**		37 :	Transportation Equipment
28 :	Chemicals and Allied Products		38 :	Instruments and Related Products
29 :	Petroleum and Coal Products		39 :	Miscellaneous Manufactures*

The main data are collected in two datafiles: `ec781.invcur` has data about fixed nonresidential private investment, capital stock (net of capital consumption allowance), and gross national product by industry in current dollars. The file `ec781.invcon` has

the corresponding data in constant 1982 dollars (has missing values for some data for industry 27, that is why industry has a double star). Furthermore, the dataset `ec781.invmisc` has additional data which might be interesting. Among those are capacity utilization data for the industries 20, 22, 26, 28, 29, 30, 32, 33, 34, 35, 36, 37, and 38 (all industries for which there are no capacity utilization data have at least one star) and profit rates for all industries. The profit rate data are constructed as follows from current dollar data: numerator= corporate profits before tax + corporate inventory valuation adjustment + noncorporate income + noncorporate inventory valuation adjustment + government subsidies + net interest. Denominator: capital stock + inventories (the inventories come in part from the NIPAs, in part from the census). It also has the prime rate (short term lending interest rate), and the 10 year treasury note interest rate, and the consumer price index. Note that the interest rates are in percent, for most applications you will have to divide them by 100. The profit rate is not in percent, it is a decimal fraction.

All three datasets have the year as one of the variable, and they go from 1947–85, with often some of the data for the beginning and the end of that period missing. These datasets will be available on your `d`-disk, SAS should find them if you just call them up by the name given here.

CHAPTER 51

Distinguishing Random Variables from Variables Created by a Deterministic Chaotic Process

Dynamical systems are either described as recursive functions (discrete time) or as differential equations.

With discrete time, recursive functions (recursive functions are difference equations, discrete analog of differential equations), one can easily get chaotic behavior. E.g., the tent map or logistic function.

The problem is: how to distinguish the output of such a process from a randomly generated output.

The same problem can also happen in the continuous case. First-order differential equations can be visualized as vector fields.

An *attractor* A is a compact set which has a neighborhood U such that A is the *limit set* of all trajectories starting in U. That means, every trajectory starting in U comes arbitrarily close to each point of the attractor.

In $\mathbb{R}^2$, there are three different types of attractors: fixed points, limit cycles, and saddle loops. But in $\mathbb{R}^3$ and higher, chaos can occur, i.e., the trajectory can have a "strange attractor." Example: Lorenz attractor.

There is no commonly accepted definition of a strange attractor, it is an attractor that is neither a point nor a closed curve, and trajectories attracted by it take vastly different courses after a short time.

Now fractal dimensions: first the Hausdorff dimension as $\lim_{\varepsilon\to 0}\frac{\log \mathsf{N}(\varepsilon)}{\log(1/\varepsilon)}$, indicating the exponent with which the number of covering pieces $\mathsf{N}(\varepsilon)$ increases as the diameter of the pieces diminishes.

Examples with integer dimensions: for points we have $\mathsf{N}(\varepsilon) = 1$ always, therefore dimension is 0. For straight lines of length L, $\mathsf{N}(\varepsilon) = L/\varepsilon$, therefore we get $\lim_{\varepsilon\to 0}\frac{\log(L/\varepsilon)}{\log(1/\varepsilon)} = 1$, and for an area with surface S it is $\lim_{\varepsilon\to 0}\frac{\log(S/\varepsilon^2)}{\log(1/\varepsilon)} = 2$.

Famous example of set with fractal dimension is the Cantor set: start with unit interval, take middle third out, then take middle third of the two remaining segments out, etc. For $\varepsilon = 1/3$ one gets $\mathsf{N}(\varepsilon) = 2$, for $\varepsilon = 1/9$ one gets $\mathsf{N}(\varepsilon) = 4$,

and generally, for $\varepsilon = (1/3)^m$ one gets $\mathsf{N}(\varepsilon) = 2^m$. Therefore the dimension is $\lim_{m\to\infty}\frac{\log 2^m}{\log 3^m} = \frac{\log 2}{\log 3} = 0.63$.

A concept related to the Hausdorff dimension is the correlation dimension. To compute this one needs $C(\varepsilon)$, the fraction of the total number of points that are within the Euclidian distance ε of a given point. (This $C(\varepsilon)$ is a quotient of two infinite numbers, but in finite samples it is a quotient of two large but finite numbers, this is why it is more tractable than the Hausdorff dimension.) Example again with straight line and area, using sup norm: line: $C(\varepsilon) = 2\varepsilon/L$, area: $C(\varepsilon) = 4\varepsilon^2/S$. Then the correlation dimension is $\lim_{\varepsilon\to 0}\frac{\log C(\varepsilon)}{\log\varepsilon}$, again indicating how this count varies with the distance.

To compute it, use $\log C_M(\varepsilon)$, which is the sample analog of $\log C(\varepsilon)$ for a sample of size M, and plot it against $\log\varepsilon$. To get this sample analog, look at all pairs of different points, and count those which are less than ε apart, and divide by total number of pairs of different points $\mathsf{N}(N-1)/2$.

Clearly, if ε is too small, it falls through between the points, and if it is too large, it extends beyond the boundaries of the set. Therefore one cannot look at the slope in the origin but must look at the slope of a straight line segment near the origin. Another reason for not looking at too small ε is that there may be a measurement error.)

It seems the correlation dimension is close to and cannot exceed the Hausdorff dimension. What one really wants is apparently the Hausdorff dimension, but the correlation dimension is a numerically convenient surrogate.

Importance of fractal dimensions: If an attractor has a fractal dimension, then it is likely to be a strange attractor (although strictly speaking it is neither necessary nor sufficient). E.g. it seems to me the precise Hausdorff dimension of the Lorentz attractor is not known, but the correlation dimension is around 2.05.

51.1. Empirical Methods: Grassberger-Procaccia Plots.

With conventional statistical means, it is hard to distinguish chaotic deterministic from random timeseries. In a timeseries generated by a tent map, one obtains for almost all initial conditions a time series whose the autocorrelation function is zero for all lags. We need sophisticated results from chaos theory to be able to tell them apart.

Here is the first such result: Assume there is a time series of n-dimensional vectors $\mathbf{x}_t$ having followed a deterministic chaotic motion for a long time, so that for all practical purposes it has arrived at its strange attractor, but at every time point t you only observe the jth component x_j_t. Then an *embedding of dimension* m is an artificial dynamical system formed by the m-histories of this jth component. Takens proved that if $\mathbf{x}_t$ lies on a strange attractor, and the embedding dimension

$m > 2n-1$ then the embedding is topologically equivalent to the original time series. In particular this means that it has the same correlation dimension.

This has important implications: if a time series is part of a deterministic system also including other time series, then one can draw certain conclusions about the attractor without knowing the other time series.

Next point: the correlation dimension of this embedding is $\lim_{\varepsilon\to 0} \frac{\log C(\varepsilon,m)}{\log \varepsilon}$, where the embedding dimension m is added as second argument into the function C. If the system is deterministic, the correlation dimension settles to a stationary value as the embedding dimension m increases; for a random system it keeps increasing, in the i.i.d. case it is m. (In the special case that this i.i.d. distribution is the uniform one, the m-histories are uniformly distributed on the m-dimensional unit cube, and it follows immediately, like our examples above.) Therefore the Grassberger-Procaccia plots show for each m one curve, plotting $\log C(\varepsilon, m)$ against $\log \varepsilon$.

For ε small, i.e., $\log \varepsilon$ going towards $-\infty$, the plots of the true C's become asymptotically a straight line emanating from the origin with a given slope which indicates the dimension. Now one cannot make ε very small for two reasons: (1) there are only finitely many data points, and (2) there is also a measurement error whose effect disappears if ε becomes bigger than a few standard deviations of this measurement error. Therefore one looks at the slope for values of ε that are not too small.

One method to see whether there is a deterministic structure is to compare this sample correlation dimension with that of "scrambled" data and see whether the slopes of the original data do not become steeper while those of the scrambled data still become steeper. Scrambling means: fit an autocorrelation and then randomly draw the residuals.

This is a powerful tool for distinguishing random noise from a deterministic system.

CHAPTER 52

Instrumental Variables

Compare here [**DM93**, chapter 7] and [**Gre97**, Section 6.7.8]. Greene first introduces the simple instrumental variables estimator and then shows that the generalized one picks out the best linear combinations for forming simple instruments. I will follow [**DM93**] and first introduce the generalized instrumental variables estimator, and then go down to the simple one.

In this chapter, we will discuss a sequence of models $\mathbf{y}_n = \mathbf{X}_n\boldsymbol{\beta} + \boldsymbol{\varepsilon}_n$, where $\boldsymbol{\varepsilon}_n \sim (\boldsymbol{o}_n, \sigma^2\boldsymbol{I}_n)$, and $\mathbf{X}_n$ are $n \times k$-matrices of random regressors, and the number of observations $n \to \infty$. We do not make the assumption $\operatorname{plim} \frac{1}{n}\mathbf{X}_n^\top\boldsymbol{\varepsilon}_n = \boldsymbol{o}$ which would ensure consistency of the OLS estimator (compare Problem 394). Instead, a sequence of $n \times m$ matrices of (random or nonrandom) "instrumental variables" $\boldsymbol{W}_n$ is available which satisfies the following three conditions:

$$\operatorname{plim} \frac{1}{n}\boldsymbol{W}_n^\top\boldsymbol{\varepsilon}_n = \boldsymbol{o} \tag{52.0.1}$$

$$\operatorname{plim} \frac{1}{n}\boldsymbol{W}_n^\top\boldsymbol{W}_n = \boldsymbol{Q} \quad \text{exists, is nonrandom and nonsingular} \tag{52.0.2}$$

$$\operatorname{plim} \frac{1}{n}\boldsymbol{W}_n^\top\mathbf{X}_n = \boldsymbol{D} \quad \text{exists, is nonrandom and has full column rank} \tag{52.0.3}$$

Full column rank in (52.0.3) is only possible if $m \geq k$.

In this situation, regression of $\mathbf{y}$ on $\mathbf{X}$ is inconsistent. But if one regresses $\mathbf{y}$ on the projection of $\mathbf{X}$ on $\mathrm{R}[\boldsymbol{W}]$, the column space of $\boldsymbol{W}$, one obtains a consistent estimator. This is called the instrumental variables estimator.

If $\boldsymbol{x}_i$ is the ith column vector of $\mathbf{X}$, then $\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top\boldsymbol{x}_i$ is the projection of $\boldsymbol{x}_i$ on the space spanned by the columns of $\boldsymbol{W}$. Therefore the matrix $\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top\mathbf{X}$ consists of the columns of $\mathbf{X}$ projected on $\mathrm{R}[\boldsymbol{W}]$. This is what we meant by the projection of $\mathbf{X}$ on $\mathrm{R}[\boldsymbol{W}]$. With these projections as regressors, the vector of regression coefficients becomes the "generalized instrumental variables estimator"

$$\tilde{\boldsymbol{\beta}} = \Big(\mathbf{X}^\top\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top\mathbf{X}\Big)^{-1}\mathbf{X}^\top\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top\mathbf{y} \tag{52.0.4}$$

Problem 460. *3 points We are in the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ and we have a matrix $\boldsymbol{W}$ of "instrumental variables" which satisfies the following three conditions:*

$\operatorname{plim} \frac{1}{n}\boldsymbol{W}^\top\boldsymbol{\varepsilon} = \boldsymbol{o}$, $\operatorname{plim} \frac{1}{n}\boldsymbol{W}^\top\boldsymbol{W} = \boldsymbol{Q}$ exists, is nonrandom and positive definite, and $\operatorname{plim} \frac{1}{n}\boldsymbol{W}^\top\mathbf{X} = \boldsymbol{D}$ exists, is nonrandom and has full column rank. Show that the instrumental variables estimator

$$\tilde{\boldsymbol{\beta}} = \Big(\mathbf{X}^\top\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top\mathbf{X}\Big)^{-1}\mathbf{X}^\top\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top\mathbf{y} \tag{52.0.5}$$

is consistent. Hint: Write $\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta} = \mathbf{B}_n \cdot \frac{1}{n}\boldsymbol{W}^\top\boldsymbol{\varepsilon}$ and show that the sequence of matrices $\mathbf{B}_n$ has a plim.

Answer. Write it as

$$\begin{aligned}\tilde{\boldsymbol{\beta}}_n &= \Big(\mathbf{X}^\top\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top\mathbf{X}\Big)^{-1}\mathbf{X}^\top\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \\ &= \boldsymbol{\beta} + \Big(\mathbf{X}^\top\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top\mathbf{X}\Big)^{-1}\mathbf{X}^\top\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top\boldsymbol{\varepsilon} \\ &= \boldsymbol{\beta} + \Big((\frac{1}{n}\mathbf{X}^\top\boldsymbol{W})(\frac{1}{n}\boldsymbol{W}^\top\boldsymbol{W})^{-1}(\frac{1}{n}\boldsymbol{W}^\top\mathbf{X})\Big)^{-1}(\frac{1}{n}\mathbf{X}^\top\boldsymbol{W})(\frac{1}{n}\boldsymbol{W}^\top\boldsymbol{W})^{-1}\frac{1}{n}\boldsymbol{W}^\top\boldsymbol{\varepsilon},\end{aligned}$$

i.e., the $\mathbf{B}_n$ and $\boldsymbol{B}$ of the hint are as follows:

$$\begin{aligned}\mathbf{B}_n &= \Big((\frac{1}{n}\mathbf{X}^\top\boldsymbol{W})(\frac{1}{n}\boldsymbol{W}^\top\boldsymbol{W})^{-1}(\frac{1}{n}\boldsymbol{W}^\top\mathbf{X})\Big)^{-1}(\frac{1}{n}\mathbf{X}^\top\boldsymbol{W})(\frac{1}{n}\boldsymbol{W}^\top\boldsymbol{W})^{-1} \\ \boldsymbol{B} &= \operatorname{plim}\mathbf{B}_n = (\boldsymbol{D}^\top\boldsymbol{Q}^{-1}\boldsymbol{D})^{-1}\boldsymbol{D}^\top\boldsymbol{Q}^{-1}\end{aligned}$$

□

Problem 461. *Assume $\operatorname{plim} \frac{1}{n}\mathbf{X}^\top\mathbf{X}$ exists, and $\operatorname{plim} \frac{1}{n}\mathbf{X}^\top\boldsymbol{\varepsilon}$ exists. (We only need the existence, not that the first is nonsingular and the second zero). Show that σ^2 can be estimated consistently by $s^2 = \frac{1}{n}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})^\top(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})$.*

Answer. $\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} - \mathbf{X}\tilde{\boldsymbol{\beta}} = \boldsymbol{\varepsilon} - \mathbf{X}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})$. Therefore

$$\frac{1}{n}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})^\top(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) = \frac{1}{n}\boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon} - \frac{2}{n}\boldsymbol{\varepsilon}^\top\mathbf{X}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) + (\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top\Big(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\Big)(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}).$$

All summands have plims, the plim of the first is σ^2 and those of the other two are zero.

□

Problem 462. *In the situation of Problem 460, add the stronger assumption $\frac{1}{\sqrt{n}}\boldsymbol{W}^\top\boldsymbol{\varepsilon} \to \mathsf{N}(\boldsymbol{o}, \sigma^2\boldsymbol{Q})$, and show that $\sqrt{n}(\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \to \mathsf{N}(\boldsymbol{o}, \sigma^2(\boldsymbol{D}^\top\boldsymbol{Q}^{-1}\boldsymbol{D})^{-1})$*

Answer. $\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta} = \mathbf{B}_n\frac{1}{n}\boldsymbol{W}_n^\top\boldsymbol{\varepsilon}_n$, therefore $\sqrt{n}(\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) = \mathbf{B}_n n^{-1/2}\boldsymbol{W}_n^\top\boldsymbol{\varepsilon}_n \to \boldsymbol{B}\mathsf{N}(\boldsymbol{o}, \sigma^2\boldsymbol{Q}) = \mathsf{N}(\boldsymbol{o}, \sigma^2\boldsymbol{B}\boldsymbol{Q}\boldsymbol{B}^\top)$. Since $\boldsymbol{B} = (\boldsymbol{D}^\top\boldsymbol{Q}^{-1}\boldsymbol{D})^{-1}\boldsymbol{D}^\top\boldsymbol{Q}^{-1}$, the result follows. □

From Problem 462 follows that for finite samples approximately $\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta} \sim N\big(\boldsymbol{o}, \frac{\sigma^2}{n}(\boldsymbol{D}^\top\boldsymbol{Q}^{-1}\boldsymbol{D})^{-1}\big)$. Since $\frac{1}{n}(\boldsymbol{D}^\top\boldsymbol{Q}^{-1}\boldsymbol{D})^{-1} = (n\boldsymbol{D}^\top(n\boldsymbol{Q})^{-1}n\boldsymbol{D})^{-1}$, $\mathcal{MSE}[\tilde{\boldsymbol{\beta}}; \boldsymbol{\beta}]$ can be estimated by $\mathrm{s}^2\Big(\mathbf{X}^\top\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top\mathbf{X}\Big)^{-1}$

The estimator (52.0.4) is sometimes called the two stages least squares estimate, because the projection of $\mathbf{X}$ on the column space of $\boldsymbol{W}$ can be considered the predicted

values if one regresses every column of $\mathbf{X}$ on $\boldsymbol{W}$. I.e., instead of regressing $\mathbf{y}$ on $\mathbf{X}$ one regresses $\mathbf{y}$ on those linear combinations of the columns of $\boldsymbol{W}$ which best approximate the columns of $\mathbf{X}$. Here is more detail: the matrix of estimated coefficients in the first regression is $\hat{\boldsymbol{\Pi}} = (\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top\mathbf{X}$, and the predicted values in this regression are $\hat{\mathbf{X}} = \boldsymbol{W}\hat{\boldsymbol{\Pi}} = \boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top\mathbf{X}$. The second regression, which regresses $\mathbf{y}$ on $\hat{\mathbf{X}}$, gives the coefficient vector

$$\tilde{\boldsymbol{\beta}} = (\hat{\mathbf{X}}^\top\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}^\top\mathbf{y}. \tag{52.0.6}$$

If you plug this in you see this is exactly (52.0.4) again.

Now let's look at the geometry of instrumental variable regression of one variable $\mathbf{y}$ on one other variable $\mathbf{x}$ with $\boldsymbol{w}$ as an instrument. The specification is $\mathbf{y} = \mathbf{x}\beta + \boldsymbol{\varepsilon}$. On p. 437 we visualized the asymptotic results if $\boldsymbol{\varepsilon}$ is asymptotically orthogonal to $\mathbf{x}$. Now let us assume $\boldsymbol{\varepsilon}$ is asymptotically not orthogonal to $\mathbf{x}$. One can visualize this as three vectors, again normalized by dividing by $\sqrt{n}$, but now even in the asymptotic case the $\boldsymbol{\varepsilon}$-vector is not orthogonal to $\mathbf{x}$. (Draw $\boldsymbol{\varepsilon}$ vertically, and make $\mathbf{x}$ long enough that $\beta < 1$.) We assume n is large enough so that the asymptotic results hold for the sample already (or, perhaps better, that the difference between the sample and its plim is only infinitesimal). Therefore the OLS regression, with estimates β by $\mathbf{x}^\top\mathbf{y}/\mathbf{x}^\top\mathbf{x}$, is inconsistent. Let O be the origin, A the point on the $\mathbf{x}$-vector where $\boldsymbol{\varepsilon}$ branches off (i.e., the end of $\mathbf{x}\beta$), furthermore let B be the point on the $\mathbf{x}$-vector where the orthogonal projection of $\mathbf{y}$ comes down, and C the end of the $\mathbf{x}$-vector. Then $\mathbf{x}^\top\mathbf{y} = \bar{OC}\,\bar{OB}$ and $\mathbf{x}^\top\mathbf{x} = \bar{OC}^2$, therefore $\mathbf{x}^\top\mathbf{y}/\mathbf{x}^\top\mathbf{x} = \bar{OB}/\bar{OC}$, which would be the β if the errors were orthogonal. Now introduce a new variable $\boldsymbol{w}$ which is orthogonal to the errors. (Since $\boldsymbol{\varepsilon}$ is vertical, $\boldsymbol{w}$ is on the horizontal axis.) Call D the projection of $\mathbf{y}$ on $\boldsymbol{w}$, which is the prolongation of the vector $\boldsymbol{\varepsilon}$, and call E the end of the $\boldsymbol{w}$-vector, and call F the projection of $\mathbf{x}$ on $\boldsymbol{w}$. Then $\boldsymbol{w}^\top\mathbf{y} = \bar{OE}\,\bar{OD}$, and $\boldsymbol{w}^\top\mathbf{x} = \bar{OE}\,\bar{OF}$. Therefore $\boldsymbol{w}^\top\mathbf{y}/\boldsymbol{w}^\top\mathbf{x} = (\bar{OE}\,\bar{OD})(\bar{OE}\,\bar{OF}) = \bar{OD}/\bar{OF} = \bar{OA}/\bar{OC} = \beta$. Or geometrically it is obvious that the regression of $\mathbf{y}$ on the projection of $\mathbf{x}$ on $\boldsymbol{w}$ will give the right $\hat{\beta}$. One also sees here why the s^2 based on this second regression is inconsistent.

If I allow two instruments, the two instruments must be in the horizontal plane perpendicular to the vector $\boldsymbol{\varepsilon}$ which is assumed still vertical. Here we project $\mathbf{x}$ on this horizontal plane and then regress the $\mathbf{y}$, which stays where it is, on this $\mathbf{x}$. In this way the residuals have the right direction!

What if there is one instrument, but it does not not lie in the same plane as $\mathbf{x}$ and $\mathbf{y}$? This is the most general case as long as there is only one regressor and one instrument. This instrument $\boldsymbol{w}$ must lie somewhere in the horizontal plane. We have to project $\mathbf{x}$ on it, and then regress $\mathbf{y}$ on this projection. Look at it this way: take the plane orthogonal to $\boldsymbol{w}$ which goes through point C. The projection of $\mathbf{x}$ on $\boldsymbol{w}$ is the intersection of the ray generated by $\boldsymbol{w}$ with this plane. Now move this

plane parallel until it intersects point A. Then the intersection with the $\boldsymbol{w}$-ray is the projection of $\mathbf{y}$ on $\boldsymbol{w}$. But this latter plane contains $\boldsymbol{\varepsilon}$, since $\boldsymbol{\varepsilon}$ is orthogonal to $\boldsymbol{w}$. This makes sure that the regression gives the right results.

Problem 463. *4 points The asymptotic $\mathcal{MSE}$ matrix of the instrumental variables estimator with $\mathbf{W}$ as matrix of instruments is $\sigma^2 \operatorname{plim}\left(\mathbf{X}^\top\mathbf{W}(\mathbf{W}^\top\mathbf{W})^{-1}\mathbf{W}^\top\mathbf{X}\right)^{-1}$ Show that if one adds more instruments, then this asymptotic $\mathcal{MSE}$-matrix can only decrease. It is sufficient to show that the inequality holds before going over to the* plim*, i.e., if $\mathbf{W} = \begin{bmatrix}\mathbf{U} & \mathbf{V}\end{bmatrix}$, then*

$$\left(\mathbf{X}^\top\mathbf{U}(\mathbf{U}^\top\mathbf{U})^{-1}\mathbf{U}^\top\mathbf{X}\right)^{-1} - \left(\mathbf{X}^\top\mathbf{W}(\mathbf{W}^\top\mathbf{W})^{-1}\mathbf{W}^\top\mathbf{X}\right)^{-1} \tag{52.0.7}$$

is nonnegative definite. Hints: (1) Use theorem A.5.5 in the Appendix (proof is not required). (2) Note that $\mathbf{U} = \boldsymbol{WG}$ for some $\boldsymbol{G}$. Can you write this $\boldsymbol{G}$ in partitioned matrix form? (3) Show that, whatever $\mathbf{W}$ and $\boldsymbol{G}$, $\mathbf{W}(\mathbf{W}^\top\mathbf{W})^{-1}\mathbf{W}^\top - \boldsymbol{WG}(\boldsymbol{G}^\top\mathbf{W}^\top\boldsymbol{WG})^{-1}\boldsymbol{G}^\top\mathbf{W}^\top$ is idempotent.

Answer.

$$\mathbf{U} = \begin{bmatrix}\mathbf{U} & \mathbf{V}\end{bmatrix}\begin{bmatrix}\boldsymbol{I}\\ \boldsymbol{O}\end{bmatrix} = \boldsymbol{WG} \quad \text{where} \quad \boldsymbol{G} = \begin{bmatrix}\boldsymbol{I}\\ \boldsymbol{O}\end{bmatrix}. \tag{52.0.8}$$

□

Problem 464. *2 points Show: if a matrix $\boldsymbol{D}$ has full column rank and is square, then it has an inverse.*

Answer. Here you need that column rank is row rank: if $\boldsymbol{D}$ has full column rank it also has full row rank. And to make the proof complete you need: if $\boldsymbol{A}$ has a left inverse $\boldsymbol{L}$ and a right inverse $\boldsymbol{R}$, then $\boldsymbol{L}$ is the only left inverse and $\boldsymbol{R}$ the only right inverse and $\boldsymbol{L} = \boldsymbol{R}$. Proof: $\boldsymbol{L} = \boldsymbol{L}(\boldsymbol{AR}) = (\boldsymbol{LA})\boldsymbol{R} = \boldsymbol{R}$. □

Problem 465. *2 points If $\mathbf{W}^\top\mathbf{X}$ is square and has full column rank, then it is nonsingular. Show that in this case* (52.0.4) *simplifies to the "simple" instrumental variables estimator:*

$$\tilde{\boldsymbol{\beta}} = (\mathbf{W}^\top\mathbf{X})^{-1}\mathbf{W}^\top\mathbf{y} \tag{52.0.9}$$

Answer. In this case the big inverse can be split into three:

$$\tilde{\boldsymbol{\beta}} = \left(\mathbf{X}^\top\mathbf{W}(\mathbf{W}^\top\mathbf{W})^{-1}\mathbf{W}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{W}(\mathbf{W}^\top\mathbf{W})^{-1}\mathbf{W}^\top\mathbf{y} = \tag{52.0.10}$$

$$= (\mathbf{W}^\top\mathbf{X})^{-1}\mathbf{W}^\top\mathbf{W}(\mathbf{X}^\top\mathbf{W})^{-1}\mathbf{X}^\top\mathbf{W}(\mathbf{W}^\top\mathbf{W})^{-1}\mathbf{W}^\top\mathbf{y} \tag{52.0.11}$$

□

PROBLEM 466. *We only have one regressor with intercept, i.e.,* $\mathbf{X} = \begin{bmatrix} \boldsymbol{\iota} & \mathbf{x} \end{bmatrix}$, *and we have one instrument* $\boldsymbol{w}$ *for* $\mathbf{x}$ *(while the constant term is its own instrument), i.e.,* $\boldsymbol{W} = \begin{bmatrix} \boldsymbol{\iota} & \boldsymbol{w} \end{bmatrix}$. *Show that the instrumental variables estimators for slope and intercept are*

$$\tilde{\beta} = \frac{\sum(w_t - \bar{w})(y_t - \bar{y})}{\sum(w_t - \bar{w})(x_t - \bar{x})} \tag{52.0.12}$$

$$\tilde{\alpha} = \bar{y} - \tilde{\beta}\bar{x} \tag{52.0.13}$$

Hint: the math is identical to that in question 238.

PROBLEM 467. *2 points Show that, if there are as many instruments as there are observations, then the instrumental variables estimator* (52.0.4) *becomes identical to OLS.*

ANSWER. In this case $\boldsymbol{W}$ has an inverse, therefore the projection on $\mathrm{R}[\boldsymbol{W}]$ is the identity. Staying in the algebraic paradigm, $(\boldsymbol{W}^\top\boldsymbol{W})^{-1} = \boldsymbol{W}^{-1}(\boldsymbol{W}^\top)^{-1}$. □

An implication of Problem 467 is that one must be careful not to include too many instruments if one has a small sample. Asymptotically it is better to have more instruments, but for $n = m$, the instrumental variables estimator is equal to OLS, i.e., the sequence of instrumental variables estimators starts at the (inconsistent) OLS. If one uses fewer instruments, then the asymptotic $\mathcal{MSE}$ matrix is not so good, but one may get a sequence of estimators which moves away from the inconsistent OLS more quickly.

CHAPTER 53

Errors in Variables

53.1. The Simplest Errors-in-Variables Model

We will explain here the main principles of errors in variables by the example of simple regression, in which y is regressed on one explanatory variable with a constant term. Assume the explanatory variable is a random variable, called x^*, and the disturbance term in the regression, which is a zero mean random variable independent of x^*, will be called v. In other words, we have the following relationship between random variables:

$$\mathsf{y} = \alpha + \mathsf{x}^*\beta + \mathsf{v}. \tag{53.1.1}$$

If n observations of the variables y and x^* are available, one can obtain estimates of α and β and predicted values of the disturbances by running a regression of the vector of observations $\boldsymbol{y}$ on $\boldsymbol{x}^*$:

$$\boldsymbol{y} = \boldsymbol{\iota}\alpha + \boldsymbol{x}^*\beta + \boldsymbol{v}. \tag{53.1.2}$$

But now let us assume that x^* can only be observed with a random error. I.e., we observe $\mathsf{x} = \mathsf{x}^* + \mathsf{u}$. The error u is assumed to have zero mean, and to be independent of x^* and v. Therefore we have the model with the "latent" variable x^*:

$$\mathsf{y} = \alpha + \mathsf{x}^*\beta + \mathsf{v} \tag{53.1.3}$$

$$\mathsf{x} = \mathsf{x}^* + \mathsf{u} \tag{53.1.4}$$

This model is sometimes called "regression with both variables subject to error." It is symmetric between the dependent and the explanatory variable, because one can also write it as

$$\mathsf{y}^* = \alpha + \mathsf{x}^*\beta \tag{53.1.5}$$

$$\mathsf{x} = \mathsf{x}^* + \mathsf{u} \tag{53.1.6}$$

$$\mathsf{y} = \mathsf{y}^* + \mathsf{v} \tag{53.1.7}$$

and, as long as $\beta \neq 0$, $\mathsf{y}^* = \alpha + \mathsf{x}^*\beta$ is equivalent to $\mathsf{x}^* = -\alpha/\beta + \mathsf{y}^*/\beta$.

What happens if this is the true model and one regresses $\boldsymbol{y}$ on $\boldsymbol{x}$? Plug $\boldsymbol{x}^* = \boldsymbol{x} - \boldsymbol{u}$ into (53.1.2):

$$\boldsymbol{y} = \boldsymbol{\iota}\alpha + \boldsymbol{x}\beta + \underbrace{(\boldsymbol{v} - \boldsymbol{u}\beta)}_{\boldsymbol{\varepsilon}} \tag{53.1.8}$$

The problem is that the disturbance term ε is correlated with the explanatory variable:

$$\operatorname{cov}[\mathsf{x}, \varepsilon] = \operatorname{cov}[\mathsf{x}^* + \mathsf{u}, \mathsf{v} - \mathsf{u}\beta] = -\beta \operatorname{var}[\mathsf{u}]. \tag{53.1.9}$$

Therefore OLS will give inconsistent estimates of α and β:

$$\hat{\beta}_{OLS} = \frac{\sum(y_i - \bar{y})(x_i - \bar{x})}{\sum(x_i - \bar{x})^2} \tag{53.1.10}$$

$$\operatorname{plim} \hat{\beta}_{OLS} = \frac{\operatorname{cov}[\mathsf{y}, \mathsf{x}]}{\operatorname{var}[\mathsf{x}]} = \beta\Big(1 - \frac{\operatorname{var}[\mathsf{u}]}{\operatorname{var}[\mathsf{x}]}\Big). \tag{53.1.11}$$

Since $\operatorname{var}[\mathsf{u}] \leq \operatorname{var}[\mathsf{x}]$, $\hat{\beta}_{OLS}$ will have the right sign in the plim, but its absolute value will understimate the true β.

Problem 468. *1 point* [**SM86**, A3.2/3] *Assume the variance of the measurement error σ^2_{u} is 10% of the variance of the unobserved exogenous variable $\sigma^2_{\mathsf{x}^*}$. By how many percent will then the OLS estimator $\hat{\beta}_{OLS}$ asymptotically underestimate the absolute value of the true parameter β?*

Answer. $1 - \operatorname{var}[\mathsf{u}]/\operatorname{var}[\mathsf{x}] = 1 - 0.1/1.1 = 0.90909$, which is 9.09% below 1. □

Although $\hat{\beta}_{OLS}$ is not a consistent estimator of the underlying parameter β, it nevertheless converges towards a meaningful magnitude, namely, the best linear predictor of y on the basis of x, which characterizes the *empirical relation* between x and y in the above model.

What is the difference between the underlying *structural* relationship between the two variables and their *empirical* relationship? Assume for a moment that x is observed and y is not observed, but one knows the means $\begin{bmatrix}\operatorname{E}[\mathsf{x}]\\ \operatorname{E}[\mathsf{y}]\end{bmatrix}$ and the covariance matrix $\begin{bmatrix}\operatorname{var}[\mathsf{x}] & \operatorname{cov}[\mathsf{x},\mathsf{y}]\\ \operatorname{cov}[\mathsf{x},\mathsf{y}] & \operatorname{var}[\mathsf{y}]\end{bmatrix}$. Then the best linear predictor of y based on the observation of x is

$$\mathsf{y}^* = \operatorname{E}[\mathsf{y}] + \frac{\operatorname{cov}[\mathsf{y},\mathsf{x}]}{\operatorname{var}[\mathsf{x}]}(\mathsf{x} - \operatorname{E}[\mathsf{x}]) = \Big(\operatorname{E}[\mathsf{y}] - \frac{\operatorname{cov}[\mathsf{y},\mathsf{x}]}{\operatorname{var}[\mathsf{x}]}\operatorname{E}[\mathsf{x}]\Big) + \frac{\operatorname{cov}[\mathsf{y},\mathsf{x}]}{\operatorname{var}[\mathsf{x}]}\mathsf{x} \tag{53.1.12}$$

This is a linear transformation of x, whose slope and intercept are not necessarity equal to the "underlying" α and β. One sees that the slope is exactly what $\hat{\beta}_{OLS}$ converges to, and question 469 shows that the intercept is the plim of $\hat{\alpha}_{OLS}$.

PROBLEM 469. *Compute* $\operatorname{plim}\hat{\alpha}_{OLS}$.

Is it possible to estimate the true underlying parameters α and β consistently? Not if they are jointly normal. For this look at the following two scenarios:

$$\begin{aligned} \mathsf{y}^* &= 2\mathsf{x}^* - 100 & & \mathsf{y}^* = \mathsf{x}^* \\ \mathsf{x}^* &\sim \mathsf{N}(100, 100) & & \mathsf{x}^* \sim \mathsf{N}(100, 200) \\ \mathsf{v} &\sim \mathsf{N}(0, 200) & \text{versus} & \mathsf{v} \sim \mathsf{N}(0, 400) \\ \mathsf{u} &\sim \mathsf{N}(0, 200) & & \mathsf{u} \sim \mathsf{N}(0, 100) \end{aligned} \tag{53.1.13}$$

They lead to identical joint distributions of x and y, although the underlying parameters are different. Therefore the model is unidentified.

PROBLEM 470. *Compute means, variances, and the correlation coefficient of* x *and* y *in both versions of* (53.1.13).

ANSWER. First the joint distributions of y^* and x^*:

$$\begin{bmatrix} \mathsf{y}^* \\ \mathsf{x}^* \end{bmatrix} \sim N\left(\begin{bmatrix} 100 \\ 100 \end{bmatrix}, \begin{bmatrix} 400 & 200 \\ 200 & 100 \end{bmatrix}\right) \quad \text{versus} \quad \begin{bmatrix} \mathsf{y}^* \\ \mathsf{x}^* \end{bmatrix} \sim N\left(\begin{bmatrix} 100 \\ 100 \end{bmatrix}, \begin{bmatrix} 200 & 200 \\ 200 & 200 \end{bmatrix}\right). \tag{53.1.14}$$

Add to this the independent

$$\begin{bmatrix} \mathsf{v} \\ \mathsf{u} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 200 & 0 \\ 0 & 200 \end{bmatrix}\right) \quad \text{versus} \quad \begin{bmatrix} \mathsf{v} \\ \mathsf{u} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 400 & 0 \\ 0 & 100 \end{bmatrix}\right). \tag{53.1.15}$$

□

PROBLEM 471. *Compute a third specification of the underlying relationship between* x^* *and* y^**, the mean and variance of* x^**, and the error variances, which leads again to the same joint distribution of* x *and* y*, and under which the OLS estimate is indeed a consistent estimate of the underlying relationship.*

53.1.1. Three Restrictions on the True Parameteres. The lack of identification means that the mean vector and dispersion matrix of the observed variables are compatible with many different values of the underlying parameters. But this lack of identification is not complete; the data give three important restrictions for the true parameters.

Equation (53.1.5) implies for the means

$$\begin{bmatrix} \mu_{\mathsf{y}} & \mu_{\mathsf{x}} \end{bmatrix} = \begin{bmatrix} \mu_{\mathsf{y}^*} & \mu_{\mathsf{x}^*} \end{bmatrix} = \begin{bmatrix} \alpha + \mu_{\mathsf{x}^*}\beta & \mu_{\mathsf{x}^*} \end{bmatrix}, \tag{53.1.16}$$

and variances and covariances satisfy, due to (53.1.6) and (53.1.7),

$$\begin{aligned} \begin{bmatrix} \sigma^2_{\mathsf{y}} & \sigma_{\mathsf{xy}} \\ \sigma_{\mathsf{xy}} & \sigma^2_{\mathsf{x}} \end{bmatrix} &= \begin{bmatrix} \sigma^2_{\mathsf{y}^*} & \sigma_{\mathsf{x}^*\mathsf{y}^*} \\ \sigma_{\mathsf{x}^*\mathsf{y}^*} & \sigma^2_{\mathsf{x}^*} \end{bmatrix} + \begin{bmatrix} \sigma^2_{\mathsf{v}} & 0 \\ 0 & \sigma^2_{\mathsf{u}} \end{bmatrix} \\ &= \begin{bmatrix} \beta^2\sigma^2_{\mathsf{x}^*} & \beta\sigma^2_{\mathsf{x}^*} \\ \beta\sigma^2_{\mathsf{x}^*} & \sigma^2_{\mathsf{x}^*} \end{bmatrix} + \begin{bmatrix} \sigma^2_{\mathsf{v}} & 0 \\ 0 & \sigma^2_{\mathsf{u}} \end{bmatrix}. \end{aligned} \tag{53.1.17}$$

We know five moments of the observed variables: μ_{y}, μ_{x}, σ^2_{y}, σ_{xy}, and σ^2_{x}; but there are six independent parameters of the model: α, β, μ_{x^*}, $\sigma^2_{\mathsf{x}^*}$, σ^2_{v}, σ^2_{u}. It is therefore no wonder that the parameters cannot be determined uniquely from the knowledge of means and variances of the observed variables, as shown by counterexample (53.1.13). However α and β cannot be chosen arbitrarily either. The above equations imply three constraints on these parameters.

The first restriction on the parameters comes from equation (53.1.16) for the means: From $\mu_{\mathsf{y}^*} = \alpha + \beta\mu_{\mathsf{x}^*}$ follows, since $\mu_{\mathsf{y}} = \mu_{\mathsf{y}^*}$ and $\mu_{\mathsf{x}} = \mu_{\mathsf{x}^*}$, that

$$\mu_{\mathsf{y}} = \alpha + \beta\mu_{\mathsf{x}}, \tag{53.1.18}$$

i.e., all true underlying relationships compatible with the means and variances of the observed variables go through the same point $\begin{bmatrix} \mu_{\mathsf{x}} \\ \mu_{\mathsf{y}} \end{bmatrix}$.

If $\sigma_{\mathsf{xy}} = 0$, this is the only restriction on the parameter vectors. To see this, remember $\sigma_{\mathsf{xy}} = \beta\sigma^2_{\mathsf{x}^*}$. This product is zero if either $\sigma^2_{\mathsf{x}^*} = 0$, or $\sigma^2_{\mathsf{x}^*} \neq 0$ and $\beta = 0$. If $\sigma^2_{\mathsf{x}^*} = 0$, then x^* and therefore also y^* are constants. Any two constants satisfy infinitely many affine relationships, and all α and β which satisfy the first constraint are possible parameter vectors which all describe the same affine relationship between x^* and y^*. In the other case, if $\sigma^2_{\mathsf{x}^*} \neq 0$ and $\beta = 0$, then the linear relation underlying the observations has coefficient zero, they are noisy observations of two linearly unrelated variables.

In the regular case $\sigma_{\mathsf{xy}} \neq 0$, condition (53.1.17) for the dispersion matrices gives two more restrictions on the parameter vectors. From $\sigma_{\mathsf{xy}} = \beta\sigma^2_{\mathsf{x}^*}$ follows the second restriction on the parameters:

$$\beta \text{ must have the same sign as } \sigma_{\mathsf{xy}}. \tag{53.1.19}$$

And here is a derivation of the third restriction (53.1.23): from

$$0 \leq \sigma^2_{\mathsf{u}} = \sigma^2_{\mathsf{x}} - \sigma^2_{\mathsf{x}^*} \quad \text{and} \quad 0 \leq \sigma^2_{\mathsf{v}} = \sigma^2_{\mathsf{y}} - \beta^2\sigma^2_{\mathsf{x}^*} \tag{53.1.20}$$

follows

$$\sigma^2_{\mathsf{x}^*} \leq \sigma^2_{\mathsf{x}} \quad \text{and} \quad \beta^2\sigma^2_{\mathsf{x}^*} \leq \sigma^2_{\mathsf{y}}. \tag{53.1.21}$$

Multiply the first inequality by $|\beta|$ and substitute in both inequalities σ_{xy} for $\beta\sigma^2_{\mathsf{x}^*}$:

$$|\sigma_{\mathsf{xy}}| \leq |\beta|\,\sigma^2_{\mathsf{x}} \quad \text{and} \quad |\beta|\,|\sigma_{\mathsf{xy}}| \leq \sigma^2_{\mathsf{y}} \tag{53.1.22}$$

or

$$\frac{|\sigma_{\mathsf{xy}}|}{\sigma^2_{\mathsf{x}}} \leq |\beta| \leq \frac{\sigma^2_{\mathsf{y}}}{|\sigma_{\mathsf{xy}}|}. \tag{53.1.23}$$

The lower bound is the absolute value of the plim of the regression coefficient if one regresses the observations of y on those of x, and the reciprocal of the upper bound is the absolute value of the plim of the regression coefficient if one regresses the observed values of x on those of y.

PROBLEM 472. *We have seen that the data generated by the two processes* (53.1.13) *do not determine the underlying relationship completely. What restrictions do these data impose on the parameters α and β of the underlying relation $\mathsf{y}^* = \alpha + \beta\mathsf{x}^*$?*

PROBLEM 473. *The model is $\mathsf{y} = \alpha + \mathsf{x}^*\beta + \mathsf{v}$, but x^* is not observed; one can only observe $\mathsf{x} = \mathsf{x}^* + \mathsf{u}$. The errors u and v have zero expected value and are independent of each other and of x^*. You have lots of data available, and for the sake of the argument we assume that the joint distribution of x and y is known precisely: it is*

$$\begin{bmatrix}\mathsf{y}\\ \mathsf{x}\end{bmatrix} \sim N\Bigl(\begin{bmatrix}1\\ -1\end{bmatrix}, \begin{bmatrix}6 & -2\\ -2 & 3\end{bmatrix}\Bigr). \tag{53.1.24}$$

• a. *3 points What does the information about y and x given in equation* (53.1.24) *imply about α and β?*

ANSWER. (53.1.18) gives $\alpha - \beta = 1$, (53.1.19) gives $\beta \leq 0$, and (53.1.23) $2/3 \leq |\beta| \leq 3$. □

• b. *3 points Give the* plim*s of the OLS estimates of α and β in the regression of y on x.*

ANSWER. $\operatorname{plim}\hat\beta = \operatorname{cov}[\mathsf{x},\mathsf{y}]/\operatorname{var}[\mathsf{x}] = -\frac{2}{3}$, $\operatorname{plim}\hat\alpha = \operatorname{E}[\mathsf{y}] - \operatorname{E}[\mathsf{x}]\operatorname{plim}\hat\beta = \frac{1}{3}$. □

• c. *3 points Now assume it is known that $\alpha = 0$. What can you say now about β, σ^2_{u}, and σ^2_{v}? If β is identified, how would you estimate it?*

ANSWER. From $\mathsf{y} = (\mathsf{x} - \mathsf{u})\beta + \mathsf{v}$ follows, by taking expectations, $\operatorname{E}[\mathsf{y}] = \operatorname{E}[\mathsf{x}]\beta$ (i.e., the true relationship still goes through the means), therefore $\beta = -1$, and a consistent estimate would be $\bar{\mathsf{y}}/\bar{\mathsf{x}}$. Now if one knows β one gets $\operatorname{var}[\mathsf{x}^*]$ from $\operatorname{cov}[\mathsf{x},\mathsf{y}] = \operatorname{cov}[\mathsf{x}^*+\mathsf{u}, \beta\mathsf{x}^*+\mathsf{v}] = \beta\operatorname{var}[\mathsf{x}^*]$, i.e., $\operatorname{var}[\mathsf{x}^*] = 2$. Then one can get $\operatorname{var}[\mathsf{u}] = \operatorname{var}[\mathsf{x}] - \operatorname{var}[\mathsf{x}^*] = 3 - 2 = 1$, and $\operatorname{var}[\mathsf{v}] = \operatorname{var}[\mathsf{y}] - \operatorname{var}[\mathsf{y}^*] = 6 - 2 = 4$. Luckily, those variances came out to be positive; otherwise the restriction $\alpha = 0$ would not be compatible with (53.1.24). □

53.2. General Definition of the EV Model

Given a $n \times k$ matrix $\mathbf{X}$ whose columns represent the observed variables. These observations are generated by a linear *EV* model if the following holds:

$$\mathbf{X} = \mathbf{X}^* + \mathbf{U}, \tag{53.2.1}$$

$$\boldsymbol{X}^*\boldsymbol{B} = \boldsymbol{O} \tag{53.2.2}$$

$\mathbf{X}^*$ is an $n \times k$ matrix of the values of the unobserved "systematic" or "latent" variables. We assume that $\boldsymbol{Q}^* = \operatorname{plim}_{n\to\infty}\frac{1}{n}\mathbf{X}^{*\top}\mathbf{X}^*$ exists. If the systematic variables

are independent observations from the same joint distribution, this means that first and second order moments exist. If the systematic variables are nonrandom, the plim becomes the ordinary limit. These two special cases are called the "structural variant" and the "functional variant."

$\mathbf{U}$ is the $n \times k$ matrix of the values of the unobserved "errors" or "statistical disturbances." These errors are assumed to be random; they have zero expectations, the rows of $\mathbf{U}$ are independent and identically distributed with covariance matrix $\widetilde{\boldsymbol{Q}}$. If the systematic variables are random as well, then we assume that the errors are independent of them.

The letter $\boldsymbol{B}$ in (53.2.2) is an upper-case Greek $\boldsymbol{\beta}$, the columns of $\boldsymbol{B}$ will therefore be written $\boldsymbol{\beta}_i$. Every such column constitutes a linear relation between the systematic variables. $\boldsymbol{B}$ is assumed to be exhaustive in the sense that for any vector $\boldsymbol{\gamma}$ which satisfies $\mathbf{X}^*\boldsymbol{\gamma} = \boldsymbol{o}$ there is a vector $\boldsymbol{q}$ so that $\boldsymbol{\gamma} = \boldsymbol{B}\boldsymbol{q}$. The rank of $\boldsymbol{B}$ is denoted by q. If $q = 1$, then only one linear relation holds, and the model is called a univariate EV model, otherwise it is a multivariate EV model.

In this specification, there are therefore one or several exact linear relations between the true variables $\mathbf{X}^*$, but $\mathbf{X}^*$ can only be observed with error. The task of getting an estimate of these linear relations has been appropriately called by Kalman "identification of linear relations from noisy data" [**Kal83**, p. 119], compare also the title of [**Kal82**]. One can say, among the columns of $\mathbf{X}$ there is both a stochastic relationship and a linear relationship, and one wants to extract the linear relationship from this mixture.

The above are the minimum assumptions which we will make of each of the models below. From these assumptions follows that

$$\operatorname*{plim}_{n\to\infty}\frac{1}{n}\mathbf{X}^{*\top}\mathbf{U} = \boldsymbol{O} \tag{53.2.3}$$

Proof: First we prove it for the case that the systematic variables are nonrandom, in which case we write them $\boldsymbol{X}^*$. Since the expected value $\operatorname{E}[\mathbf{U}] = \boldsymbol{O}$, also $\operatorname{E}[\frac{1}{n}\mathbf{U}^\top\boldsymbol{X}^*] = \boldsymbol{O}$. For (53.2.3) it is sufficient to show that the variances of these arithmetic means converge to zero: if their expected value is zero and their variance converges to zero, their plim is zero as well. The i,k element of $\boldsymbol{X}^{*\top}\mathbf{U}$ is $\sum_{j=1}^n x^*{}_{ji}\mathsf{u}_{jk}$, or it can also be written as $\boldsymbol{x}^*{}_i^\top\mathbf{u}_k$, it is the scalar product of the ith column of $\boldsymbol{X}^*$ with the kth column of $\mathbf{U}$. Since all the elements in the kth column of $\mathbf{U}$ have same variance, namely, $\operatorname{var}[\mathsf{u}_{jk}] = \tilde{q}_{kk}$, and since u_{jk} is independent of u_{mk} for $j \neq m$, it follows

$$\operatorname{var}\Bigl[\frac{1}{n}\sum_j x^*{}_{ji}\mathsf{u}_{jk}\Bigr] = \frac{\tilde{q}_{ii}}{n}\frac{1}{n}\sum_n x^{*2}_{ji}. \tag{53.2.4}$$

Since $\operatorname{plim}\frac{1}{n}\sum_j x^{*2}_{ji}$ exists (it is $q^*{}_{jj}$), (53.2.4) converges toward zero.

Given this special case, the general case follows by an argument of the form: since this plim exists and is the same conditionally on any realization of $\mathbf{X}^*$, it also exists unconditionally.

Other assumptions, made frequently, are: the covariance matrix of the errors $\widetilde{\boldsymbol{Q}}$ is p.d. and/or diagonal.

There may also be linear restrictions on $\boldsymbol{B}$, and restrictions on the elements of $\widetilde{\boldsymbol{Q}}$.

An important extension is the following: If the columns of $\mathbf{X}^*$ and $\mathbf{U}$ are realizations of (weakly) stationary stochastic processes, and/or $\mathbf{X}^*$ contains lagged variables, then one speaks of a dynamic EV model. Here the rows of $\mathbf{U}$ are no longer independent.

53.3. Particular Forms of EV Models

The matrix of parameters is not uniquely determined. In order to remove this ambiguity, one often requires that it have the form $\begin{bmatrix} -\boldsymbol{I} \\ \boldsymbol{B} \end{bmatrix}$, where $\boldsymbol{I}$ is a $q \times q$ identity matrix. Such a form can always be achieved by rearranging the variables and/or going over to linear combinations. Any symmetric EV model is equivalent to one in which the parameter matrix has this form, after an appropriate linear transformation of the variables. Partitioning the vectors of systematic variables and errors conformably, one obtains the following form of the EV model:

$$\begin{aligned} \mathbf{Y}^* &= \mathbf{X}^*\boldsymbol{B} \\ \mathbf{Y} &= \mathbf{Y}^* + \mathbf{V} \\ \mathbf{X} &= \mathbf{X}^* + \mathbf{U} \end{aligned} \quad \text{i.e.,} \quad \begin{aligned} \begin{bmatrix} \mathbf{Y}^* & \mathbf{X}^* \end{bmatrix} \begin{bmatrix} -\boldsymbol{I} \\ \boldsymbol{B} \end{bmatrix} &= \boldsymbol{O} \\ \begin{bmatrix} \mathbf{Y} & \mathbf{X} \end{bmatrix} &= \begin{bmatrix} \mathbf{Y}^* & \mathbf{X}^* \end{bmatrix} + \begin{bmatrix} \mathbf{V} & \mathbf{U} \end{bmatrix} \end{aligned} \tag{53.3.1}$$

The OLS model is a special case of the errors in variables model. Using the definition $\mathbf{y}^* = \mathbf{X}^*\boldsymbol{\beta}$, i.e., $\mathbf{y}^*$ is the vector which $\hat{\mathbf{y}}$ estimates, one can write the regression model in the form

$$\begin{aligned} \mathbf{y}^* &= \mathbf{X}^*\boldsymbol{\beta} \\ \mathbf{y} &= \mathbf{y}^* + \boldsymbol{\nu} \\ \mathbf{X} &= \mathbf{X}^* \end{aligned} \quad \text{or in the symmetric form} \quad \begin{aligned} \begin{bmatrix} \mathbf{y}^* & \mathbf{X}^* \end{bmatrix} \begin{bmatrix} -1 \\ \boldsymbol{\beta} \end{bmatrix} &= \boldsymbol{o} \\ \begin{bmatrix} \mathbf{y} & \mathbf{X} \end{bmatrix} &= \begin{bmatrix} \mathbf{y}^* & \mathbf{X}^* \end{bmatrix} + \begin{bmatrix} \boldsymbol{\nu} & \boldsymbol{O} \end{bmatrix}. \end{aligned}$$

If there is a single "bad" variable, say it is $\mathbf{x}$, and we will call the matrix of the "good" variables $\mathbf{Z}$, then the univariate EV model has the form

$$\begin{aligned} \mathbf{y}^* &= \mathbf{x}^*\beta + \mathbf{Z}\boldsymbol{\gamma} \\ \mathbf{x} &= \mathbf{x}^* + \mathbf{u} \\ \mathbf{y} &= \mathbf{y}^* + \boldsymbol{\nu} \\ \mathbf{Z} &= \mathbf{Z}^* \end{aligned} \quad \text{or} \quad \begin{aligned} \begin{bmatrix} \mathbf{y}^* & \mathbf{x}^* & \mathbf{Z}^* \end{bmatrix} \begin{bmatrix} -1 \\ \beta \\ \boldsymbol{\gamma} \end{bmatrix} &= \boldsymbol{o} \\ \begin{bmatrix} \mathbf{y} & \mathbf{x} & \mathbf{Z} \end{bmatrix} &= \begin{bmatrix} \mathbf{y}^* & \mathbf{x}^* & \mathbf{Z}^* \end{bmatrix} + \begin{bmatrix} \boldsymbol{\nu} & \boldsymbol{w} & \boldsymbol{O} \end{bmatrix}. \end{aligned}$$

This model is discussed in [**Gre97**, p. 439/40].

The constant term in the univariate EV model can be considered the coefficient of the "pseudovariable" $\boldsymbol{\iota}$, which has the value 1 in all observations, and which is observed without error. Using $\mathbf{y}^* = \boldsymbol{\iota}\alpha + \mathbf{X}^*\boldsymbol{\beta}$, its general form is

$$\begin{bmatrix} \mathbf{y}^* & \boldsymbol{\iota} & \mathbf{X}^* \end{bmatrix} \begin{bmatrix} -1 \\ \alpha \\ \boldsymbol{\beta} \end{bmatrix} = 0 \tag{53.3.2}$$

$$\begin{bmatrix} \mathbf{y} & \boldsymbol{\iota} & \mathbf{X} \end{bmatrix} = \begin{bmatrix} \mathbf{y}^* & \boldsymbol{\iota} & \mathbf{X}^* \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon} + \boldsymbol{\nu} & \boldsymbol{o} & \mathbf{U} \end{bmatrix}. \tag{53.3.3}$$

Some well-known models, which are not usually considered EV models, are in fact special cases of the above specification.

A *Simultaneous Equations System*, as used often in econometrics, has the form

$$\mathbf{Y}\boldsymbol{\Gamma} = \mathbf{X}^*\boldsymbol{B} + \boldsymbol{E} \tag{53.3.4}$$

where $\mathbf{Y}$ (the endogenous variables) and $\mathbf{X}^*$ (the exogenous variables) are observed. $\boldsymbol{E}$ is independent of $\mathbf{X}^*$ (it characterizes the exogenous variables that they are independent of the errors). $\boldsymbol{B}$ and $\boldsymbol{\Gamma}$ are matrices of nonrandom but unknown parameter vectors, $\boldsymbol{\Gamma}$ is assumed to be nonsingular. Defining $\mathbf{Y}^* = \mathbf{X}^*\boldsymbol{B}\boldsymbol{\Gamma}^{-1}$ and $\mathbf{V} = \boldsymbol{E}\boldsymbol{\Gamma}^{-1}$, one can put this into the EV-form

$$\begin{bmatrix} \mathbf{Y}^* & \mathbf{X}^* \end{bmatrix} \begin{bmatrix} -\boldsymbol{\Gamma} \\ \boldsymbol{B} \end{bmatrix} = \boldsymbol{O} \tag{53.3.5}$$

$$\begin{bmatrix} \mathbf{Y} & \mathbf{X} \end{bmatrix} = \begin{bmatrix} \mathbf{Y}^* & \mathbf{X}^* \end{bmatrix} + \begin{bmatrix} \mathbf{V} & \boldsymbol{O} \end{bmatrix}. \tag{53.3.6}$$

If one assumes that also $\mathbf{X}^*$ is observed with errors, one obtains a simultaneous equations model with errors in the variables.

The main difference between simultaneous equations systems and EV models is that in the former, identification is usually achieved by linear constraints on the parameters, and in the latter by restrictions on the covariance matrix of the errors.

The *Factor Analytical* Model assumes the observed variables $\mathbf{X}$ are linear combinations of a small number of unobserved "factors" $\boldsymbol{\Psi}$ plus an error term which has diagonal covariance matrix, i.e.,

$$\mathbf{X} = \boldsymbol{\Psi}\boldsymbol{C} + \mathbf{U}. \tag{53.3.7}$$

There is a bijection between EV models and FA models for which we need the matrix theoretical concept of a *deficiency matrix*, which is discussed in more detail in section A.4 of the appendix. Here is only a brief overview:

Definition: We say a matrix $\boldsymbol{C}$ is a left deficiency matrix of $\boldsymbol{B}$, in symbols, $\boldsymbol{C} \perp \boldsymbol{B}$, iff $\boldsymbol{C}\boldsymbol{B} = \boldsymbol{O}$, and for all $\boldsymbol{Q}$ with $\boldsymbol{Q}\boldsymbol{B} = \boldsymbol{O}$ there is an $\boldsymbol{X}$ with $\boldsymbol{Q} = \boldsymbol{X}\boldsymbol{C}$.

This is an antisymmetric relation between the two matrices in the sense that from $\boldsymbol{C} \perp \boldsymbol{B}$ follows $\boldsymbol{B}^\top \perp \boldsymbol{C}^\top$. $\boldsymbol{C} \perp \boldsymbol{B}$ means therefore also that for all $\boldsymbol{R}$ with

$\boldsymbol{CR} = \boldsymbol{O}$ there is a $\boldsymbol{Y}$ with $\boldsymbol{R} = \boldsymbol{BY}$. We can therefore also say that $\boldsymbol{B}$ is a right deficiency matrix of $\boldsymbol{C}$.

$\boldsymbol{C} \perp \boldsymbol{B}$ simply means that the row vectors of $\boldsymbol{C}$ span the vector space orthogonal to the vector space spanned by the column vectors of $\boldsymbol{B}$. If therefore $\boldsymbol{B}$ is $k \times q$ and has rank q, then $\boldsymbol{C}$ can be chosen $(k-q) \times k$ with rank $k-q$.

Start with an EV model

$$\mathbf{X}^*\boldsymbol{B} = \boldsymbol{O} \tag{53.3.8}$$

$$\mathbf{X} = \mathbf{X}^* + \mathbf{U} \tag{53.3.9}$$

where $\boldsymbol{B}$ is a $k \times q$ matrix with rank q, and choose a $(k-q) \times k$ matrix $\boldsymbol{C} \perp \boldsymbol{B}$. Then there exists a $n \times (k-q)$ matrix $\boldsymbol{\Psi}$ with $\mathbf{X}^* = \boldsymbol{\Psi C}$, and therefore one obtains the factor analytical model $\mathbf{X} = \boldsymbol{\Psi C} + \mathbf{U}$ with $k-q$ factors. Conversely, from such a factor analytical model one can, by choosing a right deficiency matrix of $\boldsymbol{C}$, obtain an EV model. [**Mul72**], [**Kim**], [**KM78**]

A model which violates one of the assumptions of the EV model is the Berkson model. Let us discuss the simplest case

$$\mathfrak{y}^* = \alpha + \mathfrak{x}^*\beta \tag{53.3.10}$$

$$\mathfrak{y} = \mathfrak{y}^* + \nu \tag{53.3.11}$$

$$\mathfrak{x} = \mathfrak{x}^* + \mathfrak{u}. \tag{53.3.12}$$

While usually $\mathfrak{u}$ is indepenent of $\mathfrak{x}^*$, now we assume $\mathfrak{u}$ is indepenent of $\mathfrak{x}$. If this is the case, then the regression of $\mathfrak{y}$ on $\mathfrak{x}$ is unbiased and efficient:

$$\mathfrak{y} = \mathfrak{y}^* + \nu = \mathfrak{x}^*\beta + \nu = \mathfrak{x}\beta + \nu - \mathfrak{u}\beta \tag{53.3.13}$$

Here the error terms are independent of the explanatory variable.

How can this happen? Use the example with $\mathfrak{y}^*$ is voltage in volts, $\mathfrak{x}^*$ is current in amperes, and β is the resistance in Ohm. Here is a circuit diagram: the experimenter adjusts the current until the ampere meter shows, for instance, one ampere, and then he reads the voltage in volts, which are his estimate of the resistance.

53.4. The Identification Problem

In this section we will consider estimation under the hypothetical situation that we have so many observations that the sample means and variances of the columns of $\mathbf{X}$, $\mathbf{X}^*$, and $\mathbf{U}$ are exactly equal to their plims. We will seek to generalize the "three restrictions on the parameters" (53.1.18), (53.1.19), and (53.1.23) from the simple regression with errors in dependent and independent variables to the general EV model.

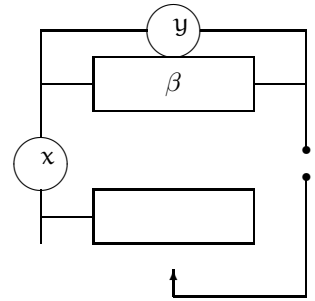


FIGURE 1. Circuit Diagram

53.4.1. The Frisch Problem. Start with the general EV model

$$\mathbf{X}^*\boldsymbol{B} = \boldsymbol{O} \tag{53.4.1}$$

$$\mathbf{X} = \mathbf{X}^* + \mathbf{U}. \tag{53.4.2}$$

Since we have infinitely many observations, equation (53.2.3), $\frac{1}{n}\mathbf{X}^{*\top}\mathbf{U} = \boldsymbol{O}$, which was asymptotically true in the general model, holds precisely. Therefore

$$\boldsymbol{Q} := \frac{1}{n}\mathbf{X}^\top\mathbf{X} \tag{53.4.3}$$

$$= \frac{1}{n}(\mathbf{X}^* + \mathbf{U})^\top(\mathbf{X}^* + \mathbf{U}) \tag{53.4.4}$$

$$= \frac{1}{n}\mathbf{X}^{*\top}\mathbf{X}^* + \frac{1}{n}\mathbf{U}^\top\mathbf{U} \tag{53.4.5}$$

$$= \boldsymbol{Q}^* + \widetilde{\boldsymbol{Q}}. \tag{53.4.6}$$

For any matrix $\boldsymbol{B}$, $\mathbf{X}^*\boldsymbol{B} = \boldsymbol{O}$ is equivalent with $\frac{1}{n}\mathbf{X}^{*\top}\mathbf{X}^*\boldsymbol{B} = \boldsymbol{O}$, i.e., $\boldsymbol{Q}^*\boldsymbol{B} = \boldsymbol{O}$.

If one assumes that all errors have nonzero variances and are uncorrelated, i.e., if $\frac{1}{n}\mathbf{U}^\top\mathbf{U}$ is in the plim diagonal and nonsingular, then the identification problem of errors in the variables can be reduced to the following "Frisch Problem":

Given a positive definite symmetric matrix $\boldsymbol{Q}$, how many ways are there to split it up as a sum

$$\boldsymbol{Q} = \boldsymbol{Q}^* + \widetilde{\boldsymbol{Q}} \tag{53.4.7}$$

where $\boldsymbol{Q}^*$ is singular and nonnegative definite, and $\widetilde{\boldsymbol{Q}}$ diagonal and positive definite?

This is a surprisingly difficult problem which has not yet been resolved in general. Here is one partial result:

Theorem ("Elementary Regression Theorem"): Assume the limit moment matrix of the observations, $\boldsymbol{Q}$, has an inverse $\boldsymbol{Q}^{-1}$ all elements of which are positive. Then

the EV problem is necessarily univariate, and $\boldsymbol{\beta}$ is a solution if and only if it can be written in the form $\boldsymbol{\beta} = \boldsymbol{Q}^{-1}\boldsymbol{\gamma}$ where $\boldsymbol{\gamma} > \boldsymbol{o}$.

Interpretation of the result: The ith column of $\boldsymbol{Q}^{-1}$ is proportional to the regression coefficients of the ith "elementary regression," in which the observations of the ith variable $\boldsymbol{x}_i$ are regressed on all the other variables. Therefore this theorem is a direct generalization of the result obtained in two dimensions, but it is only valid if all elementary regressions give positive parameters or can be made to give positive parameters by sign changes. If this is the case, the feasible parameter vectors are located in the convex set spanned by the plims of all elementary regression coefficients.

Proof of Theorem: Assume $\boldsymbol{Q}$ is positive definite, $\widetilde{\boldsymbol{Q}}$ is diagonal and positive definite, and $\boldsymbol{Q} - \widetilde{\boldsymbol{Q}}$ nonnegative definite and singular. Singularity means that there exists a vector $\boldsymbol{\beta}$, which is not the null vector, with $(\boldsymbol{Q} - \widetilde{\boldsymbol{Q}})\boldsymbol{\beta} = \boldsymbol{o}$. This can also be expressed as: $\boldsymbol{\beta}$ is eigenvector of $\boldsymbol{Q}^{-1}\widetilde{\boldsymbol{Q}}$ with 1 as eigenvalue.

First we will take any such eigenvalue and show that it can be written in the form as required. For this we will show first that every eigenvector $\boldsymbol{\alpha}$ of $\boldsymbol{Q}^{-1}\widetilde{\boldsymbol{Q}}$, which does not satisfy $(\boldsymbol{Q} - \widetilde{\boldsymbol{Q}})\boldsymbol{\alpha} = \boldsymbol{o}$ has an eigenvector smaller than 1. Call this eigenvalue λ. Then

$$\boldsymbol{Q}^{-1}\widetilde{\boldsymbol{Q}}\boldsymbol{\alpha} = \boldsymbol{\alpha}\lambda \tag{53.4.8}$$

$$\widetilde{\boldsymbol{Q}}\boldsymbol{\alpha} = \boldsymbol{Q}\boldsymbol{\alpha}\lambda \tag{53.4.9}$$

$$\boldsymbol{Q}\boldsymbol{\alpha} - \widetilde{\boldsymbol{Q}}\boldsymbol{\alpha} = \boldsymbol{Q}\boldsymbol{\alpha}(1-\lambda) \tag{53.4.10}$$

$$\underbrace{\boldsymbol{\alpha}^\top(\boldsymbol{Q} - \widetilde{\boldsymbol{Q}})\boldsymbol{\alpha}}_{>0} = \underbrace{\boldsymbol{\alpha}^\top\boldsymbol{Q}\boldsymbol{\alpha}}_{>0}(1-\lambda). \tag{53.4.11}$$

Therefore $1 - \lambda > 0$, or $\lambda < 1$.

Now we need the assumption $\boldsymbol{Q}^{-1} > \boldsymbol{O}$ and therefore also $\boldsymbol{Q}^{-1}\widetilde{\boldsymbol{Q}} > \boldsymbol{O}$. According to the Perron-Frobenius theorem, the maximal eigenvalue of a positive matrix is simple, and the eigenvector belonging to it is positive. Therefore we know that $\boldsymbol{\beta}$ is simple, i.e., the systematic variables satisfy only one linear relation, the errors in variables problem is univariate, and $\boldsymbol{\beta} > \boldsymbol{o}$. Since $\boldsymbol{\beta} = \boldsymbol{Q}^{-1}\widetilde{\boldsymbol{Q}}\boldsymbol{\beta} = \boldsymbol{Q}^{-1}\boldsymbol{\gamma}$ with $\boldsymbol{\gamma} > \boldsymbol{o}$, it can be written in the form as stated in the theorem.

In order to prove the converse, take any vector $\boldsymbol{\beta}$ that can be written as $\boldsymbol{\beta} = \boldsymbol{Q}^{-1}\boldsymbol{\gamma}$ with $\boldsymbol{\gamma} > \boldsymbol{o}$. For this $\boldsymbol{\beta}$ take that matrix $\widetilde{\boldsymbol{Q}}$ whose diagonal elements are $\tilde{q}_i^2 = \gamma_i/\beta_i$. This gives $\boldsymbol{\gamma} = \widetilde{\boldsymbol{Q}}\boldsymbol{\beta}$ and therefore $\boldsymbol{Q}\boldsymbol{\beta} = \widetilde{\boldsymbol{Q}}\boldsymbol{\beta}$. This shows that $\boldsymbol{\beta}$ is a possible solution.

This completes the proof of the theorem. But we still need to prove its interpretation. If one regresses the first variable on all others, i.e., estimates the equation

$$\mathbf{x}_1 = \begin{bmatrix}\mathbf{x}_2 & \cdots & \mathbf{x}_K\end{bmatrix}\boldsymbol{\beta} + \boldsymbol{\varepsilon} =: \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{53.4.12}$$

the solution is $\hat{\boldsymbol{\beta}} = (\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\mathbf{x}_1$. Note that the elements from which $\hat{\boldsymbol{\beta}}$ is formed are partitions of $\boldsymbol{Q}$. (Note that $\boldsymbol{Q}$ is not the dispersion matrix but the matrix of uncentered moments of $\mathbf{X}$.)

$$\boldsymbol{Q} = \frac{1}{n}\begin{bmatrix}\mathbf{x}_1^\top\mathbf{x}_1 & \mathbf{x}_1^\top\mathbf{Z} \\ \mathbf{Z}^\top\mathbf{x}_1 & \mathbf{Z}^\top\mathbf{Z}\end{bmatrix}. \tag{53.4.13}$$

Postmultiplication of $\boldsymbol{Q}$ by $\begin{bmatrix}1 \\ -\hat{\boldsymbol{\beta}}\end{bmatrix}$ gives therefore

$$\frac{1}{n}\begin{bmatrix}\mathbf{x}_1^\top\mathbf{x}_1 & \mathbf{x}_1^\top\mathbf{Z} \\ \mathbf{Z}^\top\mathbf{x}_1 & \mathbf{Z}^\top\mathbf{Z}\end{bmatrix}\begin{bmatrix}1 \\ -(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\mathbf{x}_1\end{bmatrix} = \frac{1}{n}\begin{bmatrix}\mathbf{x}_1^\top\mathbf{x}_1 - \mathbf{x}_1^\top\mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\mathbf{x}_1 \\ \boldsymbol{o}\end{bmatrix}. \tag{53.4.14}$$

In other words, $\begin{bmatrix}1 \\ -\hat{\boldsymbol{\beta}}\end{bmatrix}$ is proportional to the first column of $\boldsymbol{Q}^{-1}$.

Once the mathematical tools are better developed, it will be feasible to take the following approach to estimation: first solve the Frisch problem in order to get an estimate of the feasible parameter region compatible with the data, and then use additional information, not coming from the data, to narrow down this region to a single point. The emphasis on the Frisch problem is due to Kalman, see [**Kal82**]. Also look at [**HM89**].

53.4.2. "Sweeping Out" of Variables Measured Without Errors. The example of the simple EV model does not quite fit under this umbrella since the pseudo-variable $\boldsymbol{\iota}$ consisting of ones only is "observed" without error, while our treatment of the Frisch Problem assumed that all variables are observed with errors. This section will demonstrate that any variables in the model which are observed without error can be "swept out" by regression, thus reducing the number of variables which may pose an identification problem.

Go back to the general EV model in its symmetric form, but assume that some of the variables are observed without error, i.e., the model reads

$$\begin{bmatrix}\mathbf{X}^* & \mathbf{Z}^*\end{bmatrix}\begin{bmatrix}\boldsymbol{B} \\ \boldsymbol{\Gamma}\end{bmatrix} = \boldsymbol{O} \tag{53.4.15}$$

$$\begin{bmatrix}\mathbf{X} & \mathbf{Z}\end{bmatrix} = \begin{bmatrix}\mathbf{X}^* & \mathbf{Z}^*\end{bmatrix} + \begin{bmatrix}\mathbf{U} & \boldsymbol{O}\end{bmatrix}. \tag{53.4.16}$$

The moment matrices associated with this satisfy the following Frisch decomposition:

$$\begin{bmatrix}\boldsymbol{Q}_{XX} & \boldsymbol{Q}_{XZ} \\ \boldsymbol{Q}_{ZX} & \boldsymbol{Q}_{ZZ}\end{bmatrix} = \begin{bmatrix}\boldsymbol{Q}_{XX} & \boldsymbol{Q}_{XZ} \\ \boldsymbol{Q}_{ZX} & \boldsymbol{Q}_{ZZ}^*\end{bmatrix} + \begin{bmatrix}\boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \widetilde{\boldsymbol{Q}}_{ZZ}\end{bmatrix} \tag{53.4.17}$$

Reformulate this: Now how does one get variables whose moment matrix is the one in (53.4.20)? By regressing every variable in $\mathbf{Z}$ on all variables in $\mathbf{X}$ and taking

the residuals in this regression. Write the estimated regression equation and the residuals as

$$\mathsf{Z} = \mathsf{X}\boldsymbol{\Pi}^* + \boldsymbol{E}^* \qquad \text{where} \quad \boldsymbol{\Pi}^* = (\mathsf{X}^\top\mathsf{X})^{-1}\mathsf{X}^\top\mathsf{Z}. \tag{53.4.18}$$

Therefore $\boldsymbol{E}^* = (\boldsymbol{I} - \mathsf{X}(\mathsf{X}^\top\mathsf{X})^{-1}\mathsf{X}^\top)\mathsf{Z}$, hence

$$\operatorname{plim}\frac{1}{n}\boldsymbol{E}^{*\top}\boldsymbol{E}^* = \operatorname{plim}\frac{1}{n}(\mathsf{Z}^\top\mathsf{Z} - \mathsf{Z}^\top\mathsf{X}(\mathsf{X}^\top\mathsf{X})^{-1}\mathsf{X}^\top\mathsf{Z}) = \boldsymbol{Q}_{ZZ} - \boldsymbol{Q}_{ZX}\boldsymbol{Q}_{XX}^{-1}\boldsymbol{Q}_{XZ}. \tag{53.4.19}$$

Define $\boldsymbol{Q}_{ZZ.X} := \boldsymbol{Q}_{ZZ} - \boldsymbol{Q}_{ZX}\boldsymbol{Q}_{XX}^{-1}\boldsymbol{Q}_{XZ}$. Claim: there is a one-to-one correspondence between decompositions of the kind (53.4.17) (in which the error variances of the first partition are constrained to be zero) and unconstrained Frisch decompositions of $\boldsymbol{Q}_{ZZ.X}$. In this bijection, (53.4.17) corresponds to the decomposition

$$\boldsymbol{Q}_{ZZ.X} = (\boldsymbol{Q}_{ZZ}^* - \boldsymbol{Q}_{ZX}\boldsymbol{Q}_{XX}^{-1}\boldsymbol{Q}_{XZ}) + \widetilde{\boldsymbol{Q}}_{ZZ}. \tag{53.4.20}$$

Proof: given that $\boldsymbol{Q} = \begin{bmatrix}\boldsymbol{Q}_{XX} & \boldsymbol{Q}_{XZ}\\ \boldsymbol{Q}_{ZX} & \boldsymbol{Q}_{ZZ}\end{bmatrix}$ is nonnegative definite, we have to show that $\boldsymbol{Q}^* = \begin{bmatrix}\boldsymbol{Q}_{XX} & \boldsymbol{Q}_{XZ}\\ \boldsymbol{Q}_{ZX} & \boldsymbol{Q}_{ZZ}^*\end{bmatrix}$ is nonnegative definite if and only if $\boldsymbol{Q}_{ZZ.X} = \boldsymbol{Q}_{ZZ}^* - \boldsymbol{Q}_{ZX}\boldsymbol{Q}_{XX}^{-1}\boldsymbol{Q}_{XZ}$ is nonnegative definite. By theorem A.5.11, $\boldsymbol{Q}^*$ is nonnegative definite if and only if $\boldsymbol{Q}_{XX}$ is nnd, $\boldsymbol{Q}_{XZ} = \boldsymbol{Q}_{XX}\boldsymbol{Q}_{XX}^{-}\boldsymbol{Q}_{XZ}$, and $\boldsymbol{Q}_{ZZ.X}$ is nnd. The first two conditions follow from $\boldsymbol{Q}$ being nnd. Therefore the third condition is an iff condition.

Furthermore, the mapping

$$\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{\gamma}\end{bmatrix} \mapsto \boldsymbol{\gamma} \mapsto \begin{bmatrix}-\boldsymbol{Q}_{XX}^{-1}\boldsymbol{Q}_{XZ}\boldsymbol{\gamma}\\ \boldsymbol{\gamma}\end{bmatrix} \tag{53.4.21}$$

is a bijection between vectors that annul $\begin{bmatrix}\boldsymbol{Q}_{XX} & \boldsymbol{Q}_{XZ}\\ \boldsymbol{Q}_{ZX} & \boldsymbol{Q}_{ZZ}^*\end{bmatrix}$, which is the moment matrix of the systematic variables in (53.4.17), and vectors that annul $\boldsymbol{Q}_{ZZ}^* - \boldsymbol{Q}_{ZX}\boldsymbol{Q}_{XX}^{-1}\boldsymbol{Q}_{XZ}$, which is the moment matrix of the systematic variables in (53.4.20).

Proof: Assume I have a solution

$$\begin{bmatrix}\boldsymbol{Q}_{XX} & \boldsymbol{Q}_{XZ}\\ \boldsymbol{Q}_{ZX} & \boldsymbol{Q}_{ZZ}^*\end{bmatrix}\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{\gamma}\end{bmatrix} = \begin{bmatrix}\boldsymbol{o}\\ \boldsymbol{o}\end{bmatrix} \iff \begin{aligned}\boldsymbol{Q}_{XX}\boldsymbol{\beta} + \boldsymbol{Q}_{XZ}\boldsymbol{\gamma} &= \boldsymbol{o}\\ \boldsymbol{Q}_{ZX}\boldsymbol{\beta} + \boldsymbol{Q}_{ZZ}\boldsymbol{\gamma} &= \boldsymbol{o}\end{aligned}. \tag{53.4.22}$$

Since the nonsingularity of $\boldsymbol{Q}$ implies the nonsingularity of $\boldsymbol{Q}_{XX}$, the first equation implies $\boldsymbol{\beta} = -\boldsymbol{Q}_{XX}^{-1}\boldsymbol{Q}_{XZ}\boldsymbol{\gamma}$, and plugging this into the second equation gives $\boldsymbol{Q}_{ZZ}^* - \boldsymbol{Q}_{ZX}\boldsymbol{Q}_{XX}^{-1}\boldsymbol{Q}_{XZ}\boldsymbol{\gamma} = \boldsymbol{o}$. On the other hand, starting with a $\boldsymbol{\beta}$ annulling the systematic moment matrix of the compacted problem $(\boldsymbol{Q}_{ZZ}^* - \boldsymbol{Q}_{ZX}\boldsymbol{Q}_{XX}^{-1}\boldsymbol{Q}_{XZ})\boldsymbol{\beta} = \boldsymbol{o}$,

this implies

$$\begin{bmatrix}\boldsymbol{Q}_{XX} & \boldsymbol{Q}_{XZ}\\ \boldsymbol{Q}_{ZX} & \boldsymbol{Q}_{ZZ}^*\end{bmatrix}\begin{bmatrix}-\boldsymbol{Q}_{XX}^{-1}\boldsymbol{Q}_{XZ}\boldsymbol{\beta}\\ \boldsymbol{\beta}\end{bmatrix} = \begin{bmatrix}\boldsymbol{o}\\ \boldsymbol{o}\end{bmatrix} \tag{53.4.23}$$

If $\mathsf{X} = \boldsymbol{\iota}$, then going over to the residuals simply means that one has to take deviations from the means. In this case, (53.4.20) is the decomposition of the covariance matrix of the variables. In other words, if there is a constant term in the regressions, then $\boldsymbol{Q}$ and $\widetilde{\boldsymbol{Q}}$ should not be considered to be the moments of the observed and systematic variables about the origin, but their covariance matrices. We will use this rule extensively in the following examples.

53.5. Properties of Ordinary Least Squares in the EV model

In the estimation of a univariate EV model it is customary to single out one variable which is known to have a nonzero coefficient in the linear relation to be estimated, and to normalize its coefficient to be -1. Writing this variable as the first variable, the symmetric form reads

$$\begin{aligned}\begin{bmatrix}\mathsf{y}^* & \mathsf{X}^*\end{bmatrix}\begin{bmatrix}-1\\ \boldsymbol{\beta}\end{bmatrix} &= \boldsymbol{o}\\ \begin{bmatrix}\mathsf{y} & \mathsf{X}\end{bmatrix} &= \begin{bmatrix}\mathsf{y}^* & \mathsf{X}^*\end{bmatrix} + \begin{bmatrix}\boldsymbol{\nu} & \mathsf{U}\end{bmatrix};\end{aligned} \tag{53.5.1}$$

but the usual way of writing this is, of course,

$$\begin{aligned}\mathsf{y}^* &= \mathsf{X}^*\boldsymbol{\beta}\\ \mathsf{y} &= \mathsf{y}^* + \boldsymbol{\nu}\\ \mathsf{X} &= \mathsf{X}^* + \mathsf{U}\end{aligned} \tag{53.5.2}$$

In this situation it is tempting to write

$$\mathsf{y} = \mathsf{X}\boldsymbol{\beta} + \underbrace{\boldsymbol{\nu} - \mathsf{U}\boldsymbol{\beta}}_{\boldsymbol{\varepsilon}} \tag{53.5.3}$$

and to regress y on X. This gives a biased and inconsistent estimator of $\boldsymbol{\beta}$, since $\boldsymbol{\varepsilon}$ is correlated with X. It is still worthwhile to look at this regression, since the coefficient towards which b_{OLS} converges is an estimate of the "empirical relation" among the variables, i.e., an estimate of the conditional mean of y given x_i.

$$\mathsf{b}_{OLS} - \boldsymbol{\beta} = (\frac{1}{n}\mathsf{X}^\top\mathsf{X})^{-1}\frac{1}{n}\mathsf{X}^\top\boldsymbol{E} \tag{53.5.4}$$

$$= -(\frac{1}{n}\mathsf{X}^\top\mathsf{X})^{-1}\frac{1}{n}\mathsf{X}^\top\mathsf{U}\boldsymbol{\beta} + (\frac{1}{n}\mathsf{X}^\top\mathsf{X})^{-1}\frac{1}{n}\mathsf{X}^\top\boldsymbol{\nu}. \tag{53.5.5}$$

Since $\mathbf{X}^\top\mathbf{U} = \mathbf{X}^{*\top}\mathbf{U} + \mathbf{U}^\top\mathbf{U}$ and $\operatorname{plim}\frac{1}{n}\mathbf{X}^{*\top}\mathbf{U} = \boldsymbol{O}$, this becomes in the plim

$$\operatorname{plim}\mathbf{b}_{OLS} - \boldsymbol{\beta} = -\boldsymbol{Q}^{-1}\widetilde{\boldsymbol{Q}}\boldsymbol{\beta} + \boldsymbol{Q}^{-1}\boldsymbol{\sigma}_{\mathbf{U}\boldsymbol{\nu}}. \tag{53.5.6}$$

This, under the additional assumption that $\boldsymbol{\nu}$ and $\mathbf{U}$ are in the plim uncorrelated, i.e., that $\boldsymbol{\sigma}_{\mathbf{U}\boldsymbol{\nu}} = \boldsymbol{o}$, is [**Gre97**, (9.28) on p. 439]. Greene says "this is a mixture of all the parameters in the model," implying that it is hopeless to get information about these parameters out of this. However, if one looks for inequalities instead of equalities, some information is available.

For instance one can show that the sample variance of the residuals remains between the variance of $\boldsymbol{\varepsilon}$ and the variance of $\boldsymbol{\nu}$, i.e.,

$$\sigma^2_{\boldsymbol{\nu}} \leq \operatorname{plim}\frac{1}{n}\boldsymbol{e}^\top\boldsymbol{e} \leq \sigma^2_{\boldsymbol{\varepsilon}}. \tag{53.5.7}$$

For this start with

$$\boldsymbol{e} = (\boldsymbol{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)\boldsymbol{\varepsilon} \tag{53.5.8}$$

therefore

$$\operatorname{plim}\frac{1}{n}\boldsymbol{e}^\top\boldsymbol{e} = \sigma^2_{\boldsymbol{\varepsilon}} - \operatorname{plim}\frac{1}{n}\boldsymbol{\varepsilon}^\top\mathbf{X}(\frac{1}{n}\mathbf{X}^\top\mathbf{X})^{-1}\frac{1}{n}\mathbf{X}^\top\boldsymbol{\varepsilon}. \tag{53.5.9}$$

Since $\mathbf{X}^\top\mathbf{X}^{-1}$ is nonnegative definite, this shows the second half of (53.5.7). For the first half use

$$\boldsymbol{\varepsilon} = -\mathbf{U}\boldsymbol{\beta} + \boldsymbol{\nu}, \quad \text{hence} \tag{53.5.10}$$

$$\boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon} = \boldsymbol{\beta}^\top\mathbf{U}^\top\mathbf{U}\boldsymbol{\beta} - 2\boldsymbol{\beta}^\top\mathbf{U}^\top\boldsymbol{\nu} + \boldsymbol{\nu}^\top\boldsymbol{\nu} \quad \text{and} \tag{53.5.11}$$

$$\sigma^2_{\boldsymbol{\varepsilon}} = \boldsymbol{\beta}^\top\widetilde{\boldsymbol{Q}}\boldsymbol{\beta} + \sigma^2_{\boldsymbol{\nu}} \tag{53.5.12}$$

and

$$\operatorname{plim}\frac{1}{n}\mathbf{X}^\top\boldsymbol{\varepsilon} = \operatorname{plim}\frac{1}{n}(\mathbf{X}^* + \mathbf{U})^\top(-\mathbf{U}\boldsymbol{\beta} + \boldsymbol{\nu}) = -\widetilde{\boldsymbol{Q}}\boldsymbol{\beta}. \tag{53.5.13}$$

Plugging (53.5.12) and (53.5.13) into (53.5.9) gives

$$\operatorname{plim}\frac{1}{n}\boldsymbol{e}^\top\boldsymbol{e} = \sigma^2_{\boldsymbol{\nu}} + \boldsymbol{\beta}^\top(\widetilde{\boldsymbol{Q}} - \widetilde{\boldsymbol{Q}}\boldsymbol{Q}^{-1}\widetilde{\boldsymbol{Q}})\boldsymbol{\beta}. \tag{53.5.14}$$

Since $\widetilde{\boldsymbol{Q}} - \widetilde{\boldsymbol{Q}}\boldsymbol{Q}^{-1}\widetilde{\boldsymbol{Q}}$ is nonnegative definite, this proves the first half of the inequality.

Problem 474. *Assuming $\boldsymbol{\sigma}_{\mathbf{U}\boldsymbol{\nu}} = \boldsymbol{o}$, show that in the* plim,

$$\mathbf{b}^\top_{OLS}\boldsymbol{Q}\mathbf{b}_{OLS} \leq \boldsymbol{\beta}^\top\boldsymbol{Q}\boldsymbol{\beta}. \tag{53.5.15}$$

You will need Problem 583 for this.

Answer.

$$\boldsymbol{\beta}^\top\boldsymbol{Q}\boldsymbol{\beta} - \operatorname{plim}\mathbf{b}^\top_{OLS}\boldsymbol{Q}\mathbf{b}_{OLS} \tag{53.5.16}$$

$$= \boldsymbol{\beta}^\top\big(\boldsymbol{Q} - (\boldsymbol{I} - \widetilde{\boldsymbol{Q}}\boldsymbol{Q}^{-1})\boldsymbol{Q}\boldsymbol{Q}^{-1}\boldsymbol{Q}(\boldsymbol{I} - \boldsymbol{Q}^{-1}\widetilde{\boldsymbol{Q}})\big)\boldsymbol{\beta} \tag{53.5.17}$$

$$= \boldsymbol{\beta}^\top\big(\boldsymbol{Q} - (\boldsymbol{Q} - \widetilde{\boldsymbol{Q}})\boldsymbol{Q}^{-1}(\boldsymbol{Q} - \widetilde{\boldsymbol{Q}})\big)\boldsymbol{\beta} \tag{53.5.18}$$

$$= \boldsymbol{\beta}^\top\big(\widetilde{\boldsymbol{Q}} + \boldsymbol{Q}^* - \boldsymbol{Q}^*\boldsymbol{Q}^{-1}\boldsymbol{Q}^*\big)\boldsymbol{\beta}. \tag{53.5.19}$$

By Problem 583, $\boldsymbol{Q}^* - \boldsymbol{Q}^*\boldsymbol{Q}^{-1}\boldsymbol{Q}^*$ is nonnegative definite. □

Problem 475. *Assume the data $\mathbf{X}$ and $\mathbf{y}$ can be modeled as a univariate EV model:*

$$\mathbf{y}^* = \mathbf{X}^*\boldsymbol{\beta} \tag{53.5.20}$$

$$\mathbf{X} = \mathbf{X}^* + \mathbf{U} \tag{53.5.21}$$

$$\mathbf{y} = \mathbf{y}^* + \boldsymbol{\nu} \tag{53.5.22}$$

Let $\mathbf{x}_i{}^\top$, $\mathbf{x}_i^{*\top}$, and $\mathbf{u}_i{}^\top$ be the ith rows of $\mathbf{X}$, $\mathbf{X}^*$ and $\mathbf{U}$. Assume they are distributed $\mathbf{x}_i^* \sim \mathrm{NID}(\boldsymbol{o}, \boldsymbol{Q}^*)$, $\mathbf{u}_i \sim \mathrm{NID}(\boldsymbol{o}, \widetilde{\boldsymbol{Q}})$, and $\nu_i \sim \mathrm{NID}(0, \sigma^2)$, and all three are independent of each other. Define $\boldsymbol{Q} = \boldsymbol{Q}^* + \widetilde{\boldsymbol{Q}}$. Therefore

$$\begin{bmatrix}\mathbf{x}_i \\ \mathbf{y}_i\end{bmatrix} \sim N\Big(\begin{bmatrix}\boldsymbol{o} \\ 0\end{bmatrix}, \begin{bmatrix}\boldsymbol{Q} & \boldsymbol{Q}^*\boldsymbol{\beta} \\ \boldsymbol{\beta}^\top\boldsymbol{Q}^* & \sigma^2 + \boldsymbol{\beta}^\top\boldsymbol{Q}^*\boldsymbol{\beta}\end{bmatrix}\Big) \tag{53.5.23}$$

Compute $\mathrm{E}[\mathbf{y}_i|\mathbf{x}_i]$ and $\operatorname{var}[\mathbf{y}_i|\mathbf{x}_i]$. (Since $\mathbf{y}_i$ is a linear function of $\mathbf{x}_i$, you can use the formulas for best linear predictors here.)

Answer.

$$\mathrm{E}[\mathbf{y}_i|\mathbf{x}_i] = \boldsymbol{\beta}^\top\boldsymbol{Q}^*\boldsymbol{Q}^{-1}\mathbf{x}_i = \boldsymbol{\beta}^\top(\boldsymbol{Q} - \widetilde{\boldsymbol{Q}})\boldsymbol{Q}^{-1}\mathbf{x}_i = \boldsymbol{\beta}^\top\mathbf{x}_i - \boldsymbol{\beta}^\top\widetilde{\boldsymbol{Q}}\boldsymbol{Q}^{-1}\mathbf{x}_i \tag{53.5.24}$$

$$\operatorname{var}[\mathbf{y}_i|\mathbf{x}_i] = \sigma^2 + \boldsymbol{\beta}^\top\boldsymbol{Q}^*\boldsymbol{\beta} - \boldsymbol{\beta}^\top\widetilde{\boldsymbol{Q}}\boldsymbol{Q}^{-1}\widetilde{\boldsymbol{Q}}\boldsymbol{\beta} \tag{53.5.25}$$

□

• a. *It is possible to build an alternative model*

$$\mathbf{y}_i = \mathbf{x}_i{}^\top\boldsymbol{\gamma} + \nu_i \qquad \text{or} \quad \mathbf{y} = \mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\nu}; \qquad \mathbf{V} \sim (\boldsymbol{o}, \boldsymbol{\Xi}) \tag{53.5.26}$$

with $\mathbf{X}$ and $\boldsymbol{\nu}$ independent *(and again for simplicity all variables having zero mean), which gives the same joint distribution of $\mathbf{X}$ and $\mathbf{y}$ as the above specification. Compute $\boldsymbol{\gamma}$, $\mathcal{V}[\boldsymbol{x}_i]$, and $\operatorname{var}[\mathbf{y}_i]$ in terms of the structural data of the above specification.*

Answer.

$$\mathbf{y}^* = \mathbf{X}^*\big(\boldsymbol{\beta} - \boldsymbol{Q}^{-1}\widetilde{\boldsymbol{Q}}\boldsymbol{\beta}\big) \tag{53.5.27}$$

$$\mathbf{X} = \mathbf{X}^* \tag{53.5.28}$$

$$\mathbf{y} = \mathbf{y}^* + \boldsymbol{\nu} \tag{53.5.29}$$

where $\mathbf{x}^*{}_i = \mathbf{x}_i \sim \mathrm{N}(\boldsymbol{o}, \boldsymbol{Q})$ and $\operatorname{var}[\nu_i] = \sigma^2 + \boldsymbol{\beta}^\top\boldsymbol{Q}^*\boldsymbol{\beta} - \boldsymbol{\beta}^\top\widetilde{\boldsymbol{Q}}\boldsymbol{Q}^{-1}\widetilde{\boldsymbol{Q}}\boldsymbol{\beta}$. As one sees, $\boldsymbol{\gamma} = \boldsymbol{\beta} - \boldsymbol{Q}^{-1}\widetilde{\boldsymbol{Q}}\boldsymbol{\beta}$. In this latter model, OLS is appropriate, and these are therefore the plims of the OLS estimates.

Note that $\mathcal{C}[\mathbf{x}_i, \mathrm{y}_i] = \mathcal{C}[\mathbf{x}_i, \mathbf{x}_i{}^\top\boldsymbol{\gamma} + \nu_i] = \mathcal{C}[\mathbf{x}_i, \boldsymbol{\gamma}^\top\mathbf{x}_i] = \mathcal{V}[\mathbf{x}_i]\boldsymbol{\gamma} = \boldsymbol{Q}(\boldsymbol{\beta} - \boldsymbol{Q}^{-1}\widetilde{\boldsymbol{Q}}\boldsymbol{\beta}) = \boldsymbol{Q}^*\boldsymbol{\beta}$, and $\operatorname{var}[\mathrm{y}^*] = \operatorname{var}[\boldsymbol{\gamma}^\top\mathbf{x}^*] = \boldsymbol{\gamma}^\top\boldsymbol{Q}\boldsymbol{\gamma} = (\boldsymbol{\beta}^\top - \boldsymbol{\beta}^\top\widetilde{\boldsymbol{Q}}\boldsymbol{Q}^{-1})\boldsymbol{Q}(\boldsymbol{\beta} - \boldsymbol{Q}^{-1}\widetilde{\boldsymbol{Q}}\boldsymbol{\beta}) = b^\top\boldsymbol{Q}b - 2b^\top\widetilde{\boldsymbol{Q}}b + b^\top\widetilde{\boldsymbol{Q}}\boldsymbol{Q}^{-1}\widetilde{\boldsymbol{Q}}b$. Adding $\operatorname{var}[\nu_i] = \sigma^2 + \boldsymbol{\beta}^\top\boldsymbol{Q}^*\boldsymbol{\beta} - \boldsymbol{\beta}^\top\widetilde{\boldsymbol{Q}}\boldsymbol{Q}^{-1}\widetilde{\boldsymbol{Q}}\boldsymbol{\beta}$ to this gives, if everything is right, $\sigma^2 + \boldsymbol{\beta}^\top\boldsymbol{Q}^*\boldsymbol{\beta}$. About the form of the alternative coefficient vector compare their theorem 4.1, and about the residual variance compare their theorem 4.3. □

53.6. Kalman's Critique of Malinvaud

PROBLEM 476. *In Malinvaud's econometrics textbook* [**Mal78**] *and* [**Mal70**, pp. 17–31 and 211–221], *the following data about the French economy are used (all amounts in billion nouveaux francs, at 1959 prices):*

	imports	*gdp*	*invchge*	*hhconsum*
1949	15.9	149.3	4.2	108.1
1950	16.4	161.2	4.1	114.8
1951	19.0	171.5	3.1	123.2
1952	19.1	175.5	3.1	126.9
1953	18.8	180.8	1.1	132.1
1954	20.4	190.7	2.2	137.7
1955	22.7	202.1	2.1	146.0
1956	26.5	212.4	5.6	154.1
1957	28.1	226.1	5.0	162.3
1958	27.6	231.9	5.1	164.3
1959	26.3	239.0	0.7	167.6
1960	31.1	258.0	5.6	176.8
1961	33.3	269.8	3.9	186.6
1962	37.0	288.4	3.1	199.7
1963	43.3	304.5	4.6	213.9
1964	49.0	323.4	7.0	223.8
1965	50.3	336.8	1.2	232.0
1966	56.6	353.9	4.5	242.9

This dataset is also discussed in [**Mad88**, pp. 239, 382] *and* [**CP77**, pp. 152, 164], *but the following exercise follows* [**Kal84**]. *If you have the* R*-library* `ecmet` *installed, then the data can be made available by the command* `data(malvaud)`. *You can also download them from www.econ.utah.edu/ehrbar/data/malvaud.txt.*

• a. *Run the three elementary regressions for the whole period, then choose at least two subperiods and run them for those. Plot all regression coefficients as points in a plane, using different colors for the different subperiods (you have to normalize them in a special way that they all fit on the same plot).*

ANSWER. Assume you have downloaded the data and put them into the SAS dataset `malvaud`. The command for one of the regressions over the whole period is


```
proc reg data=malvaud;
    model imports=hhconsum;
run;
```

For regression over subperiods you must first form a dataset which only contains the subperiod:

```
data fifties;
    set ec781.malvaud;
    if 1950<=year<=1959;
run;
proc reg data=fifties;
    model imports=hhconsum;
run;
```

You can run several regressions at once by including several `model` statements with different models. □

• b. *The elementary regressions give you three fitted equations of the form*

$$\textit{imports} = \hat\alpha_1 + \hat\beta_{12}\,\textit{gdp} + \hat\beta_{13}\,\textit{hhconsum} + \textit{residual}_1 \tag{53.6.1}$$

$$\textit{gdp} = \hat\alpha_2 + \hat\beta_{21}\,\textit{imports} + \hat\beta_{23}\,\textit{hhconsum} + \textit{residual}_2 \tag{53.6.2}$$

$$\textit{hhconsum} = \hat\alpha_3 + \hat\beta_{31}\,\textit{imports} + \hat\beta_{32}\,\textit{gdp} + \textit{residual}_3. \tag{53.6.3}$$

In order to compare the slope parameters of the second regression to the ones obtained in the first, solve (53.6.2) *for* `imports`,

$$\textit{imports} = -\frac{\hat\alpha_2}{\hat\beta_{21}} + \frac{1}{\hat\beta_{21}}\,\textit{gdp} - \frac{\hat\beta_{23}}{\hat\beta_{21}}\,\textit{hhconsum} - \frac{\textit{residual}_2}{\hat\beta_{21}} \tag{53.6.4}$$

and compare $\hat\beta_{12}$ *with* $1/\hat\beta_{21}$ *and* $\hat\beta_{13}$ *with* $-\hat\beta_{23}/\hat\beta_{21}$. *In the same way compare the results of the third regression with the ones of the first. This comparison is conveniently done in table 1. Fill in the values for the whole period and also for several*

	Slope of `imports` *with respect to* `gdp`	*Slope of* `imports` *with respect to* `hhconsum`
Regression of `imports` *on* `gdp` *and* `hhconsum`		
Regression of `gdp` *on* `imports` *and* `hhconsum`		
Regression of `hhconsum` *on* `imports` *and* `gdp`		

TABLE 1. Comparison of Coefficients in Elementary Regressions

sample subperiods. Make a scatter plot of the contents of this table, i.e., represent

each regression result as a point in a plane, using different colors for different sample periods.

• c. *You will probably find that these points form a very narrow but often quite long triangle. The triangles for different subperiods lie on the same stable line. This indicates that the data should be modeled as observations with errors of systematic data which satisfy two linear relationships at once. Using the plots of the different regression coefficients, compute approximately the coefficients of these two linear relationships.*

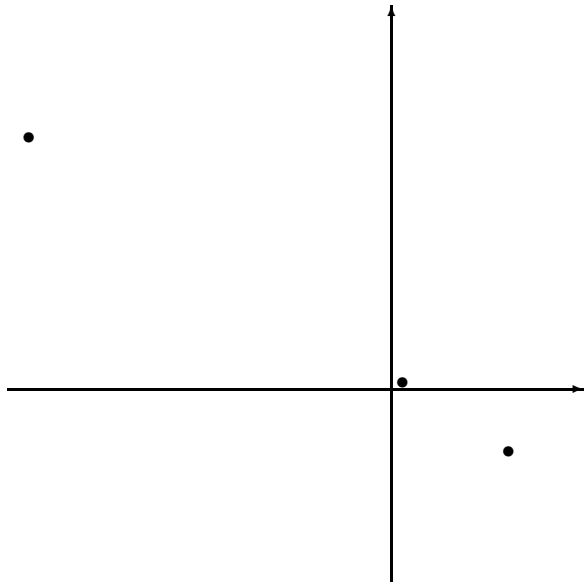


FIGURE 2. Coefficients of `hhconsum` (horizontal) and `gdp` (vertical), dependent variable `imports`. 1949–1966

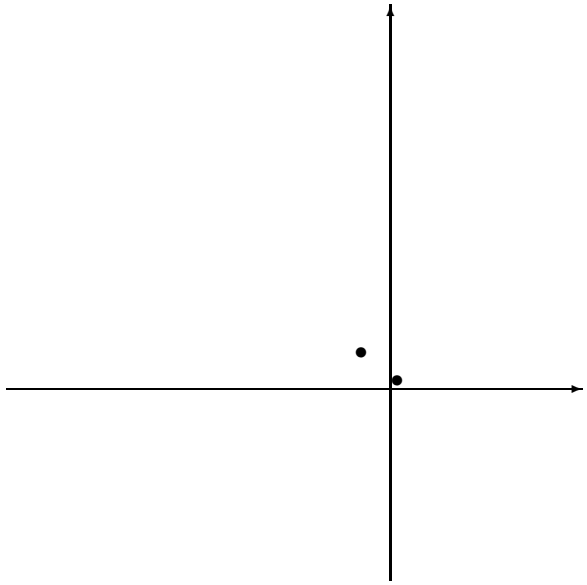


FIGURE 3. Coefficients of `hhconsum` (horizontal) and `gdp` (vertical), dependent variable `imports`. 1949–54, third point out of bounds

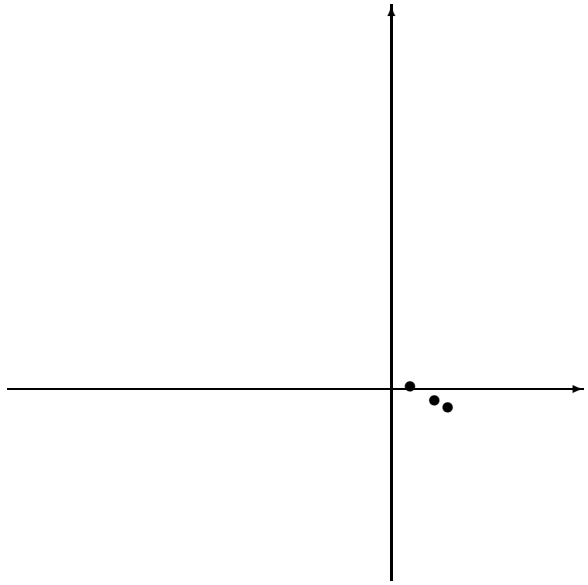


FIGURE 4. Coefficients of `hhconsum` (horizontal) and `gdp` (vertical), dependent variable `imports`. 1955–60

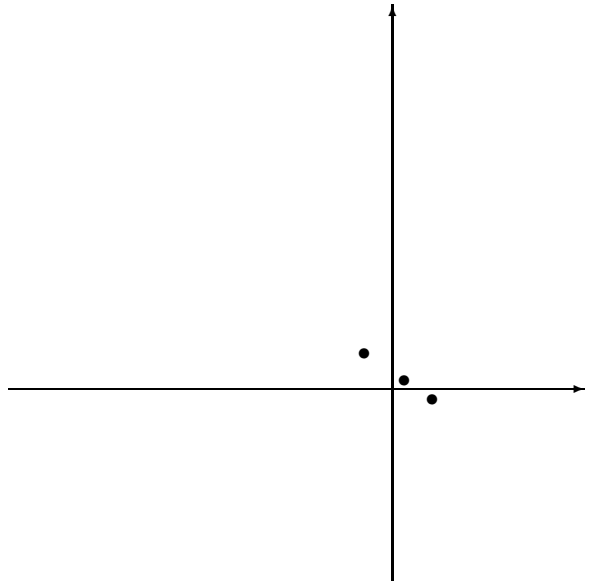


FIGURE 5. Coefficients of `hhconsum` (horizontal) and `gdp` (vertical), dependent variable `imports`. 1961–66

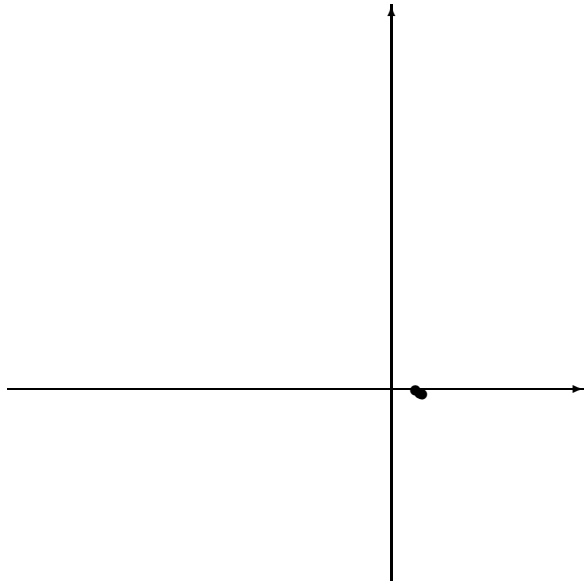


FIGURE 6. Coefficients of `hhconsum` (horizontal) and `gdp` (vertical), dependent variable `imports`. 1953–57

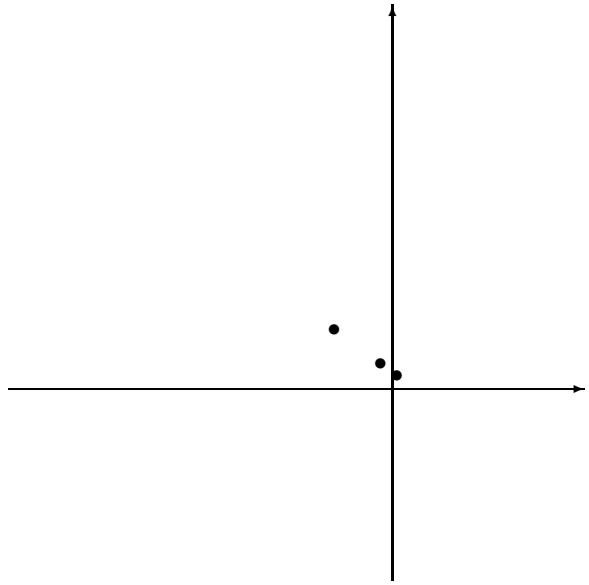


FIGURE 7. Coefficients of `hhconsum` (horizontal) and `gdp` (vertical), dependent variable `imports`. 1960–64

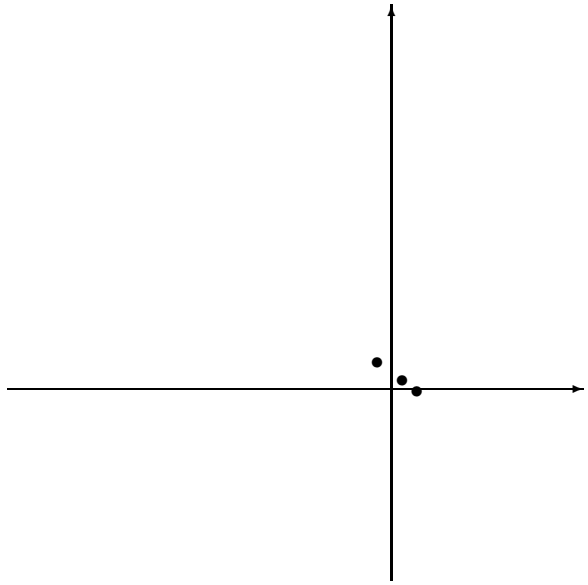


FIGURE 8. Coefficients of `hhconsum` (horizontal) and `gdp` (vertical), dependent variable `imports`. 1959–63

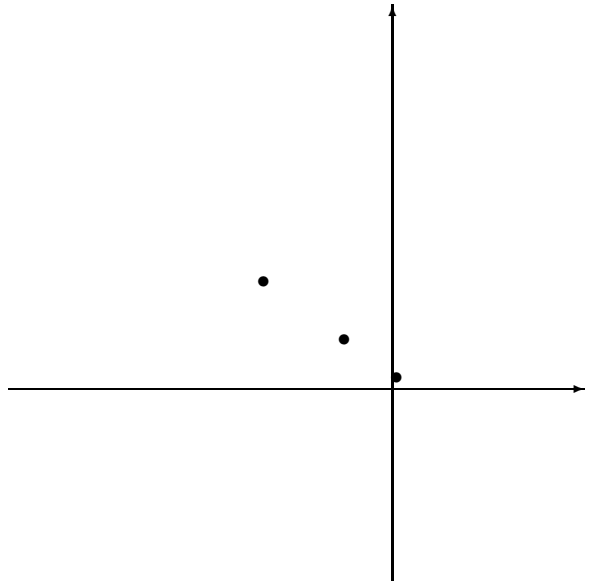


FIGURE 9. Coefficients of `hhconsum` (horizontal) and `gdp` (vertical), dependent variable `imports`. 1949–57

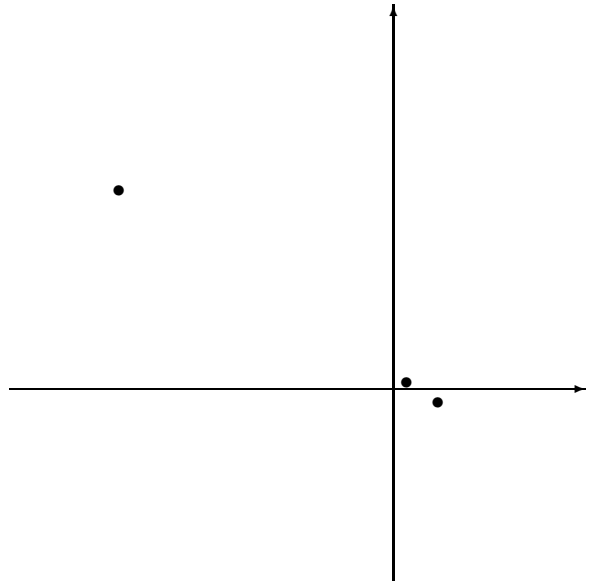


FIGURE 10. Coefficients of `hhconsum` (horizontal) and `gdp` (vertical), dependent variable `imports`. 1958–66

53.7. Estimation if the EV Model is Identified

If there is a small number of parameters in relation to the number of relations to be estimated, the EV model may be identified after all. One of the simplest cases is Friedman's model in which x^* denotes permanent income, y^* permanent consumption, u and v transitory income and consumption, and x and y the observed actual income and consumption. Friedman's hypothesis is

$$y^* = x^*\beta \tag{53.7.1}$$

$$y = y^* + v \tag{53.7.2}$$

$$x = x^* + u \tag{53.7.3}$$

This has one parameter less than the previous simple regression model, since $\alpha = 0$. Therefore the parameters are determined by (53.1.16) and (53.1.17).

$$\beta = \frac{\mu_y}{\mu_x} \tag{53.7.4}$$

$$\sigma^2_{x^*} = \frac{\sigma_{xy}}{\beta} = \frac{\sigma_{xy}\mu_x}{\mu_y} \tag{53.7.5}$$

$$\sigma^2_u = \sigma^2_x - \sigma^2_{x^*} = \sigma^2_x - \frac{\sigma_{xy}\mu_x}{\mu_y} \tag{53.7.6}$$

$$\sigma^2_v = \sigma^2_y - \beta^2\sigma^2_{x^*} = \sigma^2_y - \frac{\sigma_{xy}\mu_y}{\mu_x}. \tag{53.7.7}$$

Replacing these by sample moments gives consistent estimates.

Here is an example of a bivariate *EV* model that is identified. Assume one has three different measurement instruments all of which measure the same quantity x^*. Then the readings of these instruments, which we will denote x, y, and z, are usually modeled to be noisy linear transformations of the true value x^*:

$$x = x^* + u \tag{53.7.8}$$

$$y = \alpha + \beta x^* + v \tag{53.7.9}$$

$$z = \gamma + \delta x^* + w. \tag{53.7.10}$$

The measurement errors u, v, and w are assumed independent of each other. The first instrument is called the "standard instrument" since the origin and scale of the true variable x^* are assumed to be identical to the origin and scale of this instrument; the other two instruments have different origins and scales.

Here are the formulas for the method of moments estimates:

$$\hat{\beta} = \frac{s_{yz}}{s_{xz}} \tag{53.7.11}$$

$$\hat{\delta} = \frac{s_{yz}}{s_{xy}} \tag{53.7.12}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \tag{53.7.13}$$

$$\hat{\gamma} = \bar{z} - \hat{\gamma}x \tag{53.7.14}$$

$$\hat{\sigma}^2_u = s^2_x - \frac{s_{xy}s_{xz}}{s_{yz}} \tag{53.7.15}$$

$$\hat{\sigma}^2_v = s^2_y - \frac{s_{xy}s_{yz}}{s_{xz}} \tag{53.7.16}$$

$$\hat{\sigma}^2_w = s^2_x - \frac{s_{xz}s_{yz}}{s_{xy}}. \tag{53.7.17}$$

Here is another example related with the permanent income hypothesis. If one has several categories of consumption C_j, such as food, housing, education and entertainment, etc., then the permanent income hypothesis says

$$Y = Y^p + Y^t \tag{53.7.18}$$

$$C_j = \alpha_j + \beta_j Y^p + C^t_j \tag{53.7.19}$$

If all the C^t_j are independent of each other, then this system is identified.

PROBLEM 477. *Given a bivariate problem with three variables all of which have zero mean. (This is the model apparently appropriate to the* **malvaud** *data after taking out the means.) Call the observed variables* x, y, *and* z, *with underlying systematic variables* x^*, y^*, *and* z^*, *and error variables* u, v, *and* w. *Write this model in the form* (53.3.1).

ANSWER.

$$\begin{bmatrix} x^* & y^* & z^* \end{bmatrix} \begin{bmatrix} -1 & 0 \\ 0 & -1 \\ \beta & \gamma \end{bmatrix} = \boldsymbol{O} \tag{53.7.20}$$

$$\begin{bmatrix} x & y & z \end{bmatrix} = \begin{bmatrix} x^* & y^* & z^* \end{bmatrix} + \begin{bmatrix} u & v & w \end{bmatrix}$$

or

$$\begin{aligned} x^* &= \beta z^* \\ y^* &= \gamma z^* \\ x &= x^* + u \\ y &= y^* + v \\ z &= z^* + w. \end{aligned}$$

□

• a. *The moment matrix of the systematic variables can be written fully in terms of* $\sigma^2_{z^*}$ *and the unknown parameters. Write out the moment matrix and therefore the Frisch decomposition.*

ANSWER.

$$\begin{bmatrix} \sigma^2_x & \sigma_{xy} & \sigma_{xz} \\ \sigma_{xy} & \sigma^2_y & \sigma_{yz} \\ \sigma_{xz} & \sigma_{yz} & \sigma^2_z \end{bmatrix} = \sigma^2_{z^*} \begin{bmatrix} \beta^2 & \beta\gamma & \beta \\ \beta\gamma & \gamma^2 & \gamma \\ \beta & \gamma & 1 \end{bmatrix} + \begin{bmatrix} \sigma^2_u & 0 & 0 \\ 0 & \sigma^2_v & 0 \\ 0 & 0 & \sigma^2_w \end{bmatrix}. \tag{53.7.21}$$

□

• b. *Show that the unknown parameters are identified, and derive estimates of all parameters of the model.*

ANSWER. Solving the Frisch equations one gets

$$\begin{aligned} \beta &= \frac{\sigma_{xy}}{\sigma_{yz}} & \sigma^2_u &= \sigma^2_x - \frac{\sigma_{xz}\sigma_{xy}}{\sigma_{yz}} \\ \gamma &= \frac{\sigma_{xy}}{\sigma_{xz}} & \sigma^2_v &= \sigma^2_y - \frac{\sigma_{xy}\sigma_{yz}}{\sigma_{xz}} \\ \sigma^2_{z^*} &= \frac{\sigma_{xz}\sigma_{yz}}{\sigma_{xy}} & \sigma^2_w &= \sigma^2_z - \frac{\sigma_{yz}\sigma_{xz}}{\sigma_{xy}} \end{aligned} \tag{53.7.22}$$

If you replace the true moments by the sample moments, you see that β and γ are estimated by instrumental variables. □

• c. *Compare these estimates with OLS estimates. Derive equations for the bias of OLS.*

ANSWER.

$$\operatorname{plim} \hat{\beta}_{OLS} - \beta = \frac{\sigma_{xz}}{\sigma_{zz}} - \frac{\sigma_{xy}}{\sigma_{yz}} = \frac{\begin{vmatrix} \sigma_{zz} & \sigma_{xz} \\ \sigma_{yz} & \sigma_{xy} \end{vmatrix}}{\sigma_{zz}\sigma_{yz}}. \tag{53.7.23}$$

□

• d. *As an application, estimate the* **malvaud** *data according to this method, and compare your results with Kalman's results*

$$\textit{imports} = (0.3)\textit{hhconsum} \tag{53.7.24}$$

$$\textit{gdp} = (1.5)\textit{hhconsum}. \tag{53.7.25}$$

Here are the SAS commands to get the coefficient of **imports** *on* **hhconsum**:

```
proc syslin data=malvaud 2sls;
    model imports=hhconsum;
    endogenous imports hhconsum;
    instruments gdp;
run;
```

• e. *3 points Now run regressions with only one explanatory variable. Are the results close to the relations which you would expect from the result of the previous step?*

53.8. P-Estimation

A type of prior information that can be handled well mathematically is that $\widetilde{\boldsymbol{Q}}$ is known except for a constant multiple, i.e., one knows a $\boldsymbol{\Lambda}$ so that there is a $\kappa \neq 0$ with $\widetilde{\boldsymbol{Q}} = \kappa\boldsymbol{\Lambda}$.

In the simple EV model in which the errors of the $\mathbf{x}$ and $\mathbf{y}$ variables are independent this means that one knows a κ with

$$\sigma_{\mathsf{v}}^2 = \kappa\sigma_{\mathsf{u}}^2. \tag{53.8.1}$$

This equation, together with the Frisch equations (after elimination of the constant term α)

$$\sigma_{\mathsf{x}}^2 = \sigma_{\mathsf{x}^*}^2 + \sigma_{\mathsf{u}}^2 \tag{53.8.2}$$

$$\sigma_{\mathsf{y}}^2 = \beta^2\sigma_{\mathsf{x}^*}^2 + \sigma_{\mathsf{v}}^2 \tag{53.8.3}$$

$$\sigma_{\mathbf{xy}} = \beta\sigma_{\mathsf{x}^*}^2 \tag{53.8.4}$$

allows identification of all parameters as follows: In (53.8.2) and (53.8.3), replace $\sigma_{\mathsf{x}^*}^2$ by $\sigma_{\mathbf{xy}}/\beta$ and σ_{v}^2 by $\kappa\sigma_{\mathsf{u}}^2$, and put σ_{u}^2 on the lefthand side:

$$\sigma_{\mathsf{u}}^2 = \sigma_{\mathsf{x}}^2 - \frac{\sigma_{\mathbf{xy}}}{\beta} \tag{53.8.5}$$

$$\sigma_{\mathsf{u}}^2 = \frac{1}{\kappa}\big(\sigma_{\mathsf{y}}^2 - \beta\sigma_{\mathbf{xy}}\big). \tag{53.8.6}$$

Setting those equal and multiplying by $\beta\kappa$ gives the quadratic equation

$$\beta^2\sigma_{\mathbf{xy}} + \beta(\kappa\sigma_{\mathsf{x}}^2 - \sigma_{\mathsf{y}}^2) - \kappa\sigma_{\mathbf{xy}} = 0, \tag{53.8.7}$$

which has the solutions

$$\beta_{1|2} = \frac{\sigma_{\mathsf{y}}^2 - \kappa\sigma_{\mathsf{x}}^2}{2\sigma_{\mathbf{xy}}} \pm \sqrt{\kappa + \Big(\frac{\sigma_{\mathsf{y}}^2 - \kappa\sigma_{\mathsf{x}}^2}{2\sigma_{\mathbf{xy}}}\Big)^2}. \tag{53.8.8}$$

Since β must have the same sign as $\sigma_{\mathbf{xy}}$, only one of these solutions is valid, which can be written as

$$\beta = \frac{1}{2\sigma_{\mathbf{xy}}}\Big[\sigma_{\mathsf{y}}^2 - \kappa\sigma_{\mathsf{x}}^2 + \sqrt{4\kappa\sigma_{\mathbf{xy}}^2 + (\sigma_{\mathsf{y}}^2 - \kappa\sigma_{\mathsf{x}}^2)^2}\Big]. \tag{53.8.9}$$

By replacing the true moments of the observed variables by the sample moments one obtains an estimate which we will denote with $\hat{\beta}_P$. The Frisch equations will then also yield estimates of the other parameters.

Problem 478. [**SM86**, A 3.3/11] *Show that* (53.8.9) *can also be written as*

$$\beta = \frac{2\sigma_{\mathbf{xy}}}{b + \sqrt{4\sigma_{\mathbf{xy}}^2/\kappa + b^2}} \qquad \textit{where} \quad b = \sigma_{\mathsf{x}}^2 - \sigma_{\mathsf{y}}^2/\kappa. \tag{53.8.10}$$

Answer. Write $a := -\kappa b$; then

$$\beta = \frac{1}{2\sigma_{\mathbf{xy}}}\Big[a + \sqrt{4\kappa\sigma_{\mathbf{xy}}^2 + a^2}\Big] \tag{53.8.11}$$

$$= \frac{1}{2\sigma_{\mathbf{xy}}}\Big[\frac{a^2 - (4\kappa\sigma_{\mathbf{xy}}^2 + a^2)}{a - \sqrt{4\kappa\sigma_{\mathbf{xy}}^2 + a^2}}\Big] \tag{53.8.12}$$

$$= \frac{1}{2\sigma_{\mathbf{xy}}}\Big[\frac{-4\kappa\sigma_{\mathbf{xy}}^2}{a - \sqrt{4\kappa\sigma_{\mathbf{xy}}^2 + a^2}}\Big]. \tag{53.8.13}$$

From this (53.8.10) follows. □

Problem 479. *Show that*

$$\Big|\hat{\beta}_{OLS}\Big| \leq \Big|\hat{\beta}_P\Big| \leq \Big|\hat{\beta}_{ROLS}\Big| \tag{53.8.14}$$

where $\hat{\beta}_{ROLS}$ *is the parameter obtained by reversed OLS (i.e., by regressing* $\mathbf{x}$ *on* $\mathbf{y}$*).*

Answer. For this one needs (53.8.9) and (53.8.10). □

Now we will show that the same estimate can be obtained by minimizing the weighted sum

$$\frac{1}{\sigma_{\mathsf{u}}^2}\sum(\mathsf{x}_i - \mathsf{x}^*{}_i)^2 + \frac{1}{\sigma_{\mathsf{v}}^2}\sum(\mathsf{y}_i - \alpha - \beta\mathsf{x}^*{}_i)^2 \tag{53.8.15}$$

with respect to α, β, and $\mathsf{x}^*{}_i$.

This minimization is done in three steps. In the first step, we ask: given α and β, what are the best $\mathsf{x}^*{}_i$? (Here we see that this alternative approach to P-estimation also gives us predictions of the systematic variables.) Since each $\mathsf{x}^*{}_i$ occurs only in one summand of each of the two sums, we can minimize these individual summands separately. For the ith summands we minimize

$$\frac{1}{\sigma_{\mathsf{u}}^2}(\mathsf{x}_i - \mathsf{x}^*{}_i)^2 + \frac{1}{\sigma_{\mathsf{v}}^2}(\mathsf{y}_i - \alpha - \beta\mathsf{x}^*{}_i)^2 \tag{53.8.16}$$

with respect to $\mathsf{x}^*{}_i$. The partial derivative is

$$\frac{2}{\sigma_{\mathsf{u}}^2}(\mathsf{x}_i - \mathsf{x}^*{}_i)(-1) + \frac{2}{\sigma_{\mathsf{v}}^2}(\mathsf{y}_i - \alpha - \beta\mathsf{x}^*{}_i)(-\beta) \tag{53.8.17}$$

Setting this zero gives

$$\sigma_{\mathsf{v}}^2(\mathsf{x}_i - \mathsf{x}^*{}_i) + \beta\sigma_{\mathsf{u}}^2(\mathsf{y}_i - \alpha - \beta\mathsf{x}^*{}_i) = 0 \tag{53.8.18}$$

or

$$\mathsf{x}^*{}_i = \frac{\sigma_{\mathsf{v}}^2\mathsf{x}_i + \beta\sigma_{\mathsf{u}}^2(\mathsf{y}_i - \alpha)}{\beta^2\sigma_{\mathsf{u}}^2 + \sigma_{\mathsf{v}}^2}. \tag{53.8.19}$$

If one plugs these $\mathsf{x}^*{}_i$ into the objective function one ends up with a surprisingly simple form:

$$\mathsf{x}_i - \mathsf{x}^*{}_i = -\beta\sigma_{\mathsf{u}}^2 \frac{\mathsf{y}_i - \alpha - \beta\mathsf{x}_i}{\beta^2\sigma_{\mathsf{u}}^2 + \sigma_{\mathsf{v}}^2} \tag{53.8.20}$$

$$\mathsf{y}_i - \alpha - \beta\mathsf{x}^*{}_i = \sigma_{\mathsf{v}}^2 \frac{\mathsf{y}_i - \alpha - \beta\mathsf{x}_i}{\beta^2\sigma_{\mathsf{u}}^2 + \sigma_{\mathsf{v}}^2} \tag{53.8.21}$$

$$\frac{(\mathsf{x}_i - \mathsf{x}^*{}_i)^2}{\sigma_{\mathsf{u}}^2} + \frac{(\mathsf{y}_i - \alpha - \beta\mathsf{x}^*{}_i)^2}{\sigma_{\mathsf{v}}^2} = \frac{(\mathsf{y}_i - \alpha - \beta\mathsf{x}_i)^2}{\beta^2\sigma_{\mathsf{u}}^2 + \sigma_{\mathsf{v}}^2}. \tag{53.8.22}$$

The objective function is the sum of this over i:

$$\frac{1}{\beta^2\sigma_{\mathsf{u}}^2 + \sigma_{\mathsf{v}}^2} \sum_i (\mathsf{y}_i - \alpha - \beta\mathsf{x}_i)^2 \tag{53.8.23}$$

Note: with the true values of α and β, the numerator in (53.8.23) is $\sum_i \mathsf{c}_i^2$ where $\mathsf{c}_i = \mathsf{y}_i - \alpha - \beta\mathsf{x}_i$, and the denominator σ_{c}^2. This form of the objective function can also be derived by a geometrical argument. From $\mathsf{y}^*{}_i = \alpha + \beta\mathsf{x}^*{}_i$ follows

$$\frac{\mathsf{y}^*{}_i}{\sigma_{\mathsf{v}}} = \frac{\alpha}{\sigma_{\mathsf{v}}} + \beta\frac{\sigma_{\mathsf{u}}}{\sigma_{\mathsf{v}}}\frac{\mathsf{x}^*{}_i}{\sigma_{\mathsf{u}}}; \tag{53.8.24}$$

and if one plots $\mathsf{y}_i/\sigma_{\mathsf{v}}$ against $\mathsf{x}_i/\sigma_{\mathsf{u}}$, then the objective function is the squared orthogonal distance of the points from the straight line as in Figure 11. The line has slope $\beta\frac{\sigma_{\mathsf{u}}}{\sigma_{\mathsf{v}}}$. Define $\mathsf{c}_i := \mathsf{y}_i - \alpha - \beta\mathsf{x}_i$ as the vertical distance of the observation from the line if one does not divide the y_i by σ_{v}. And call the orthogonal distance in the normalized plot d_i and the complement of the orthogonal distance a_i. Then $\frac{\mathsf{a}_i}{\mathsf{d}_i} = \beta\frac{\sigma_{\mathsf{u}}}{\sigma_{\mathsf{v}}}$.

By Phytagoras,

$$\mathsf{d}_i^2 + \beta^2\frac{\sigma_{\mathsf{u}}^2}{\sigma_{\mathsf{v}}^2}\mathsf{d}_i^2 = \frac{\mathsf{c}_i^2}{\sigma_{\mathsf{v}}^2} \tag{53.8.25}$$

$$\mathsf{d}_i^2(\sigma_{\mathsf{v}}^2 + \beta^2\sigma_{\mathsf{u}}^2) = \mathsf{c}_i^2 \tag{53.8.26}$$

from which the above simple form of the objective function follows.

Given this objective function, the second step is to minimize out α. Since α appears only in the numerator of the objective function, differentiation is easy:

$$\frac{\partial}{\partial\alpha}\frac{\sum_i(\mathsf{y}_i - \alpha - \beta\mathsf{x}_i)^2}{\beta^2\sigma_{\mathsf{u}}^2 + \sigma_{\mathsf{v}}^2} = \frac{(-2)\sum_i(\mathsf{y}_i - \alpha - \beta\mathsf{x}_i)}{\beta^2\sigma_{\mathsf{u}}^2 + \sigma_{\mathsf{v}}^2} \tag{53.8.27}$$

Setting this zero gives $\sum(\mathsf{y}_i - \alpha - \beta\mathsf{x}_i) = 0$ or $\alpha = \bar{\mathsf{y}} - \beta\bar{\mathsf{x}}$. Plugging this α into the objective function gives

$$\frac{\sum_i(\mathsf{y}_i - \bar{\mathsf{y}} - \beta\mathsf{x}_i + \beta\bar{\mathsf{x}})^2}{\beta^2\sigma_{\mathsf{u}}^2 + \sigma_{\mathsf{v}}^2} =: \frac{\sum_i(\mathsf{y}_i^* - \beta\mathsf{x}_i^*)^2}{\beta^2\sigma_{\mathsf{u}}^2 + \sigma_{\mathsf{v}}^2}. \tag{53.8.28}$$

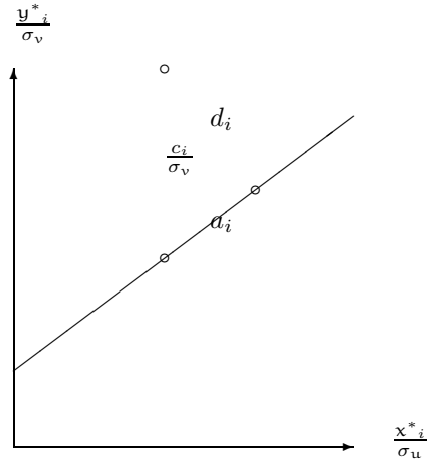


FIGURE 11. Geometric Interpretation of Objective Function

The final step minimizes this with respect to β. Use $(u/v)' = (u'v - uv')/v^2$ to get the partial with respect to β:

$$\frac{2\sum_i(\mathsf{y}_i^* - \beta\mathsf{x}_i^*)(-\mathsf{x}_i^*)(\beta^2\sigma_{\mathsf{u}}^2 + \sigma_{\mathsf{v}}^2) - \sum_i(\mathsf{y}_i^* - \beta\mathsf{x}_i^*)^2 2\beta\sigma_{\mathsf{u}}^2}{(\beta^2\sigma_{\mathsf{u}}^2 + \sigma_{\mathsf{v}}^2)^2}. \tag{53.8.29}$$

Setting this zero gives

$$\sum_i(\mathsf{y}_i^* - \beta\mathsf{x}_i^*)\mathsf{x}_i^*(\beta^2\sigma_{\mathsf{u}}^2 + \sigma_{\mathsf{v}}^2) + \sum_i(\mathsf{y}_i^* - \beta\mathsf{x}_i^*)^2\beta\sigma_{\mathsf{u}}^2 = 0. \tag{53.8.30}$$

Using the sample moments $\boldsymbol{S}_{\mathsf{xy}} = \frac{1}{n}\sum_i \mathbf{x}_i^*\mathbf{y}_i^*$ etc., one obtains

$$(\boldsymbol{S}_{\mathsf{xy}} - \beta\boldsymbol{S}_{\mathsf{x}}^2)(\beta^2\sigma_{\mathsf{u}}^2 + \sigma_{\mathsf{v}}^2) + (\boldsymbol{S}_{\mathsf{y}}^2 - 2\beta\boldsymbol{S}_{\mathsf{xy}} + \beta^2\boldsymbol{S}_{\mathsf{x}}^2)\beta\sigma_{\mathsf{u}}^2 = 0 \tag{53.8.31}$$

or

$$\beta^2\sigma_{\mathsf{u}}^2\boldsymbol{S}_{\mathsf{xy}} + \beta(\boldsymbol{S}_{\mathsf{y}}^2\sigma_{\mathsf{u}}^2 - \boldsymbol{S}_{\mathsf{x}}^2\sigma_{\mathsf{v}}^2) + \sigma_{\mathsf{v}}^2\boldsymbol{S}_{\mathsf{xy}} = 0. \tag{53.8.32}$$

Dividing by σ_{x}^2 and using $\kappa = \sigma_{\mathsf{v}}^2/\sigma_{\mathsf{u}}^2$ one obtains exactly (53.8.7) with the true moments of the observed variables replaced by their sample moments.

53.8.1. Multiple P-Estimation. The Frisch problem in several dimensions reads: given a positive definite matrix $\boldsymbol{Q}$ and a nonnegative definite matrix $\boldsymbol{\Lambda} \neq \boldsymbol{O}$, find a κ so that $\boldsymbol{Q} - \kappa\boldsymbol{\Lambda}$ is positive semidefinite (i.e., nonnegative definite and singular). This κ always exists and is uniquely determined: it is the smallest κ for which $\boldsymbol{Q} - \kappa\boldsymbol{\Lambda}$ is singular, i.e., the smallest root of the equation $\det(\boldsymbol{Q} - \kappa\boldsymbol{\Lambda}) = 0$.

Proof: This is true when $\boldsymbol{Q}$ is the identity matrix and $\boldsymbol{\Lambda}$ is diagonal, then κ is the inverse of the largest diagonal element of $\boldsymbol{\Lambda}$. The general case can always be transformed into this by a nonsingular transformation (see Rao, *Linear Statistical Inference and Its Applications*, p. 41): given a positive definite symmetric $\boldsymbol{Q}$ and a symmetric $\boldsymbol{\Lambda}$, there is always a nonsingular $\boldsymbol{R}$ and a diagonal $\boldsymbol{\Gamma}$ so that $\boldsymbol{Q} = \boldsymbol{R}^\top\boldsymbol{R}$ and $\boldsymbol{\Lambda} = \boldsymbol{R}^\top\boldsymbol{\Gamma}\boldsymbol{R}$. Therefore $\boldsymbol{Q} - \kappa\boldsymbol{\Lambda} = \boldsymbol{R}^\top(\boldsymbol{I} - \kappa\boldsymbol{\Gamma})\boldsymbol{R}$, which is positive semidefinite if and only if $\boldsymbol{I} - \kappa\boldsymbol{\Gamma}$ is. Once one has κ, it is no problem to get all those vectors that annul $\boldsymbol{Q}^*$.

Here is an equivalent procedure which gets $\boldsymbol{\beta}^*$ and κ simultaneously. We will use the following mathematical fact: Given a symmetric positive definite $\boldsymbol{Q}$ and a symmetric nonnebagive definite matrix $\boldsymbol{\Lambda}$. Then the vector $\boldsymbol{\gamma}^*$ annulls $\boldsymbol{Q} - \kappa\boldsymbol{\Lambda}$ iff it is a scalar multiple of a $\boldsymbol{\beta}^*$ which has the minimum property that

(53.8.33)

$$\boxed{\boldsymbol{\beta} = \boldsymbol{\beta}^* \qquad \text{minimizes} \qquad \boldsymbol{\beta}^\top\boldsymbol{Q}\boldsymbol{\beta} \qquad \text{s. t.} \qquad \boldsymbol{\beta}^\top\boldsymbol{\Lambda}\boldsymbol{\beta} = 1,}$$

and κ is the minimum value in this minimization problem. Alternatively one can say that $\boldsymbol{\gamma}^*$ itself has the following maximum property:

(53.8.34)

$$\boxed{\boldsymbol{\gamma} = \boldsymbol{\gamma}^* \qquad \text{maximizes} \qquad \frac{\boldsymbol{\gamma}^\top\boldsymbol{\Lambda}\boldsymbol{\gamma}}{\boldsymbol{\gamma}^\top\boldsymbol{Q}\boldsymbol{\gamma}}}$$

and the maximum value is $1/\kappa$.

Proof: Assume we have κ and $\boldsymbol{\gamma}^* \neq \boldsymbol{o}$ with $\boldsymbol{Q} - \kappa\boldsymbol{\Lambda}$ nonnegative definite (call it $\boldsymbol{Q}^*$) and $(\boldsymbol{Q} - \kappa\boldsymbol{\Lambda})\boldsymbol{\gamma}^* = \boldsymbol{o}$. Since $\boldsymbol{Q}$ is positive definite, $\boldsymbol{\Lambda}\boldsymbol{\gamma}^* \neq \boldsymbol{o}$ and we can define $\boldsymbol{\beta}^* = \boldsymbol{\gamma}^*/\sqrt{\boldsymbol{\gamma}^{*\top}\boldsymbol{\Lambda}\boldsymbol{\gamma}^*}$. Like $\boldsymbol{\gamma}^*$, also $\boldsymbol{\beta}^*$ satisfies $\boldsymbol{\beta}^{*\top}(\boldsymbol{Q} - \kappa\boldsymbol{\Lambda})\boldsymbol{\beta}^* = 0$, but since also $\boldsymbol{\beta}^{*\top}\boldsymbol{\Lambda}\boldsymbol{\beta}^* = 1$, it follows $\boldsymbol{\beta}^{*\top}\boldsymbol{Q}\boldsymbol{\beta}^* = \kappa$. Any other vector $\boldsymbol{\beta}$ with $\boldsymbol{\beta}^\top\boldsymbol{\Lambda}\boldsymbol{\beta} = 1$ satisfies $\boldsymbol{\beta}^\top\boldsymbol{Q}\boldsymbol{\beta} = \boldsymbol{\beta}^\top(\boldsymbol{Q}^* + \kappa\boldsymbol{\Lambda})\boldsymbol{\beta} = \boldsymbol{\beta}^\top\boldsymbol{Q}^*\boldsymbol{\beta} + \kappa \geq \kappa$; in other words, $\boldsymbol{\beta}^*$ is a solution of the minimum problem. This proves the "only if" part, and at the same time shows that the minimum value in (53.8.33) is κ.

For the "if" part assume that $\boldsymbol{\beta}^*$ solves (53.8.33), and $\boldsymbol{\beta}^{*\top}\boldsymbol{Q}\boldsymbol{\beta}^* = \kappa$. Then $\boldsymbol{\beta}^{*\top}(\boldsymbol{Q} - \kappa\boldsymbol{\Lambda})\boldsymbol{\beta}^* = 0$. To show that $\boldsymbol{Q} - \kappa\boldsymbol{\Lambda}$ is nonnegative definite, we will assume there is a $\boldsymbol{\gamma}$ with $\boldsymbol{\gamma}^\top(\boldsymbol{Q} - \kappa\boldsymbol{\Lambda})\boldsymbol{\gamma} < 0$ and derive a contradiction from this. By the same argument as above one can construct a scalar multiple $\boldsymbol{\gamma}^*$ with $\boldsymbol{\gamma}^{*\top}\boldsymbol{\Lambda}\boldsymbol{\gamma}^* = 1$ which still satisfies $\boldsymbol{\gamma}^{*\top}(\boldsymbol{Q} - \kappa\boldsymbol{\Lambda})\boldsymbol{\gamma}^* < 0$. Hense $\boldsymbol{\gamma}^{*\top}\boldsymbol{Q}\boldsymbol{\gamma}^* < \kappa$, which contradicts $\boldsymbol{\beta}^*$ being a minimum value.

If $\boldsymbol{\Lambda}$ has rank one, i.e., there exists a $\mathbf{u}$ with $\boldsymbol{\Lambda} = \mathbf{u}\mathbf{u}^\top$, then the constraint in (53.8.33) reads $(\mathbf{u}^\top\boldsymbol{\beta})^2 = 1$, or $\mathbf{u}^\top\beta = \pm 1$. Since we are looking for all scalar multiples of the solution, we can restrict ourselves to $\mathbf{u}^\top\beta = 1$, i.e., we are back to a linearly constrained problem. For $\boldsymbol{Q}$ positive definite I get the solution formula

$\boldsymbol{\beta}^* = \boldsymbol{Q}^{-1}\mathbf{u}(\mathbf{u}^\top\boldsymbol{Q}^{-1}\mathbf{u})^{-1}$. and the minimum value is $1/(\mathbf{u}^\top\boldsymbol{Q}^{-1}\mathbf{u})$. This can be written in a much neater and simpler form for problem (53.8.34); its solution is any $\boldsymbol{\gamma}^*$ that is a scalar multiple of $\boldsymbol{Q}^{-1}\mathbf{u}$, and the maximum value is $\mathbf{u}^\top\boldsymbol{Q}^{-1}\mathbf{u}$.

If one applies this construction to the sample moments instead of the true limiting moments, one obtains estimates of κ and $\boldsymbol{\beta}$, and also of $\boldsymbol{Q}^*$ and $\widetilde{\boldsymbol{Q}}$. The estimate of $\boldsymbol{Q}^*$ is positive semidefinite by construction, and since $\kappa > 0$ (otherwise $\boldsymbol{Q} - \kappa\boldsymbol{\Lambda}$ would not be singular) also the estimate of $\widetilde{\boldsymbol{Q}} = \kappa\boldsymbol{\Lambda}$ is nonnegative definite. In P-estimation, therefore, the estimates cannot lead to negative variance estimation as in V-estimation.

53.8.2. The P-Estimator as MLE. One can show that the P-estimator is MLE in the structural as well as in the functional variant. We will give the proofs only for the simple EV model.

In the structural variant, in which $\mathbf{x}^*{}_i$ and $\mathbf{y}^*{}_i$ are independent observations from a jointly normal distribution, the same argument applies that we used for the V-estimator: (53.8.9) expresses β as a function of the true moments of the observed variables which are jointly normal; the MLE of these true moments are therefore the sample moments, and the MLE of a function of these true moments is the same function of the sample moments. In this case, not only $\hat{\beta}_P$ but also the estimates for $\sigma^2_{\mathbf{u}}$ etc. derived from the Frisch equations are MLE.

In the functional variant, the $\mathbf{x}^*{}_i$ and $\mathbf{y}^*{}_i$ are nonstochastic and must be maximized over ("incidental parameters") together with the structural parameters of interest α, β, $\sigma^2_{\mathbf{u}}$, and $\sigma^2_{\mathbf{v}}$. We will discuss this functional variant in the slightly more general case of *heteroskedastic errors*, which is a violation of the assumptions made in the beginning, but which occurs frequently, especially with replicated observations which we will discuss next. In the functional variant, we have

$$\mathbf{x}_i \sim \mathsf{N}(\mathbf{x}^*{}_i, \sigma^2_{\mathbf{u}_i}) \qquad f_{\mathbf{x}_i}(\mathrm{x}_i) = \frac{1}{\sqrt{2\pi\sigma^2_{\mathbf{u}_i}}}e^{-(\mathrm{x}_i - \mathbf{x}^*{}_i)^2/2\sigma^2_{\mathbf{u}_i}}; \tag{53.8.35}$$

$$\mathbf{y}_i \sim \mathsf{N}(\alpha + \beta\mathbf{x}^*{}_i, \sigma^2_{\mathbf{v}_i}) \qquad f_{\mathbf{y}_i}(\boldsymbol{Y}_i) = \frac{1}{\sqrt{2\pi\sigma^2_{\mathbf{v}_i}}}e^{-(\boldsymbol{Y}_i - \alpha - \beta\mathbf{x}^*{}_i)^2/2\sigma^2_{\mathbf{v}_i}}. \tag{53.8.36}$$

The likelihood function is therefore

$$(2\pi)^{-n}\Big(\prod_i \sigma^2_{\mathbf{u}_i}\sigma^2_{\mathbf{v}_i}\Big)^{-1/2}\exp -\frac{1}{2}\sum\Big(\frac{(\mathbf{x}_i - \mathbf{x}^*{}_i)^2}{\sigma^2_{\mathbf{u}_i}} + \frac{(\mathbf{y}_i - \alpha - \beta\mathbf{x}^*{}_i)^2}{\sigma^2_{\mathbf{v}_i}}\Big) \tag{53.8.37}$$

Since the parameters $\mathbf{x}^*{}_i$, α, and β only appear in the exponents, their maximum likelihood estimates can be obtained by minimizing the exponent only (and for this, the $\sigma^2_{\mathbf{u}_i}$ and $\sigma^2_{\mathbf{v}_i}$ must be known only to a joint multiplicative factor). If these variances do not depend on i, i.e., in the case of homoskedasticity, one is back to the weighted least squares discussed above.

Also in the case of heteroskedasticity, it is convenient to use the three steps outlined above. Step 1 always goes through, and one ends up with an objective function of the form

$$\sum_i \frac{(\boldsymbol{Y}_i - \alpha - \beta \mathrm{x}_i)^2}{\beta^2 \sigma^2_{\mathbf{u}_i} + \sigma^2_{\mathbf{v}_i}} = \sum_i g_i (\boldsymbol{Y}_i - \alpha - \beta \mathrm{x}_i)^2 \quad \text{with} \quad g_i := \frac{1}{\beta^2 \sigma^2_{\mathbf{u}_i} + \sigma^2_{\mathbf{v}_i}}. \tag{53.8.38}$$

Step 2 establishes an identity involving $\hat{\alpha}$, β, and the weighted means of the observations:

$$\hat{\alpha} = \bar{\bar{\mathbf{y}}} - \beta \bar{\bar{\mathbf{x}}} \quad \text{where} \quad \bar{\bar{\mathbf{x}}} := \frac{\sum g_i \mathbf{x}_i}{\sum g_i} \qquad \bar{\bar{\mathbf{y}}} := \frac{\sum g_i \mathbf{y}_i}{\sum g_i}. \tag{53.8.39}$$

In the general case, step 3 leads to very complicated formulas for $\hat{\beta}$ because $\bar{\bar{\mathbf{x}}}$ and $\bar{\bar{\mathbf{y}}}$ depend on β through the g_i. But there is one situation in which this is not the case: this is when one knows a λ so that $\sigma^2_{\mathbf{v}_i} = \lambda \sigma^2_{\mathbf{u}_i}$ for all i. Then $g_i = \frac{1}{(\beta^2+\lambda)\sigma^2_{\mathbf{u}_i}}$ and one can ignore the constant factor $1/(\beta^2 + \lambda)$ in all g_i to get

$$\bar{\bar{\mathbf{x}}} = \frac{\sum \sigma^2_{\mathbf{u}_i} \mathbf{x}_i}{\sum \sigma^2_{\mathbf{u}_i}} \qquad \bar{\bar{\mathbf{y}}} = \frac{\sum \sigma^2_{\mathbf{u}_i} \mathbf{y}_i}{\sum \sigma^2_{\mathbf{u}_i}} \tag{53.8.40}$$

and the whole procedure goes through as in weighted least squares.

Although we showed that in this case the P estimator of α and β is a maximum likelihood estimate, one can show the interesting and surprising fact that the MLE of $\sigma^2_{\mathbf{u}}$ and $\sigma^2_{\mathbf{v}}$ are by a factor 2 smaller than the P estimators, i.e., they are inconsistent.

53.9. Estimation When the Error Covariance Matrix is Exactly Known

If the error covariance matrix is known exactly, not only up to a constant factor κ, one can use either V estimation or P estimation. However P estimation is better, since it is equal to the MLE of α and β even with this additional information. The knowledge of κ helps improving the variance estimates, but not the estimates of α and β.

53.9.1. Examples. Age determination by Beta decay of Rb^{87} into Sr^{87} [**CCCM81**] This decay follows the equation

$$\frac{dRb^{87}}{dt} = -\lambda Rb^{87} \qquad \lambda = 1.42 \cdot 10^{-11} yr^{-1} \tag{53.9.1}$$

which has the solution $Rb^{87} = Rb^{87}_0 e^{-\lambda t}$. The amount of Sr^{87} is the amount present at time 0 plus the amount created by the decay of Sr^{87} since then:

$$Sr^{87} = Sr^{87}_0 + \overbrace{Rb^{87}_0}^{=Rb^{87}e^{\lambda t}} (1 - e^{-\lambda t}) \tag{53.9.2}$$

$$= Sr^{87}_0 + Rb^{87}(e^{\lambda t} - 1). \tag{53.9.3}$$

If one divides by the stable reference isotope Sr^{86} which does not change over time:

$$\frac{Sr^{87}}{Sr^{86}} = \left(\frac{Sr^{87}}{Sr^{86}}\right)_0 + \frac{Rb^{87}}{Sr^{86}}(e^{\lambda t} - 1) \tag{53.9.4}$$

then this equation is valid not only for one rock sample but for several co-genetic rock samples which have equal starting values of Sr^{87}/Sr^{86} but may have had different concentrations of Rb^{87}/Sr^{86} at the beginning, i.e., one can write this last equation as

$$\mathbf{y}^* = \alpha + \mathbf{x}^* \beta. \tag{53.9.5}$$

The observations of the Sr^{87}/Sr^{86} and Rb^{87}/Sr^{86} ratios do therefore lie on a straight line which is called an isochrone.

In some situations, other errors are introduced by geological accidents during the lifetime of the rock, against which the measurement errors are negligible: lack of chemical isolation of the rock samples, or differences in original Sr^{87}/Sr^{86} contents. Approach: make up several scenarios of what might have happened to the rock, then decide from looking at the scatter plot of the samples which scenario applied, and estimate the data according to the appropriate model. Also: one can use the additional knowledge that all minerals on the earth come from the crust of the earth solidifying a certain known number of years ago, in order to determine one point on every isochrone; then one fits the additional data to a line through this point.

In Biometrics: $\mathbf{y}^*{}_i$ is the log of the wood increase of the ith apple tree in a certain orchard, and $\mathbf{x}^*{}_i$ the log of the increase of the girth. The hypothesis is

$$\mathbf{y}^* = \alpha + \beta \mathbf{x}^*. \tag{53.9.6}$$

However the the actually observed increases in wood and girth are

$$\mathbf{y} = \mathbf{y}^* + \boldsymbol{\nu} \tag{53.9.7}$$

$$\mathbf{x} = \mathbf{x}^* + \mathbf{u} \tag{53.9.8}$$

The errors $\boldsymbol{\nu}$ and $\mathbf{u}$ are assumed to be correlated, but the variances and covariances of these errors can be estimated by repeated measurements of the same trees in the same year, and the ML estimation is only slightly more complicated with this correlation. These variances depend on the ages of the trees (heteroskedasticity), but in order to prevent autocorrelation of the errors every year different trees have to be measured.

CHAPTER 54

Dynamic Linear Models

This chapter draws on the monograph [**WH97**]. The authors are Bayesians, but they attribute the randomness of the parameters not only to the ignorance of the observer, but also to shifts of the true underlying parameters.

54.1. Specification and Recursive Solution

The dynamic linear model (DLM) is a regression with a random parameters $\boldsymbol{\beta}_t$. Unlike the model in chapter 61, the coefficients are not independent drawings from some unchanging distribution, but they are assumed to follow a random walk. In addition to random parameters, the model also has an observational disturbance, and it is an important estimation issue to distinguish the observation noise from the random shifts of the parameters.

We will write the model equations observation by observation. As usual, $\mathbf{y}$ is the n-vector of dependent variables, and $\boldsymbol{X}$ the $n \times k$-matrix of independent variables. But the coefficient vector is different in every period. For all t with $1 \leq t \leq n$, the unobserved underlying $\boldsymbol{\beta}_t$ and the observation y_t obey the familiar regression relationship ("observation equation"):

$$\mathsf{y}_t = \boldsymbol{x}_t^\top \boldsymbol{\beta}_t + \varepsilon_t \qquad \varepsilon_t \sim (0, \sigma^2 u_t) \tag{54.1.1}$$

Here $\boldsymbol{x}_t^\top$ is the tth row of $\boldsymbol{X}$. The "system equation" models the evolution over time of the underlying $\boldsymbol{\beta}_t$:

$$\boldsymbol{\beta}_t = \boldsymbol{G}_t \boldsymbol{\beta}_{t-1} + \boldsymbol{\omega}_t \qquad \boldsymbol{\omega}_t \sim (\boldsymbol{o}, \tau^2 \boldsymbol{\Xi}_t). \tag{54.1.2}$$

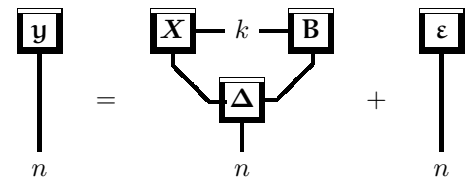
Finally, the model can also handle the following initial information

$$\boldsymbol{\beta}_0 \sim (\boldsymbol{b}_0, \tau^2 \boldsymbol{\Psi}_0) \tag{54.1.3}$$

but it can also be estimated if no prior information is given ("reference model"). The scalar disturbance terms ε_t and the disturbance vectors $\boldsymbol{\omega}_t$ are mutually independent. We know the values of all u_t and $\boldsymbol{\Xi}_t$ and $\kappa^2 = \sigma^2/\tau^2$ (which can be considered the inverse of the signal-to-noise ratio) and, if applicable, $\boldsymbol{\Psi}_0$ and $\boldsymbol{b}_0$, but σ^2 and τ^2 themselves are unknown.

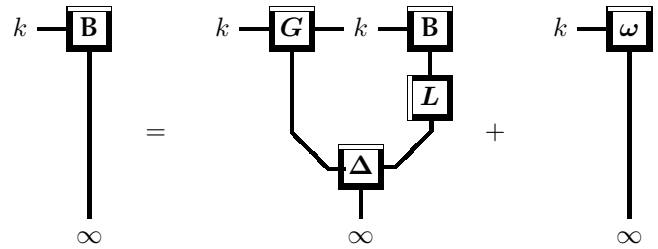
In tiles, the observation equation is

(54.1.4)



and if $n = \infty$ the system equation can be written as

(54.1.5)



where $\boldsymbol{L}$ is the lag operator.

Notation: If y_i are observed for $i = 1, \ldots, t$, then we will use the symbols $\mathbf{b}_t$ for the best linear predictor of $\boldsymbol{\beta}_t$ based on this information, and $\tau^2 \boldsymbol{\Psi}_t$ for its $\mathcal{MSE}$-matrix.

The model with prior information is mathematically easier than that without, because the formulas for $\mathbf{b}_{t+1}$ and $\boldsymbol{\Psi}_{t+1}$ can be derived from those for $\mathbf{b}_t$ and $\boldsymbol{\Psi}_t$ using the following four steps:

(1) The best linear predictor of $\boldsymbol{\beta}_{t+1}$ still with the old information, i.e., the y_i are observed only until $i = 1, \ldots, t$, is simply $\boldsymbol{G}_{t+1}\mathbf{b}_t$. This predictor is unbiased and its $\mathcal{MSE}$-matrix is

$$\mathcal{MSE}[\boldsymbol{G}_{t+1}\mathbf{b}_t; \boldsymbol{\beta}_{t+1}] = \tau^2(\boldsymbol{G}_{t+1}\boldsymbol{\Psi}_t\boldsymbol{G}_{t+1}^\top + \boldsymbol{\Xi}_{t+1}) = \tau^2 \boldsymbol{R}_{t+1} \tag{54.1.6}$$

where we use the abbreviation $\boldsymbol{R}_t = \boldsymbol{G}_t\boldsymbol{\Psi}_{t-1}\boldsymbol{G}_t^\top + \boldsymbol{\Xi}_t$. This formula encapsules the prior information about $\boldsymbol{\beta}_{t+1}$.

(2) The best linear predictor of y_{t+1}, i.e., the one-step-ahead forecast, is $\hat{\mathsf{y}}_{t+1} = \boldsymbol{x}_{t+1}\boldsymbol{G}_{t+1}\mathbf{b}_t$. This predictor is again unbiased and its MSE is

(54.1.7)

$$\mathrm{MSE}[\hat{\mathsf{y}}_{t+1}; \mathsf{y}_{t+1}] = \tau^2 \boldsymbol{x}_{t+1}(\boldsymbol{G}_{t+1}\boldsymbol{\Psi}_t\boldsymbol{G}_{t+1}^\top + \boldsymbol{\Xi}_{t+1})\boldsymbol{x}_{t+1}^\top + \sigma^2 u_t = \tau^2 \boldsymbol{x}_{t+1}\boldsymbol{R}_{t+1}\boldsymbol{x}_{t+1}^\top + \sigma^2 u_t$$

This formula is an encapsulation of the prior information regarding y_{t+1} available before y_{t+1} was observed.

(3) The joint $\mathcal{MSE}$-matrix of $\boldsymbol{G}_{t+1}\mathbf{b}_t$ and $\boldsymbol{x}_{t+1}^\top\boldsymbol{G}_{t+1}\mathbf{b}_t$ as best linear predictors of $\boldsymbol{\beta}_{t+1}$ and y_{t+1} based on all observations up to and including time t is

$$\mathcal{MSE}\left[\begin{bmatrix}\boldsymbol{x}_{t+1}^\top\boldsymbol{G}_{t+1}\mathbf{b}_t\\ \boldsymbol{G}_{t+1}\mathbf{b}_t\end{bmatrix};\begin{bmatrix}\mathsf{y}_{t+1}\\ \boldsymbol{\beta}_{t+1}\end{bmatrix}\right]=\tau^2\begin{bmatrix}\boldsymbol{x}_{t+1}^\top\boldsymbol{R}_{t+1}\boldsymbol{x}_{t+1}+\kappa^2u_t & \boldsymbol{x}_{t+1}^\top\boldsymbol{R}_{t+1}\\ \boldsymbol{R}_{t+1}\boldsymbol{x}_{t+1} & \boldsymbol{R}_{t+1}.\end{bmatrix}\tag{54.1.8}$$

An inverse of this $\mathcal{MSE}$ matrix is

$$\frac{1}{\sigma^2}\begin{bmatrix}u_t^{-1} & u_t^{-1}\boldsymbol{x}_{t+1}^\top\\ \boldsymbol{x}_{t+1}u_t^{-1} & \kappa^2\boldsymbol{R}_{t+1}^{-1}+\boldsymbol{x}_{t+1}u_t^{-1}\boldsymbol{x}_{t+1}^\top\end{bmatrix}\tag{54.1.9}$$

(4) Now we are in the situation of Problem 327. After observation of y_{t+1} we can use the best linear prediction formula (27.1.15) to get the "posterior" predictor of $\boldsymbol{\beta}_{t+1}$ as

$$\mathbf{b}_{t+1}=\Big(\boldsymbol{x}_{t+1}u_t^{-1}\boldsymbol{x}_{t+1}^\top+\kappa^2\boldsymbol{R}_{t+1}^{-1}\Big)^{-1}\Big(\boldsymbol{x}_{t+1}u_t^{-1}\mathsf{y}_{t+1}+\kappa^2\boldsymbol{R}_{t+1}^{-1}\boldsymbol{G}_{t+1}\mathbf{b}_t\Big)\tag{54.1.10}$$

which has $\mathcal{MSE}$-matrix

$$\tau^2\Big(\boldsymbol{x}_{t+1}u_t^{-1}\boldsymbol{x}_{t+1}^\top+\kappa^2\boldsymbol{R}_{t+1}^{-1}\Big)^{-1}\tag{54.1.11}$$

Let's look at (54.1.10). It can also be written as

$$\mathbf{b}_{t+1}=\Big(\frac{1}{\sigma^2u_t}\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^\top+\frac{1}{\tau^2}\boldsymbol{R}_{t+1}^{-1}\Big)^{-1}\Big(\frac{1}{\sigma^2u_t}\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^\top(\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^\top)^-\boldsymbol{x}_{t+1}\mathsf{y}_{t+1}+\frac{1}{\tau^2}\boldsymbol{R}_{t+1}^{-1}\boldsymbol{G}_{t+1}\mathbf{b}_t\Big)\tag{54.1.12}$$

It is a matrix weighted average between $(\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^\top)^-\boldsymbol{x}_{t+1}\mathsf{y}_{t+1}$ and $\boldsymbol{G}_{t+1}\mathbf{b}_t$. The second term is the prior information about $\boldsymbol{\beta}_{t+1}$, and the weight is the inverse of the $\mathcal{MSE}$-matrix. The first term can be considered the information flowing back to $\boldsymbol{\beta}_{t+1}$ from the observation of y_{t+1}: from $\mathsf{y}_{t+1}=\boldsymbol{x}^\top\boldsymbol{\beta}_{t+1}+\varepsilon_{t+1}$ with error variance σ^2u_t one would get, by a naive application of the OLS-formula $\hat{\boldsymbol{\beta}}_{t+1}=(\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^\top)^{-1}\boldsymbol{x}_{t+1}\mathsf{y}_{t+1}$ and the covariance matrix would be $\sigma^2u_t(\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^\top)^{-1}$. Now these inverses do not exist, therefore the formula here uses the g-inverse, and the weight is again an analog of the inverse of the covariance matrix.

54.2. Locally Constant Model

The simplest special case of the dynamic linear model is the "locally constant" model:

$$\mathsf{y}_t=\beta_t+\varepsilon_t\qquad\varepsilon_t\sim\mathsf{NID}(0,\sigma^2)\tag{54.2.1}$$

$$\beta_t=\beta_{t-1}+\omega_t\qquad\omega_t\sim\mathsf{NID}(0,\tau^2)\tag{54.2.2}$$

$$\beta_0\sim(b_0,\tau^2\psi_0)\tag{54.2.3}$$

All ε_s and all ω_t and β_0 are mutually independent. $\kappa^2=\sigma^2/\tau^2$ is known but σ^2 and τ^2 separately are not.

Table 1. Locally Constant as Special Case of Dynamic Model

General	k	$\boldsymbol{\beta}_t$	$\boldsymbol{x}_t$	$\boldsymbol{b}_0$	$\mathbf{b}_t$	$\boldsymbol{\Psi}_0$	$\boldsymbol{\Psi}_t$	$\boldsymbol{G}_t$	$\boldsymbol{\omega}_t$	$\boldsymbol{\Xi}_t$	u_t
Locally const.	1	β_t	1	b_0	b_t	ψ_0	ψ_t	1	ω_t	1	1

We can use the general solution formulas derived earlier inserting the specific values listed in Table 1, but it is more instructive to derive these formulas from scratch.

Problem 480. *The BLUE of β_{t+1} based on the observations $\mathsf{y}_1,\dots,\mathsf{y}_{t+1}$ is the optimal combination of the following two unbiased estimators of β_{t+1}.*

• a. *1 point The estimator is the BLUE of β_t before y_{t+1} was available; call this estimator b_t. For the purposes of this recursion b_t is known, it was computed in the previous iteration, and $\mathrm{MSE}[\mathsf{b}_t;\beta_t]=\tau^2\psi_t$ is known for the same reason. b_t is not only the BLUE of β_t based on the observations $\mathsf{y}_1,\dots,\mathsf{y}_t$, but it is also the BLUE of β_{t+1} based on the observations $\mathsf{y}_1,\dots,\mathsf{y}_t$. Compute $\mathrm{MSE}[\mathsf{b}_t;\beta_{t+1}]$ as a function of $\tau^2\psi_t$.*

Answer. Since $\beta_{t+1}=\beta_t+\omega_{t+1}$ where $\omega_{t+1}\sim(0,\tau^2)$ is independent of b_t, b_t can also serve as a predictor of β_{t+1}, with $\mathrm{MSE}[\mathsf{b}_t;\beta_{t+1}]=\tau^2(\psi_t+1)$. □

• b. *1 point The second unbiased estimator is the new observation y_{t+1}. What is its* MSE *as a estimator of β_{t+1}?*

Answer. Since $\mathsf{y}_{t+1}=\beta_{t+1}+\varepsilon_{t+1}$ where $\varepsilon_{t+1}\sim(0,\sigma^2)$, clearly $\mathrm{MSE}[\mathsf{y}_{t+1};\beta_{t+1}]=\sigma^2$. □

• c. *3 points The estimation errors of the two unbiased estimators, y_{t+1} and b_t are independent of each other. Therefore use problem 206 to compute the best linear combination of these estimators and the* MSE *of this combination?*

Answer. We have to take their weighted average, with weights proportional to the inverses of the MSE's.

$$\mathsf{b}_{t+1}=\frac{\frac{1}{\tau^2(\psi_t+1)}\mathsf{b}_t+\frac{1}{\sigma^2}\mathsf{y}_{t+1}}{\frac{1}{\tau^2(\psi_t+1)}+\frac{1}{\sigma^2}}=\frac{\kappa^2\mathsf{b}_t+(\psi_t+1)\mathsf{y}_{t+1}}{\kappa^2+\psi_t+1}\tag{54.2.4}$$

(the second formula is obtained from the first by multiplying numerator and denominator by $\sigma^2(\psi_t+1)$). The MSE of this pooled estimator is $\mathrm{MSE}[\mathsf{b}_{t+1};\beta_{t+1}]=\tau^2\psi_{t+1}$ where

$$\psi_{t+1}=\frac{\kappa^2(\psi_t+1)}{\kappa^2+\psi_t+1}.\tag{54.2.5}$$

(54.2.4) and (54.2.5) are recursive formulas, which allow to compute ψ_{t+1} from ψ_t, and b_{t+1} from b_t and ψ_t.

- □

In every recursive step we first compute ψ_t and use this to get the weights of y_{t+1} and b_{t+1} in their weighted average. These two steps can be combined since the weight of y_{t+1} is exactly $a_{t+1} = \psi_t/\kappa^2$, which is called the "adaptive coefficient": (**??**) can be written as

$$\mathsf{b}_{t+1} = a_{t+1}\mathsf{y}_{t+1} + (1 - a_{t+1})\mathsf{b}_t = \mathsf{b}_t + a_{t+1}(\mathsf{y}_{t+1} - \mathsf{b}_t), \tag{54.2.6}$$

and (**??**) gives the following recursive formula for a_t:

$$a_{t+1} = \frac{a_t + \frac{1}{\kappa^2}}{a_t + \frac{1}{\kappa^2} + 1}. \tag{54.2.7}$$

For a_1 it is more convenient to use

$$a_1 = \frac{\psi_0 + 1}{\kappa^2 + \psi_0 + 1} \tag{54.2.8}$$

Problem 481. *Write a program for the locally constant model in the programming language of your choice.*

Answer. Here is the R-function `dlm.loconst` which is in the `ecmet` package, but the more general function `dlm.origin` has the same functionality except the `discount` computation:

```
dlm.loconst <- function(y, kappasqr=1, priormean, priorvar)
{
  ##locally constant dynamic linear model for y
  lngth <- length(y)
  kappinv <-  1/kappasqr
  ##first initialize bvec and avec to their full length
  bvec <- vector(mode="numeric",length=lngth)
  avec <- vector(mode="numeric",length=lngth)
  avec[[1]] <- (priorvar+1)/(kappasqr+priorvar+1)
  bvec[[1]] <- avec[[1]]*y[[1]]+(1-avec[[1]])*priormean
  for (i in 2:lngth)
    {avec[[i]] <- (avec[[i-1]]+kappinv)/(avec[[i-1]]+kappinv+1);
     bvec[[i]] <- avec[[i]]*y[[i]]+(1-avec[[i]])*bvec[[i-1]];
   }
  ##For the computation of the one-step-ahead prediction mse
  ##note that y[-1] is vector y with first observation dropped
  ##and bvec[-lngth] is bvec with last component dropped
  ##value returned:
  list(coefficients=bvec,
       adaptive=avec,
       residuals=y-bvec,
       mse=sum((y[-1]-bvec[-lngth])^2)/(lngth-1),
       discount=1-sqrt(kappinv*(1+0.25*kappinv))+0.5*kappinv)
}
```

□

The limit value a satisfies

$$a = \frac{a + \frac{1}{\kappa^2}}{a + \frac{1}{\kappa^2} + 1} \tag{54.2.9}$$

i.e., it depends on κ^2 alone. This quadratic equation has one nonnegative solution

$$a = \sqrt{\frac{1}{\kappa^2} + \frac{1}{4\kappa^4}} - \frac{1}{2\kappa^2} \tag{54.2.10}$$

Problem 482. *Solve the quadratic equation* (54.2.9).

Answer. Multiply out to get $a^2 + \frac{1}{\kappa^2}a = \frac{1}{\kappa^2}$; complete the square on lefthand side $(a + \frac{1}{2\kappa^2})^2 = a^2 + \frac{1}{\kappa^2}a + \frac{1}{4\kappa^4} = \frac{1}{\kappa^2} + \frac{1}{4\kappa^4}$; therefore $a + \frac{1}{2\kappa^2} = \pm\sqrt{\frac{1}{\kappa^2} + \frac{1}{4\kappa^4}}$. Since $\sqrt{\frac{1}{\kappa^2} + \frac{1}{4\kappa^4}} \geq \sqrt{\frac{1}{4\kappa^4}} = \frac{1}{2\kappa^2}$, only the + sign gives a positive a. One should also mention that $\kappa^2 = 0$ gives $a = 1$, while $1/\kappa^2 = 0$ gives $a = 0$. □

The pre-limit values also depend on the initial value a_1 and can be written (here $d = 1 - a$ is the "discount factor")

$$a_t = a\frac{(1 - d^{2t-2})a + (d + d^{2t-2})a_1}{(1 + d^{2t-1})a + (d - d^{2t-1})a_1} \tag{54.2.11}$$

Problem 483. *Simulate and plot time series following a locally constant model with various values of κ^2 so that you become familiar with the forms of behavior such series can display.*

54.3. The Reference Model

The reference prior estimator of $\boldsymbol{\beta}_t$ in the dynamic model, i.e., the best linear unbiased estimator using the observations $\mathsf{y}_1, \ldots, \mathsf{y}_{t-1}$, has the form $\boldsymbol{H}_t^{-1}\mathbf{h}_t$, with $\mathcal{MSE}$-matrix $\tau^2\boldsymbol{H}_t^{-1}$, and the posterior estimator, using the observations $\mathsf{y}_1, \ldots, \mathsf{y}_t$, has the form $\boldsymbol{K}_t^{-1}\mathbf{k}_t$, with $\mathcal{MSE}$-matrix $\tau^2\boldsymbol{K}_t^{-1}$, where the k-vectors $\mathbf{h}_t$ and $\mathbf{k}_t$ and the $k \times k$ matrices $\boldsymbol{H}_t$ and $\boldsymbol{K}_t$ are constructed as follows:

Starting values are $\mathbf{h}_1 = \boldsymbol{o}$ and $\boldsymbol{H}_1 = \boldsymbol{O}$ because in the reference model there is no information prior to the data, therefore the estimator is an indeterminate vector $\boldsymbol{O}^{-1}\boldsymbol{o}$ with zero precision matrix. Then define

$$\mathbf{k}_t = \mathbf{h}_t + \boldsymbol{x}_t(\kappa^2 u_t)^{-1}\mathsf{y}_t \qquad \boldsymbol{K}_t = \boldsymbol{H}_t + \boldsymbol{x}_t(\kappa^2 u_t)^{-1}\boldsymbol{x}_t^\top. \tag{54.3.1}$$

And starting from $\mathbf{k}_t$ and $\boldsymbol{K}_t$ define $\mathbf{h}_{t+1}$ and $\boldsymbol{H}_{t+1}$:

$$\mathbf{h}_{t+1} = \boldsymbol{\Xi}_{t+1}^{-1}\boldsymbol{G}_{t+1}(\boldsymbol{G}_{t+1}^\top\boldsymbol{\Xi}_{t+1}^{-1}\boldsymbol{G}_{t+1} + \boldsymbol{K}_t)^{-1}\mathbf{k}_t \tag{54.3.2}$$

$$\boldsymbol{H}_{t+1} = \boldsymbol{\Xi}_{t+1}^{-1} - \boldsymbol{\Xi}_{t+1}^{-1}\boldsymbol{G}_{t+1}(\boldsymbol{G}_{t+1}^\top\boldsymbol{\Xi}_{t+1}^{-1}\boldsymbol{G}_{t+1} + \boldsymbol{K}_t)^{-1}\boldsymbol{G}_{t+1}^\top\boldsymbol{\Xi}_{t+1}^{-1}. \tag{54.3.3}$$

These formulas are taken from [**WH97**, p. 129–131].

Here are elements of a proof, which must still be worked out better. "Conditionally" on $\mathsf{y}_1, \cdots \mathsf{y}_{t-1}$, and now I am leaving the subscripts t out: $\boldsymbol{\beta} \sim \boldsymbol{H}^{-1}\boldsymbol{h}, \tau^2\boldsymbol{H}^{-1}$ and $\mathsf{y} = \boldsymbol{x}^\top\boldsymbol{\beta} + \varepsilon$ with $\boldsymbol{\varepsilon}$ uncorrelated with $\boldsymbol{\beta}$, variance $\sigma^2 u$. Therefore

$$\begin{bmatrix}\mathsf{y}\\ \boldsymbol{\beta}\end{bmatrix} \sim \begin{bmatrix}\boldsymbol{x}^\top\boldsymbol{H}^{-1}\boldsymbol{h}\\ \boldsymbol{H}^{-1}\boldsymbol{h}\end{bmatrix}, \tau^2\begin{bmatrix}\boldsymbol{x}^\top\boldsymbol{H}^{-1}\boldsymbol{x}+\kappa^2 u & \boldsymbol{x}^\top\boldsymbol{H}^{-1}\\ \boldsymbol{H}^{-1}\boldsymbol{x} & \boldsymbol{H}^{-1}\end{bmatrix} \tag{54.3.4}$$

The inverse of this covariance matrix is

$$\frac{1}{\tau^2}\begin{bmatrix}(\kappa^2 u)^{-1} & -(\kappa^2 u)^{-1}\boldsymbol{x}^\top\\ -\boldsymbol{x}(\kappa^2 u)^{-1} & \boldsymbol{H}+\boldsymbol{x}(\kappa^2 u)^{-1}\boldsymbol{x}^\top\end{bmatrix} \tag{54.3.5}$$

Therefore after observing y_t one gets the best predictor

$$\mathsf{b} = \boldsymbol{H}^{-1}\boldsymbol{h} + \big(\boldsymbol{H}+\boldsymbol{x}(\kappa^2 u)^{-1}\boldsymbol{x}^\top\big)^{-1}\boldsymbol{x}(\kappa^2 u)^{-1}(\mathsf{y}_t - \boldsymbol{x}^\top\boldsymbol{H}^{-1}\boldsymbol{h}) \tag{54.3.6}$$

$$= \Big(\boldsymbol{I} - \big(\boldsymbol{H}+\boldsymbol{x}(\kappa^2 u)^{-1}\boldsymbol{x}^\top\big)^{-1}\boldsymbol{x}(\kappa^2 u)^{-1}\boldsymbol{x}^\top\Big)\boldsymbol{H}^{-1}\boldsymbol{h} \tag{54.3.7}$$

$$+ \big(\boldsymbol{H}+\boldsymbol{x}(\kappa^2 u)^{-1}\boldsymbol{x}^\top\big)^{-1}\boldsymbol{x}(\kappa^2 u)^{-1}\mathsf{y}_t \tag{54.3.8}$$

now use $\big(\boldsymbol{H}+\boldsymbol{x}(\kappa^2 u)^{-1}\boldsymbol{x}^\top\big)^{-1}\boldsymbol{H} + \big(\boldsymbol{H}+\boldsymbol{x}(\kappa^2 u)^{-1}\boldsymbol{x}^\top\big)^{-1}\boldsymbol{x}(\kappa^2 u)^{-1}\boldsymbol{x}^\top = \boldsymbol{I}$:

$$= \big(\boldsymbol{H}+\boldsymbol{x}(\kappa^2 u)^{-1}\boldsymbol{x}^\top\big)^{-1}\boldsymbol{H}\boldsymbol{H}^{-1}\boldsymbol{h} + \big(\boldsymbol{H}+\boldsymbol{x}(\kappa^2 u)^{-1}\boldsymbol{x}^\top\big)^{-1}\boldsymbol{x}(\kappa^2 u)^{-1}\mathsf{y}_t \tag{54.3.9}$$

$$= \big(\boldsymbol{H}+\boldsymbol{x}(\kappa^2 u)^{-1}\boldsymbol{x}^\top\big)^{-1}(\boldsymbol{h}+\boldsymbol{x}(\kappa^2 u)^{-1}\mathsf{y}_t) = \boldsymbol{K}^{-1}\mathsf{k} \tag{54.3.10}$$

where $\boldsymbol{K}$ and k are defined in (54.3.1). The $\mathcal{MSE}$-matrix of b is the inverse of the lower right partition of the inverse covariance matrix (54.3.5), which is $\tau^2\boldsymbol{K}^{-1}$.

For the proof of equations (54.3.2) and (54.3.3) note first that

$$\mathcal{MSE}[\boldsymbol{G}\mathsf{b}_t;\boldsymbol{\beta}_{t+1}] = \mathcal{E}\big[(\boldsymbol{G}\mathsf{b}_t - \boldsymbol{\beta}_{t+1})(\boldsymbol{G}\mathsf{b}_t - \boldsymbol{\beta}_{t+1})^\top\big] \tag{54.3.11}$$

$$= \mathcal{E}\big[(\boldsymbol{G}\mathsf{b}_t - \boldsymbol{G}\boldsymbol{\beta}_t - \boldsymbol{\omega}_{t+1})(\boldsymbol{G}\mathsf{b}_t - \boldsymbol{G}\boldsymbol{\beta}_t - \boldsymbol{\omega}_{t+1})^\top\big] \tag{54.3.12}$$

$$= \mathcal{E}\big[(\boldsymbol{G}\mathsf{b}_t - \boldsymbol{G}\boldsymbol{\beta}_t)(\boldsymbol{G}\mathsf{b}_t - \boldsymbol{G}\boldsymbol{\beta}_t)^\top\big] + \mathcal{E}\big[\boldsymbol{\omega}_{t+1}\boldsymbol{\omega}_{t+1}^\top\big] \tag{54.3.13}$$

$$= \boldsymbol{G}\,\mathcal{MSE}[\mathsf{b}_t;\boldsymbol{\beta}_t]\boldsymbol{G}^\top + \mathcal{V}[\boldsymbol{\omega}_{t+1}] = \tau^2\boldsymbol{G}\boldsymbol{K}_t^{-1}\boldsymbol{G}^\top + \tau^2\boldsymbol{\Xi}_{t+1}, \tag{54.3.14}$$

By problem 484, (54.3.3) is the inverse of this $\mathcal{MSE}$-matrix.

Problem 484. *Verify that*

$$\big(\boldsymbol{G}\boldsymbol{K}_t^{-1}\boldsymbol{G}^\top + \boldsymbol{\Xi}_{t+1}\big)^{-1} = \boldsymbol{\Xi}_{t+1}^{-1} - \boldsymbol{\Xi}_{t+1}^{-1}\boldsymbol{G}_{t+1}(\boldsymbol{G}_{t+1}^\top\boldsymbol{\Xi}_{t+1}^{-1}\boldsymbol{G}_{t+1} + \boldsymbol{K}_t)^{-1}\boldsymbol{G}_{t+1}^\top\boldsymbol{\Xi}_{t+1}^{-1}. \tag{54.3.15}$$

Answer. Multiply the matrix with its alleged inverse and see whether you get $\boldsymbol{I}$:

$$\begin{aligned}
&\big(\boldsymbol{G}\boldsymbol{K}_t^{-1}\boldsymbol{G}^\top + \boldsymbol{\Xi}_{t+1}\big)\big(\boldsymbol{\Xi}_{t+1}^{-1} - \boldsymbol{\Xi}_{t+1}^{-1}\boldsymbol{G}_{t+1}(\boldsymbol{G}_{t+1}^\top\boldsymbol{\Xi}_{t+1}^{-1}\boldsymbol{G}_{t+1} + \boldsymbol{K}_t)^{-1}\boldsymbol{G}_{t+1}^\top\boldsymbol{\Xi}_{t+1}^{-1}\big) =\\
&= \boldsymbol{G}\boldsymbol{K}_t^{-1}\boldsymbol{G}^\top\boldsymbol{\Xi}_{t+1}^{-1} - \boldsymbol{G}\boldsymbol{K}_t^{-1}\boldsymbol{G}^\top\boldsymbol{\Xi}_{t+1}^{-1}\boldsymbol{G}_{t+1}(\boldsymbol{G}_{t+1}^\top\boldsymbol{\Xi}_{t+1}^{-1}\boldsymbol{G}_{t+1} + \boldsymbol{K}_t)^{-1}\boldsymbol{G}_{t+1}^\top\boldsymbol{\Xi}_{t+1}^{-1} +\\
&\quad + \boldsymbol{\Xi}_{t+1}\boldsymbol{\Xi}_{t+1}^{-1} - \boldsymbol{\Xi}_{t+1}\boldsymbol{\Xi}_{t+1}^{-1}\boldsymbol{G}_{t+1}(\boldsymbol{G}_{t+1}^\top\boldsymbol{\Xi}_{t+1}^{-1}\boldsymbol{G}_{t+1} + \boldsymbol{K}_t)^{-1}\boldsymbol{G}_{t+1}^\top\boldsymbol{\Xi}_{t+1}^{-1} =\\
&= \boldsymbol{I} + \big(\boldsymbol{G}\boldsymbol{K}_t^{-1} - \boldsymbol{G}\boldsymbol{K}_t^{-1}\boldsymbol{G}^\top\boldsymbol{\Xi}_{t+1}^{-1}\boldsymbol{G}_{t+1}(\boldsymbol{G}_{t+1}^\top\boldsymbol{\Xi}_{t+1}^{-1}\boldsymbol{G}_{t+1} + \boldsymbol{K}_t)^{-1} -\\
&\quad - \boldsymbol{G}_{t+1}(\boldsymbol{G}_{t+1}^\top\boldsymbol{\Xi}_{t+1}^{-1}\boldsymbol{G}_{t+1} + \boldsymbol{K}_t)^{-1}\big)\boldsymbol{G}_{t+1}^\top\boldsymbol{\Xi}_{t+1}^{-1} =\\
&= \boldsymbol{I} + \boldsymbol{G}\boldsymbol{K}_t^{-1}\big(\boldsymbol{I} - \boldsymbol{G}^\top\boldsymbol{\Xi}_{t+1}^{-1}\boldsymbol{G}_{t+1}(\boldsymbol{G}_{t+1}^\top\boldsymbol{\Xi}_{t+1}^{-1}\boldsymbol{G}_{t+1} + \boldsymbol{K}_t)^{-1} -\\
&\quad - \boldsymbol{K}_t(\boldsymbol{G}_{t+1}^\top\boldsymbol{\Xi}_{t+1}^{-1}\boldsymbol{G}_{t+1} + \boldsymbol{K}_t)^{-1}\big)\boldsymbol{G}_{t+1}^\top\boldsymbol{\Xi}_{t+1}^{-1} = \boldsymbol{I}
\end{aligned}$$

□

Now let us see what this looks like in the simplest case where $\boldsymbol{X} = \boldsymbol{x}$ has only one column, and $\boldsymbol{\Xi}_t = 1$ and $u_t = 1$. For clarity I am using here capital letters for certain scalars

$$\mathsf{k}_t = \mathsf{h}_t + \frac{x_t\mathsf{y}_t}{\kappa^2} \tag{54.3.16}$$

$$K_t = H_t + \frac{x_t^2}{\kappa^2} \tag{54.3.17}$$

$$\mathsf{h}_{t+1} = (1+K_t)^{-1}\mathsf{k}_t = \frac{\mathsf{k}_t}{1+K_t} \tag{54.3.18}$$

$$H_{t+1} = 1 - (1+K_t)^{-1} = \frac{K_t}{1+K_t} \tag{54.3.19}$$

Starting values are $\mathsf{h}_1 = 0$ and $H_1 = 0$; then

$$\mathsf{k}_1 = \frac{x_1\mathsf{y}_1}{\kappa^2} \qquad K_1 = \frac{x_1^2}{\kappa^2}$$

$$\mathsf{h}_2 = \frac{\mathsf{k}_1}{1+K_1} \qquad H_2 = \frac{K_1}{1+K_1}$$

etc.

Problem 485. *Write a program for the dynamic regression line through the origin, reference model, in the programming language of your choice, or write a macro in the spreadsheet of your choice.*

Answer. The R-code is in Table 2. If argument `x` is missing, the locally constant model will be estimated. Note that `y[-1]` is vector `y` with first observation dropped, and since `lngth` is the length of the vectors, `bvec[-lngth]` is `bvec` with last component dropped. The last line is the expression returned by the function.

□

```
dlmref.origin <- function(y, x = rep(1,length(y)), kappasqr)
{
  lngth <- length(y)
  Hvec <-  hvec <-  Kvec <- kvec <- vector(mode="numeric",length=lngth)
  ##The next statement commented out, redundant in R which initializes Hvec and
  ##hvec by default as 0; but perhaps needed in other programming languages:
  ##Hvec[[1]] <- hvec[[1]] <- 0
  kvec[[1]] <- x[[1]]*y[[1]]/kappasqr
  Kvec[[1]] <- x[[1]]*x[[1]]/kappasqr
  for (i in 2:lngth)
    { hvec[[i]] <- kvec[[i-1]]/(1+Kvec[[i-1]])
      Hvec[[i]] <- Kvec[[i-1]]/(1+Kvec[[i-1]])
      kvec[[i]] <- hvec[[i]]+x[[i]]*y[[i]]/kappasqr
      Kvec[[i]] <- Hvec[[i]]+x[[i]]*x[[i]]/kappasqr
    }
  bvec <- kvec/Kvec
  ##Now computation of the one-step-ahead prediction mse
  onestep <- y[-1]-x[-1]*bvec[-lngth]
  mse <- sum(onestep^2)/(lngth-1)
  list(coefficients=bvec, residuals=y-x*bvec,
       onestep=c(NA, onestep), mse=mse)
}
```

TABLE 2. Dynamic Regression Line through Origin, Reference Model

54.4. Exchange Rate Forecasts

PROBLEM 486. *8 points The daily levels of the exchange rate of the Pound Sterling (£) against the US $, taken at noon in New York City, from 1990 to the present, are published at*

www.federalreserve.gov/releases/H10/hist/dat96_uk.txt

and the data from 1971–1989 are at

www.federalreserve.gov/releases/H10/hist/dat89_uk.txt

Download these data, make a high quality plot, import the plot into your wordprocessor, and write a short essay describing what you see.

ANSWER. Figure 1 plots the daily levels of the exchange rate. It has a lot of detail which one can only see if one magnifies the plot on the pdf-reader. Similar graphs are in [**WH97**, p. 67] and [**Gut94**, Figure 14.1 on p. 370].

The following description of what you see in an exchange rate graph borrows heavily from [**Gut94**, p. 369].

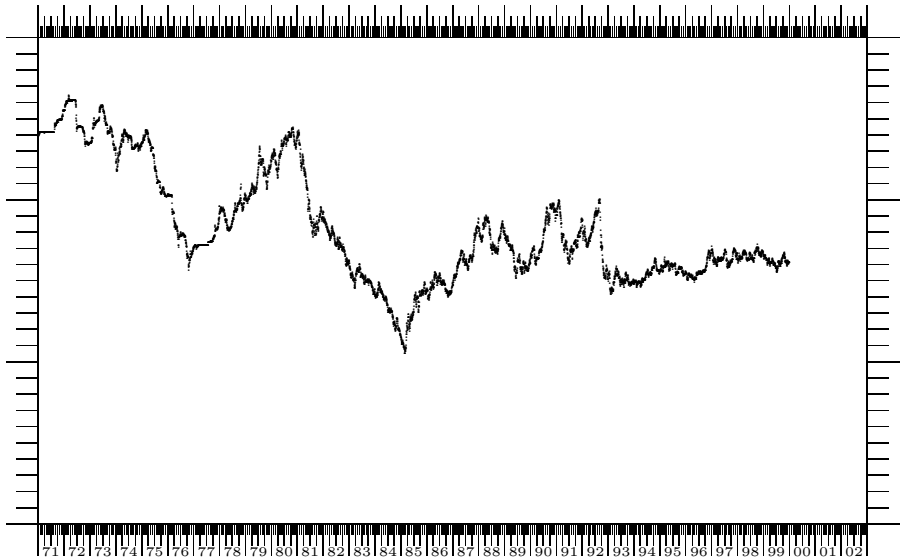


FIGURE 1. Exchange Rate of Pound in terms of Dollar

Secular trend: In long run exchange rates reflect a country's competitiveness in the international hierarchy of nations. When a country manages to strengthen its competitive position in the world market, its external accounts improve and its currency appreciates (e.g. Germany, Japan). The reverse happens if a country faces gradual erosion (Great Britain, United States). Therefore you see steady runs of gradual linear advances or declines over many years.

Business cycle: Woven around this secular trend are cycles of 4–7 years. "This pattern suggests that exchange-rate movements trigger counteracting adjustments in goods and assets markets. But these effects take time to unfold, and in the meantime foreign-exchange markets overshoot. The overshooting sets the stage for the next phase of the cycle, when it has finally begun to turn around such economic fundamentals as inflation and the direction of macroeconomic policy."

Is there an even shorter cycle due to inventories and the time it takes to find new suppliers?

Shorter-term exchange rate fluctuations lasting a few weeks or months are due to expectations and speculation: "At times expectational biases are widely differentiated, and the markets move sideways. But most of the time we can see pronounced price movements in one direction reflecting widely shared market sentiments. These speculative "runs" usually last a few weeks or months before being temporarily interrupted by even shorter countermovements. Because runs outweigh corrections, they reinforce whatever phase of the currency cycle we are in."

Daily variability: Despite these regularities, exchange rates are very volatile in the short run: they often fluctuate 1–2% per day.

Then there are several complete changes in regime which are due to institutional changes in the monetary system. In August 1971, when Nixon abolished the convertibility of the dollar, at the beginning of 1993, with increasing European monetary integration, and at the beginning of 1999, with the introduction of the euro.

□

y_t is the observed market exchange rate, which fluctuates around an underlying level β_t determined by slow-moving "fundamentals." The movement of the underlying level displays a certain inertia: if it has been moving up chances are it will keep moving up at the same rate, and likewise if it has been moving down. This is a situation in which the first differences might follow a dynamic "locally constant" regression model. Here it can be argued that the ε_t are *objectively* random: they are a superposition of many different influences which cannot be predicted. However the ω_t are *subjective* probabilities. The rate of increase or decrease of the exchange rate does not really follow a random walk; instead, it moves slowly as the underlying fundamentals move. But in this model, the researcher is not trying to model the law governing the movement of the fundamentals; the only assumption here is that the movement of the fundamentals is slower than the movement of the disturbance term. The authors of [**WH97**, p. 38] say: "Significant changes over longer periods of time are expected but the zero-mean and independent nature of the $[\omega_t]$ series imply that the modeller does not wish to anticipate the form of this longer term variation, merely describing it as a purely stochastic process."

PROBLEM 487. *Simulate and plot time series whose first differences follow a locally constant model with various values of κ^2 so that you become familiar with the forms of behavior such series can display.*

PROBLEM 488. *The dataset* `dolperpd` *has two variables; the first variable,* `dates`*, is the date coded as an integer, namely the number of days since midnight January 1, 1970 (i.e., January 1, 1970 is day number 1). The second variable* `dolperpd` *is the exchange rate of the British Pound in terms of the dollar at noon of that day in New York City. The data go from 1971 until 1989, they are a subset of the data plotted in figure 1. These data are included in* R*-library* `ecmet`*, and a text file with the data is available on the web at* `www.econ.utah.edu/ehrbar/data/dolperpd.txt`*.*

• a. *Compute the dynamic linear reference model for various values of κ^2 and look at the average squared forecasting error. For which value of κ^2 is it lowest? Interpret your result.*

ANSWER. If one takes the daily data, the SSE becomes lowest as $\kappa^2 \to 0$. This means, one gets best forecasting performance if one treats the data as a random walk. The most recent observation is the best predictor of the future, there is no such thing as an "underlying level." □

• b. *Now take the weekly averages of these data, and the monthly averages, and see which κ^2 minimizes the* SSE*.*

ANSWER. Still the SSE declines as $\kappa^2 \to 0$.

□

• c. *Now take the* first differences *of the weekly and monthly averages, and see which κ^2 minimizes the forecasting error.*

ANSWER. Now a nonzero κ^2 minimizes the SSE. What does this mean? Instead of fluctuating around slowly varying levels, the data fluctuate around slowly changing straight lines. □

• d. *Make several simulations of datasets which have the same length as the datasets which you started out with, using the optimal κ^2. What differences do you see between your original and the simulated datasets?*

ANSWER. On the one hand, the original data are bounded; they move in a fixed range, while the simulated data wander off randomly also in the long run. Therefore the plots can be misleading; if the range is large, it is compressed and the data look much smoother than they are if the range happens to be comparable to that of the economic data. On the other, the economic data have different behavior if the series goes up than if it goes down. □

54.5. Company Market Share

Year	Company Sales y_t Quarter				Total Market x_t Quarter			
	1	2	3	4	1	2	3	4
1975	71.2	52.7	44.0	64.5	161.7	126.4	105.5	150.7
1976	70.2	52.3	45.2	66.8	162.1	124.2	107.2	156.0
1977	72.4	55.1	48.9	64.8	165.8	130.8	114.3	152.4
1978	73.3	56.5	50.0	66.8	166.7	132.8	115.8	155.6
1979	80.2	58.8	51.1	67.9	183.0	138.3	119.1	157.3
1980	73.8	55.9	49.8	66.6	169.1	128.6	112.2	149.5
1981	70.0	54.8	48.7	67.7	156.9	123.4	108.8	153.3
1982	70.4	52.7	49.1	64.8	158.3	119.5	107.7	145.0
1983	70.0	55.3	50.1	65.6	155.3	123.1	109.2	144.8
1984	72.7	55.2	51.5	66.2	160.6	119.1	109.5	144.8
1985	75.5	58.5			165.8	127.4		

TABLE 3. Market Share Data [**WH97**, p. 83]

The data in Table 3 are included in the R-library `ecmet` under the name `mktshare`, and a text file with the data is available on the web at `www.econ.utah.edu/ehrbar/data/mktshare.txt`. A scatterplot of the data as for instance [**WH97**, Figure 3.4 on p. 77], "seems to support a simple, essentially static straight line regression with, from the nature of the data, zero origin." But if one tries to forecast next year's data using a fixed straight line based on this year and all past data, one gets poor results, because the line is slowly moving over time.

PROBLEM 489. *Plot both* `sales` *and* `market` *against time, plot them against each other, and plot their ratio* `sales/market` *against time. What do you see?*

With β_t being the company's market share, this movement of the line can be modeled as a dynamic zero-intercept regression model:

$$\mathsf{y}_t = x_t\beta_t + \varepsilon_t \qquad \varepsilon_t \sim \mathrm{NID}(0,\sigma^2) \tag{54.5.1}$$

$$\beta_t = \beta_{t-1} + \omega_t \qquad \omega_t \sim \mathrm{NID}(0,\tau^2) \tag{54.5.2}$$

$$\beta_0 \sim (b_0, \tau^2\psi_0) \tag{54.5.3}$$

All ε_s are independent of b_0 and all ω_t. $\kappa^2 = \sigma^2/\tau^2$ is known.

Let's compute again the best estimate of β_{t+1} given the observation of $\mathsf{y}_1, \ldots, \mathsf{y}_{t+1}$. We have two pieces of information about β_{t+1}. On the one hand, $\beta_{t+1} = \beta_t + \omega_{t+1}$ where an estimator b_t of β_t is available with $\mathcal{MSE}[\mathsf{b}_t;\beta_t] = \tau^2\psi_t$, and $\omega_{t+1} \sim (0,\tau^2)$ is independent of this estimator. Therefore $\mathrm{MSE}[\mathsf{b}_t;\beta_{t+1}] = \tau^2(\psi_t+1)$. On the other hand, $\mathsf{y}_{t+1} = x_{t+1}\beta_{t+1} + \varepsilon_{t+1}$ where $\varepsilon_{t+1} \sim (0,\sigma^2)$, therefore $\mathrm{MSE}[\frac{\mathsf{y}_{t+1}}{x_{t+1}};\beta_{t+1}] = \frac{\sigma^2}{x_{t+1}^2}$. These two pieces of information are independent of each other. To combine them optimally, take their weighted average, with weights proportional to the inverses of the MSE's.

$$\mathsf{b}_{t+1} = \frac{\frac{1}{\tau^2(\psi_t+1)}\mathsf{b}_t + \frac{x_{t+1}^2}{\sigma^2}\frac{\mathsf{y}_{t+1}}{x_{t+1}}}{\frac{1}{\tau^2(\psi_t+1)} + \frac{x_{t+1}^2}{\sigma^2}} = \frac{\kappa^2\mathsf{b}_t + x_{t+1}\mathsf{y}_{t+1}(\psi_t+1)}{\kappa^2 + x_{t+1}^2(\psi_t+1)} = \frac{\kappa^2\mathsf{b}_t + (\psi_t+1)\mathsf{y}_{t+1}}{\kappa^2+\psi_t+1} \tag{54.5.4}$$

and the MSE of this pooled estimator is $\mathrm{MSE}[\mathsf{b}_{t+1};\beta_{t+1}] = \tau^2\psi_{t+1}$ where

$$\psi_{t+1} = \frac{\kappa^2(\psi_t+1)}{\kappa^2+\psi_t+1}. \tag{54.5.5}$$

Here it is convenient to define $a_t = x_t\psi_t/\kappa^2$ so that

$$\mathsf{b}_{t+1} = a_{t+1}\mathsf{y}_{t+1} + (1 - x_{t+1}a_{t+1})\mathsf{b}_t = \mathsf{b}_t + a_{t+1}(\mathsf{y}_{t+1} - x_{t+1}\mathsf{b}_t) \tag{54.5.6}$$

with the recursive relation

$$a_{t+1} = \frac{x_{t+1}\left(\frac{a_t}{x_t} + \frac{1}{\kappa^2}\right)}{1 + x_{t+1}^2\left(\frac{a_t}{x_t} + \frac{1}{\kappa^2}\right)}. \tag{54.5.7}$$

For a_1 is is more convenient to use

$$a_1 = x_1\frac{\psi_0+1}{\kappa^2 + x_1^2(\psi_0+1)} \tag{54.5.8}$$

These two formulas are used in the R-function `dlm.origin` in the `ecmet` package, see Table 4. Again, if the `x`-argument is missing, the locally constant model will be estimated.


```
dlm.origin <- function(y, x, kappasqr, priormean, priorvar)
{
  lngth <- length(y)
  kappinv <- 1/kappasqr
  if (missing(x))  { x <- rep(1,lngth) }
  avec <- bvec <- auxvec <- vector(mode="numeric",length=lngth)
  avec[[1]] <- x[[1]]*(priorvar+1)/(kappasqr+x[[1]]^2*(priorvar+1))
  bvec[[1]] <- priormean + avec[[1]]*(y[[1]]-x[[1]]*priormean)
  for (i in 2:lngth)
    { auxvec[[i]] <- x[[i]]*(avec[[i-1]]/x[[i-1]]+kappinv)
      avec[[i]] <- auxvec[[i]]/(1+x[[i]]*auxvec[[i]])
      bvec[[i]] <- bvec[[i-1]]+avec[[i]]*(y[[i]]-x[[i]]*bvec[[i-1]])
    }
  residuals <- y - x*bvec
  list(coefficients=bvec, adaptive=avec, residuals=residuals)
}
```

TABLE 4. Dynamic Line through Origin, Prior Information

Now let us contrast this with the model in which $\omega_t = 0$ for all t, i.e., the regression line itself does not move, but since the data arrive sequentially one updates the estimate with each data point. In analogy with the dynamic model, we write it as follows

$$\mathsf{y}_t = x_t\beta + \varepsilon_t \qquad \varepsilon_t \sim \mathrm{NID}(0,\sigma^2) \tag{54.5.9}$$

$$\beta_0 \sim (b_0, \sigma^2\psi_0). \tag{54.5.10}$$

Here β_0 is the prior estimate of β before any data are available. All ε_s are assumed independent of β_0.

Let's compute again recursively the best estimate of β given the observation of $\mathsf{y}_1, \ldots, \mathsf{y}_{t+1}$. We have two pieces of information about β. On the one hand, b_t is the best estimator of β; it is unbiased with $\mathrm{MSE}[\mathsf{b}_t;\beta_{t+1}] = \sigma^2\psi_t$. On the other hand, $\mathsf{y}_{t+1} = x_{t+1}\beta + \varepsilon_{t+1}$ where $\varepsilon_{t+1} \sim (0,\sigma^2)$, therefore $\mathrm{MSE}[\frac{\mathsf{y}_{t+1}}{x_{t+1}};\beta] = \frac{\sigma^2}{x_{t+1}^2}$. These two pieces of information are independent of each other. To combine them optimally, take their weighted average, with weights proportional to the inverses of the MSE's.

$$\mathsf{b}_{t+1} = \frac{\frac{1}{\sigma^2\psi_t}\mathsf{b}_t + \frac{x_{t+1}^2}{\sigma^2}\frac{\mathsf{y}_{t+1}}{x_{t+1}}}{\frac{1}{\sigma^2\psi_t} + \frac{x_{t+1}^2}{\sigma^2}} = \frac{\mathsf{b}_t + x_{t+1}\mathsf{y}_{t+1}\psi_t}{1 + x_{t+1}^2\psi_t} \tag{54.5.11}$$

The MSE of this pooled measure is

$$\sigma^2\psi_{t+1} = \mathrm{MSE}[\mathsf{b}_{t+1};\beta] = \frac{1}{\frac{1}{\sigma^2\psi_t}+\frac{x_{t+1}^2}{\sigma^2}} = \frac{\sigma^2}{\frac{1}{\psi_t}+x_{t+1}^2} = \frac{\sigma^2\psi_t}{1+x_{t+1}^2\psi_t}. \tag{54.5.12}$$

Here it is convenient to define $a_t = x_t\psi_t$ so that

$$\mathsf{b}_{t+1} = a_{t+1}\mathsf{y}_{t+1} + (1-x_{t+1}a_{t+1})\mathsf{b}_t = \mathsf{b}_t + a_{t+1}(\mathsf{y}_{t+1} - x_{t+1}\mathsf{b}_t) \tag{54.5.13}$$

with the recursive relation

$$a_{t+1} = \frac{\frac{x_{t+1}a_t}{x_t}}{1+\frac{x_{t+1}^2a_t}{x_t}} = \frac{x_{t+1}a_t}{x_t+x_{t+1}^2a_t}. \tag{54.5.14}$$

For a_1 it is more convenient to use

$$a_1 = \frac{x_1\psi_0}{1+x_1^2\psi_0} \tag{54.5.15}$$

The other extreme, in which $\varepsilon = 0$, leads to the estimates $\beta_t = \frac{\mathsf{y}_t}{x_t}$ which is based on one point only. If you run `ecmet.script(mktshare)` you will see these different estimations (the reference model was used in all cases). The static line through the origin is ok in the first 3 years, but then it hopelessly underpredicts. Even if you allow the regression line to move a tiny little bit, making $\kappa^2 = 100,000$, i.e., the standard deviation of ω is one third of one percent of that of ε, the predicted trajectory stays much closer to the observed one, although it still underpredicts. A $\kappa^2 = 10,000$ gives a better fit, but now there is very little smoothing going on, and there is still underprediction. The problem is that the dynamic line through the origin allows the line to move, but assumes that there will be zero movement even if the line has been moving in the same direction for a long time. Apparently there is some momentum in the movement of the market share: the product's market share is trending upwards and the predictions should take this into consideration. The linear growth model (54.6.1) – (54.6.3) does exactly this, and if one looks at the one-step-ahead forecasts now, finally there is no longer underprediction, but one sees a very clear seasonal pattern which should be tackled next.

West and Harrison [**WH97**, Section 3.4.2 on pp. 84–91] use a prior not only for the means but also for the variance and estimate the variance from the data too. We are using a simpler model, therefore our results and theirs are a little different.

54.6. Productivity in Milk Production

Problem 490. *One of the examples on* [**WH97**, p. 75] *is the milk production data series reproduced in Table 5 and shown in Figure 2.* y_t *is the annual milk production in the United States in* 10^9 *lbs., over a 13-year period, and* x_t *is the number of milk cows in millions. In* R *with the* `ecmet` *package loaded, the command*

`data(milkprod)` *makes these data available. These data can also be downloaded as a text file from* `www.econ.utah.edu/ehrbar/data/milkprod.txt`.

• a. *Plot both* `milk` *and* `cows` *against time, plot them against each other, and plot their ratio* `milk/market` *against time. What do you see?*

• b. *West and Harrison compare the forecasting performance of their dynamic straight line through the origin with that of a static straight line, and say that the dynamic model, although not perfect, is much to be preferred. One of the criteria of a good model is its forecasting ability. Plot the one-step ahead forecasting errors in both of these models into the same figure.*

Table 5. Annual Milk Production and Milk Cows

Year t	1970	1971	1972	1973	1974	1975	1976
Milk y_t	117.0	118.6	120.0	115.5	115.6	115.4	120.2
Cows x_t	12.0	11.8	11.7	11.4	11.2	11.1	11.0
Year t		1977	1978	1979	1980	1981	1982
Milk y_t		122.7	121.5	123.4	128.5	130.0	135.8
Cows x_t		11.0	10.8	10.7	10.8	10.9	11.0

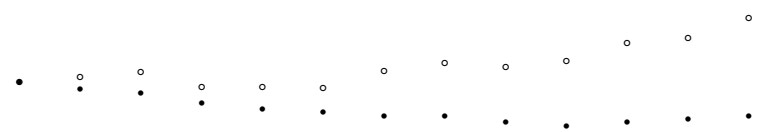


Figure 2. Annual Milk Production (hollow dots) and Milk Cows (filled dots)

The plots just made show that the dynamic linear model is still not satisfactory. A better fitting dynamic linear model with the same data is estimated in [**NT92**].

$$\mathfrak{y}_t = x_t\beta_t + \varepsilon_t \qquad \varepsilon_t \sim (0, \kappa^2\tau^2) \tag{54.6.1}$$

$$\beta_t = \beta_{t-1} + \delta_t + \mathfrak{u}_t \qquad \mathfrak{u}_t \sim (0, \tau^2) \tag{54.6.2}$$

$$\delta_t = \delta_{t-1} + \nu_t \qquad \nu_t \sim (0, \lambda^2\tau^2) \tag{54.6.3}$$

This can be considered a linear growth curve model. β_t is the annual milk output per cow in year t, i.e., it measures productivity. This productivity increases from year to year. The productivity increase between year $t-1$ and year t is $\delta_t + \mathfrak{u}_t$. The first component is persistent; it does not change much between consecutive years but follows a random walk. The second component is transitory with zero expected value. I.e., the three error terms in this model have three different meanings: ν_t are persistent random shocks in the yearly productivity increases, $\mathfrak{u}_t$ are transient annual fluctuations in productivity, and ε_t represents the discrepancy between actual output and productivity-determined normal output. All three are assumed independent. We will use the notation $\operatorname{var}[\nu_t] = \tau^2$ $\operatorname{var}[\varepsilon] = \sigma^2 = \kappa^2\tau^2$, and $\operatorname{var}[\mathfrak{u}_t] = \lambda^2\tau^2$. There are not enough data to estimate the relative variances of the different error terms as in the exchange rate example; here prior information enters the model.

Problem 491. [**NT92**] *This is an exercise about the growth model* (54.6.1) – (54.6.3).

• a. *3 points Describe the intuitive meaning of* β_t, δ_t, *the three disturbances, and the two parameters* κ *and* λ.

• b. *2 points Show how this model can be fitted in the framework of the dynamic linear model as defined here. Note that in this framework the unobserved random parameters linearly depend on their* lagged *values, while in equation* (54.6.2) β_t *depends on* δ_t *instead of* δ_{t-1}. *But there is a trick to get around this.*

Answer. The trick is to replace δ_t in equation (54.6.2) with $\delta_{t-1} + \nu_t$:

$$\beta_t = \beta_{t-1} + \delta_{t-1} + \nu_t + \mathfrak{u}_t = \beta_{t-1} + \delta_{t-1} + \omega_t \tag{54.6.4}$$

so that the whole system reads

$$\mathfrak{y}_t = x_t\beta_t + \varepsilon_t \tag{54.6.5}$$

$$\beta_t = \beta_{t-1} + \delta_{t-1} + \omega_t \tag{54.6.6}$$

$$\delta_t = \delta_{t-1} + \nu_t \tag{54.6.7}$$

where $\omega_t = \nu_t + \mathfrak{u}_t$. The new disturbances ω_t and ν_t are no longer independent. From the original

$$\varepsilon_t \sim \mathrm{IID}(0, \sigma^2) \qquad \begin{bmatrix}\mathfrak{u}_t\\ \nu_t\end{bmatrix} \sim \mathrm{IID}(\begin{bmatrix}0\\0\end{bmatrix}, \tau^2\begin{bmatrix}1 & 0\\ 0 & \lambda^2\end{bmatrix}) \tag{54.6.8}$$

follows

$$\varepsilon_t \sim \mathrm{IID}(0, \sigma^2) \qquad \begin{bmatrix}\omega_t\\ \nu_t\end{bmatrix} \sim \mathrm{IID}(\begin{bmatrix}0\\0\end{bmatrix}, \tau^2\begin{bmatrix}1+\lambda^2 & \lambda^2\\ \lambda^2 & \lambda^2\end{bmatrix}) \tag{54.6.9}$$

Table 6. Growth Model as Special Case of Dynamic Linear Model

General	$\boldsymbol{\beta}_t$	$\boldsymbol{x}_t$	$\boldsymbol{G}_t$	$\boldsymbol{\omega}_t$	$\boldsymbol{\Xi}_t$	u_t
Growth	$\begin{bmatrix}\beta_t\\ \delta_t\end{bmatrix}$	$\begin{bmatrix}1\\0\end{bmatrix}$	$\begin{bmatrix}1 & 1\\ 0 & 1\end{bmatrix}$	$\begin{bmatrix}\omega_t\\ \nu_t\end{bmatrix}$	$\begin{bmatrix}1+\lambda^2 & \lambda^2\\ \lambda^2 & \lambda^2\end{bmatrix}$	1

□

• c. *Plug the matrices in Table 6 into the formulas for the reference estimator* (54.3.1), (54.3.2), *and* (54.3.3), *and develop simple formulas without matrix notation, which can then be programmed in a spreadsheet or other application.*

CHAPTER 55

Numerical Minimization

Literature: [**Thi88**, p. 199–219] or [**KG80**, pp. 425–475].

Regarding numerical methods, the books are classics, and they are available on-line for free at `lib-www.lanl.gov/numerical/`

Assume $\boldsymbol{\theta} \mapsto f(\boldsymbol{\theta})$ is a scalar function of a vector argument, with continuous first and second derivatives, which has a global minimum, i.e., there is an argument $\hat{\boldsymbol{\theta}}$ with $f(\hat{\boldsymbol{\theta}}) \leq f(\boldsymbol{\theta})$ for all $\boldsymbol{\theta}$.

The numerical methods to find this minimum argument are usually recursive: the computer is given a starting value $\boldsymbol{\theta}_0$, uses it to compute $\boldsymbol{\theta}_1$, then it uses $\boldsymbol{\theta}_1$ to compute $\boldsymbol{\theta}_2$, and so on, constructing a sequence $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots$ that converges towards a minimum argument. If convergence occurs, this minimum is usually a local minimum, and often one is not sure whether there is not another, better, local minimum somewhere else.

At every step, the computer makes two decisions, which can be symbolized as

$$\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i + \alpha_i \boldsymbol{d}_i. \tag{55.0.10}$$

Here $\boldsymbol{d}_i$, a vector, is the step direction, and α_i, a scalar, is the step size. The choice of the step direction is the main characteristic of the program. Most programs (notable exception: simulated annealing) always choose directions at every step along which the objective function slopes downward, so that one will get lower values of the objective function for small increments in that direction. The step size is then chosen such that the objective function actually decreases. In elaborate cases, the step size is chosen to be that traveling distance in the step direction which gives the best improvement in the objective function, but it is not always efficient to spend this much time on the step size.

Let us take a closer look how to determine the step direction. If $\boldsymbol{g}_i^\top = (\boldsymbol{g}(\boldsymbol{\theta}_i))^\top$ is the Jacobian of f at $\boldsymbol{\theta}_i$, i.e., the row vector consisting of the partial derivatives of f, then the objective function will slope down along direction $\boldsymbol{d}_i$ if the scalar product $\boldsymbol{g}_i^\top \boldsymbol{d}_i$ is negative. In determining the step direction, the following fact is useful: All vectors $\boldsymbol{d}_i$ for which $\boldsymbol{g}_i^\top \boldsymbol{d}_i < 0$ can be obtained by premultiplying the transpose of the negative Jacobian, i.e., the negative *gradient vector* $-\boldsymbol{g}_i$, by an appropriate positive definite matrix $\boldsymbol{R}_i$.

Problem 492. *4 points Here is a proof for those who are interested in this issue: Prove that $\boldsymbol{g}^\top \boldsymbol{d} < 0$ if and only if $\boldsymbol{d} = -\boldsymbol{R}\boldsymbol{g}$ for some positive definite symmetric matrix $\boldsymbol{R}$. Hint: to prove the "only if" part use $\boldsymbol{R} = \boldsymbol{I} - \boldsymbol{g}\boldsymbol{g}^\top/(\boldsymbol{g}^\top\boldsymbol{g}) - \boldsymbol{d}\boldsymbol{d}^\top/(\boldsymbol{d}^\top\boldsymbol{g})$. This formula is from* [**Bar74**, p. 86]. *To prove that $\boldsymbol{R}$ is positive definite, note that $\boldsymbol{R} = \boldsymbol{Q} + \boldsymbol{S}$ with both $\boldsymbol{Q} = \boldsymbol{I} - \boldsymbol{g}\boldsymbol{g}^\top/(\boldsymbol{g}^\top\boldsymbol{g})$ and $\boldsymbol{S} = -\boldsymbol{d}\boldsymbol{d}^\top/(\boldsymbol{d}^\top\boldsymbol{g})$ nonnegative definite. It is therefore sufficient to show that any $\boldsymbol{x} \neq \boldsymbol{o}$ for which $\boldsymbol{x}^\top\boldsymbol{Q}\boldsymbol{x} = 0$ satisfies $\boldsymbol{x}^\top\boldsymbol{S}\boldsymbol{x} > 0$.*

Answer. If $\boldsymbol{R}$ is positive definite, then $\boldsymbol{d} = -\boldsymbol{R}\boldsymbol{g}$ clearly satisfies $\boldsymbol{d}^\top\boldsymbol{g} < 0$. Conversely, for any $\boldsymbol{d}$ satisfying $\boldsymbol{d}^\top\boldsymbol{g} < 0$, define $\boldsymbol{R} = \boldsymbol{I} - \boldsymbol{g}\boldsymbol{g}^\top/(\boldsymbol{g}^\top\boldsymbol{g}) - \boldsymbol{d}\boldsymbol{d}^\top/(\boldsymbol{d}^\top\boldsymbol{g})$. One immediately checks that $\boldsymbol{d} = -\boldsymbol{R}\boldsymbol{g}$. To prove that $\boldsymbol{R}$ is positive definite, note that $\boldsymbol{R}$ is the sum of two nonnegative definite matrices $\boldsymbol{Q} = \boldsymbol{I} - \boldsymbol{g}\boldsymbol{g}^\top/(\boldsymbol{g}^\top\boldsymbol{g})$ and $\boldsymbol{S} = -\boldsymbol{d}\boldsymbol{d}^\top/(\boldsymbol{d}^\top\boldsymbol{g})$. It is therefore sufficient to show that any $\boldsymbol{x} \neq \boldsymbol{o}$ for which $\boldsymbol{x}^\top\boldsymbol{Q}\boldsymbol{x} = 0$ satisfies $\boldsymbol{x}^\top\boldsymbol{S}\boldsymbol{x} > 0$. Indeed, if $\boldsymbol{x}^\top\boldsymbol{Q}\boldsymbol{x} = 0$, then already $\boldsymbol{Q}\boldsymbol{x} = \boldsymbol{o}$, which means $\boldsymbol{x} = \frac{\boldsymbol{g}\boldsymbol{g}^\top\boldsymbol{x}}{\boldsymbol{g}^\top\boldsymbol{g}}$. Therefore

$$\boldsymbol{x}^\top\boldsymbol{S}\boldsymbol{x} = -\frac{\boldsymbol{x}\boldsymbol{g}\boldsymbol{g}^\top}{\boldsymbol{g}^\top\boldsymbol{g}}\frac{\boldsymbol{d}\boldsymbol{d}^\top}{\boldsymbol{d}^\top\boldsymbol{g}}\frac{\boldsymbol{g}\boldsymbol{g}^\top\boldsymbol{x}}{\boldsymbol{g}^\top\boldsymbol{g}} = -(\boldsymbol{g}^\top\boldsymbol{x}/\boldsymbol{g}^\top\boldsymbol{g})^2\boldsymbol{d}^\top\boldsymbol{g} > 0. \tag{55.0.11}$$

□

Many important numerical methods, the so-called *gradient methods* [**KG80**, p. 430] use exactly this principle: they find the step direction $\boldsymbol{d}_i$ by premultiplying $-\boldsymbol{g}_i$ by some positive definite $\boldsymbol{R}_i$, i.e., they use the recursion equation

$$\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i - \alpha_i \boldsymbol{R}_i \boldsymbol{g}_i \tag{55.0.12}$$

The most important ingredient here is the choice of $\boldsymbol{R}_i$. We will discuss two "natural" choices.

The choice which immediately comes to mind is to set $\boldsymbol{R}_i = \boldsymbol{I}$, i.e., $\boldsymbol{d}_i = -\alpha_i \boldsymbol{g}_i$. Since the gradient vector shows into the direction where the slope is steepest, this is called the method of steepest descent. However this choice is not as natural as one might first think. There is no benefit to finding the steepest direction, since one can easily increase the step length. It is much more important to find a direction which allows one to go down for a long time—and for this one should also consider how the gradient is changing. The fact that the direction of steepest descent changes if one changes the scaling of the variables, is another indication that selecting the steepest descent is not a natural criterion.

The most "natural" choice for $\boldsymbol{R}_i$ is the inverse of the "Hessian matrix" $\boldsymbol{G}(\boldsymbol{\theta}_i)$, which is the matrix of second partial derivatives of f, evaluated at $\boldsymbol{\theta}_i$. This is called the Newton-Raphson method. If the inverse Hessian is positive definite, the Newton Raphson method amounts to making a Taylor development of f around the so far best point $\boldsymbol{\theta}_i$, breaking this Taylor development off after the quadratic term (so that one gets a quadratic function which at point $\boldsymbol{\theta}_i$ has the same first and second derivatives as the given objective function), and choosing $\boldsymbol{\theta}_{i+1}$ to be the minimum point of this quadratic approximation to the objective function.

Here is a proof that one accomplishes all this if $\boldsymbol{R}_i$ is the inverse Hessian. The quadratic approximation (second order Taylor development) of f around $\boldsymbol{\theta}_i$ is

$$f(\boldsymbol{\theta}) \approx f(\boldsymbol{\theta}_i) + \big(\boldsymbol{g}(\boldsymbol{\theta}_i)\big)^\top(\boldsymbol{\theta}-\boldsymbol{\theta}_i) + \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_i)^\top \boldsymbol{G}(\boldsymbol{\theta}_i)(\boldsymbol{\theta}-\boldsymbol{\theta}_i). \tag{55.0.13}$$

By theorem 55.0.1, the minimum argument of this quadratic approximation is

$$\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i - \big(\boldsymbol{G}(\boldsymbol{\theta}_i)\big)^{-1}\boldsymbol{g}(\boldsymbol{\theta}_i), \tag{55.0.14}$$

which is the above procedure with step size 1 and $\boldsymbol{R}_i = (\boldsymbol{G}(\boldsymbol{\theta}_i))^{-1}$.

THEOREM 55.0.1. *Let $\boldsymbol{G}$ be a $n \times n$ positive definite matrix, and $\boldsymbol{g}$ a n-vector. Then the minimum argument of the function*

$$q:\ \boldsymbol{z} \mapsto \boldsymbol{g}^\top\boldsymbol{z} + \frac{1}{2}\boldsymbol{z}^\top\boldsymbol{G}\boldsymbol{z} \qquad \textit{is} \quad \boldsymbol{x} = -\boldsymbol{G}^{-1}\boldsymbol{g}. \tag{55.0.15}$$

Proof: Since $\boldsymbol{G}\boldsymbol{x} = -\boldsymbol{g}$, it follows for any $\boldsymbol{z}$ that

$$\boldsymbol{z}^\top\boldsymbol{g} + \frac{1}{2}\boldsymbol{z}^\top\boldsymbol{G}\boldsymbol{z} = -\boldsymbol{z}^\top\boldsymbol{G}\boldsymbol{x} + \frac{1}{2}\boldsymbol{z}^\top\boldsymbol{G}\boldsymbol{z} = \tag{55.0.16}$$

$$= \frac{1}{2}\boldsymbol{x}^\top\boldsymbol{G}\boldsymbol{x} - \boldsymbol{z}^\top\boldsymbol{G}\boldsymbol{x} + \frac{1}{2}\boldsymbol{z}^\top\boldsymbol{G}\boldsymbol{z} - \frac{1}{2}\boldsymbol{x}^\top\boldsymbol{G}\boldsymbol{x} \tag{55.0.17}$$

$$= \frac{1}{2}(\boldsymbol{x}-\boldsymbol{z})^\top\boldsymbol{G}(\boldsymbol{x}-\boldsymbol{z}) - \frac{1}{2}\boldsymbol{x}^\top\boldsymbol{G}\boldsymbol{x} \tag{55.0.18}$$

This is minimized by $\boldsymbol{z} = \boldsymbol{x}$.

The Newton-Raphson method requires the Hessian matrix. [**KG80**] recommend to establish mathematical formulas for the derivatives which are then evaluated at $\boldsymbol{\theta}_i$, since it is very tricky and unprecise to compute derivatives and the Hessian numerically. The analytical derivatives, on the other hand, are time consuming and the computation of these derivatives may be subject to human error. However there are computer programs which automatically compute such derivatives. `Splus`, for instance, has the `deriv` function which automatically constructs functions which are the derivatives or gradients of given functions.

The main drawback of the Newton-Raphson method is that $\boldsymbol{G}(\boldsymbol{\theta}_i)$ is only positive definite if the function is strictly convex. This will be the case when $\boldsymbol{\theta}_i$ is close to a minimum, but if one starts too far away from a minimum, the Newton-Raphson method may not converge.

There are many modifications of the Newton-Raphson method which get around computing the Hessian and inverting it at every step and at the same time ensure that the matrix $\boldsymbol{R}_i$ is always positive definite by using an updating formula for $\boldsymbol{R}_i$, which turns $\boldsymbol{R}_i$, after sufficiently many steps into the inverse Hessian. These are probably the most often used methods. A popular one used by the `gauss` software is the *Davidson-Fletcher-Powell* algorithm.

One drawback of all these methods using matrices is the fact that the size of the matrix $\boldsymbol{R}_i$ increases with the square of the number of variables. For problems with large numbers of variables, memory limitations in the computer make it necessary to use methods which do without such a matrix. A method to do this is the "conjugate gradient method." If it is too difficult to compute the gradient vector, the "conjugate direction method" may also compare favorably with computing the gradient numerically.

CHAPTER 56

Nonlinear Least Squares

This chapter ties immediately into chapter 55 about Numerical Minimization. The notation is slightly different; what we called f is now called SSE, and what we called $\boldsymbol{\theta}$ is now called $\boldsymbol{\beta}$. A much more detailed discussion of all this is given in [**DM93**, Chapter 6], which uses the notation $\boldsymbol{x}(\boldsymbol{\beta})$ instead of our $\boldsymbol{\eta}(\boldsymbol{\beta})$. [**Gre97**, Chapter 10] defines the vector function $\boldsymbol{\eta}(\boldsymbol{\beta})$ by $\eta_t(\boldsymbol{\beta}) = h(\boldsymbol{x}_t, \boldsymbol{\beta})$, i.e., all elements of the vector function $\boldsymbol{\eta}$ have the same functional form h but differ by the values of the additional arguments $\boldsymbol{x}_t$. [**JHG+88**, Chapter (12.2.2)] set it up in the same way as [**Gre97**], but they call the function f instead of h.

An additional important "natural" choice for $\boldsymbol{R}_i$ is available if the objective function has the nonlinear least squares form

$$\mathsf{SSE}(\boldsymbol{\beta}) = \big(\mathbf{y} - \boldsymbol{\eta}(\boldsymbol{\beta})\big)^{\top}\big(\mathbf{y} - \boldsymbol{\eta}(\boldsymbol{\beta})\big), \tag{56.0.19}$$

where $\mathbf{y}$ is a given vector of observations and $\boldsymbol{\eta}(\boldsymbol{\beta})$ is a vector function of a vector argument, i.e., it consists of n scalar functions of k scalar arguments each:

$$\boldsymbol{\eta}(\boldsymbol{\beta}) = \begin{bmatrix} \eta_1(\beta_1, \beta_2, \cdots, \beta_k) \\ \eta_2(\beta_1, \beta_2, \cdots, \beta_k) \\ \vdots \\ \eta_n(\beta_1, \beta_2, \cdots, \beta_k) \end{bmatrix} \tag{56.0.20}$$

Minimization of this objective function is an obvious and powerful estimation method whenever the following *nonlinear least squares* model specification holds:

$$\mathbf{y} = \boldsymbol{\eta}(\boldsymbol{\beta}) + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim (\boldsymbol{o}, \sigma^2 \boldsymbol{I}) \tag{56.0.21}$$

If the errors are normally distributed, then nonlinear least squares is equal to the maximum likelihood estimator. (But this is only true as long as the covariance matrix is spherical as assumed here.)

Instead of the linear least squares model

$$\mathsf{y}_1 = x_{11}\beta_1 + x_{12}\beta_2 + \cdots + x_{1k}\beta_k + \varepsilon_1 \tag{56.0.22}$$

$$\mathsf{y}_2 = x_{21}\beta_1 + x_{22}\beta_2 + \cdots + x_{2k}\beta_k + \varepsilon_2 \tag{56.0.23}$$

$$\vdots \qquad \vdots \tag{56.0.24}$$

$$\mathsf{y}_n = x_{n1}\beta_1 + x_{n2}\beta_2 + \cdots + x_{nk}\beta_k + \varepsilon_n \tag{56.0.25}$$

we have here its nonlinear generalization

$$\mathsf{y}_1 = \eta_1(\beta_1, \beta_2, \cdots, \beta_k) + \varepsilon_1 \tag{56.0.26}$$

$$\mathsf{y}_2 = \eta_2(\beta_1, \beta_2, \cdots, \beta_k) + \varepsilon_2 \tag{56.0.27}$$

$$\vdots \qquad \vdots \tag{56.0.28}$$

$$\mathsf{y}_n = \eta_n(\beta_1, \beta_2, \cdots, \beta_k) + \varepsilon_n \tag{56.0.29}$$

Usually there are other independent variables involved in $\boldsymbol{\eta}$ which are not shown here explicitly because they are not needed for the results proved here.

Problem 493. *4 points* [**Gre97**, 10.1 on p. 494] *Describe as precisely as you can how you would estimate the model*

$$\mathsf{y}_i = \alpha \mathsf{x}_i^{\beta} + \varepsilon_i, \tag{56.0.30}$$

and how you would get estimates of the standard deviations of the parameter estimates.

Answer. You want to minimize the nonlinear LS objective function $\mathsf{SSE} = \sum(\mathsf{y}_i - \alpha\mathsf{x}_i^{\beta})^2$. First order conditions you have to set zero $\frac{\partial \mathsf{SSE}}{\partial \alpha} = -2\sum \mathsf{x}_i^{\beta}(\mathsf{y}_i - \alpha\mathsf{x}_i^{\beta})$ and $\frac{\partial \mathsf{SSE}}{\partial \beta} = -2\alpha\sum \log(\mathsf{x}_i)\mathsf{x}_i^{\beta}(\mathsf{y}_i - \alpha\mathsf{x}_i^{\beta})$. There are only two parameters to minimize over, and for every given β it is easy to get the α which minimizes the objective function with the given β fixed, namely, this is

$$\alpha = \frac{\sum \mathsf{x}_i^{\beta}\mathsf{y}_i}{\sum \mathsf{x}_i^{2\beta}} \tag{56.0.31}$$

Plug this α into the objective function gives you the concentrated objective function, then plot this concentrated objective function and make a grid search for the best β. The concentrated objective function can also be obtained by running the regression for every β and getting the SSE from the regression.

After you have the point estimates $\hat{\alpha}$ and $\hat{\beta}$ write $\mathsf{y}_t = \eta_t + \varepsilon_t$ and construct the pseudoregressors $\partial\eta_t/\partial\alpha = \mathsf{x}_t^{\hat{\beta}}$ and $\partial\eta_t/\partial\beta = \alpha(\log \mathsf{x}_t)\mathsf{x}_t^{\hat{\beta}}$. If you regress the residuals on the pseudoregressors you will get parameter estimates zero (if the estimates $\hat{\alpha}$ and $\hat{\beta}$ are good), but you will get the right standard errors.

□

Next we will derive the first-order conditions, and then describe how to run the linearized Gauss-Newton regression. For this we need some notation. For an arbitrary but fixed vector $\boldsymbol{\beta}_i$ (below it will be the ith approximation to the nonlinear least squares parameter estimate) we will denote the Jacobian matrix of the function $\boldsymbol{\eta}$ evaluated at $\boldsymbol{\beta}_i$ with the symbol $\boldsymbol{X}(\boldsymbol{\beta}_i)$, i.e., $\boldsymbol{X}(\boldsymbol{\beta}_i) = \partial\boldsymbol{\eta}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}^\top(\boldsymbol{\beta}_i)$. $\boldsymbol{X}(\boldsymbol{\beta}_i)$ is called the matrix of pseudoregressors at $\boldsymbol{\beta}_i$. The mh-th element of $\boldsymbol{X}(\boldsymbol{\beta}_i)$ is

$$x_{mh}(\boldsymbol{\beta}_i) = \frac{\partial\eta_m}{\partial\beta_h}(\boldsymbol{\beta}_i), \tag{56.0.32}$$

i.e., $\boldsymbol{X}(\boldsymbol{\beta}_i)$ is the matrix of partial derivatives evaluated at $\boldsymbol{\beta}_i$

$$\boldsymbol{X}(\boldsymbol{\beta}_i) = \begin{bmatrix} \frac{\partial\eta_1}{\partial\beta_1}(\boldsymbol{\beta}_i) & \cdots & \frac{\partial\eta_1}{\partial\beta_k}(\boldsymbol{\beta}_i) \\ \vdots & & \\ \frac{\partial\eta_n}{\partial\beta_1}(\boldsymbol{\beta}_i) & \cdots & \frac{\partial\eta_n}{\partial\beta_k}(\boldsymbol{\beta}_i) \end{bmatrix}, \tag{56.0.33}$$

but $\boldsymbol{X}(\boldsymbol{\beta}_i)$ should first and foremost be thought of as the coefficient matrix of the best linear approximation of the function $\boldsymbol{\eta}$ at the point $\boldsymbol{\beta}_i$. In other words, it is the matrix which appears in the Taylor expansion of $\boldsymbol{\eta}(\boldsymbol{\beta})$ around $\boldsymbol{\beta}_i$:

$$\boldsymbol{\eta}(\boldsymbol{\beta}) = \boldsymbol{\eta}(\boldsymbol{\beta}_i) + \boldsymbol{X}(\boldsymbol{\beta}_i)(\boldsymbol{\beta} - \boldsymbol{\beta}_i) + \text{higher order terms}. \tag{56.0.34}$$

Now let us compute the Jacobian of the objective function itself

$$\mathsf{SSE} = (\mathbf{y} - \boldsymbol{\eta}(\boldsymbol{\beta}))^\top(\mathbf{y} - \boldsymbol{\eta}(\boldsymbol{\beta})) = \hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}} \qquad \text{where} \quad \hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \boldsymbol{\eta}(\boldsymbol{\beta}). \tag{56.0.35}$$

This Jacobian is a row vector because the objective function is a scalar function. We need the chain rule (C.1.23) to compute it. In the present situation it is useful to break our function into three pieces and apply the chain rule for three steps:

$$\partial\mathsf{SSE}/\partial\boldsymbol{\beta}^\top = \partial\mathsf{SSE}/\partial\hat{\boldsymbol{\varepsilon}}^\top\cdot\partial\hat{\boldsymbol{\varepsilon}}/\partial\boldsymbol{\eta}^\top\cdot\partial\boldsymbol{\eta}/\partial\boldsymbol{\beta}^\top = 2\hat{\boldsymbol{\varepsilon}}^\top\cdot(-\boldsymbol{I})\cdot\boldsymbol{X}(\boldsymbol{\beta}) = -2(\mathbf{y}-\boldsymbol{\eta}(\boldsymbol{\beta}))^\top\boldsymbol{X}(\boldsymbol{\beta}). \tag{56.0.36}$$

Problem 494. *3 points Compute the Jacobian of the nonlinear least squares objective function*

$$\mathsf{SSE} = (\mathbf{y} - \boldsymbol{\eta}(\boldsymbol{\beta}))^\top(\mathbf{y} - \boldsymbol{\eta}(\boldsymbol{\beta})) \tag{56.0.37}$$

where $\boldsymbol{\eta}(\boldsymbol{\beta})$ is a vector function of a vector argument. Do not use matrix differentiation but compute it element by element and then verify that it is the same as equation (56.0.36).

Answer.

$$\mathsf{SSE} = \sum_{t=1}^{n}(\mathsf{y}_t - \eta_t(\boldsymbol{\beta}))^2 \tag{56.0.38}$$

$$\frac{\partial\mathsf{SSE}}{\partial\beta_h} = \sum_{t=1}^{n} 2(\mathsf{y}_t - \eta_t(\boldsymbol{\beta}))\cdot(-\frac{\partial\eta_t}{\partial\beta_h}) \tag{56.0.39}$$

$$= -2\sum_{t}(\mathsf{y}_t - \eta_t(\boldsymbol{\beta}))\frac{\partial\eta_t}{\partial\beta_h} \tag{56.0.40}$$

$$= -2\Big((\mathsf{y}_1 - \eta_1(\boldsymbol{\beta}))\frac{\partial\eta_1}{\partial\beta_h} + \cdots + (\mathsf{y}_n - \eta_n(\boldsymbol{\beta}))\frac{\partial\eta_n}{\partial\beta_h}\Big) \tag{56.0.41}$$

$$= -2\begin{bmatrix}\mathsf{y}_1 - \eta_1(\boldsymbol{\beta}) & \cdots & \mathsf{y}_n - \eta_n(\boldsymbol{\beta})\end{bmatrix}\begin{bmatrix}\frac{\partial\eta_1}{\partial\beta_h}\\ \vdots \\ \frac{\partial\eta_n}{\partial\beta_h}\end{bmatrix} \tag{56.0.42}$$

Therefore

$$\begin{bmatrix}\frac{\partial\mathsf{SSE}}{\partial\beta_1} & \cdots & \frac{\partial\mathsf{SSE}}{\partial\beta_k}\end{bmatrix} = -2\begin{bmatrix}\mathsf{y}_1 - \eta_1(\boldsymbol{\beta}) & \cdots & \mathsf{y}_n - \eta_n(\boldsymbol{\beta})\end{bmatrix}\begin{bmatrix}\frac{\partial\eta_1}{\partial\beta_1} & \cdots & \frac{\partial\eta_1}{\partial\beta_k}\\ \vdots & & \\ \frac{\partial\eta_n}{\partial\beta_1} & \cdots & \frac{\partial\eta_n}{\partial\beta_k}\end{bmatrix}. \tag{56.0.43}$$

□

The gradient vector is the transpose of (56.0.36):

$$\boldsymbol{g}(\boldsymbol{\beta}) = -2\boldsymbol{X}^\top(\boldsymbol{\beta})(\mathbf{y} - \boldsymbol{\eta}(\boldsymbol{\beta})) \tag{56.0.44}$$

Setting this zero gives the first order conditions.

$$\boldsymbol{X}^\top(\boldsymbol{\beta})\boldsymbol{\eta}(\boldsymbol{\beta}) = \boldsymbol{X}^\top(\boldsymbol{\beta})\mathbf{y} \tag{56.0.45}$$

It is a good idea to write down these first order conditions and to check whether some of them can be solved for the respective parameter, i.e., whether some parameters can be concentrated out.

Plugging (56.0.34) into (56.0.21) and rearranging gives the regression equation

$$\mathbf{y} - \boldsymbol{\eta}(\boldsymbol{\beta}_i) = \boldsymbol{X}(\boldsymbol{\beta}_i)(\boldsymbol{\beta} - \boldsymbol{\beta}_i) + \text{error term} \tag{56.0.46}$$

Here the lefthand side is known (it can be written $\hat{\boldsymbol{\varepsilon}}_i$, the residual associated with the vector $\boldsymbol{\beta}_i$), we observe $\mathbf{y}$, and $\boldsymbol{\beta}_i$ is the so far best approximation to the minimum argument. The matrix of "pseudoregressors" $\boldsymbol{X}(\boldsymbol{\beta}_i)$ is known, but the coefficent $\boldsymbol{\delta}_i = \boldsymbol{\beta} - \boldsymbol{\beta}_i$ is not known (because we do not know $\boldsymbol{\beta}$) and must be estimated. The error term contains the higher order terms in (56.0.34) plus the vector of random disturbances in (56.0.21). This regression is called the Gauss-Newton regression (GNR) at $\boldsymbol{\beta}_i$. [**Gre97**, (10-8) on p, 452] writes it as

$$\mathbf{y} - \boldsymbol{\eta}(\boldsymbol{\beta}_i) + \boldsymbol{X}(\boldsymbol{\beta}_i)\boldsymbol{\beta}_i = \boldsymbol{X}(\boldsymbol{\beta}_i)\boldsymbol{\beta} + \text{error term} \tag{56.0.47}$$

PROBLEM 495. *6 points* [**DM93**, p. 178], *which is very similar to* [**Gre97**, (10-2) on p. 450]*: You are estimating by nonlinear least squares the model*

$$\mathfrak{y}_t = \alpha + \beta x_t + \gamma z_t^\delta + \varepsilon_t \qquad \text{or} \quad \boldsymbol{\mathfrak{y}} = \alpha\boldsymbol{\iota} + \beta\boldsymbol{x} + \gamma\boldsymbol{z}^\delta + \boldsymbol{\varepsilon} \tag{56.0.48}$$

You are using the iterative Newton-Raphson algorithm.

• a. *In the ith step you have obtained the vector of estimates*

$$\hat{\boldsymbol{\beta}}_i = \begin{bmatrix} \hat\alpha \\ \hat\beta \\ \hat\gamma \\ \hat\delta \end{bmatrix}. \tag{56.0.49}$$

Write down the matrix $\boldsymbol{X}$ of pseudoregressors, the first order conditions, the Gauss-Newton regression at the given parameter values, and the updated estimate $\hat{\boldsymbol{\beta}}_{i+1}$.

ANSWER. The matrix of pseudoregressors is, column by column,

$$\boldsymbol{X} = \begin{bmatrix} \partial\boldsymbol{\eta}/\partial\alpha & \partial\boldsymbol{\eta}/\partial\beta & \partial\boldsymbol{\eta}/\partial\gamma & \partial\boldsymbol{\eta}/\partial\delta \end{bmatrix} \tag{56.0.50}$$

where $\boldsymbol{\eta}(\alpha,\beta,\gamma,\delta) == \alpha\boldsymbol{\iota}+\beta\boldsymbol{x}+\gamma\boldsymbol{z}^\delta$. From $\partial\eta_t/\partial\alpha = 1$ follows $\partial\boldsymbol{\eta}/\partial\alpha = \boldsymbol{\iota}$; from $\partial\eta_t/\partial\beta = x_t$ follows $\partial\boldsymbol{\eta}/\partial\beta = \boldsymbol{x}$; from $\partial\eta_t/\partial\gamma = z_t^\delta$ follows $\partial\boldsymbol{\eta}/\partial\gamma = \boldsymbol{z}^\delta$ (which is the vector taken to the δth power element by element). And from $\partial\eta_t/\partial\delta = \frac{\partial}{\partial\delta}\gamma z_t^\delta = \frac{\partial}{\partial\delta}\gamma\exp(\delta\log(z_t)) = \gamma\log(z_t)\exp(\delta(\log z_t)) = \gamma\log(z_t)z_t^\delta$ follows $\partial\boldsymbol{\eta}/\partial\delta = \gamma\log(\boldsymbol{z}) * \boldsymbol{z}^\delta$ where $*$ denotes the Hadamard product of two matrices (their element-wise multiplication). Putting it together gives $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{\iota} & \boldsymbol{x} & \boldsymbol{z}^\delta & \gamma\log(\boldsymbol{z}) * \boldsymbol{z}^\delta \end{bmatrix}$.

Write the first order conditions (56.0.45) in the form $\boldsymbol{X}^\top(\boldsymbol{\beta})(\boldsymbol{\mathfrak{y}} - \boldsymbol{\eta}(\boldsymbol{\beta})) = \boldsymbol{o}$ which gives here

$$\begin{bmatrix} \boldsymbol{\iota}^\top \\ \boldsymbol{x}^\top \\ \boldsymbol{z}^{\top\delta} \\ \gamma\log(\boldsymbol{z}^\top) * \boldsymbol{z}^{\top\delta} \end{bmatrix} (\boldsymbol{\mathfrak{y}} - \boldsymbol{\iota}\alpha - \boldsymbol{x}\beta - \boldsymbol{z}^\delta\gamma) = \boldsymbol{o} \tag{56.0.51}$$

or, element by element,

$$\sum_t (\mathfrak{y}_t - \alpha - \beta x_t - \gamma z_t^\delta) = 0 \tag{56.0.52}$$

$$\sum_t x_t(\mathfrak{y}_t - \alpha - \beta x_t - \gamma z_t^\delta) = 0 \tag{56.0.53}$$

$$\sum_t z_t^\delta(\mathfrak{y}_t - \alpha - \beta x_t - \gamma z_t^\delta) = 0 \tag{56.0.54}$$

$$\gamma\sum_t \log(z_t)z_t^\delta(\mathfrak{y}_t - \alpha - x_t\beta - z_t^\delta\gamma) = 0 \tag{56.0.55}$$

which is very similar to [**Gre97**, Example 10.1 on p. 451]. These element-by-element first order conditions can also be easily derived as the partial derivatives of $\mathsf{SSE} = \sum_t(\mathfrak{y}_t - \alpha - \beta x_t - \gamma z_t^\delta)^2$.

The Gauss-Newton regression (56.0.46) is the regression of the residuals on the columns of the Jacobian.

$$\mathfrak{y}_t - \hat\alpha - \hat\beta x_t - \hat\gamma z_t^{\hat\delta} = a + bx_t + cz_t^{\hat\delta} + d\hat\gamma\log(z_t)z_t^{\hat\delta} + \text{error term} \tag{56.0.56}$$

and from this one gets an updated estimate as

$$\begin{bmatrix} \hat\alpha + \hat a \\ \hat\beta + \hat b \\ \hat\gamma + \hat c \\ \hat\delta + \hat d \end{bmatrix}. \tag{56.0.57}$$

One can also write it in the form (56.0.47), which in the present model happens to be even a little simpler (because the original regression is almost linear) and gives the true regression coefficient:

$$\mathfrak{y}_t + \hat\gamma\hat\delta\log(z_t)z_t^{\hat\delta} = a + bx_t + cz_t^{\hat\delta} + d\hat\gamma\log(z_t)z_t^{\hat\delta} + \text{error term} \tag{56.0.58}$$

□

• b. *How would you obtain the starting value for the Newton-Raphson algorithm?*

ANSWER. One possible set of starting values would be to set $\hat\delta = 1$ and to get $\hat\alpha$, $\hat\beta$, and $\hat\gamma$ from the linear regression. □

The Gauss-Newton algorithm runs this regression and uses the OLS estimate $\hat{\boldsymbol{\delta}}_i$ of $\boldsymbol{\delta}_i$ to define $\boldsymbol{\beta}_{i+1} = \boldsymbol{\beta}_i + \hat{\boldsymbol{\delta}}_i$. The recursion formula is therefore

$$\boldsymbol{\beta}_{i+1} = \boldsymbol{\beta}_i + \hat{\boldsymbol{\delta}}_i = \boldsymbol{\beta}_i + ((\boldsymbol{X}(\boldsymbol{\beta}_i))^\top\boldsymbol{X}(\boldsymbol{\beta}_i))^{-1}(\boldsymbol{X}(\boldsymbol{\beta}_i))^\top(\boldsymbol{\mathfrak{y}} - \boldsymbol{\eta}(\boldsymbol{\beta}_i)). \tag{56.0.59}$$

The notation $(\boldsymbol{\eta}(\boldsymbol{\beta}))^\top = \boldsymbol{\eta}^\top(\boldsymbol{\beta})$ and $(\boldsymbol{X}(\boldsymbol{\beta}))^\top = \boldsymbol{X}^\top(\boldsymbol{\beta})$ makes this perhaps a little easier to read:

$$\boldsymbol{\beta}_{i+1} = \boldsymbol{\beta}_i + (\boldsymbol{X}^\top(\boldsymbol{\beta}_i)\boldsymbol{X}(\boldsymbol{\beta}_i))^{-1}\boldsymbol{X}^\top(\boldsymbol{\beta}_i)(\boldsymbol{\mathfrak{y}} - \boldsymbol{\eta}(\boldsymbol{\beta}_i)). \tag{56.0.60}$$

This is [**Gre97**, last equation on p. 455].

A look at (56.0.44) shows that (56.0.60) is again a special case of the general principle (55.0.12), i.e., $\boldsymbol{\beta}_{i+1} = \boldsymbol{\beta}_i - \alpha_i\boldsymbol{R}_i\boldsymbol{g}_i$, with $\boldsymbol{R}_i = \left(\boldsymbol{X}^\top(\boldsymbol{\beta}_i)\boldsymbol{X}(\boldsymbol{\beta}_i)\right)^{-1}$ and $\alpha_i = 1/2$.

About the bad convergence properties of Gauss-Newton see [**Thi88**, p. 213–215].

Although the Gauss-Newton regression had initially been introduced as a numerical tool, it was soon realized that this regression is very important. Read [**DM93**, Chapter 6] about this.

If one runs the GNR at the minimum argument of the nonlinear least squares objective function, then the coefficients estimated by the GNR are zero, i.e., the adjustments to the minimum argument $\hat{\boldsymbol{\delta}} = \boldsymbol{\beta}_{i+1} - \boldsymbol{\beta}_i$ are zero.

How can a regression be useful whose outcome we know beforehand? Several points: If the estimated parameters turn out not to be zero after all, then $\boldsymbol{\beta}^*$ was not really a minimum. I.e., the GNR serves as a check of the minimization procedure which one has used. One can also use regression diagnostics on the GNR in order to identify influential observations. The covariance matrix produced by the regression printout is an asymptotic covariance matrix of the NLS estimator. One can check for collinearity. If $\boldsymbol{\beta}^*$ is a restricted NLS estimate, then the GNR yields Lagrange

multiplier tests for the restriction, or tests whether more variables should be added, or specification tests.

Properties of NLS estimators: If $\boldsymbol{X}_0$ is the matrix of pseudoregressors computed at the true parameter values, one needs the condition $\operatorname{plim} \frac{1}{n}\boldsymbol{X}_0^\top \boldsymbol{X}_0 = \boldsymbol{Q}_0$ exists and is positive definite. For consistency we need $\operatorname{plim} \frac{1}{n}\boldsymbol{X}_0^\top \boldsymbol{\varepsilon} = 0$, and for asymptotic normality $\frac{1}{\sqrt{n}}\boldsymbol{X}_0^\top \boldsymbol{\varepsilon} \to \mathsf{N}(\boldsymbol{o}, \sigma^2 \boldsymbol{Q}_0)$. $\hat{\sigma}^2 = \mathsf{SSE}(\hat{\boldsymbol{\beta}})/n$ is a consistent estimate of σ^2 (a degree of freedom correction, i.e., dividing by $n-k$ instead of n, has no virtue here since the results are valid only asymptotically). Furthermore, $\hat{\sigma}^2\big(\boldsymbol{X}^\top(\boldsymbol{\beta})\boldsymbol{X}(\boldsymbol{\beta})\big)^{-1}$ is a consistent estimate of the asymptotic covariance matrix.

56.1. The J Test

Start out with two non-nested hypotheses about the data:

$$H_0 : \mathbf{y} = \boldsymbol{\eta}_0(\boldsymbol{\beta}) + \boldsymbol{\varepsilon}_0 \tag{56.1.1}$$

$$H_1 : \mathbf{y} = \boldsymbol{\eta}_1(\boldsymbol{\gamma}) + \boldsymbol{\varepsilon}_1 \tag{56.1.2}$$

A model which has these two hypotheses artificially nested is:

$$\mathbf{y} = (1-\alpha)\boldsymbol{\eta}_0(\boldsymbol{\beta}) + \alpha\boldsymbol{\eta}_1(\boldsymbol{\gamma}) + \boldsymbol{\varepsilon} \tag{56.1.3}$$

The problem here is that often it is not possible to estimate α, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$ together. For instance, in the linear case

$$\mathbf{y} = (1-\alpha)\boldsymbol{X}_0\boldsymbol{\beta} + \alpha\boldsymbol{X}_1\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \tag{56.1.4}$$

every change in α can be undone by counteracting changes in $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. Therefore the idea is to estimate $\boldsymbol{\gamma}$ from model 1, call this estimate $\hat{\hat{\boldsymbol{\gamma}}}$, and get the predicted value of $\mathbf{y}$ from model 1 $\hat{\hat{\mathbf{y}}}_1 = \boldsymbol{\eta}_1(\hat{\hat{\boldsymbol{\gamma}}})$, and plug this into this model, i.e., one estimates α and $\boldsymbol{\beta}$ in the model

$$\mathbf{y} = (1-\alpha)\boldsymbol{\eta}_0(\boldsymbol{\beta}) + \alpha\hat{\hat{\mathbf{y}}}_1 + \boldsymbol{\varepsilon} \tag{56.1.5}$$

This is called the J test. A mathematical simplification, called the P-test, would be to get an estimate $\hat{\hat{\boldsymbol{\beta}}}$ of $\boldsymbol{\beta}$ from the first model, and use the linearized version of $\boldsymbol{\eta}_0$ around $\hat{\hat{\boldsymbol{\beta}}}$, i.e., replace $\boldsymbol{\eta}_0$ in the above regression by

$$\boldsymbol{\eta}_0(\hat{\hat{\boldsymbol{\beta}}}) + \boldsymbol{X}_0(\hat{\hat{\boldsymbol{\beta}}})(\boldsymbol{\beta} - \hat{\hat{\boldsymbol{\beta}}}). \tag{56.1.6}$$

If one does this, one gets the linear regression

$$\mathbf{y} - \hat{\hat{\mathbf{y}}}_0 = \boldsymbol{X}\boldsymbol{\delta} + \alpha(\hat{\hat{\mathbf{y}}}_1 - \hat{\hat{\mathbf{y}}}_0) \tag{56.1.7}$$

where $\hat{\hat{\mathbf{y}}}_0 = \boldsymbol{\eta}_0(\hat{\hat{\boldsymbol{\beta}}})$, and $\boldsymbol{\delta} = (1-\alpha)(\boldsymbol{\beta} - \hat{\hat{\boldsymbol{\beta}}})$, and one simply has to test for $\alpha = 0$.

Problem 496. *Computer Assignment: The data in table 10.1 are in the file* `/home/econ/ehrbar/ec781/consum.txt` *and they will also be sent to you by e-mail. Here are the commands to enter them into SAS:*

```
    libname ec781 '/home/econ/ehrbar/ec781/sasdata';
   filename consum '/home/econ/ehrbar/ec781/consum.txt';
data ec781.consum;
infile consum;
input year y c;
run;
```

Use them to re-do the estimation of the consumption function in Table 10.2. In SAS this can be done with the procedure NLIN, described in the SAS Users Guide Statistics, [**SAS85**]. *Make a scatter plot of the data and plot the fitted curve into this same plot.*

```
    libname ec781 '/home/econ/ehrbar/ec781/sasdata';

  proc nlin data=ec781.consum
      maxiter=200;
      parms a1=11.1458
          b1=0.89853
          g1=1;

      model c=a1+b1*exp(g1 * log(y));
      der.a1=1;
      der.b1=exp(g1 * log(y));
      der.g1=b1*(exp(g1*log(y)))*log(y);

      run;
```

56.2. Nonlinear instrumental variables estimation

If instrumental variables are necessary, one minimizes, instead of $(\mathbf{y}-\boldsymbol{\eta}(\boldsymbol{\beta}))^\top(\mathbf{y}-\boldsymbol{\eta}(\boldsymbol{\beta}))$, the following objective function:

$$(\mathbf{y} - \boldsymbol{\eta}(\boldsymbol{\beta}))^\top \boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top(\mathbf{y} - \boldsymbol{\eta}(\boldsymbol{\beta})) \tag{56.2.1}$$

(As before, $\boldsymbol{\eta}$ contains $\boldsymbol{X}$ although we are not making this explicit.)

Example: If one uses instrumental variables on the consumption function, one gets this time quite different estimates than from the nonlinear least squares.

If one transforms the dependent variable as well, [**Gre97**, p. 473] recommends maximum likelihood estimation, and Problem 497 shows why it is sometimes necessary. [**CR88**, p. 21–23] compare ML and NLS estimation. On paper, ML is more efficient, but if there are only slight deviations from normality, not visible to the eye, then NLS may be more efficient. NLS is more robust in the case of mis-specification. Some statisticians also believe that even under ideal circumstances the MLE attains its asymptotic properties more slowly than NLS.

PROBLEM 497. *This is* [**DM93**, pp. 243 and 284]. *The model is*

$$\mathsf{y}_t^\gamma = \alpha + \beta x_t + \varepsilon_t \tag{56.2.2}$$

with $\boldsymbol{\varepsilon} \sim \mathsf{N}(\boldsymbol{o}, \sigma^2 \boldsymbol{I})$. $\mathsf{y}_i > 0$.

• a. *1 point Why can this model not be estimated with nonlinear least squares?*

ANSWER. If all the y's are greater than unity, then the SSE can be made arbitrarily small by letting γ tend to $-\infty$ and setting α and β zero. $\gamma = 1$, $\alpha = 1$, and $\beta = 0$ leads to a zero SSE as well. The idea behind LS is to fit the curve to the data. If γ changes, the data points themselves move. We already saw when we discussed the R^2 that there is no good way to compare SSE's for different $\mathbf{y}$'s. (How about the information matrix: is it block-diagonal? Are the Kmenta-Oberhofer conditions applicable?) □

• b. *3 points Show that the log likelihood function is*

$$-\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=1}^{n}\left(\mathsf{y}_t^\gamma - \alpha - \beta x_t\right)^2 + n\log|\gamma| + (\gamma-1)\sum_{t=1}^{n}\log(\mathsf{y}_t) \tag{56.2.3}$$

ANSWER. This requires the transformation theorem for densities. $\varepsilon_t = \mathsf{y}_t^\gamma - \alpha - \beta x_t$; therefore $\partial\varepsilon_t/\partial\mathsf{y}_t = \gamma\mathsf{y}_t^{\gamma-1}$ and $\partial\varepsilon_t/\partial\mathsf{y}_s = 0$ for $s \neq t$. The Jacobian has this in the diagonal and 0 in the off-diagonal, therefore the determinant is $J = \gamma^n(\prod \mathsf{y}_t)^{\gamma-1}$ and $|J| = |\gamma|^n(\prod \mathsf{y}_t)^{\gamma-1}$. This gives the above formula: which I assume is right, it is from [**DM93**], but somewhere [**DM93**] has a typo. □

• c. *2 points Concentrate the log likelihood function with respect to* σ^2. *(Write the precise value of the constant.)*

ANSWER. $\frac{\partial\log\ell}{\partial\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{t=1}^{n}\left(\mathsf{y}_t^\gamma - \alpha - \beta x_t\right)^2$. This gives the usual estimate $\tilde{\sigma}^2 = \frac{1}{n}\sum_{t=1}^{n}\left(\mathsf{y}_t^\gamma - \alpha - \beta x_t\right)^2$. If one plugs this into the log likelihood function and extracts the constants and those parts which depend on n, one gets the following:

$$\frac{n}{2}(\ln n - \ln 2\pi - 1) - \frac{n}{2}\log\sum_{t=1}^{n}\left(\mathsf{y}_t^\gamma - \alpha - \beta x_t\right)^2 + n\log|\gamma| + (\gamma-1)\sum_{t=1}^{n}\log(\mathsf{y}_t) \tag{56.2.4}$$

In [**DM93**], the constant is not given explicitly; in this way I can check if they have understood it. □

CHAPTER 57

Applications of GLS with Nonspherical Covariance Matrix

In most cases in which the covariance matrix is nonspherical, $\boldsymbol{\Psi}$ contains unknown parameters, which must be estimated before formula (26.0.2) can be applied. Of course, if *all* entries of $\boldsymbol{\Psi}$ are unknown, such estimation is impossible, since one needs $n(n+1)/2-1$ parameters to specify a symmetric matrix up to a multiplicative factor, but with n observations only n unrelated parameters can be estimated consistently. Only in a few exceptional cases, $\boldsymbol{\Psi}$ is known, and in some even more exceptional cases, there are unknown parameters in $\boldsymbol{\Psi}$ but (26.0.2) does not depend on them. We will discuss such examples first: heteroskedastic disturbances with known relative variances, and some examples involving equicorrelated disturbances.

57.1. Cases when OLS and GLS are identical

Problem 498. *From $\mathbf{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim (\boldsymbol{o}, \sigma^2\boldsymbol{I})$ follows $\boldsymbol{P}\mathbf{y} = \boldsymbol{PX\beta} + \boldsymbol{P\varepsilon}$ with $\boldsymbol{P\varepsilon} \sim (\boldsymbol{o}, \sigma^2\boldsymbol{PP}^\top)$. Which conditions must $\boldsymbol{P}$ satisfy so that the generalized least squares regression of $\boldsymbol{P}\mathbf{y}$ on $\boldsymbol{PX}$ with covariance matrix $\boldsymbol{PP}^\top$ gives the same result as the original regression?*

Problem 499. *We are in the model $\mathbf{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \sigma^2\boldsymbol{\Psi}$. As always, we assume $\boldsymbol{X}$ has full column rank, and $\boldsymbol{\Psi}$ is nonsingular. We will discuss the special situation here in which $\boldsymbol{X}$ and $\boldsymbol{\Psi}$ are such that $\boldsymbol{\Psi X} = \boldsymbol{XA}$ for some $\boldsymbol{A}$.*

• a. *3 points Show that the requirement $\boldsymbol{\Psi X} = \boldsymbol{XA}$ is equivalent to the requirement that $\mathrm{R}[\boldsymbol{\Psi X}] = \mathrm{R}[\boldsymbol{X}]$. Here $\mathrm{R}[\boldsymbol{B}]$ is the range space of a matrix $\boldsymbol{B}$, i.e., it is the vector space consisting of all vectors that can be written in the form $\boldsymbol{Bc}$ for some $\boldsymbol{c}$. Hint: For $\Rightarrow$ show first that $\mathrm{R}[\boldsymbol{\Psi X}] \subset \mathrm{R}[\boldsymbol{X}]$, and then show that $\mathrm{R}[\boldsymbol{\Psi X}]$ has the same dimension as $\mathrm{R}[\boldsymbol{X}]$.*

Answer. $\Rightarrow$: Clearly $\mathrm{R}[\boldsymbol{\Psi X}] \subset \mathrm{R}[\boldsymbol{X}]$ since $\boldsymbol{\Psi X} = \boldsymbol{XA}$ and every $\boldsymbol{XAc}$ has the form $\boldsymbol{Xd}$ with $\boldsymbol{d} = \boldsymbol{Ac}$. And since $\boldsymbol{\Psi}$ is nonsingular, and the range space is the space spanned by the column vectors, and the columns of $\boldsymbol{\Psi X}$ are the columns of $\boldsymbol{X}$ premultiplied by $\boldsymbol{\Psi}$, it follows that the range space of $\boldsymbol{\Psi X}$ has the same dimension as that of $\boldsymbol{X}$. $\Leftarrow$: The ith column of $\boldsymbol{\Psi X}$ lies in $\mathrm{R}[\boldsymbol{X}]$, i.e., it can be written in the form $\boldsymbol{Xa}_i$ for some $\boldsymbol{a}_i$. $\boldsymbol{A}$ is the matrix whose columns are all the $\boldsymbol{a}_i$. □

• b. *2 points Show that $\boldsymbol{A}$ is nonsingular.*

Answer. $\boldsymbol{A}$ is square, since $\boldsymbol{XA} = \boldsymbol{\Psi X}$, i.e., $\boldsymbol{XA}$ has has as many columns as $\boldsymbol{X}$. Now assume $\boldsymbol{Ac} = \boldsymbol{o}$. Then $\boldsymbol{XAc} = \boldsymbol{o}$ or $\boldsymbol{\Psi Xc} = \boldsymbol{o}$, and since $\boldsymbol{\Psi}$ is nonsingular this gives $\boldsymbol{Xc} = \boldsymbol{o}$, and since $\boldsymbol{X}$ has full column rank, this gives $\boldsymbol{c} = \boldsymbol{o}$. □

• c. *2 points Show that $\boldsymbol{XA}^{-1} = \boldsymbol{\Psi}^{-1}\boldsymbol{X}$.*

Answer. $\boldsymbol{X} = \boldsymbol{\Psi}^{-1}\boldsymbol{\Psi X} = \boldsymbol{\Psi}^{-1}\boldsymbol{XA}$, and now postmultiply by $\boldsymbol{A}^{-1}$. □

• d. *2 points Show that in this case $(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$, i.e., the OLS is BLUE ("Kruskal's theorem").*

Answer. $(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1} = \left((\boldsymbol{A}^{-1})^\top\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}(\boldsymbol{A}^{-1})^\top\boldsymbol{X}^\top = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{A}^\top(\boldsymbol{A}^{-1})^\top\boldsymbol{X}^\top = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$ □

57.2. Heteroskedastic Disturbances

Heteroskedasticity means: error terms are independent, but their variances are not equal. $\boldsymbol{\Psi}$ is diagonal, with positive diagonal elements. In a few rare cases the relative variances are known. The main example is that the observations are means of samples from a homoskedastic population with varying but known sizes.

This is a plausible example of a situation in which the relative variances are known to be proportional to an observed (positive) nonrandom variable $\boldsymbol{z}$ (which may or may not be one of the explanatory variables in the regression). Here $\mathcal{V}[\boldsymbol{\varepsilon}] = \sigma^2\boldsymbol{\Psi}$ with a known diagonal

(57.2.1)
$$\boldsymbol{\Psi} = \begin{bmatrix} z_1 & 0 & \cdots & 0 \\ 0 & z_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & z_n \end{bmatrix}. \qquad \text{Therefore} \quad \boldsymbol{P} = \begin{bmatrix} 1/\sqrt{z_1} & 0 & \cdots & 0 \\ 0 & 1/\sqrt{z_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/\sqrt{z_n} \end{bmatrix},$$

i.e., one divides every observation by the appropriate factor so that after the division the standard deviations are equal. Note: this means that this transformed regression usually no longer has a constant term, and therefore also R^2 loses its meaning.

Problem 500. *3 points The specification is*

$$\mathrm{y}_t = \beta_1 + \beta_2 x_t + \beta_3 x_t^2 + \varepsilon_t, \tag{57.2.2}$$

with $\mathrm{E}[\varepsilon_t] = 0$, $\mathrm{var}[\varepsilon_t] = \sigma^2 x_t^2$ for some unknown $\sigma^2 > 0$, and the errors are uncorrelated. Someone runs the OLS regression

$$\frac{\mathrm{y}_t}{x_t} = \gamma_1 + \gamma_2\frac{1}{x_t} + \gamma_3 x_t + v_t \tag{57.2.3}$$

and you have the estimates $\hat\gamma_1$, $\hat\gamma_2$, and $\hat\gamma_3$ from this regression. Compute estimates of β_1, β_2, and β_3 using the $\hat\gamma_i$. What properties do your estimates of the β_i have?

Answer. Divide the original specification by x_t to get

$$\frac{\mathsf{y}_t}{x_t} = \beta_2 + \beta_1 \frac{1}{x_t} + \beta_3 x_t + \frac{\varepsilon_t}{x_t}. \tag{57.2.4}$$

Therefore $\hat{\gamma}_2$ is the BLUE of β_1, $\hat{\gamma}_1$ that of β_2, and $\hat{\gamma}_3$ that of β_3. Note that the constant terms of the old and new regression switch places! □

Now let us look at a random parameter model $\mathsf{y}_t = x_t\gamma_t$, or in vector notation, using $*$ for element-by-element multiplication of two vectors, $\boldsymbol{\mathsf{y}} = \boldsymbol{x} * \boldsymbol{\gamma}$. Here $\gamma_t \sim \text{IID}(\beta, \sigma^2)$, one can also write it $\gamma_t = \beta + \delta_t$ or $\boldsymbol{\gamma} = \boldsymbol{\iota}\beta + \boldsymbol{\delta}$ with $\boldsymbol{\delta} \sim (\boldsymbol{o}, \sigma^2\boldsymbol{I})$.

This model can be converted into a heteroskedastic Least Squares model if one defines $\boldsymbol{\varepsilon} = \boldsymbol{x} * \boldsymbol{\delta}$. Then $\boldsymbol{\mathsf{y}} = \boldsymbol{x}\beta + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim (\boldsymbol{o}, \sigma^2\boldsymbol{\Psi})$ where

$$\boldsymbol{\Psi} = \begin{bmatrix} x_1^2 & 0 & \cdots & 0 \\ 0 & x_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & x_n^2 \end{bmatrix}. \tag{57.2.5}$$

Since $\boldsymbol{x}^\top\boldsymbol{\Psi}^{-1} = \boldsymbol{x}^{-1\top}$ (taking the inverse element by element), and therefore $\boldsymbol{x}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{x} = n$, one gets $\hat{\beta} = \frac{1}{n}\sum\frac{\mathsf{y}_t}{x_t}$ and $\text{var}[\hat{\beta}] = \sigma^2/n$. On the other hand, $\boldsymbol{x}^\top\boldsymbol{\Psi}\boldsymbol{x} = \sum x^4$, therefore $\text{var}[\hat{\beta}_{OLS}] = \sigma^2\frac{\sum x^4}{(\sum x^2)^2}$. Assuming that the x_t are independent drawings of a random variable x with zero mean and finite fourth moments, it follows

$$\text{plim}\,\frac{\text{var}[\hat{\boldsymbol{\beta}}_{OLS}]}{\text{var}[\hat{\boldsymbol{\beta}}]} = \text{plim}\,\frac{n\sum x^4}{(\sum x^2)^2} = \frac{\text{plim}\,\frac{1}{n}\sum x^4}{(\text{plim}\,\frac{1}{n}\sum x^2)^2} = \frac{\text{E}[\mathsf{x}^4]}{(\text{E}[\mathsf{x}^2])^2} \tag{57.2.6}$$

This is the kurtosis (without subtracting the 3). Theoretically it can be anything ≥ 1, the Normal distribution has kurtosis 3, and the economics time series usually have a kurtosis between 2 and 4.

57.3. Equicorrelated Covariance Matrix

Problem 501. *Assume $\mathsf{y}_i = \mu + \varepsilon_i$, where μ is nonrandom, $\text{E}[\varepsilon_i] = 0$, $\text{var}[\varepsilon_i] = \sigma^2$, and $\text{cov}[\varepsilon_i, \varepsilon_j] = \rho\sigma^2$ for $i \neq j$ (i.e., the ε_i are equicorrelated).*

$$\mathcal{V}[\boldsymbol{\varepsilon}] = \sigma^2 \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}. \tag{57.3.1}$$

If $\rho \geq 0$, then these error terms could have been obtained as follows: $\boldsymbol{\varepsilon} = \boldsymbol{z} + \boldsymbol{\iota}\mathsf{u}$ where $\boldsymbol{z} \sim (\boldsymbol{o}, \tau^2\boldsymbol{I})$ and $\mathsf{u} \sim (0, \omega^2)$ independent of $\boldsymbol{z}$.

• a. *1 point Show that the covariance matrix of $\boldsymbol{\varepsilon}$ is $\mathcal{V}[\boldsymbol{\varepsilon}] = \tau^2\boldsymbol{I} + \omega^2\boldsymbol{\iota}\boldsymbol{\iota}^\top$.*

Answer. $\mathcal{V}[\boldsymbol{\iota}\mathsf{u}] = \boldsymbol{\iota}\,\text{var}[\mathsf{u}]\boldsymbol{\iota}^\top$, add this to $\mathcal{V}[\boldsymbol{z}]$. □

• b. *1 point What are the values of τ^2 and ω^2 so that $\boldsymbol{\varepsilon}$ has the above covariance structure?*

Answer. To write it in the desired form, the following identities must hold: for the off-diagonal elements $\sigma^2\rho = \omega^2$, which gives the desired formula for ω^2 and for the diagonal elements $\sigma^2 = \tau^2 + \omega^2$. Solving this for τ^2 and plugging in the formula for ω^2 gives $\tau^2 = \sigma^2 - \omega^2 = \sigma^2(1-\rho)$, □

• c. *3 points Using matrix identity* (A.8.20) *(for ordinary inverses, not for g-inverses) show that the generalized least squares formula for the BLUE in this model is equivalent to the ordinary least squares formula. In other words, show that the sample mean $\bar{\mathsf{y}}$ is the BLUE of μ.*

Answer. Setting $\gamma = \tau^2/\sigma^2$, we want to show that

$$\left(\boldsymbol{\iota}^\top(\boldsymbol{I} + \frac{\boldsymbol{\iota}\boldsymbol{\iota}^\top}{\gamma})^{-1}\boldsymbol{\iota}\right)^{-1}\boldsymbol{\iota}^\top(\boldsymbol{I} + \frac{\boldsymbol{\iota}\boldsymbol{\iota}^\top}{\gamma})^{-1}\boldsymbol{\mathsf{y}} = \left(\boldsymbol{\iota}^\top\boldsymbol{I}^{-1}\boldsymbol{\iota}\right)^{-1}\boldsymbol{\iota}^\top\boldsymbol{I}^{-1}\boldsymbol{\mathsf{y}}. \tag{57.3.2}$$

This is even true for arbitrary $\boldsymbol{h}$ and $\boldsymbol{A}$:

$$\boldsymbol{h}^\top(\boldsymbol{A} + \frac{\boldsymbol{h}\boldsymbol{h}^\top}{\gamma})^{-1} = \boldsymbol{h}^\top\boldsymbol{A}^{-1}\frac{\gamma}{\gamma + \boldsymbol{h}^\top\boldsymbol{A}^{-1}\boldsymbol{h}}; \tag{57.3.3}$$

$$\left(\boldsymbol{h}^\top(\boldsymbol{A} + \frac{\boldsymbol{h}\boldsymbol{h}^\top}{\gamma})^{-1}\boldsymbol{h}\right)^{-1} = \frac{\gamma + \boldsymbol{h}^\top\boldsymbol{A}^{-1}\boldsymbol{h}}{\gamma\boldsymbol{h}^\top\boldsymbol{A}^{-1}\boldsymbol{h}} = \frac{1}{\boldsymbol{h}^\top\boldsymbol{A}^{-1}\boldsymbol{h}} + \frac{1}{\gamma}; \tag{57.3.4}$$

Now multiply the left sides and the righthand sides (use middle term in (57.3.4))

$$\left(\boldsymbol{h}^\top(\boldsymbol{A} + \frac{\boldsymbol{h}\boldsymbol{h}^\top}{\gamma})^{-1}\boldsymbol{h}\right)^{-1}\boldsymbol{h}^\top(\boldsymbol{A} + \frac{\boldsymbol{h}\boldsymbol{h}^\top}{\gamma})^{-1} = \left(\boldsymbol{h}^\top\boldsymbol{A}^{-1}\boldsymbol{h}\right)^{-1}\boldsymbol{h}^\top\boldsymbol{A}^{-1}. \tag{57.3.5}$$

□

• d. *3 points* [**Gre97**, Example 11.1 on pp. 499/500]*: Show that* $\text{var}[\bar{\mathsf{y}}]$ *does not converge to zero as $n \to \infty$ while ρ remains constant.*

Answer. By (57.3.4),

$$\text{var}[\bar{\mathsf{y}}] = \tau^2(\frac{1}{n} + \frac{1}{\gamma}) = \sigma^2(\frac{1-\rho}{n} + \rho) = \frac{\tau^2}{n} + \omega^2 \tag{57.3.6}$$

As $n \to \infty$ this converges towards ω^2, not to 0. □

Problem 502. [**Chr87**, pp. 361–363] *Assume there are 1000 families in a certain town, and denote the income of family k by z_k. Let $\mu = \frac{1}{1000}\sum_{k=1}^{1000} z_k$ be the* population average *of all 1000 incomes in this finite population, and let $\sigma^2 = \frac{1}{1000}\sum_{k=1}^{1000}(z_k - \mu)^2$ be the* population variance *of the incomes. For the purposes of this question, the z_k are nonrandom, therefore μ and σ^2 are nonrandom as well.*

You pick at random 20 families without replacement, ask them what their income is, and you want to compute the BLUE of μ on the basis of this random sample. Call the incomes in the sample $\mathsf{y}_1, \ldots, \mathsf{y}_{20}$. We are using the letters y_i instead of z_i for this sample, because y_1 is not necessarily z_1, i.e., the income of family 1, but it may be, e.g., z_{258}. The y_i are random. The process of taking the sample of y_i is represented by a 20×1000 matrix of random variables q_{ik} ($i = 1, \ldots, 20$, $k = 1, \ldots, 1000$) with: $\mathsf{q}_{ik} = 1$ if family k has been picked as ith family in the sample, and 0 otherwise. In other words, $\mathsf{y}_i = \sum_{k=1}^{1000} \mathsf{q}_{ik} z_k$ or $\mathsf{y} = \mathbf{Q}\boldsymbol{z}$.

• a. *Let $i \neq j$ and $k \neq l$. Is q_{ik} independent of q_{il}? Is q_{ik} independent of q_{jk}? Is q_{ik} independent of q_{jl}?*

Answer. q_{ik} is not independent of q_{il}: if $\mathsf{q}_{ik} = 1$, this means that family ik as been selected as the jth family in the sample. Since only one family can be selected as the ith family in the sample, this implies $\mathsf{q}_{il} = 0$ for all $l \neq k$. q_{ik} is dependent of q_{jk}, because sampling is without replacement: if family k has been selected as the ith family in the sample, then it cannot be selected again as the jth family of the sample. Is q_{ik} independent of q_{jl}? I think it is. □

• b. *Show that the first and second moments are*

(57.3.7)
$$\mathrm{E}[\mathsf{q}_{ik}] = 1/1000, \qquad \textit{and} \quad \mathrm{E}[\mathsf{q}_{ik}\mathsf{q}_{jl}] = \begin{cases} 1/1000 & \textit{if } i = j \textit{ and } k = l \\ 1/(1000 \cdot 999) & \textit{if } i \neq j \textit{ and } k \neq l \\ 0 & \textit{otherwise.} \end{cases}$$

For these formulas you need the rules how to take expected values of discrete random variables.

Answer. Since q_{ik} is a zero-one variable, $\mathrm{E}[\mathsf{q}_{ik}] = \Pr[\mathsf{q}_{ik} = 1] = 1/1000$. This is obvious if $i = 1$, and one can use a symmetry argument that it should not depend on i. And since for a zero-one variable, $\mathsf{q}_{ik}^2 = \mathsf{q}_{ik}$, it follows $\mathrm{E}[\mathsf{q}_{ik}^2] = 1/1000$ too. Now for $i \neq j, k \neq l$, $\mathrm{E}[\mathsf{q}_{ik}\mathsf{q}_{jl}] = \Pr[\mathsf{q}_{ik} = 1 \cap \mathsf{q}_{jl} = 1] = (1/1000)(1/999)$. Again this is obvious for $i = 1$ and $j = 2$, and can be extended by symmetry to arbitrary pairs $i \neq j$. For $i \neq j$, $\mathrm{E}[\mathsf{q}_{ik}\mathsf{q}_{jk}] = 0$ since z_k cannot be chosen twice, and for $k \neq l$, $\mathrm{E}[\mathsf{q}_{ik}\mathsf{q}_{il}] = 0$ since only one z_k can be chosen as the ith element in the sample. □

• c. *Since $\sum_{k=1}^{1000} \mathsf{q}_{ik} = 1$ for all i, one can write*

$$\mathsf{y}_i = \mu + \sum_{k=1}^{1000} \mathsf{q}_{ik}(z_k - \mu) = \mu + \varepsilon_i \tag{57.3.8}$$

where $\varepsilon_i = \sum_{k=1}^{1000} \mathsf{q}_{ik}(z_k - \mu)$. Show that

$$\mathrm{E}[\varepsilon_i] = 0 \qquad \mathrm{var}[\varepsilon_i] = \sigma^2 \qquad \mathrm{cov}[\varepsilon_i, \varepsilon_j] = -\sigma^2/999 \quad \textit{for } i \neq j \tag{57.3.9}$$

Hint: *For the covariance note that from $0 = \sum_{k=1}^{1000}(z_k - \mu)$ follows*

(57.3.10)
$$0 = \sum_{k=1}^{1000}(z_k - \mu)\sum_{l=1}^{1000}(z_l - \mu) = \sum_{k \neq l}(z_k - \mu)(z_l - \mu) + \sum_{k=1}^{1000}(z_k - \mu)^2 = \sum_{k \neq l}(z_k - \mu)(z_l - \mu) + 1000\sigma^2.$$

Answer.

$$\mathrm{E}[\varepsilon_i] = \sum_{k=1}^{1000}(z_k - \mu)\,\mathrm{E}[\mathsf{q}_{ik}] = \sum_{k=1}^{1000}\frac{z_k - \mu}{1000} = 0 \tag{57.3.11}$$

$$\mathrm{var}[\varepsilon_i] = \mathrm{E}[\varepsilon_i^2] = \sum_{k,l=1}^{1000}(z_k - \mu)(z_l - \mu)\,\mathrm{E}[\mathsf{q}_{ik}\mathsf{q}_{il}] = \sum_{k=1}^{1000}\frac{(z_k - \mu)^2}{1000} = \sigma^2 \tag{57.3.12}$$

and for $i \neq j$ follows, using the hint for the last equal-sign

(57.3.13)
$$\mathrm{cov}[\varepsilon_i, \varepsilon_j] = \mathrm{E}[\varepsilon_i\varepsilon_j] = \sum_{k,l=1}^{1000}(z_k - \mu)(z_l - \mu)\,\mathrm{E}[\mathsf{q}_{ik}\mathsf{q}_{jl}] = \sum_{k \neq l}\frac{(z_k - \mu)(z_l - \mu)}{1000 \cdot 999} = -\sigma^2/999.$$

□

With $\boldsymbol{\iota}_{20}$ being the 20×1 column vector consisting of ones, one can therefore write in matrix notation

$$\mathsf{y} = \boldsymbol{\iota}_{20}\mu + \boldsymbol{\varepsilon} \qquad \mathrm{E}[\boldsymbol{\varepsilon}] = \boldsymbol{o} \qquad \mathcal{V}[\boldsymbol{\varepsilon}] = \sigma^2\boldsymbol{\Psi}$$

where

$$\boldsymbol{\Psi} = \begin{bmatrix} 1 & -1/999 & \cdots & -1/999 \\ -1/999 & 1 & \cdots & -1/999 \\ \vdots & \vdots & \ddots & \vdots \\ -1/999 & -1/999 & \cdots & 1 \end{bmatrix} \tag{57.3.14}$$

From what we know about GLS with equicorrelated errors (question 501) follows therefore that the sample mean $\bar{\mathsf{y}}$ is the BLUE of μ. (This last part was an explanation of the relevance of the question, you are not required to prove it.)

CHAPTER 58

Unknown Parameters in the Covariance Matrix

If $\boldsymbol{\Psi}$ depends on certain unknown parameters which are not, at the same time, components of $\boldsymbol{\beta}$ or functions thereof, and if a consistent estimate of these parameters is available, then GLS with this estimated covariance matrix, called "feasible GLS," is usually asymptotically efficient. This is an important result: one does not not need an efficient estimate of the covariance matrix to get efficient estimates of $\boldsymbol{\beta}$! In this case, all the results are asymptotically valid, with $\hat{\boldsymbol{\Psi}}$ in the formulas instead of $\boldsymbol{\Psi}$. These estimates are sometimes even unbiased!

58.1. Heteroskedasticity

Heteroskedasticity means: error terms are independent, but do not have equal variances. There are not enough data to get consistent estimates of all error variances, therefore we need additional information.

The simplest kind of additional information is that the sample can be partitioned into two different subsets, each subset corresponding to a different error variance, with the relative variances known. Write the model as

$$\begin{bmatrix}\mathbf{y}_1\\ \mathbf{y}_2\end{bmatrix} = \begin{bmatrix}\boldsymbol{X}_1\\ \boldsymbol{X}_2\end{bmatrix}\boldsymbol{\beta} + \begin{bmatrix}\boldsymbol{\varepsilon}_1\\ \boldsymbol{\varepsilon}_2\end{bmatrix}; \qquad \mathcal{V}[\begin{bmatrix}\boldsymbol{\varepsilon}_1\\ \boldsymbol{\varepsilon}_2\end{bmatrix}] = \sigma^2\begin{bmatrix}\kappa_1^2\boldsymbol{I} & \boldsymbol{O}\\ \boldsymbol{O} & \kappa_2^2\boldsymbol{I}\end{bmatrix} = \boldsymbol{\Phi}. \tag{58.1.1}$$

Assume $\mathbf{y}_1$ has n_1 and $\mathbf{y}_2$ n_2 observations. The GLSE is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{\Phi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Phi}^{-1}\mathbf{y} = \Big(\frac{\boldsymbol{X}_1{}^\top\boldsymbol{X}_1}{\kappa_1^2} + \frac{\boldsymbol{X}_2{}^\top\boldsymbol{X}_2}{\kappa_2^2}\Big)^{-1}\Big(\frac{\boldsymbol{X}_1{}^\top\mathbf{y}_1}{\kappa_1^2} + \frac{\boldsymbol{X}_2{}^\top\mathbf{y}_2}{\kappa_2^2}\Big). \tag{58.1.2}$$

To make this formula operational, we have to replace the κ_i^2 by estimates. The simplest way (if each subset has at least $k+1$ observations) is to use the unbiased estimates s_i^2 $(i = 1, 2)$ from the OLS regressions on the two subsets separately. Associated with this estimation is also an easy test, the Goldfeld Quandt test [**Gre97**, 551/2]. simply use an F-test on the ratio $\mathrm{s}_2^2/\mathrm{s}_1^2$; but reject if it is too big or too small. If we don't have the lower significance points, check $\mathrm{s}_1^2/\mathrm{s}_2^2$ if it is > 1 and $\mathrm{s}_2^2/\mathrm{s}_1^2$ otherwise.

Problem 503. *3 points In the model*

$$\begin{bmatrix}\mathbf{y}_1\\ \mathbf{y}_2\end{bmatrix} = \begin{bmatrix}\boldsymbol{X}_1\\ \boldsymbol{X}_2\end{bmatrix}\boldsymbol{\beta} + \begin{bmatrix}\boldsymbol{\varepsilon}_1\\ \boldsymbol{\varepsilon}_2\end{bmatrix}; \qquad \mathcal{V}[\begin{bmatrix}\boldsymbol{\varepsilon}_1\\ \boldsymbol{\varepsilon}_2\end{bmatrix}] = \begin{bmatrix}\sigma_1^2\boldsymbol{I} & \boldsymbol{O}\\ \boldsymbol{O} & \sigma_2^2\boldsymbol{I}\end{bmatrix} \tag{58.1.3}$$

in which $\boldsymbol{X}_1$ is a 10×5 and $\boldsymbol{X}_2$ a 20×5 matrix, you run the two regressions separately and you get $\mathrm{s}_1^2 = 500$ and $\mathrm{s}_2^2 = 100$. Can you reject at the 5% significance level that these variances are equal? Can you reject it at the 1% level? The enclosed tables are from [**Sch59**, pp. 424–33].

Answer. The distribution of the ratio of estimated variances is $s_2^2/s_1^2 \sim F_{15,5}$, but since its observed value is smaller than 1, use instead $s_1^2/s_2^2 \sim F_{5,15}$. The upper significance points for 0.005% $F_{(5,15;0.005)} = 5.37$ (which gives a two-sided 1% significance level), for 1% it is $F_{(5.15;0.01)} = 4.56$ (which gives a two-sided 2% significance level), for 2.5% $F_{(5,15;0.025)} = 3.58$ (which gives a two-sided 5% significance level), and for 5% it is $F_{(5,15;0.05)} = 2.90$ (which gives a two-sided 10% significance level). A table can be found for instance in [**Sch59**, pp. 428/9]. To get the upper 2.5% point one can also use the `Splus`-command `qf(1-5/200,5,15)`. One can also get the lower significance points simply by the command `qf(5/200,5,15)`. The test is therefore significant at the 5% level but not significant at the 1% level. □

Since the so-called Kmenta-Oberhofer conditions are satisfied, i.e., since $\boldsymbol{\Psi}$ does not depend on $\boldsymbol{\beta}$, the following iterative procedure converges to the maximum likelihood estimator:

(1) start with some initial estimate of κ_1^2 and κ_2^2. [**Gre97**, p. 516] proposes to start with the assumption of homoskedasticity, i.e., $\kappa_1^2 = \kappa_2^2 = 1$, but if each group has enough observations to make separate estimates then I think a better starting point would be the s_i^2 of the separate regressions.

(2) Use those κ_i^2 to get the feasible GLSE.

(3) use this feasible GLSE to get a new set $\kappa_i^2 = \mathrm{s}_i^2$ (but divide by n_i, not $n_i - k$).

(4) Go back to (2).

Once the maximum likelihood estimates of $\boldsymbol{\beta}$, σ^2, and κ_i^2 are computed (actually σ^2 and κ_i^2 cannot be identified separately, therefore one conventionally imposes a condition like $\sigma^2 = 1$ or $\sum_i \kappa_i^2 = n$ to identify them), then it is easy to test for homoskedasticity by the LR test. In order to get the maximum value of the likelihood function it saves us some work to start with the concentrated likelihood functions, therefore we start with (35.0.17):

$$\log f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\beta},\boldsymbol{\Psi}) = -\frac{n}{2}(1+\ln 2\pi - \ln n) - \frac{n}{2}\ln(\mathbf{y}-\boldsymbol{X}\boldsymbol{\beta})^\top\boldsymbol{\Psi}^{-1}(\mathbf{y}-\boldsymbol{X}\boldsymbol{\beta}) - \frac{1}{2}\ln\det[\boldsymbol{\Psi}] \tag{58.1.4}$$

Since $\hat{\sigma}^2 = \frac{1}{n}(\mathbf{y}-\boldsymbol{X}\boldsymbol{\beta})^\top\boldsymbol{\Psi}^{-1}(\mathbf{y}-\boldsymbol{X}\boldsymbol{\beta})$ and $\det[k\boldsymbol{\Psi}] = k^n\det[\boldsymbol{\Psi}]$ one can rewrite (35.0.17) as

$$\log f_{\mathbf{y}}(\mathbf{y};\boldsymbol{\beta},\boldsymbol{\Psi}) = -\frac{n}{2}(1+\ln 2\pi) - \frac{1}{2}\ln\det[\hat{\sigma}^2\boldsymbol{\Psi}] \tag{58.1.5}$$

Now in the constrained case, with homoskedasticity assumed, $\boldsymbol{\Psi} = \boldsymbol{I}$ and we will write the OLS estimator as $\hat{\hat{\boldsymbol{\beta}}}$ and $\hat{\hat{\sigma}}^2 = (\hat{\hat{\boldsymbol{\varepsilon}}}^\top \hat{\hat{\boldsymbol{\varepsilon}}})/n$. Then $\ln\det[\hat{\hat{\sigma}}^2 \boldsymbol{I}] = n\ln[\hat{\hat{\sigma}}^2]$. Let $\hat{\boldsymbol{\beta}}$ be the unconstrained MLE, and

$$\hat{\boldsymbol{\Psi}} = \begin{bmatrix} \hat{\sigma}_1^2 \boldsymbol{I} & \boldsymbol{O} \\ \boldsymbol{O} & \hat{\sigma}_2^2 \boldsymbol{I} \end{bmatrix} \tag{58.1.6}$$

there $\hat{\sigma}_i^2 = \hat{\boldsymbol{\varepsilon}}_i^\top \hat{\boldsymbol{\varepsilon}}_i / n_i$. The LR statistic is therefore (compare [**Gre97**, p. 516])

$$\lambda = 2(\log f_{\text{constrained}} - \log f_{\text{unconstrained}}) = n\ln\hat{\hat{\sigma}}^2 - \sum n_i \ln\hat{\sigma}_i^2 \tag{58.1.7}$$

In this particular case, the feasible GLSE is so simple that its finite sample properties are known. Therefore [**JHG+88**] use it as a showcase example to study the question: Should one use the feasible GLSE always, or should one use a pre-test estimator, i.e., test whether the variances are equal, and use the feasible GLS only if this test can be rejected, otherwise use OLS? [**JHG+88**, figure 9.2 on p. 364] gives the trace of the MSE-matrix for several possibilities.

58.1.1. Logarithm of Error Variances Proportional to Unknown Linear Combination of Explanatory Variables. When we discussed heteroskedasticity with known relative variances, the main example was the prior knowledge that the error variances were proportional to some observed $\boldsymbol{z}$. To generalize this procedure, [**Har76**] proposes the following specification:

$$\ln\sigma_t^2 = \boldsymbol{z}_t^\top \boldsymbol{\alpha}, \tag{58.1.8}$$

where $\boldsymbol{\alpha}$ is a vector of unknown nonrandom parameters, and $\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{z}_1^\top \\ \vdots \\ \boldsymbol{z}_n^\top \end{bmatrix}$ consists of observations of m nonrandom explanatory variables which include the constant "variable" $\boldsymbol{\iota}$. The variables in $\boldsymbol{Z}$ are often functions of certain variables in $\boldsymbol{X}$, but this is not necessary for the derivation that follows.

A special case of this specification is $\sigma_t^2 = \sigma^2 x_t^p$ or, after taking logarithms, $\ln\sigma_t^2 = \ln\sigma^2 + p\ln x_t$. Here $\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{\iota} & \ln\boldsymbol{x} \end{bmatrix}$ and $\boldsymbol{\alpha}^\top = \begin{bmatrix} \ln\sigma^2 & p \end{bmatrix}$.

Write (58.1.8) as $0 = \boldsymbol{z}_t^\top \boldsymbol{\alpha} - \ln\sigma_t^2$ and add $\ln\varepsilon_t^2$ to both sides to get

$$\ln\varepsilon_t^2 = \boldsymbol{z}_t^\top \boldsymbol{\alpha} + \ln(\varepsilon_t^2/\sigma_t^2). \tag{58.1.9}$$

This can be considered a regression equation with $\ln(\varepsilon_t^2/\sigma_t^2)$ as the disturbance term. The assumption is that $\operatorname{var}[\ln(\varepsilon_t^2/\sigma_t^2)]$ does not depend on t, which is the case if the ε_t/σ_t are i.i.d. The lefthand side of (58.1.9) is not observed, but one can take the OLS residuals $\hat{\varepsilon}_t$; usually $\ln\hat{\varepsilon}_t^2 \to \ln\varepsilon_t^2$ in the probability limit.

There is only one hitch: the disturbances in regression (58.1.9) do not have zero expected value. Their expected value is an unknown constant. If one ignores

that and runs a regression on (58.1.9), one gets an inconsistent estimate of the element of $\boldsymbol{\alpha}$ which is the coefficient of the constant term in $\boldsymbol{Z}$. This estimate really estimates the sum of the constant term plus the expected value of the disturbance. As a consequence of this inconsistency, the vector $\exp(\boldsymbol{Z}\boldsymbol{\alpha})$ estimates the vector of variances only up to a joint multiplicative constant. I.e., this inconsistency is such that the plim of the variance estimates is not equal but nevertheless proportional to the true variances. But proportionality is all one needs for GLS; the missing multiplicative constant is then the s^2 provided by the least squares formalism.

Therefore all one has to do is: run the regression (58.1.9) (if the F test does not reject, then homoskedasticity cannot be rejected), get the (inconsistent but proportional) estimates $\hat{\sigma}_t^2 = \exp(\boldsymbol{z}_t^\top \boldsymbol{\alpha})$, divide the tth observation of the original regression by $\hat{\sigma}_t$, and re-run the original regression on the transformed data. Consistent estimates of σ_t^2 are then the s^2 from this transformed regression times the inconsistent estimates $\hat{\sigma}_t^2$.

58.1.2. Testing for heteroskedasticity: One test is the F-test in the procedure just described. Then there is the Goldfeld-Quandt test: if it is possible to order the observations in order of increasing error variance, run separate regressions on the portion of the date with low variance and that with high variance, perhaps leaving out some in the middle to increase power of the test, and then just making an F-test with $\frac{\mathsf{SSE}_{high}/\text{d.f.}}{\mathsf{SSE}_{low}/\text{d.f.}}$.

Problem 504. *Why does the Goldfeld-Quandt not use* $\mathsf{SSE}_{high} - \mathsf{SSE}_{low}$ *in the numerator?*

58.1.3. Heteroskedasticity with Unknown Pattern. For consistency of OLS one needs

$$\operatorname{plim}\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{\varepsilon} = \boldsymbol{o} \tag{58.1.10}$$

$$\boldsymbol{Q} = \operatorname{plim}\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X} \quad \text{exists and is nonsingular} \tag{58.1.11}$$

$$\boldsymbol{Q}^* = \operatorname{plim}\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{\Psi}\boldsymbol{X} \quad \text{exists and is nonsingular} \tag{58.1.12}$$

Proof:

$$\mathcal{V}[\hat{\boldsymbol{\beta}}_{OLS}] = \frac{\sigma^2}{n}\Big(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\Big)^{-1}\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{\Psi}\boldsymbol{X}\Big(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\Big)^{-1} \tag{58.1.13}$$

therefore $\operatorname{plim}\mathcal{V}[\hat{\boldsymbol{\beta}}_{OLS}] = \frac{\sigma^2}{n}\boldsymbol{Q}^{-1}\boldsymbol{Q}^*\boldsymbol{Q}^{-1}$.

Look at the following simple example from [**Gre97**, fn. 3 on p. 547:] $\mathbf{y} = \boldsymbol{x}\beta + \boldsymbol{\varepsilon}$ with $\operatorname{var}[\varepsilon_i] = \sigma^2 z_i^2$. For the variance of the OLS estimator we need

$$\boldsymbol{X}^\top \boldsymbol{\Psi} \boldsymbol{X} = \begin{bmatrix} x_1 & \dots & x_n \end{bmatrix} \begin{bmatrix} z_1^2 & 0 & \cdots & 0 \\ 0 & z_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & z_n^2 \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \sum_{i=1}^n x_i^2 z_i^2. \tag{58.1.14}$$

Then by (58.1.13) $\operatorname{var}[\hat\beta_{OLS}] = \sigma^2 \frac{\sum_i x_i^2 z_i^2}{(\sum_i x_i^2)^2}$. Now assume that x_i and z_i are independent observations of the random variables x and z with $\operatorname{E}[\mathsf{z}^2] = 1$ and $\operatorname{cov}[\mathsf{x}^2, \mathsf{z}^2] = 0$. In this case the naive regression output for the variance of $\hat\beta$, which is $\mathsf{s}_N^2 = \mathsf{s}^2 / \sum \mathsf{x}^2$, is indeed a consistent estimate of the variance.

(58.1.15)

$$\operatorname{plim} \frac{\operatorname{var}[\hat\beta_{OLS}]}{\mathsf{s}_N^2} = \operatorname{plim} \frac{\sigma^2 \sum x^2 z^2}{\mathsf{s}^2 \sum x^2} = \operatorname{plim} \frac{\sigma^2}{\mathsf{s}^2} \frac{\frac{1}{n}\sum_i \mathsf{x}_i^2 \mathsf{z}_i^2}{\frac{1}{n}\sum_i \mathsf{x}_i^2} = \frac{\operatorname{E}[\mathsf{x}^2 \mathsf{z}^2]}{\operatorname{E}[\mathsf{x}^2]} = \frac{\operatorname{cov}[\mathsf{x}^2, \mathsf{z}^2] + \operatorname{E}[\mathsf{x}^2]\,\mathrm{I}}{\operatorname{E}[\mathsf{x}^2]}$$

I.e., if one simply runs OLS in this model, then the regression printout is not misleading. On the other hand, it is clear that always $\operatorname{var}[\hat\beta_{OLS}] > \operatorname{var}[\hat\beta]$; therefore if z is observed, then one can do better than this.

Problem 505. *Someone says: the formula*

$$\mathcal{V}[\hat{\boldsymbol{\beta}}_{OLS}] = \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{\Psi} \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \tag{58.1.16}$$

is useless; if one knows $\boldsymbol{\Psi}$ then one will use GLS, and if one does not know $\boldsymbol{\Psi}$ then there are not enough data to estimate it. Comment on this.

Answer: This is a fallacy. In the above formula one does not need $\boldsymbol{\Psi}$ but $\boldsymbol{X}^\top \boldsymbol{\Psi} \boldsymbol{X}$, which is a $k \times k$ symmetric matrix, i.e., it has $k(k+1)/2$ different elements. And even an inconsistent estimate of $\boldsymbol{\Psi}$ can lead to a consistent estimate of $\boldsymbol{X}^\top \boldsymbol{\Psi} \boldsymbol{X}$. Which inconsistent estimate of $\boldsymbol{\Psi}$ shall we use? of course $\hat{\boldsymbol{\Psi}} = \begin{bmatrix} \hat\varepsilon_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \hat\varepsilon_n^2 \end{bmatrix}$. Now since

$$\boldsymbol{X}^\top \boldsymbol{\Psi} \boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_n \end{bmatrix} \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_n^2 \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_1^\top \\ \vdots \\ \boldsymbol{x}_n^\top \end{bmatrix} = \sum_i \sigma_i^2 \boldsymbol{x}_i \boldsymbol{x}_i^\top \tag{58.1.17}$$

one gets White's heteroskedastic-consistent estimator.

$$Est.Var[\hat{\boldsymbol{\beta}}_{OLS}] = \frac{\hat{\boldsymbol{\varepsilon}}^\top \hat{\boldsymbol{\varepsilon}}}{n} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} (\sum_i \hat\varepsilon_i^2 \boldsymbol{x}_i \boldsymbol{x}_i^\top) (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \tag{58.1.18}$$

This estimator has become very fashionable, since one does not have to bother with estimating the covariance structure, and since OLS is not too inefficient in these situations.

It has been observed, however, that this estimator gives too small confidence intervals in small samples. Therefore it is recommended in small samples to multiply the estimated variance by the factor $n/(n-k)$ or to use $\frac{\hat\varepsilon_i^2}{m_{ii}}$ as the estimates of σ_i^2. See [**DM93**, p. 554].

58.2. Autocorrelation

While heteroskedasticity is most often found with cross-sectional data, autocorrelation is more common with time-series.

Properties of OLS in the presence of autocorrelation. If the correlation between the observations dies off sufficiently rapidly as the observations become further apart in time, OLS is consistent and asymptotically normal, but inefficient. There is one important exception to this rule: if the regression includes lagged dependent variables and there is autocorrelation, then OLS and also GLS is inconsistent.

Problem 506. [**JHG+88**, p. 577] *and* [**Gre97**, 13.4.1]. *Assume*

$$\mathsf{y}_t = \alpha + \beta \mathsf{y}_{t-1} + \varepsilon_t \tag{58.2.1}$$

$$\varepsilon_t = \rho \varepsilon_{t-1} + \nu_t \tag{58.2.2}$$

where $\nu_t \sim \operatorname{IID}(0, \sigma_\nu^2)$ and all ν_t are independent of ε_0 and y_0, and $|\rho| < 1$ and $|\beta| < 1$.

- a. *2 points Show that ν_t is independent of all ε_s and y_s for $0 \le s < t$.*

Answer. Both proofs by induction. First independence of ν_t of ε_s: By induction assumption, ν_t independent of ε_{s-1} and since $t > s$, i.e., $t \neq s$, ν_t is also independent of ν_s, therefore ν_t independent of $\varepsilon_s = \rho\varepsilon_{s-1} + \nu_s$. Now independence of ν_t of y_s: By induction assumption, ν_t independent of y_{s-1} and since $t > s$, ν_t is also independent of ε_s, therefore ν_t independent of $\mathsf{y}_s = \alpha + \beta \mathsf{y}_{s-1} + \varepsilon_s$. □

- b. *3 points Show that $\operatorname{var}[\varepsilon_t] = \rho^{2t} \operatorname{var}[\varepsilon_0] + (1 - \rho^{2t}) \frac{\sigma_\nu^2}{1-\rho^2}$. (Hint: use induction.) I.e., since $|\rho| < 1$, $\operatorname{var}[\varepsilon_t]$ converges towards $\sigma_\varepsilon^2 = \frac{\sigma_\nu^2}{1-\rho^2}$.*

Answer. Here is the induction step. Assume that $\operatorname{var}[\varepsilon_{t-1}] = \rho^{2(t-1)} \operatorname{var}[\varepsilon_0] + (1 - \rho^{2(t-1)}) \frac{\sigma_\nu^2}{1-\rho^2}$. Since $\varepsilon_t = \rho\varepsilon_{t-1} + \nu_t$ and ν_t is independent of ε_{t-1}, it follows

(58.2.3)

$$\operatorname{var}[\varepsilon_t] = \rho^2 \operatorname{var}[\varepsilon_{t-1}] + \operatorname{var}[\nu_t] = \rho^{2t} \operatorname{var}[\varepsilon_0] + \rho^2 (1 - \rho^{2(t-1)}) \frac{\sigma_\nu^2}{1-\rho^2} + \sigma_\nu^2 = \rho^{2t} \operatorname{var}[\varepsilon_0] + (1 - \rho^{2t}) \frac{\sigma_\nu^2}{1-\rho^2}.$$

□

- c. *2 points (c) Show that $\operatorname{cov}[\varepsilon_t, \mathsf{y}_{t-1}] = \rho\beta \operatorname{cov}[\varepsilon_{t-1}, \mathsf{y}_{t-2}] + \rho \operatorname{var}[\varepsilon_{t-1}]$.*

Answer.

$$\operatorname{cov}[\varepsilon_t, \mathrm{y}_{t-1}] = \operatorname{cov}[\rho\varepsilon_{t-1} + \nu_t, \alpha + \beta\mathrm{y}_{t-2} + \varepsilon_{t-1}] \tag{58.2.4}$$

$$= \rho\beta \operatorname{cov}[\varepsilon_{t-1}, \mathrm{y}_{t-2}] + \rho \operatorname{var}[\varepsilon_{t-1}] \tag{58.2.5}$$

□

• d. *(d) 1 point Show that, if the process has had enough time to become stationary, it follows*

$$\operatorname{cov}[\varepsilon_t, \mathrm{y}_{t-1}] = \frac{\rho}{1-\rho\beta}\sigma_\varepsilon^2 \tag{58.2.6}$$

Answer. Do not yet compute $\operatorname{var}[\varepsilon_{t-1}]$ at this point, just call it σ_ε^2. Assuming stationarity, i.e., $\operatorname{cov}[\varepsilon_t, \mathrm{y}_{t-1}] = \operatorname{cov}[\varepsilon_{t-1}, \mathrm{y}_{t-2}]$, it follows

$$\operatorname{cov}[\varepsilon_t, \mathrm{y}_{t-1}](1-\rho\beta) = \rho\sigma_\varepsilon^2 \tag{58.2.7}$$

$$\operatorname{cov}[\varepsilon_t, \mathrm{y}_{t-1}] = \frac{\rho}{1-\rho\beta}\sigma_\varepsilon^2 \tag{58.2.8}$$

□

• e. *(e) 2 points Show that, again under conditions of stationarity,*

$$\operatorname{var}[\mathrm{y}_t] = \frac{1+\beta\rho}{1-\beta\rho}\frac{\sigma_\varepsilon^2}{1-\beta^2}. \tag{58.2.9}$$

Answer.

$$\operatorname{var}[\mathrm{y}_t] = \beta^2 \operatorname{var}[\mathrm{y}_{t-1}] + 2\beta \operatorname{cov}[\mathrm{y}_{t-1}, \varepsilon_t] + \operatorname{var}[\varepsilon_t] \tag{58.2.10}$$

$$(1-\beta^2)\operatorname{var}[\mathrm{y}_t] = \frac{2\beta\rho}{1-\beta\rho}\sigma_\varepsilon^2 + \sigma_\varepsilon^2 = \frac{1+\beta\rho}{1-\beta\rho}\sigma_\varepsilon^2 \tag{58.2.11}$$

$$\operatorname{var}[\mathrm{y}_{t-1}] = \operatorname{var}[\mathrm{y}_t] = \frac{1+\beta\rho}{1-\beta\rho}\frac{\sigma_\varepsilon^2}{1-\beta^2}. \tag{58.2.12}$$

□

• f. *2 points (f) Show that*

$$\operatorname{plim}\hat\beta_{OLS} = \beta + \frac{(1-\beta^2)\rho}{1+\beta\rho} \tag{58.2.13}$$

In analogy to White's heteroskedastisicty-consistent estimator one can, in the case of autocorrelation, use Newey and West's robust, consistent estimator of the $\mathcal{MSE}$-matrix of OLS. This is discussed in [**Gre97**, p. 505–5 and 590–1]. The straightforward generalization of the White estimator would be

$$Est.Var[\hat{\boldsymbol\beta}_{OLS}] = \frac{1}{n}(\boldsymbol X^\top \boldsymbol X)^{-1}(\sum_{i,j}\hat\varepsilon_i\hat\varepsilon_j \boldsymbol x_i \boldsymbol x_j^\top)(\boldsymbol X^\top \boldsymbol X)^{-1} \tag{58.2.14}$$

but this estimator does not always give a positive definite matrix. The formula which one should use is: first determine a maximum lag L beyond which the autocorrelations are small enough to ignore, and then do

(58.2.15)

$$Est.Var[\hat{\boldsymbol\beta}_{OLS}] = \frac{1}{n}(\boldsymbol X^\top \boldsymbol X)^{-1}(\sum_{j=1}^{L}\sum_{t=j+1}^{n}(1-\frac{j}{L+1})\hat\varepsilon_t\hat\varepsilon_{t-j}(\boldsymbol x_t \boldsymbol x_{t-j}^\top + \boldsymbol x_{t-j}\boldsymbol x_t^\top)(\boldsymbol X^\top \boldsymbol X)^{-1}$$

58.2.1. First-Order Autoregressive Disturbances. This is discussed in [**DM93**, Chapter 10] and [**Gre97**, 13.3.2, 13.6–13.8].

The model is $\mathrm{y}_t = \boldsymbol x_t^\top \boldsymbol\beta + \varepsilon_t$, $t = 1, \ldots, n$. For $t = 2, \ldots n$ the disturbances satisfy $\varepsilon_t = \rho\varepsilon_{t-1} + \nu_t$, where ρ is an unknown constant. and $\mathcal{V}[\boldsymbol\nu] = \sigma_\nu^2 \boldsymbol I$, and all $\boldsymbol\nu_t$ independent of ε_1, and $\operatorname{var}[\varepsilon_1]$ exists. If $|\rho| < 1$, this process becomes stationary over time. "Stationary" means: the variance of ε_t and the covariances $\operatorname{cov}[\varepsilon_t, \varepsilon_{t-j}]$ do not depend on t. First we will discuss what should be done in the hypothetical case that ρ is known.

Problem 507. *The model is*

$$\mathrm{y}_t = \beta_1 + \beta_2 x_{t2} + \cdots + \beta_k x_{tk} + \varepsilon_t \tag{58.2.16}$$

and for $t = 2, \ldots, n$ we know that $\varepsilon_t = \rho\varepsilon_{t-1} + \nu_t$, where $\nu_t \sim \mathrm{IID}(0, \sigma_\nu^2)$ and all ν_t are independent of ε_1. Assume for the sake of the argument that ρ is known.

• a. *2 points Transform the second until the nth observation in such a way that the disturbance terms in the transformed model have a spherical covariance matrix. Do not use matrix manipulations for that but do it observation by observation, and at this point do not transform the first observation yet.*

Answer. Start with

$$\mathrm{y}_t = \beta_1 + \beta_2 x_{t2} + \cdots + \beta_k x_{tk} + \varepsilon_t \tag{58.2.17}$$

Now lag by one and multiply by ρ. This can only be done for $t = 2, \ldots, n$.

$$\rho\mathrm{y}_{t-1} = \rho\beta_1 + \rho\beta_2 x_{(t-1)2} + \cdots + \rho\beta_k x_{(t-1)k} + \rho\varepsilon_{t-1} \tag{58.2.18}$$

Now subtract, to get the well-behaved disturbances.

$$\mathrm{y}_t - \rho\mathrm{y}_{t-1} = (1-\rho)\beta_1 + \beta_2(x_{t2} - \rho x_{(t-1)2}) + \cdots + \beta_k(x_{tk} - \rho x_{(t-1)k}) + \nu_t \tag{58.2.19}$$

□

• b. *2 points If $|\rho| < 1$, then the process generating the residuals converges toward a stationary process. Assuming that this stationary state has been reached, show that*

$$\operatorname{var}[\varepsilon_t] = \frac{1}{1-\rho^2}\sigma_\nu^2 \tag{58.2.20}$$

and also give a formula for $\operatorname{cov}[\varepsilon_t, \varepsilon_{t-j}]$ in terms of σ_ν^2, ρ, and j.

Answer. From the assumptions follows $\operatorname{var}[\varepsilon_{t+1}] = \rho^2 \operatorname{var}[\varepsilon_t] + \sigma_\nu^2$. Stationarity means $\operatorname{var}[\varepsilon_{t+1}] = \operatorname{var}[\varepsilon_t] = \sigma_\varepsilon^2$, say. Therefore $\sigma_\varepsilon^2 = \rho^2\sigma_\varepsilon^2 + \sigma_\nu^2$, which gives $\sigma_\varepsilon^2 = \sigma_\nu^2/(1-\rho^2)$. For the covariances one gets $\operatorname{cov}[\varepsilon_t, \varepsilon_{t-1}] = \operatorname{cov}[\rho\varepsilon_{t-1}+\nu_t, \varepsilon_{t-1}] = \rho\sigma_\varepsilon^2$; $\operatorname{cov}[\varepsilon_t, \varepsilon_{t-2}] = \operatorname{cov}[\rho\varepsilon_{t-1}+\nu_t, \varepsilon_{t-2}] = \operatorname{cov}[\rho^2\varepsilon_{t-2} + \rho\nu_{t-1} + \nu_t, \varepsilon_{t-2}] = \rho^2\sigma_\varepsilon^2$, etc. □

• c. *2 points Assuming stationarity, write down the covariance matrix of the vector* $\begin{bmatrix}\varepsilon_1 & \nu_2 & \nu_3 & \cdots & \nu_n\end{bmatrix}^\top$. *(Note that this vector has an* ε *in the first place and* ν*'s thereafter!) How can the first observation be transformed so that all transformed observations have uncorrelated and homoskedastic disturbances?*

Answer. The covariance matrix is

$$\sigma_\nu^2 \begin{bmatrix} \frac{1}{1-\rho^2} & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{58.2.21}$$

We know $\varepsilon_1 = \rho\varepsilon_0 + \nu_0$ and $\varepsilon_0 = \rho\varepsilon_{-1} + \nu_{-1}$ etc. Plugging this together gives $\varepsilon_1 = \nu_0 + \rho\nu_{-1} + \rho^2\nu_{-2} + \cdots$. The value of the very first disturbance $\varepsilon_{-\infty}$ no longer matters, since it is multiplied by basically ρ^∞. And we know the variance of the piled-up innovations $\nu_0 + \rho\nu_{-1} + \rho^2\nu_{-2} + \cdots$. It is $\sigma_\nu^2/(1-\rho^2)$. In other words, we know that the disturbance in the first observation is independent of all the later innovations, and its variance is by the factor $1/(1-\rho^2)$ higher than that of these innovations. Therefore multiply first observation by $\sqrt{1-\rho^2}$ take this together with the other differenced observations in order to get a well-behaved regression. □

In matrix notation, the intuitive procedure derived in Problem 507 looks as follows: The covariance matrix of $\boldsymbol{\varepsilon}$ can be written in the form $\mathcal{V}[\boldsymbol{\varepsilon}] = \sigma_\varepsilon^2\boldsymbol{\Psi}$, where $\boldsymbol{\Psi}$ is the correlation matrix of the error terms:

$$\boldsymbol{\Psi} = \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{bmatrix}. \tag{58.2.22}$$

The matrix-inverse of $\boldsymbol{\Psi}$ turns out to be "band-diagonal":

$$\boldsymbol{\Psi}^{-1} = \frac{1}{1-\rho^2}\begin{bmatrix} 1 & -\rho & 0 & \cdots & 0 & 0 \\ -\rho & 1+\rho^2 & -\rho & \cdots & 0 & 0 \\ 0 & -\rho & 1+\rho^2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1+\rho^2 & -\rho \\ 0 & 0 & 0 & \cdots & -\rho & 1 \end{bmatrix} \tag{58.2.23}$$

A transformation matrix $\boldsymbol{P}$ which makes the covariance matrix of the disturbances spherical, scaled such that $\mathcal{V}[\boldsymbol{P}\boldsymbol{\varepsilon}] = \sigma_\nu^2\boldsymbol{I}$, is any matrix $\boldsymbol{P}$ which satisfies $\boldsymbol{P}^\top\boldsymbol{P} = (1-\rho^2)\boldsymbol{\Psi}^{-1}$.

Problem 508. *2 points In the AR-1 model, write* $\mathcal{V}[\boldsymbol{\varepsilon}] = \sigma_\varepsilon^2\boldsymbol{\Psi}$. *Prove that the matrix* $\boldsymbol{P}$ *satisfies* $\mathcal{V}[\boldsymbol{P}\boldsymbol{\varepsilon}] = \sigma_\nu^2\boldsymbol{I}$ *if* $\boldsymbol{P}^\top\boldsymbol{P} = (1-\rho^2)\boldsymbol{\Psi}^{-1}$.

Answer. $\mathcal{V}[\boldsymbol{P}\boldsymbol{\varepsilon}] = \sigma_\nu^2\boldsymbol{I}$; $\sigma_\varepsilon^2\boldsymbol{P}\boldsymbol{\Psi}\boldsymbol{P}^\top = \sigma_\nu^2\boldsymbol{I}$; $\sigma_\varepsilon^2\boldsymbol{\Psi} = \sigma_\nu^2\boldsymbol{P}^{-1}(\boldsymbol{P}^\top)^{-1} = \sigma_\nu^2(\boldsymbol{P}^\top\boldsymbol{P}^\top)^{-1}$; $1/(1-\rho^2)\boldsymbol{\Psi} = (\boldsymbol{P}^\top\boldsymbol{P})^{-1}$; $(1-\rho^2)\boldsymbol{\Psi}^{-1} = \boldsymbol{P}^\top\boldsymbol{P}$. □

Given a nonnegative definite $n \times n$ matrix $\boldsymbol{\Sigma}$, there are usually many $n \times n$ matrices $\boldsymbol{P}$ which satisfy $\boldsymbol{P}^\top\boldsymbol{P} = \boldsymbol{\Sigma}$. But if one requires that $\boldsymbol{P}$ is a lower diagonal matrix with nonnegative elements in the diagonal, then $\boldsymbol{P}$ is unique and is called the "Cholesky root" of $\boldsymbol{\Sigma}$. The Cholesky root of $(1-\rho^2)\boldsymbol{\Psi}^{-1}$ is the following $\boldsymbol{P}$:

$$\boldsymbol{P} = \begin{bmatrix} \sqrt{1-\rho^2} & 0 & 0 & \cdots & 0 & 0 \\ -\rho & 1 & 0 & \cdots & 0 & 0 \\ 0 & -\rho & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & -\rho & 1 \end{bmatrix} \tag{58.2.24}$$

This is exactly the transformation which the procedure from Problem 507 leads to.

Problem 509. *This question is formulated in such a way that you can do each part of it independently of the others. Therefore if you get stuck, just go on to the next part. We are working in the linear regression model* $\mathrm{y}_t = \boldsymbol{x}_t^\top\boldsymbol{\beta} + \varepsilon_t$, $t = 1, \ldots, n$, *in which the following is known about the disturbances* ε_t*: For* $t = 2, \ldots, n$ *one can write* $\varepsilon_t = \rho\varepsilon_{t-1} + \nu_t$ *with an unknown nonrandom* ρ*, and the* ν_t *are well behaved, i.e., they are homoskedastic* $\nu_t \sim (0, \sigma_\nu^2)$ *and* ν_s *independent of* ν_t *for* $s \neq t$*. The first disturbance* ε_1 *has a finite variance and is independent of* $\nu_2, \ldots, \nu_n$.

• a. *1 point Show by induction that* ν_t *is independent of all* ε_s *with* $1 \leq s < t \leq n$.

Answer. ν_t $(2 \leq t \leq n)$ is independent of ε_1 by assumption. Now assume $2 \leq s \leq t-1$ and ν_t is independent of ε_{s-1}. Since $\varepsilon_s = \rho\varepsilon_{s-1} + \nu_s$, and ν_t is by assumption also independent of ν_s, it follows that ν_t is independent of ε_s. □

• b. *3 points Show by induction that, if* $|\rho| < 1$*, then*

$$\operatorname{var}\varepsilon_t = (1-\rho^{2(t-1)})\sigma_\varepsilon^2 + \rho^{2(t-1)}\operatorname{var}[\varepsilon_1] \tag{58.2.25}$$

where $\sigma_\varepsilon^2 = \sigma_\nu^2/(1-\rho^2)$*. This formula says: if* $\operatorname{var}[\varepsilon_1] = \sigma_\nu^2/(1-\rho^2)$*, then all the other* $\operatorname{var}[\varepsilon_t]$ *have the same value, and if* $\operatorname{var}[\varepsilon_1] \neq \sigma_\nu^2/(1-\rho^2)$*, then there is monotonic convergence* $\operatorname{var}[\varepsilon_t] \to \sigma_\nu^2/(1-\rho^2)$.

Answer. (58.2.25) is true by assumption for $t = 1$. Here is the induction step. Assume that $\operatorname{var}[\varepsilon_t] = \rho^{2(t-1)}\operatorname{var}[\varepsilon_1] + (1-\rho^{2(t-1)})\sigma_\nu^2/(1-\rho^2)$. Since $\varepsilon_{t+1} = \rho\varepsilon_t + \nu_{t+1}$ and ν_{t+1} is independent of ε_t, it follows

$$\operatorname{var}[\varepsilon_{t+1}] = \rho^{2t}\operatorname{var}[\varepsilon_1] + \rho^2(1-\rho^{2(t-1)})\frac{\sigma_\nu^2}{1-\rho^2} + \sigma_\nu^2 = \rho^{2t}\operatorname{var}[\varepsilon_1] + (1-\rho^{2t})\frac{\sigma_\nu^2}{1-\rho^2}. \tag{58.2.26}$$

□

• c. *1 point Assuming* $|\rho| < 1$ *and* $\operatorname{var}[\varepsilon_1] = \sigma_\nu^2/(1-\rho^2) =: \sigma_\varepsilon^2$, *compute the correlation matrix of the disturbance vector* $\boldsymbol{\varepsilon}$. *Since* $\boldsymbol{\varepsilon}$ *is homoskedastic, this is at the same time that matrix* $\boldsymbol{\Psi}$ *for which* $\mathcal{V}[\boldsymbol{\varepsilon}] = \sigma_\varepsilon^2\boldsymbol{\Psi}$. *What is a (covariance) stationary process, and do the* ε_t *form one?*

Answer. $\operatorname{cov}[\varepsilon_t, \varepsilon_{t-1}] = \operatorname{cov}[\rho\varepsilon_{t-1} + \nu_t, \varepsilon_{t-1}] = \rho\sigma_\varepsilon^2$; $\operatorname{cov}[\varepsilon_t, \varepsilon_{t-2}] = \operatorname{cov}[\rho\varepsilon_{t-1} + \nu_t, \varepsilon_{t-2}] = \operatorname{cov}[\rho^2\varepsilon_{t-2} + \rho\nu_{t-1} + \nu_t, \varepsilon_{t-2}] = \rho^2\sigma_\varepsilon^2$, etc. If we therefore write the covariance matrix of $\boldsymbol{\varepsilon}$ in the form $\mathcal{V}[\boldsymbol{\varepsilon}] = \sigma_\varepsilon^2\boldsymbol{\Psi}$, so that all elements in the diagonal of $\boldsymbol{\Psi}$ are $= 1$, which makes $\boldsymbol{\Psi}$ at the same time the correlation matrix, we get

$$\boldsymbol{\Psi} = \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{bmatrix}. \tag{58.2.27}$$

A process is covariance stationary if the expected value and the variance do not change over time, and $\operatorname{cov}[\varepsilon_s, \varepsilon_t]$ depends only on $s - t$, not on s or t separately. Yes it is a covariance stationary process. □

• d. *2 points Show that the matrix in equation* (58.2.23) *is the inverse of this correlation matrix.*

• e. *2 points Prove that the square matrix* $\boldsymbol{P}$ *satisfies* $\mathcal{V}[\boldsymbol{P}\boldsymbol{\varepsilon}] = \sigma_\nu^2\boldsymbol{I}$ *if and only if* $\boldsymbol{P}^\top\boldsymbol{P} = (1-\rho^2)\boldsymbol{\Psi}^{-1}$.

Answer. $\mathcal{V}[\boldsymbol{P}\boldsymbol{\varepsilon}] = \sigma_\nu^2\boldsymbol{I}$; $\sigma_\varepsilon^2\boldsymbol{P}\boldsymbol{\Psi}\boldsymbol{P}^\top = \sigma_\nu^2\boldsymbol{I}$; $\sigma_\varepsilon^2\boldsymbol{\Psi} = \sigma_\nu^2\boldsymbol{P}^{-1}(\boldsymbol{P}^\top)^{-1} = \sigma_\nu^2(\boldsymbol{P}^\top\boldsymbol{P}^\top)^{-1}$; $1/(1-\rho^2)\boldsymbol{\Psi} = (\boldsymbol{P}^\top\boldsymbol{P})^{-1}$; $(1-\rho^2)\boldsymbol{\Psi}^{-1} = \boldsymbol{P}^\top\boldsymbol{P}$. □

• f. *2 points Show that the* $\boldsymbol{P}$ *defined in* (58.2.24) *satisfies* $\boldsymbol{P}^\top\boldsymbol{P} = (1-\rho^2)\boldsymbol{\Psi}^{-1}$.

• g. *2 points Use* $\boldsymbol{P}$ *to show that* $\det\boldsymbol{\Psi} = (1-\rho^2)^{n-1}$.

Answer. Since $\boldsymbol{P}$ is lower diagonal, its determinant is the product of the diagonal elements, which is $\sqrt{1-\rho^2}$. Since $\boldsymbol{\Psi}^{-1} = \frac{1}{1-\rho^2}\boldsymbol{P}^\top\boldsymbol{P}$, it follows $\det[\boldsymbol{\Psi}^{-1}] = 1/(1-\rho^2)^n(\det[\boldsymbol{P}])^2 = 1/(1-\rho^2)^{n-1}$, therefore $\det\boldsymbol{\Psi} = (1-\rho^2)^{n-1}$. □

• h. *3 points Show that the general formula for the log likelihood function* (35.0.11) *reduces in our specific situation to*

$$\ln\ell(y;\beta,\rho,\sigma_\nu^2) = \textit{constant} - \frac{n}{2}\ln\sigma_\nu^2 + \frac{1}{2}\ln(1-\rho^2) - \frac{1}{2\sigma_\nu^2}\Big((1-\rho^2)\varepsilon_1^2 + \sum_{t=2}^n(\varepsilon_t - \rho\varepsilon_{t-1})^2\Big) \tag{58.2.28}$$

where $\varepsilon_t = \mathsf{y}_t - \boldsymbol{x}_t^\top\boldsymbol{\beta}$. *You will need such an expression if you have to program the likelihood function in a programming language which does not understand matrix operations. As a check on your arithmetic I want you to keep track of the value of the constant in this formula and report it.* Hint: *use* $\boldsymbol{P}$ *to evaluate the quadratic form* $(\mathsf{y} - \boldsymbol{X}\boldsymbol{\beta})^\top\boldsymbol{\Psi}^{-1}(\mathsf{y} - \boldsymbol{X}\boldsymbol{\beta})$.

Answer. The constant is $-\frac{n}{2}\ln 2\pi$. The next two terms are $-\frac{n}{2}\ln\sigma_\varepsilon^2 - \frac{1}{2}\ln|\det\boldsymbol{\Psi}| = -\frac{n}{2}\ln\sigma_\nu^2 + \frac{n}{2}\ln(1-\rho^2) - \frac{n-1}{2}\ln(1-\rho^2) = -\frac{n}{2}\ln\sigma_\nu^2 + \frac{1}{2}\ln(1-\rho^2)$. And since $\boldsymbol{\Psi}^{-1} = \frac{1}{1-\rho^2}\boldsymbol{P}^\top\boldsymbol{P}$ and $\sigma_\varepsilon^2(1-\rho^2) = \sigma_\nu^2$, the last terms coincide too, because $(\mathsf{y} - \boldsymbol{X}\boldsymbol{\beta})^\top\boldsymbol{\Psi}^{-1}(\mathsf{y} - \boldsymbol{X}\boldsymbol{\beta}) = (\mathsf{y} - \boldsymbol{X}\boldsymbol{\beta})^\top\boldsymbol{P}^\top\boldsymbol{P}(\mathsf{y} - \boldsymbol{X}\boldsymbol{\beta})$. □

• i. *4 points Show that, if one concentrates out* σ_ν^2, *i.e., maximizes this likelihood function with respect to* σ_ν^2, *taking all other parameters as given, one obtains*

$$\ln\ell_{conc.} = \textit{constant} + \frac{1}{2}\ln(1-\rho^2) - \frac{n}{2}\ln\Big((1-\rho^2)\varepsilon_1^2 + \sum_{t=2}^n(\varepsilon_t - \rho\varepsilon_{t-1})^2\Big) \tag{58.2.29}$$

Again as a check on your arithmetic, I want you to give me a formula for the constant in (58.2.29). *If you did not figure out the constant in* (58.2.28), *you may give me the constant in* (58.2.29) *as a function of the constant in* (58.2.28).

Answer. It is better to do it from scratch than to use the general formula (35.0.17): First order condition is

$$\frac{\partial}{\partial\sigma_\nu^2}\ln\ell(y;\beta,\rho,\sigma_\nu^2) = -\frac{n}{2}\frac{1}{\sigma_\nu^2} + \frac{(1-\rho^2)\varepsilon_1^2 + \sum_{t=2}^n(\varepsilon_t - \rho\varepsilon_{t-1})^2}{2\sigma_\nu^4} = 0 \tag{58.2.30}$$

which gives

$$\sigma_\nu^2 = \frac{(1-\rho^2)\varepsilon_1^2 + \sum_{t=2}^n(\varepsilon_t - \rho\varepsilon_{t-1})^2}{n} \tag{58.2.31}$$

Plugging this into the likelihood function gives (58.2.29), but this time the constant is written out:

$$\ln\ell_{conc.} = -\frac{n}{2} - \frac{n}{2}\ln 2\pi + \frac{n}{2}\ln n + \frac{1}{2}\ln(1-\rho^2) - \frac{n}{2}\ln\Big((1-\rho^2)\varepsilon_1^2 + \sum_{t=2}^n(\varepsilon_t - \rho\varepsilon_{t-1})^2\Big) \tag{58.2.32}$$

This is [**BM78**, (2) on p. 52]. □

• j. *3 points Is it possible to concentrate out further parameters from this likelihood function? What numerical procedure would you use if you had to estimate this model by maximum likelihood?*

Answer. One can either concentrate out $\boldsymbol{\beta}$ or ρ but not both. The recommended procedure is: set $\rho = 0$ and solve for $\boldsymbol{\beta}$, then use this $\boldsymbol{\beta}$ to get the best ρ, and so on until it converges. Computationally this is not much more expensive than COhrane-Orcutt, but it is much better since it gives you the maximum l;ikelihood estimator, which has much better small sample properties. This is recommended by [**BM78**]. □

As this Question shows, after concentrating out σ_ν^2 one can either concentrate out ρ or $\boldsymbol{\beta}$ but not both, and [**BM78**] propose to alternate these concentrations until it converges.

58.2.2. Prediction. To compute the BLUP for one step ahead simply predict ν_{n+1} by 0, i.e. $\varepsilon^*_{n+1} = \rho\hat{\varepsilon}_n$, hence

$$\mathsf{y}^*_{n+1} = \boldsymbol{x}^\top_{n+1}\hat{\boldsymbol{\beta}} + \rho\hat{\varepsilon}_n; \tag{58.2.33}$$

for two steps ahead it is $\mathsf{y}^*_{n+2} = \boldsymbol{x}^\top_{n+2}\hat{\boldsymbol{\beta}} + \rho^2\hat{\varepsilon}_n$, etc.

Problem 510. *3 points Use formula* (27.3.6) *to derive formula* (58.2.33).

Answer. Write the model as

$$\begin{bmatrix}\mathsf{y}\\ \mathsf{y}_{n+1}\end{bmatrix} = \begin{bmatrix}\boldsymbol{X}\\ \boldsymbol{x}^\top_{n+1}\end{bmatrix}\boldsymbol{\beta} + \begin{bmatrix}\boldsymbol{\varepsilon}\\ \varepsilon_{n+1}\end{bmatrix}; \qquad \begin{bmatrix}\boldsymbol{\varepsilon}\\ \varepsilon_{n+1}\end{bmatrix} \sim \begin{bmatrix}\boldsymbol{o}\\ 0\end{bmatrix}, \sigma^2_\varepsilon\begin{bmatrix}\boldsymbol{\Psi} & \boldsymbol{v}\\ \boldsymbol{v}^\top & 1\end{bmatrix}$$

where $\boldsymbol{v}^\top = [\rho^n, \rho^{n-1}, \dots, \rho^2, \rho]$ and $\boldsymbol{\Psi}$ is as in (58.2.22). Equation (27.3.6) gives $\mathsf{y}^*_{n+1} = \boldsymbol{x}^\top_{n+1}\hat{\boldsymbol{\beta}} + \boldsymbol{v}^\top\boldsymbol{\Psi}^{-1}(\mathsf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$. Using $\boldsymbol{\Psi}^{-1}$ from (58.2.23) one can show that

$$\boldsymbol{v}^\top\boldsymbol{\Psi}^{-1} = \frac{1}{1-\rho^2}\begin{bmatrix}\rho^n & \rho^{n-1} & \rho^{n-2} & \cdots & \rho^2 & \rho\end{bmatrix}\begin{bmatrix}1 & -\rho & 0 & \cdots & 0 & 0\\ -\rho & 1+\rho^2 & -\rho & \cdots & 0 & 0\\ 0 & -\rho & 1+\rho^2 & \cdots & 0 & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & 0 & \cdots & 1+\rho^2 & -\rho\\ 0 & 0 & 0 & \cdots & -\rho & 1\end{bmatrix}$$
$$= \begin{bmatrix}0 & 0 & 0 & \cdots & 0 & \rho\end{bmatrix}.$$

From this (58.2.33) follows. □

58.2.3. Second-Order Autoregressive Disturbances.

Problem 511. *If the error term is second order autoregressive, $\varepsilon_t = \alpha_1\varepsilon_{t-1} + \alpha_2\varepsilon_{t-2} + \nu_t$, with $\boldsymbol{\nu} \sim (\boldsymbol{o}, \sigma^2_\nu\boldsymbol{I})$ (white noise), and the model is stationary, then show that the error variance is*

$$\sigma^2_\varepsilon = \frac{1-\alpha_2}{(1+\alpha_2)\big((1-\alpha_2)^2 - \alpha_1^2\big)}\sigma^2_\nu. \tag{58.2.34}$$

• a. *The following matrices are relevant for estimation:*

(58.2.35)

$$\boldsymbol{A} = \begin{bmatrix}1 & \frac{\alpha_1}{1-\alpha_2} & \frac{\alpha_1^2+\alpha_2-\alpha_2^2}{1-\alpha_2} & \frac{\alpha_1(\alpha_1^2+2\alpha_2-\alpha_2^2)}{1-\alpha_2} & \frac{\alpha_1^4+3\alpha_1^2\alpha_2+\alpha_2^2(1-\alpha_1^2}{1-\alpha_2}\\ \frac{\alpha_1}{1-\alpha_2} & 1 & \frac{\alpha_1}{1-\alpha_2} & \frac{\alpha_1^2+\alpha_2-\alpha_2^2}{1-\alpha_2} & \frac{\alpha_1(\alpha_1^2+2\alpha_2-\alpha_2^2}{1-\alpha_2}\\ \frac{\alpha_1^2+\alpha_2-\alpha_2^2}{1-\alpha_2} & \frac{\alpha_1}{1-\alpha_2} & 1 & \frac{\alpha_1}{1-\alpha_2} & \frac{\alpha_1^2+\alpha_2-\alpha_2^2}{1-\alpha_2}\\ \frac{\alpha_1(\alpha_1^2+2\alpha_2-\alpha_2^2)}{1-\alpha_2} & \frac{\alpha_1^2+\alpha_2-\alpha_2^2}{1-\alpha_2} & \frac{\alpha_1}{1-\alpha_2} & 1 & \frac{\alpha_1}{1-\alpha_2}\\ \frac{\alpha_1^4+3\alpha_1^2\alpha_2+\alpha_2^2(1-\alpha_1^2)-\alpha_2^3}{1-\alpha_2} & \frac{\alpha_1(\alpha_1^2+2\alpha_2-\alpha_2^2)}{1-\alpha_2} & \frac{\alpha_1^2+\alpha_2-\alpha_2^2}{1-\alpha_2} & \frac{\alpha_1}{1-\alpha_2} & 1\\ \vdots & \vdots & \vdots & \vdots & \vdots\end{bmatrix}$$

(58.2.36)

$$\boldsymbol{B} = \begin{bmatrix}1 & -\alpha_1 & -\alpha_2 & 0 & \cdots & 0 & 0\\ -\alpha_1 & 1+\alpha_1^2 & -\alpha_1+\alpha_1\alpha_2 & -\alpha_2 & \cdots & 0 & 0\\ -\alpha_2 & -\alpha_1+\alpha_1\alpha_2 & 1+\alpha_1^2+\alpha_2^2 & -\alpha_1+\alpha_1\alpha_2 & \cdots & 0 & 0\\ 0 & -\alpha_2 & -\alpha_1+\alpha_1\alpha_2 & 1+\alpha_1^2+\alpha_2^2 & \cdots & 0 & 0\\ \vdots & \vdots & & & \ddots & \vdots & \vdots\\ 0 & 0 & & & \cdots & 1+\alpha_1^2 & -\alpha_1\\ 0 & 0 & & & \cdots & -\alpha_1 & 1\end{bmatrix};$$

(58.2.37)

$$\boldsymbol{C} = \begin{bmatrix}\sqrt{\frac{(1+\alpha_2)\big((1-\alpha_2)^2-\alpha_1^2\big)}{1-\alpha_2}} & 0 & 0 & 0 & \cdots & 0 & 0\\ \frac{-\alpha_1\sqrt{1-\alpha_2^2}}{1-\alpha_2} & \sqrt{1-\alpha_2^2} & 0 & 0 & \cdots & 0 & 0\\ -\alpha_2 & -\alpha_1 & 1 & 0 & \cdots & 0 & 0\\ 0 & -\alpha_2 & -\alpha_1 & 1 & \cdots & 0 & 0\\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots\\ 0 & 0 & 0 & \ddots & \ddots & 1 & 0\\ 0 & 0 & 0 & 0 & \cdots & -\alpha_1 & 1\end{bmatrix}.$$

• b. *4 points Discuss how these matrices are mathematically related to each other (perhaps one is the inverse of the other, etc.), and how they are related to the model (could it be that one is the matrix of covariances between the error terms and the explanatory variables?) A precise proof of your answer would imply tedious matrix multiplications, but you should be able to give an answer simply by carefully looking at the matrices.*

Answer. $\boldsymbol{A}$ is the correlation matrix of the errors, i.e., $\sigma^2_\varepsilon\boldsymbol{A} = \mathcal{V}[\boldsymbol{\varepsilon}]$, $\boldsymbol{B} = \sigma^2_\nu(\mathcal{V}[\boldsymbol{\varepsilon}])^{-1}$, and $\boldsymbol{C}^\top\boldsymbol{C} = \boldsymbol{B}$. □

• c. *4 points Compute the determinants of these three matrices. (Again this is easy using the structure of the matrices and their mathematical relation.) Why is this determinant important for econometrics?*

• d. *4 points In terms of these matrices, give the objective function (some matrix weighted sum of squares) which the BLUE minimizes due to the Gauss-Markov theorem, give the formula for the BLUE, and give the formula for the unbiased estimator* s_ν^2.

58.2.4. The Autoreg Procedure in SAS. This is about the "autoreg" procedure in the SAS ETS manual.

Model is $\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ or $y_t = \boldsymbol{x}_t^\top \boldsymbol{\beta} + \varepsilon$, n observations and k variables, with $\varepsilon_t = \nu_t - \alpha_1\varepsilon_{t-1} - \cdots - \alpha_p\varepsilon_{t-p}$ where $\nu_t \sim (0, \sigma_\nu^2)$ independent of each other and of $\varepsilon_{t-1}, \ldots, \varepsilon_{t-p}$, and the process is stationary.

Yule-Walker Estimation: From the equations for ε_t follows

(58.2.38)
$$\operatorname{cov}[\varepsilon_t, \varepsilon_{t-1}] = -\alpha_1 \operatorname{var}[\varepsilon_{t-1}] - \alpha_2 \operatorname{cov}[\varepsilon_{t-2}, \varepsilon_{t-1}] - \cdots - \alpha_p \operatorname{cov}[\varepsilon_{t-p}, \varepsilon_{t-1}]$$

(58.2.39)
$$\operatorname{cov}[\varepsilon_t, \varepsilon_{t-2}] = -\alpha_1 \operatorname{cov}[\varepsilon_{t-1}, \varepsilon_{t-2}] - \alpha_2 \operatorname{var}[\varepsilon_{t-2}] - \cdots - \alpha_p \operatorname{cov}[\varepsilon_{t-p}, \varepsilon_{t-2}]$$

...

(58.2.40)
$$\operatorname{cov}[\varepsilon_t, \varepsilon_{t-p}] = -\alpha_1 \operatorname{cov}[\varepsilon_{t-1}, \varepsilon_{t-p}] - \alpha_2 \operatorname{cov}[\varepsilon_{t-2}, \varepsilon_{t-p}] \cdots - \alpha_p \operatorname{var}[\varepsilon_{t-p}] - \cdots - \alpha_p \operatorname{cov}[\varepsilon_t$$

Then divide by $\operatorname{var}[\varepsilon_t] = \operatorname{var}[\varepsilon_{t-1}] = \cdots$ to get the autocorrelations $\rho_j = \operatorname{corr}[\varepsilon_t, \varepsilon_{t-j}]$:

$$\begin{bmatrix} \rho_1 \\ \rho_2 \\ \rho_3 \\ \vdots \\ \rho_p \end{bmatrix} = - \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{p-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{p-2} \\ \rho_2 & \rho_1 & 1 & \cdots & \rho_{p-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{p-1} & \rho_{p-2} & \rho_{p-3} & \cdots & 1 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_p \end{bmatrix}. \tag{58.2.41}$$

This can be used to get estimates of $\alpha_1, \ldots, \alpha_p$ from estimates of $\rho_1, \ldots, \rho_p$. How to estimate $\operatorname{cov}[\varepsilon_{t-i}, \varepsilon_{t-j}]$? In first stage use OLS residuals, and take $\frac{1}{m+j}\sum \hat\varepsilon_t \hat\varepsilon_{t-j}$ where m is the number of such products (there may be missing values). If there are no missing values, then $m + j = n$, and it is the same as multiple regression of $\hat\boldsymbol{\varepsilon}$ on the p lagged values of $\hat\boldsymbol{\varepsilon}$.

From estimates of $\alpha_1, \ldots, \alpha_p$ one can get estimates of $\boldsymbol{\Psi}$, which is up to a scalar factor the error covariance matrix.

$$\mathcal{V}[\boldsymbol{\varepsilon}] = \sigma^2\boldsymbol{\Psi} = \sigma^2 \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{n-2} \\ \rho_2 & \rho_1 & 1 & \cdots & \rho_{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{n-1} & \rho_{n-2} & \rho_{n-3} & \cdots & 1 \end{bmatrix}. \tag{58.2.42}$$

The first p of these ρ's can be obtained from the Yule-Walker equations; the other ρ's from the original difference equation: $\rho_j = -\alpha_1\rho_{j-1} - \cdots - \alpha_p\rho_{j-p}$.

58.2.5. Estimation of the Autoregressive Parameter and Testing for Zero Autoregression. If it were possible to observe the ε_t, one could regress ε_t on ε_{t-1} (lagged dependent variable, asymptotically efficient). Since the ε_t are unobserved, regress $\hat\varepsilon_t$ on $\hat\varepsilon_{t-1}$. The formula is $\hat\rho = \frac{\sum_{t=2}^n \hat\varepsilon_t\hat\varepsilon_{t-1}}{\sum_{t=2}^n \hat\varepsilon_{t-1}^2}$. Cochrane-Orcutt: iterate until it converges. (But note that the term "Cochrane-Orcutt" means different things to different people, nice discussion in [**JHG+88**, p. 393].) Problem is that ρ is underestimated. Maximum Likelihood is the preferred procedure here, and [**BM78**] has an iterative procedure which leads to the MLE and which is no more trouble than Cochrane-Orcutt.

Testing for $\rho = 0$; In the regression of $\hat\varepsilon_t$ on $\hat\varepsilon_{t-1}$, the formula for the variance (39.1.7) holds asymptotically, i.e., $\operatorname{var}[\hat\rho] = \sigma_\nu^2/\operatorname{E}[\boldsymbol{x}^\top\boldsymbol{x}]$ where $\boldsymbol{x}^\top\boldsymbol{x} = \sum_{t=2}^n \hat\varepsilon_{t-1}^2$. Asymptotically, $\boldsymbol{x}^\top\boldsymbol{x}$ has expected value $n\sigma_\varepsilon^2 = n\sigma_\nu^2/(1-\rho^2)$. Asymptotially, therefore, $\operatorname{var}[\hat\rho] = \frac{1-\rho^2}{n}$. If $\rho = 0$, it is $\operatorname{var}[\hat\rho] = 1/n$; in this case, therefore, $\sqrt{n}\hat\rho$ has asymptotic $\mathrm{N}(0,1)$ distribution.

But the most often used statistic for autoregression is the Durbin-Watson.

58.2.6. The Durbin-Watson Test Statistic. The Durbin Watson test [**DW50, DW51, DW71**] tests ε_t and ε_{t-1} are correlated in the linear model $\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ in which the conditions for hypothesis testing are satisfied (either normal disturbances or so many observations that the central limit theorem leads to normality), and in which the errors are homoskedastic with variance σ^2, and $\operatorname{cov}[\varepsilon_{t-1}, \varepsilon_t] = \rho\sigma^2$ with the same ρ for all $t = 2, \ldots, n$.

The DW test does not test the higher autocorrelations. It was found to be powerful if the overall process is an AR(1) process, but it cannot be powerful if the autocorrelation is such that ε_t is not correlated with ε_{t-1} but with higher lags. For instance, for quarterly data, Wallis [**Wal72**] argued that one should expect ε_t to be correlated with ε_{t-4} and not with ε_{t-1}, and he modified the DW test for this situation.

The test statistic is:

$$\mathsf{d} = \frac{\sum_{t=2}^{n}(\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\sum_{t=1}^{n}\hat{\varepsilon}_t^2} \tag{58.2.43}$$

(where the residuals are taken from OLS without correction for autocorrelation). This test statistic is a consistent estimator of $2 - 2\rho$ (but it has its particular form so that the distribution can be calculated). The plim can be seen as follows:

$$\mathsf{d} = \frac{\sum_{t=2}^{n}(\hat{\varepsilon}_t^2 - 2\hat{\varepsilon}_t\hat{\varepsilon}_{t-1} + \hat{\varepsilon}_{t-1}^2)}{\sum_{t=1}^{n}\hat{\varepsilon}_t^2} = \frac{\sum_{t=2}^{n}\hat{\varepsilon}_t^2}{\sum_{t=1}^{n}\hat{\varepsilon}_t^2} - 2\frac{\sum_{t=2}^{n}\hat{\varepsilon}_t\hat{\varepsilon}_{t-1}}{\sum_{t=1}^{n}\hat{\varepsilon}_t^2} + \frac{\sum_{t=2}^{n}\hat{\varepsilon}_{t-1}^2}{\sum_{t=1}^{n}\hat{\varepsilon}_t^2}. \tag{58.2.44}$$

For large n one can ignore that the sum in the numerator has one less element than the one in the denominator. Therefore the first term converges towards 1, the second towards $2\operatorname{cov}[\varepsilon_t, \varepsilon_{t-1}]/\operatorname{var}[\varepsilon_t] = 2\rho$ (note that, due to homoskedasticity, $\operatorname{var}[\varepsilon_t] = \sqrt{\operatorname{var}[\varepsilon_{t-1}]\operatorname{var}[\varepsilon_t]}$), and the third term again towards 1. d is always between 0 and 4, and is close to 2 if there is no autocorrelation, close to 0 for positive autocorrelation, and close to 4 for negative autocorrelation.

d differs from many test statistics considered so far because its distribution depends on the values taken by regressors $\boldsymbol{X}$. It is a very special situation that the distribution of the t statistic and F statistic do not depend on the $\boldsymbol{X}$. Usually one must expect that the values of $\boldsymbol{X}$ have an influence. Despite this dependence on $\boldsymbol{X}$, it is possible to give bounds for the critical values, which are tabulated as D_L (lower D) and D_U (upper D). If the alternative hypothesis is positive autocorrelation, one can reject the null hypothesis if $\mathsf{d} < D_L$ for all possible configurations of the regressors, cannot reject if $\mathsf{d} > D_U$, and otherwise the test is inconclusive, i.e., in this case it depends on $\boldsymbol{X}$ whether to reject or not, and the computer is not taking the trouble of checking which is the case.

The bounds that are usually published are calculated under the assumption that the regression has a constant term, i.e., that there is a vector $\boldsymbol{a}$ so that $\boldsymbol{X}\boldsymbol{a} = \boldsymbol{\iota}$. Tables valid if there is no constant term are given in [**Far80**]. If these tables are unavailable, [**Kme86**, p. 329/30] recommends to include the constant term into the regression before running the test, so that the usual bounds can be used. But [**JHG+88**, p. 399] says that the power of the DW test is low if there is no intercept.

On the other hand, [**Kin81**] has given sharper bounds which one can use if it is known that the regression has a trend of seasonal dummies. The computer program SHAZAM [**Whi93**] computes the exact confidence points using the available data. An approximation to the exact critical values by Durbin and Watson themselves [**DW71**] uses that affine combination of the upper bound which has the same mean and variance as the exact test statistic. This is discussed in [**Gre97**, p. 593] and Green says it is "quite accurate."

Distribution of d: The matrix version of (58.2.44) is

$$\mathsf{d} = \frac{\hat{\boldsymbol{\varepsilon}}^\top \boldsymbol{A}\hat{\boldsymbol{\varepsilon}}}{\hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}}} = \frac{\boldsymbol{\varepsilon}^\top \boldsymbol{M}\boldsymbol{A}\boldsymbol{M}\boldsymbol{\varepsilon}}{\boldsymbol{\varepsilon}^\top\boldsymbol{M}\boldsymbol{\varepsilon}} \quad \text{for} \quad \boldsymbol{A} = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & \cdots & 0 \\ -1 & 2 & -1 & 0 & 0 & \cdots & 0 \\ 0 & -1 & 2 & -1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -1 & 1 \end{bmatrix}. \tag{58.2.45}$$

Since $\boldsymbol{A}$ is symmetric and $\boldsymbol{M}\boldsymbol{A}\boldsymbol{M}$ commutes with $\boldsymbol{M}$, matrix algebra tells us, see [**Rao73**, p. 41], that an orthogonal $\boldsymbol{H}$ exists (i.e., $\boldsymbol{H}\boldsymbol{H}^\top = \boldsymbol{I}$) which simultaneously diagonalizes numerator and denominator: $\boldsymbol{H}^\top\boldsymbol{M}\boldsymbol{A}\boldsymbol{M}\boldsymbol{H}$ has the eigenvalues v_t of $\boldsymbol{M}\boldsymbol{A}$ in the diagonal (only $n-k$ of which are nonzero), and $\boldsymbol{H}\boldsymbol{M}\boldsymbol{H}$ has k zeros and $n-k$ ones in the diagonal. Then under the null hypothesis,

$$\boldsymbol{z} = \boldsymbol{H}^\top\boldsymbol{\varepsilon} \sim \mathsf{N}(\boldsymbol{o}, \sigma^2\boldsymbol{I}), \qquad \boldsymbol{\varepsilon} = \boldsymbol{H}\boldsymbol{z}, \quad \text{and} \quad \mathsf{d} = \frac{\boldsymbol{\varepsilon}^\top\boldsymbol{M}\boldsymbol{A}\boldsymbol{M}\boldsymbol{\varepsilon}}{\boldsymbol{\varepsilon}^\top\boldsymbol{M}\boldsymbol{\varepsilon}} = \frac{\boldsymbol{z}^\top\boldsymbol{H}^\top\boldsymbol{M}\boldsymbol{A}\boldsymbol{M}\boldsymbol{H}\boldsymbol{z}}{\boldsymbol{z}^\top\boldsymbol{H}\boldsymbol{M}\boldsymbol{H}\boldsymbol{z}} = \frac{\sum z_t^2 v}{\sum z_t^2} \tag{58.2.46}$$

which is a tabulated distribution. $\boldsymbol{H}$ depends on $\boldsymbol{X}$, but one can give limits for the eigenvalues, which give the upper and lower limits of the D-W test. Some computer programs (e.g., Shazam) calculate the actual significance points on basis of the given $\boldsymbol{X}$-matrix.

Robustness: D-W will detect more than just first order autoregression [**Bla73**], but not all kinds of serial correlation, e.g. not very powerful for 2nd order autoregression.

If lagged dependent variables *and* autoregression, then OLS is no longer consistent, therefore also d no longer a consistent estimate of $2 - 2\rho$ but is closer to 2 than it should be! Then the D-W has low power, accepts more often than it should. If lagged dependent variable, use Durbin's h. This is an intuitive formula, see [**JHG+88**, p. 401]. It cannot always be calculated, because of the square root which may become negative, therefore an asymptotically equivalent test is Durbin's m-test, which implies: get the OLS residuals, and regress them on all explanatory variables and the lagged residuals, and see if the coefficient on the lagged residuals is significant. This can be extended to higher order autoregression by including higher lags of the residuals [**Kme86**, p. 333]

58.3. Autoregressive Conditional Heteroskedasticity (ARCH)

An ARCH process is an error process which has covariance matrix $\sigma^2\boldsymbol{I}$ but where the errors are not independent. Question 150 gave us an example how this can be possible; the "birthday cake" distribution is a joint distribution of two variables

which have zero correlation, but which are not independent. We also saw that the conditional variance of the second variable depended on the outcome of the first.

PROBLEM 512. *The simplest ARCH process is*

$$\varepsilon_t = \mathfrak{u}_t\sqrt{\alpha_0 + \alpha_1\varepsilon_{t-1}^2} \tag{58.3.1}$$

where $\mathbf{u} \sim (\boldsymbol{o}, \boldsymbol{I})$ *is a white noise process independent of the pre-sample disturbance* ε_0*, and* $\alpha_0 > 0$ *and* $0 < \alpha_1 < 1$ *are two constants.*

• a. *2 points Show that* $\mathfrak{u}_t$ *is independent of all* ε_s *with* $0 \leq s < t$*. (Hint: use induction).*

ANSWER. By assumption, the statement is true for $s = 0$. Now assume it is true for $s - 1$, i.e., $\mathfrak{u}_t$ is independent of ε_{s-1}. Since $s < t$, therefore in particular $s \neq t$, $\mathfrak{u}_t$ is also independent of $\mathfrak{u}_s$. Therefore $\mathfrak{u}_t$ is independent of any function of ε_{s-1} and $\mathfrak{u}_s$, in particular, it is independent of $\mathfrak{u}_s\sqrt{\alpha_0 + \alpha_1\varepsilon_{s-1}^2} = \varepsilon_s$. □

• b. *2 points Show that* $\mathrm{E}[\varepsilon_t|\varepsilon_{t-1}] = 0$.

ANSWER. From (58.3.1) follows $\mathrm{E}[\varepsilon_t|\varepsilon_{t-1}] = \sqrt{\alpha_0 + \alpha_1\varepsilon_{t-1}^2}\,\mathrm{E}[\mathfrak{u}_t|\varepsilon_{t-1}] = \sqrt{\alpha_0 + \alpha_1\varepsilon_{t-1}^2}\,\mathrm{E}[\mathfrak{u}_t] = 0$. □

• c. *1 point Show that the unconditional* $\mathrm{E}[\varepsilon_t] = 0$.

ANSWER. Use law of iterated expectations $\mathrm{E}[\varepsilon_t] = \mathrm{E}\big[\mathrm{E}[\varepsilon_t|\varepsilon_{t-1}]\big] = \mathrm{E}[0] = 0$. □

• d. *1 point Show that for* $0 \leq s < t$, $\mathrm{E}[\varepsilon_t|\varepsilon_{t-1}, \varepsilon_s] = 0$.

ANSWER. This is exactly the same proof as in Part b. From (58.3.1) follows $\mathrm{E}[\varepsilon_t|\varepsilon_{t-1}, \varepsilon_s] = \sqrt{\alpha_0 + \alpha_1\varepsilon_{t-1}^2}\,\mathrm{E}[\mathfrak{u}_t|\varepsilon_{t-1}, \varepsilon_s] = \sqrt{\alpha_0 + \alpha_1\varepsilon_{t-1}^2}\,\mathrm{E}[\mathfrak{u}_t] = 0$. □

• e. *Show that in general* $\mathrm{E}[\varepsilon_t|\varepsilon_s] = 0$ *for all* $s < t$*. Hint: You are allowed to use, without proof, the following extension of the law of iterated expectations:*

$$\mathrm{E}\big[\mathrm{E}[\mathfrak{x}|\mathbf{y}, \boldsymbol{z}]\big|\mathbf{y}\big] = \mathrm{E}[\mathfrak{x}|\mathbf{y}]. \tag{58.3.2}$$

ANSWER. By (58.3.2), $\mathrm{E}[\varepsilon_t|\varepsilon_s] = \mathrm{E}\big[\mathrm{E}[\varepsilon_t|\varepsilon_{t-1}, \varepsilon_s]\big|\varepsilon_s\big] = \mathrm{E}[0|\varepsilon_s] = 0$. □

• f. *Show that* $\mathrm{cov}[\varepsilon_t, \varepsilon_s] = 0$ *for all* $s < t$*. Hint: Use Question 145.*

ANSWER. $\mathrm{cov}[\varepsilon_s, \varepsilon_t] = \mathrm{cov}\big[\varepsilon_s, \mathrm{E}[\varepsilon_t|\varepsilon_s]\big] = 0$ □

• g. *3 points Show that* $\mathrm{var}[\varepsilon_t] = \alpha_0 + \alpha_1\,\mathrm{var}[\varepsilon_{t-1}]$*. Hint: Use equation* (8.6.6).

ANSWER.

$$\begin{aligned}\mathrm{var}[\varepsilon_t] &= \mathrm{var}\big[\mathrm{E}[\varepsilon_t|\varepsilon_{t-1}]\big] + \mathrm{E}\big[\mathrm{var}[\varepsilon_t|\varepsilon_{t-1}]\big] = \mathrm{E}\big[\mathrm{var}[\varepsilon_t|\varepsilon_{t-1}]\big] = \\ &= \mathrm{E}\big[(\alpha_0 + \alpha_1\varepsilon_{t-1}^2)\,\mathrm{var}[\mathfrak{u}_t]\big] = \alpha_0 + \alpha_1\,\mathrm{E}[\varepsilon_{t-1}^2] = \alpha_0 + \alpha_1\,\mathrm{var}[\varepsilon_{t-1}]\end{aligned} \tag{58.3.3}$$

□

• h. *2 points Show that* $\mathrm{var}[\varepsilon_t] = (1 - \alpha_1^t)\frac{\alpha_0}{1-\alpha_1} + \alpha_1^t\,\mathrm{var}[\varepsilon_0]$*, in other words, the process converges towards a stationary process with variance* $\frac{\alpha_0}{1-\alpha_1}$.

ANSWER. By induction: assume it is true for $t - 1$, i.e., $\mathrm{var}[\varepsilon_{t-1}] = (1 - \alpha_1^{t-1})\frac{\alpha_0}{1-\alpha_1} + \alpha_1^{t-1}\,\mathrm{var}[\varepsilon_0]$. Then, by g,

$$\mathrm{var}[\varepsilon_t] = \alpha_0 + \alpha_1(1 - \alpha_1^{t-1})\frac{\alpha_0}{1-\alpha_1} + \alpha_1^t\,\mathrm{var}[\varepsilon_0] \tag{58.3.4}$$

$$= (1 - \alpha_1)\frac{\alpha_0}{1-\alpha_1} + (\alpha_1 - \alpha_1^t)\frac{\alpha_0}{1-\alpha_1} + \alpha_1^t\,\mathrm{var}[\varepsilon_0] \tag{58.3.5}$$

$$= (1 - \alpha_1^t)\frac{\alpha_0}{1-\alpha_1} + \alpha_1^t\,\mathrm{var}[\varepsilon_0]. \tag{58.3.6}$$

□

• i. *2 points Which kinds of economic timeseries are often modeled by ARCH processes?*

ANSWER. Processes with periods of higher turbulence wafting through randomly. □

Since the observations are no longer independent, the likelihood function is no longer the product of the one-observation likelihood functions, but the likelihood function conditional on the pre-sample values $\mathfrak{y}_0$ and $\boldsymbol{x}_0$ can be written down easily:

$$\log\ell(\mathbf{y}; \alpha_1, \alpha_2, \boldsymbol{\beta}) = -\frac{n}{2}\log 2\pi - \sum_{t=1}^{n}\log(\alpha_0 + \alpha_1(\mathfrak{y}_{t-1} - \boldsymbol{x}_{t-1}^\top\boldsymbol{\beta})^2) - \frac{1}{2}\sum_{t=1}^{n}\frac{(\mathfrak{y}_t - \boldsymbol{x}_t^\top\boldsymbol{\beta})^2}{\alpha_0 + \alpha_1(\mathfrak{y}_{t-1} - \boldsymbol{x}_{t-1}^\top\boldsymbol{\beta}} \tag{58.3.7}$$

The first-order conditions are complicated, but this can be maximized by numerical methods.

There is also a simpler feasible four-step estimation procedure available, see [**Gre97**, p. 571]. A good discussion of the ARCH processes is in [**End95**, pp. 139–165].

CHAPTER 59

Generalized Method of Moments Estimators

This follows mainly [**DM93**, Chapter 17]. A good and accessible treatment is [**Ḿ99**]. The textbook [**Hay00**] uses GMM as the organizing principle for all estimation methods except maximum likelihood.

A moment μ of a random variable y is the expected value of some function of y. Such a moment is therefore defined by the equation

$$\mathrm{E}[g(\mathsf{y}) - \mu] = 0. \tag{59.0.8}$$

The same parameter-defining function $g(\mathsf{y}) - \mu$ defines the *method of moments estimator* $\hat{\mu}$ of μ if one replaces the expected value in (59.0.8) with the sample mean of the elements of an observation vector $\mathbf{y}$ consisting of independent observations of y. In other words, $\hat{\mu}(\mathbf{y})$ is that value which satisfies $\frac{1}{n}\sum_{i=1}^{n}(g(\mathsf{y}_i) - \hat{\mu}) = 0$.

The *generalized* method of moments estimator extends this rule in several respects: the y_i no longer have to be i.i.d., the parameter-defining equations may be a system of equations defining more than one paramter at a time, there may be more parameter-defining functions than parameters (overidentification), and not only unconditional but also conditional moments are considered.

Under this definition, the OLS estimator is a GMM estimator. To show this, we will write the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ row by row as $\mathsf{y}_i = \mathbf{x}_i^\top\boldsymbol{\beta} + \varepsilon_i$, where $\mathbf{x}_i$ is, as in various earlier cases, the ith row of $\mathbf{X}$ written as a column vector. The basic property which makes least squares consistent is that the following *conditional* expectation is zero:

$$\mathrm{E}[\mathsf{y}_i - \mathbf{x}_i^\top\boldsymbol{\beta}|\mathbf{x}_i] = 0. \tag{59.0.9}$$

This is more information than just knowing that the unconditional expectation is zero. How can this additional information be used to define an estimator? From (59.0.9) follows that the unconditional expectation of the *product*

$$\mathcal{E}[\mathbf{x}_i(\mathsf{y}_i - \mathbf{x}_i^\top\boldsymbol{\beta})] = \boldsymbol{o}. \tag{59.0.10}$$

Replacing the expected value by the sample mean gives

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i(\mathsf{y}_i - \mathbf{x}_i^\top\hat{\boldsymbol{\beta}}) = \boldsymbol{o} \tag{59.0.11}$$

which can also be written as

$$\frac{1}{n}\begin{bmatrix}\mathbf{x}_1 & \cdots & \mathbf{x}_n\end{bmatrix}\begin{bmatrix}\mathsf{y}_1 - \mathbf{x}_1^\top\hat{\boldsymbol{\beta}}\\ \vdots \\ \mathsf{y}_n - \mathbf{x}_n^\top\hat{\boldsymbol{\beta}}\end{bmatrix} \equiv \frac{1}{n}\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \boldsymbol{o}. \tag{59.0.12}$$

These are exactly the OLS Normal Equations. This shows that OLS in the linear model is a GMM estimator.

Note that the rows of the $\mathbf{X}$-matrix play two different roles in this derivation: they appear in the equation $\mathsf{y}_i = \mathbf{x}_i^\top\boldsymbol{\beta} + \varepsilon_i$, and they are also the information set based on which the conditional expectation in (59.0.9) is formed. If this latter role is assumed by the rows of a different matrix of observations $\boldsymbol{W}$ then the GMM estimator becomes the Instrumental Variables Estimator.

Most maximum likelihood estimators are also GMM estimators. As long as the maxima are at the interior of the parameter region, the ML estimators solve the first order conditions, i.e., the Jacobian of the log likelihood function evaluated at these estimators is zero. But it follows from the theory of maximum likelihood estimation that the expected value of the Jacobian of the log likelihood function is zero.

Here are the general definitions and theorems, and as example their applications to the textbook example of the Gamma distribution in [**Gre97**, p. 518] and the Instrumental Variables estimator.

$\mathbf{y}$ is a vector of n observations created by a Data Generating Process (DGP) $\mu \in \mathcal{M}$. $\boldsymbol{\theta}$ is a k-vector of nonrandom parameters. A *parameter-defining function* $\boldsymbol{F}(\mathbf{y}, \boldsymbol{\theta})$ is a $n \times \ell$ matrix function with the following properties (a), (b), and (c):

(a) the ith row only depends on the ith observation y_i, i.e.,

$$\boldsymbol{F}(\mathbf{y}, \boldsymbol{\theta}) = \begin{bmatrix}\boldsymbol{f}_1^\top(\mathsf{y}_1, \boldsymbol{\theta})\\ \vdots \\ \boldsymbol{f}_n^\top(\mathsf{y}_n, \boldsymbol{\theta})\end{bmatrix} \tag{59.0.13}$$

Sometimes the $\boldsymbol{f}_i$ have identical functional form and only differ by the values of some exogenous variables, i.e., $\boldsymbol{f}_i(\mathsf{y}_i, \boldsymbol{\theta}) = g(\mathsf{y}_i, \mathbf{x}_i, \boldsymbol{\theta})$, but sometimes they have genuinely different functional forms.

In the Gamma-function example $\mathcal{M}$ is the set of all Gamma distributions, $\boldsymbol{\theta} = \begin{bmatrix} r & \lambda\end{bmatrix}^\top$ consists of the two parameters of the Gamma distribution, $\ell = k = 2$, and

the parameter-defining function has the rows

$$\boldsymbol{f}_i(\mathsf{y}_i,\boldsymbol{\theta}) = \begin{bmatrix} \mathsf{y}_i - \frac{r}{\lambda} \\ \frac{1}{\mathsf{y}_i} - \frac{\lambda}{r-1} \end{bmatrix} \qquad \text{so that} \quad \boldsymbol{F}(\mathsf{y}_i,\boldsymbol{\theta}) = \begin{bmatrix} \mathsf{y}_1 - \frac{r}{\lambda} & \frac{1}{\mathsf{y}_1} - \frac{\lambda}{r-1} \\ \vdots & \vdots \\ \mathsf{y}_n - \frac{r}{\lambda} & \frac{1}{\mathsf{y}_n} - \frac{\lambda}{r-1} \end{bmatrix}. \tag{59.0.14}$$

In the IV case, $\boldsymbol{\theta} = \boldsymbol{\beta}$ and ℓ is the number of instruments. If we split $\mathbf{X}$ and $\boldsymbol{W}$ into their rows

$$\mathbf{X} = \begin{bmatrix} \boldsymbol{x}_1^\top \\ \vdots \\ \boldsymbol{x}_n^\top \end{bmatrix} \qquad \text{and} \quad \boldsymbol{W} = \begin{bmatrix} \boldsymbol{w}_1^\top \\ \vdots \\ \boldsymbol{w}_n^\top \end{bmatrix} \tag{59.0.15}$$

then $\boldsymbol{f}_i(\mathsf{y}_i,\boldsymbol{\beta}) = \boldsymbol{w}_i(\mathsf{y}_i - \boldsymbol{x}_i^\top\boldsymbol{\beta})$. This gives

$$\boldsymbol{F}(\mathsf{y},\boldsymbol{\beta}) = \begin{bmatrix} (\mathsf{y}_1 - \boldsymbol{x}_1^\top\boldsymbol{\beta})\boldsymbol{w}_1^\top \\ \vdots \\ (\mathsf{y}_n - \boldsymbol{x}_n^\top\boldsymbol{\beta})\boldsymbol{w}_n^\top \end{bmatrix} = \operatorname{diag}(\mathsf{y} - \mathbf{X}\boldsymbol{\beta})\boldsymbol{W}. \tag{59.0.16}$$

(b) The vector functions $\boldsymbol{f}_i(\mathsf{y}_i,\boldsymbol{\theta})$ must be such that the true value of the parameter vector $\boldsymbol{\theta}_\mu$ satisfies

$$\mathcal{E}[\boldsymbol{f}_i(\mathsf{y}_i,\boldsymbol{\theta}_\mu)] = \boldsymbol{o} \tag{59.0.17}$$

for all i, while any other parameter vector $\boldsymbol{\theta} \neq \boldsymbol{\theta}_\mu$ gives $\mathcal{E}[\boldsymbol{f}_i(\mathsf{y}_i,\boldsymbol{\theta})] \neq \boldsymbol{o}$.

In the Gamma example (59.0.17) follows from the fact that the moments of the Gamma distribution are $\mathrm{E}[\mathsf{y}] = \frac{r}{\lambda}$ and $\mathrm{E}[\frac{1}{\mathsf{y}_i}] = \frac{\lambda}{r-1}$. It is also easy to see that r and λ are characterized by these two relations; given $\mathrm{E}[\mathsf{y}] = \mu$ and $\mathrm{E}[\frac{1}{\mathsf{y}_i}] = \nu$ one can solve for $r = \frac{\mu\nu}{\mu\nu-1}$ and $\lambda = \frac{\nu}{\mu\nu-1}$.

In the IV model, (59.0.17) is satisfied if the ε_i have zero expectation conditionally on $\boldsymbol{w}_i$, and uniqueness is condition (52.0.3) requiring that $\operatorname{plim}\frac{1}{n}\boldsymbol{W}_n^\top\mathbf{X}_n$ exists, is nonrandom and has full column rank. (In the 781 handout Winter 1998, (52.0.3) was equation (246) on p. 154).

Next we need a recipe how to construct an estimator from this parameter-defining function. Let us first discuss the case $k = \ell$ (exact identification). The GMM estimator $\hat{\boldsymbol{\theta}}$ defined by $\boldsymbol{F}$ satisfies

$$\frac{1}{n}\boldsymbol{F}^\top(\mathsf{y},\hat{\boldsymbol{\theta}})\boldsymbol{\iota} = \boldsymbol{o} \tag{59.0.18}$$

which can also be written in the form

$$\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{f}_i(\mathsf{y}_i,\hat{\boldsymbol{\theta}}) = \boldsymbol{o}. \tag{59.0.19}$$

Assumption (c) for a parameter-defining function is that there is only one $\hat{\boldsymbol{\theta}}$ satisfying (59.0.18).

For IV,

$$\boldsymbol{F}^\top(\mathsf{y},\tilde{\boldsymbol{\beta}})\boldsymbol{\iota} = \boldsymbol{W}^\top\operatorname{diag}(\mathsf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})\boldsymbol{\iota} = \boldsymbol{W}^\top(\mathsf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) \tag{59.0.20}$$

If there are as many instruments as explanatory variables, setting this zero gives the normal equation for the simple IV estimator $\boldsymbol{W}^\top(\mathsf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) = \boldsymbol{o}$.

In the case $\ell > k$, (59.0.17) still holds, but the system of equations (59.0.18) no longer has a solution: there are $\ell > k$ relationships for the k parameters. In order to handle this situation, we need to specify what qualifies as a *weighting matrix*. The symmetric positive definite $\ell \times \ell$ matrix $\boldsymbol{A}(\mathsf{y})$ is a weighting matrix if it has a nonrandom positive definite plim, called $\boldsymbol{A}_0(\mathsf{y}) = \operatorname{plim}_{n\to\infty}\boldsymbol{A}(\mathsf{y})$. Instead of (59.0.18), now the following equation serves to define $\hat{\boldsymbol{\theta}}$:

$$\hat{\boldsymbol{\theta}} = \operatorname{argmin}\boldsymbol{\iota}^\top\boldsymbol{F}(\mathsf{y},\hat{\boldsymbol{\theta}})\boldsymbol{A}(\mathsf{y})\boldsymbol{F}^\top(\mathsf{y},\hat{\boldsymbol{\theta}})\boldsymbol{\iota} \tag{59.0.21}$$

In this case, condition (c) for a parameter-defining equation reads that there is only one $\hat{\boldsymbol{\theta}}$ which minimizes this criterion function.

For IV, $\boldsymbol{A}(\mathsf{y})$ does not depend on y but is $\frac{1}{n}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}$. Therefore $\boldsymbol{A}_0 = \operatorname{plim}(\frac{1}{n}\boldsymbol{W}^\top\boldsymbol{W})^{-1}$, and (59.0.21) becomes $\tilde{\boldsymbol{\beta}} = \operatorname{argmin}(\mathsf{y}-\mathbf{X}^\top\boldsymbol{\beta})^\top\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top(\mathsf{y}-\mathbf{X}^\top\boldsymbol{\beta})$, which is indeed the quadratic form minimized by the generalkized instrumental variables estimator.

In order to convert the Gamma-function example into an overidentified system, we add a third relation:

$$\boldsymbol{F}(\mathsf{y}_i,\boldsymbol{\theta}) = \begin{bmatrix} \mathsf{y}_1 - \frac{r}{\lambda} & \frac{1}{\mathsf{y}_1} - \frac{\lambda}{r-1} & \mathsf{y}_1^2 - \frac{r(r+1)}{\lambda^2} \\ \vdots & \vdots & \vdots \\ \mathsf{y}_n - \frac{r}{\lambda} & \frac{1}{\mathsf{y}_n} - \frac{\lambda}{r-1} & \mathsf{y}_n^2 - \frac{r(r+1)}{\lambda^2} \end{bmatrix}. \tag{59.0.22}$$

In this case here is possible to compute the asymptotic covariance; but in real-life situations this covariance matrix is estimated using a preliminary consistent estimator of the parameters, as [**Gre97**] does it. Most GMM estimators depend on such a consistent pre-estimator.

The GMM estimator $\hat{\boldsymbol{\theta}}$ defined in this way is a particular kind of a M-estimator, and many of its properties follow from the general theory of M-estimators. We need some more definitions. Define the plim of the Jacobian of the parameter-defining mapping $\boldsymbol{D} = \operatorname{plim}\frac{1}{n}\partial\boldsymbol{F}^\top\boldsymbol{\iota}/\partial\boldsymbol{\theta}^\top$ and the plim of the covariance matrix of $\frac{1}{\sqrt{n}}\boldsymbol{F}^\top\boldsymbol{\iota}$ is $\boldsymbol{\Psi} = \operatorname{plim}\frac{1}{n}\boldsymbol{F}^\top\boldsymbol{F}$.

For IV, $\boldsymbol{D} = \operatorname{plim} \frac{1}{n} \frac{\partial \boldsymbol{W}^\top(\mathfrak{y}-\mathbf{X}\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^\top} = -\operatorname{plim}_{n\to\infty} \frac{1}{n}\boldsymbol{W}^\top\mathbf{X}$, and

$$\boldsymbol{\Psi} = \operatorname{plim}\Big(\frac{1}{n}\boldsymbol{W}^\top \operatorname{diag}(\mathfrak{y}-\mathbf{X}\boldsymbol{\beta})\operatorname{diag}(\mathfrak{y}-\mathbf{X}\boldsymbol{\beta})\boldsymbol{W}\Big) = \operatorname{plim}\frac{1}{n}\boldsymbol{W}^\top\boldsymbol{\Omega}\boldsymbol{W}$$

where $\boldsymbol{\Omega}$ is the diagonal matrix with typical element $\mathrm{E}[(\mathfrak{y}_i - \mathbf{x}_i^\top\boldsymbol{\beta})^2]$, i.e., $\boldsymbol{\Omega} = \mathcal{V}[\boldsymbol{\varepsilon}]$.

With this notation the theory of M-estimators gives us the following result: The asymptotic $\mathcal{MSE}$-matrix of the GMM is

$$(\boldsymbol{D}^\top\boldsymbol{A}_0\boldsymbol{D})^{-1}\boldsymbol{D}^\top\boldsymbol{A}_0\boldsymbol{\Psi}\boldsymbol{A}_0\boldsymbol{D}(\boldsymbol{D}^\top\boldsymbol{A}_0\boldsymbol{D})^{-1} \tag{59.0.23}$$

This gives the following expression for the plim of $\sqrt{n}$ times the sampling error of the IV estimator:

$$\operatorname{plim}(\frac{1}{n}\mathbf{X}^\top\boldsymbol{W}(\frac{1}{n}\boldsymbol{W}^\top\boldsymbol{W})^{-1}\frac{1}{n}\boldsymbol{W}^\top\mathbf{X})^{-1}\frac{1}{n}\mathbf{X}^\top\boldsymbol{W}(\frac{1}{n}\boldsymbol{W}^\top\boldsymbol{W})^{-1}\frac{1}{n}\boldsymbol{W}^\top\boldsymbol{\Omega}\boldsymbol{W}(\frac{1}{n}\boldsymbol{W}^\top\boldsymbol{W})^{-1}\frac{1}{n}\boldsymbol{W}^\top \tag{59.0.24}$$

$$= \operatorname{plim} n(\mathbf{X}^\top\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top\boldsymbol{\Omega}\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top\mathbf{X}(\mathbf{X}^\top\boldsymbol{W}(\boldsymbol{W} \tag{59.0.25}$$

The asymptotic $\mathcal{MSE}$ matrix can be obtained fom this by dividing by n. An estimate of the asymptotic covariance matrix is therefore

$$(\mathbf{X}^\top\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top\boldsymbol{\Omega}\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top\mathbf{X}(\mathbf{X}^\top\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1} \tag{59.0.26}$$

This is [**DM93**, (17.36) on p. 596].

The best choice of such a weighting matrix is $\boldsymbol{A}_0 = \boldsymbol{\Psi}^{-1}$, in which case (59.0.23) simplifies to $(\boldsymbol{D}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{D})^{-1} = (\boldsymbol{D}^\top\boldsymbol{A}_0\boldsymbol{D})^{-1}$.

The criterion function which the *optimal* IV estimator must minimize, in the presence of unknown heteroskedasticity, is therefore

$$(\mathfrak{y}-\mathbf{X}\boldsymbol{\beta})^\top\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{\Omega}\boldsymbol{W})^{-1}\boldsymbol{W}^\top(y-\mathbf{X}\boldsymbol{\beta}) \tag{59.0.27}$$

The first-order conditions are

$$\mathbf{X}^\top\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{\Omega}\boldsymbol{W})^{-1}\boldsymbol{W}^\top(y-\mathbf{X}\boldsymbol{\beta}) = \boldsymbol{o} \tag{59.0.28}$$

and the optimally weighted IVA is

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}^\top\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{\Omega}\boldsymbol{W})^{-1}\boldsymbol{W}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{\Omega}\boldsymbol{W})^{-1}\boldsymbol{W}^\top y \tag{59.0.29}$$

In this, $\boldsymbol{\Omega}$ can be replaced by an inconsistent estimate, for instance the diagonal matrix with the squared 2SLS residuals in the diagonal, this is what [**DM93**] refer to as H2SLS. In the simple IV case, this estimator is the simple IV estimator again. In other words, we need more than the minimum number of instruments to be able to take advantage of the estimated heteroskedasticity. [**Cra83**] proposes in the OLS case, i.e., $\boldsymbol{W} = \mathbf{X}$, to use the squares of the regressors etc. as additional instruments.

To show this optimality take some square nonsingular $\boldsymbol{Q}$ with $\boldsymbol{\Psi} = \boldsymbol{Q}\boldsymbol{Q}^\top$ and define $\boldsymbol{P} = \boldsymbol{Q}^{-1}$. Then

$$(\boldsymbol{D}^\top\boldsymbol{A}_0\boldsymbol{D})^{-1}\boldsymbol{D}^\top\boldsymbol{A}_0\boldsymbol{\Psi}\boldsymbol{A}_0\boldsymbol{D}(\boldsymbol{D}^\top\boldsymbol{A}_0\boldsymbol{D})^{-1} - (\boldsymbol{D}^\top\boldsymbol{A}_0\boldsymbol{D})^{-1} = \tag{59.0.30}$$

$$= (\boldsymbol{D}^\top\boldsymbol{A}_0\boldsymbol{D})^{-1}\boldsymbol{D}^\top\boldsymbol{A}_0\Big(\boldsymbol{\Psi} - \boldsymbol{D}(\boldsymbol{D}^\top\boldsymbol{A}_0\boldsymbol{D})^{-1}\boldsymbol{D}^\top\Big)\boldsymbol{A}_0\boldsymbol{D}(\boldsymbol{D}^\top\boldsymbol{A}_0\boldsymbol{D})^{-1} \tag{59.0.31}$$

Now the middle matrix can be written as $\boldsymbol{P}\Big(\boldsymbol{I} - \boldsymbol{Q}\boldsymbol{D}(\boldsymbol{D}^\top\boldsymbol{Q}^\top\boldsymbol{Q}\boldsymbol{D})^{-1}\boldsymbol{D}^\top\boldsymbol{Q}^\top\Big)\boldsymbol{P}^\top$ which is nonnegative definite because the matrix in the middle is idempotent.

The advantage of the GMM is that it is valid for many different DGP's. In this respect it is the opposite of the maximum likelihood estimator, which needs a very specific DGP. The more broadly the DGP can be defined, the better the chances are that the GMM etimator is efficient, i.e., in large samples as good as maximum likelihood.

CHAPTER 60

Bootstrap Estimators

The bootstrap method is an important general estimation principle, which can serve as as an alternative to reliance on the asymptotic properties of an estimator. Assume you have a $n \times k$ data matrix $\boldsymbol{X}$ each row of which is an independent observation from the same unknown probability distribution, characterized by the cumulative distribution function F. Using this data set you want to draw conclusions about the distribution of some statistic $\theta(\mathbf{x})$ where $\mathbf{x} \sim F$.

The "bootstrap" estimation principle is very simple: as your estimate of the distribution of $\mathbf{x}$ you use F_n, the empirical distribution of the given sample $\boldsymbol{X}$, i.e. that probability distribution which assigns probability mass $1/n$ to each of the k-dimensional observation points $\boldsymbol{x}_t$ (or, if the observation $\boldsymbol{x}_t$ occured more than once, say j times, then you assign the probability mass j/n to this point). This empirical distribution function has been called the nonparametric maximum likelihood estimate of F. And your estimate of the distribution of $\theta(\mathbf{x})$ is that distribution which derives from this empirical distribution function. Just like the maximum likelihood principle, this principle is *deceptively* simple but has some deep probability theoretic foundations.

In simple cases, this is a widely used principle; the sample mean, for instance, is the expected value of the empirical distribution, the same is true about the sample variance (divisor is n) or sample median etc. But as soon as θ becomes a little more complicated, and one wants more complex measures of its distribution, such as the standard deviation of a complicated function of $\mathbf{x}$, or some confidence intervals, an analytical expression for this bootstrap estimate is prohibitively complex.

But with the availability of modern computing power, an alternative to the analytical evaluation is feasible: draw a large random sample from the empirical distribution, evaluate $\theta(\mathbf{x})$ for each $\mathbf{x}$ in this artificially generated random sample, and use these datapoints to construct the distribution function of $\theta(\mathbf{x})$. A random sample from the empirical distribution is merely a random drawing from the given values with replacement. This requires computing power, usually one has to resample between 1,000 and 10,000 times to get accurate results, but one does not

need to do complicated math, and these so-called nonparametric bootstrap results are very close to the theoretical results wherever those are available.

So far we have been discussing the situation that all observations come from the same population. In the regression context this is not the case. In the OLS model with i.i.d. disturbances, the observations of the independent variable $\mathtt{y}_t$ have different expected values, i.e., they do not come from the same population. On the other hand, the *disturbances* come from the same population. Unfortunately, they are not observed, but it turns out that one can successfully apply bootstrap methods here by first computing the OLS residuals and then drawing from these residuals to get pseudo-datapoints and to run the regression on those. This is a surprising and strong result; but one has to be careful here that the OLS model is correctly specified. For instance, if there is heteroskedasticity which is not corrected for, then the resampling would no longer be uniform, and the bootstrap least squares estimates are inconsistent.

The jackknife is a much more complicated concept; it was originally invented and is often still introduced as a device to reduce bias, but [**Efr82**, p. 10] claims that this motivation is mistaken. It is an alternative to the bootstrap, in which random sampling is replaced by a symmetric systematic "sampling" of datasets which are by 1 observation smaller than the original one: namely, n drawings with one observation left out in each. In certain situations this is as good as bootstrapping, but much cheaper. I third concept is cross-validation.

There is a new book out, [**ET93**], for which the authors also have written bootstrap and jackknife functions for `Splus`, to be found if one does `attach("/home/econ/ehrbar/`

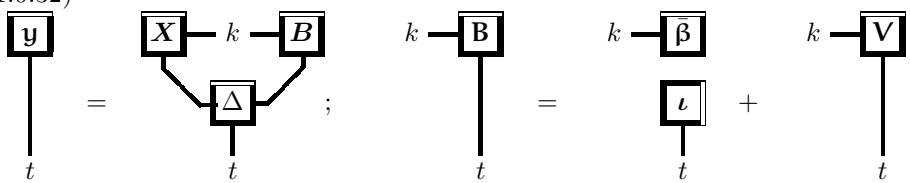
CHAPTER 61

Random Coefficients

The random coefficient model first developed in [**HH68**] cannot be written in the form $\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ because each observation has a different $\boldsymbol{\beta}$. Therefore we have to write it observation by observation: $\mathrm{y}_t = \boldsymbol{x}_t{}^\top \boldsymbol{\beta}_t$ (no separate disturbance term), where $\boldsymbol{\beta}_t = \bar{\boldsymbol{\beta}} + \boldsymbol{\nu}_t$ with $\boldsymbol{\nu}_t \sim (\boldsymbol{o}, \tau^2\boldsymbol{\Sigma})$. For $s \neq t$, $\boldsymbol{\nu}_s$ and $\boldsymbol{\nu}_t$ are uncorrelated. By re-grouping terms one gets $\mathrm{y}_t = \boldsymbol{x}_t{}^\top \bar{\boldsymbol{\beta}} + \boldsymbol{x}_t{}^\top \boldsymbol{\nu}_t = \boldsymbol{x}_t{}^\top \bar{\boldsymbol{\beta}} + \varepsilon_t$ where $\varepsilon_t = \boldsymbol{x}_t{}^\top \boldsymbol{\nu}_t$, hence $\operatorname{var}[\varepsilon_t] = \tau^2 \boldsymbol{x}_t{}^\top \boldsymbol{\Sigma} \boldsymbol{x}_t$, and for $s \neq t$, ε_s and ε_t are uncorrelated.

In tiles, this model is

(61.0.32)



Estimation under the assumption $\boldsymbol{\Sigma}$ is known: To estimate $\bar{\boldsymbol{\beta}}$ one can use the heteroskedastic model with error variances $\tau^2 \boldsymbol{x}_t{}^\top \boldsymbol{\Sigma} \boldsymbol{x}_t$, call the resulting estimate $\hat{\bar{\boldsymbol{\beta}}}$. The formula for the best linear unbiased predictor of $\boldsymbol{\beta}_t$ itself can be derived (heuristically) as follows: Assume for a moment that $\bar{\boldsymbol{\beta}}$ is known: then the model can be written as $\mathrm{y}_t - \boldsymbol{x}_t{}^\top \bar{\boldsymbol{\beta}} = \boldsymbol{x}_t{}^\top \boldsymbol{\nu}_t$. Then we can use the formula for the Best Linear Predictor, equation (**??**), applied to the situation

$$\begin{bmatrix} \boldsymbol{x}_t{}^\top \boldsymbol{\nu}_t \\ \boldsymbol{\nu}_t \end{bmatrix} \sim \begin{bmatrix} 0 \\ \boldsymbol{o} \end{bmatrix}, \tau^2 \begin{bmatrix} \boldsymbol{x}_t{}^\top \boldsymbol{\Sigma} \boldsymbol{x}_t & \boldsymbol{x}_t{}^\top \boldsymbol{\Sigma} \\ \boldsymbol{\Sigma} \boldsymbol{x}_t & \boldsymbol{\Sigma} \end{bmatrix} \tag{61.0.33}$$

where $\boldsymbol{x}_t{}^\top \boldsymbol{\nu}_t$ is observed, its value is $\mathrm{y}_t - \boldsymbol{x}_t{}^\top \bar{\boldsymbol{\beta}}$, but $\boldsymbol{\nu}_t$ is not. Note that here we predict a whole vector on the basis of *one* linear combination of its elements only. This predictor is

$$\boldsymbol{\nu}_t^* = \boldsymbol{\Sigma} \boldsymbol{x}_t (\boldsymbol{x}_t{}^\top \boldsymbol{\Sigma} \boldsymbol{x}_t)^{-1} (\mathrm{y}_t - \boldsymbol{x}_t{}^\top \bar{\boldsymbol{\beta}}) \tag{61.0.34}$$

If one adds $\bar{\boldsymbol{\beta}}$ to both sides, one obtains

$$\boldsymbol{\beta}_t^* = \bar{\boldsymbol{\beta}} + \boldsymbol{\Sigma} \boldsymbol{x}_t (\boldsymbol{x}_t{}^\top \boldsymbol{\Sigma} \boldsymbol{x}_t)^{-1} (\mathrm{y}_t - \boldsymbol{x}_t{}^\top \bar{\boldsymbol{\beta}}) \tag{61.0.35}$$

If one now replaces $\bar{\boldsymbol{\beta}}$ by $\hat{\bar{\boldsymbol{\beta}}}$, one obtains the formula for the predictor given in [**JHG+88**, p. 438]:

$$\boldsymbol{\beta}_t^* = \hat{\bar{\boldsymbol{\beta}}} + \boldsymbol{\Sigma} \boldsymbol{x}_t (\boldsymbol{x}_t{}^\top \boldsymbol{\Sigma} \boldsymbol{x}_t)^{-1} (\mathrm{y}_t - \boldsymbol{x}_t{}^\top \hat{\bar{\boldsymbol{\beta}}}). \tag{61.0.36}$$

Usually, of course, $\boldsymbol{\Sigma}$ is unknown. But if the number of observations is large enough, one can estimate the elements of the covariance matrix $\tau^2\boldsymbol{\Sigma}$. This is the fact which gives relevance to this model. Write $\tau^2 \boldsymbol{x}_t{}^\top \boldsymbol{\Sigma} \boldsymbol{x}_t = \tau^2 \operatorname{tr} \boldsymbol{x}_t{}^\top \boldsymbol{\Sigma} \boldsymbol{x}_t = \tau^2 \operatorname{tr} \boldsymbol{x}_t \boldsymbol{x}_t{}^\top \boldsymbol{\Sigma} = \boldsymbol{z}_t^\top \boldsymbol{\alpha}$, where $\boldsymbol{z}_t$ is the vector containing the unique elements of the symmetric matrix $\boldsymbol{x}_t \boldsymbol{x}_t{}^\top$ with those elements not located on the diagonal multiplied by the factor 2, since they occur twice in the matrix, and $\boldsymbol{\alpha}$ contains the corresponding unique elements of $\tau^2\boldsymbol{\Sigma}$ (but no factors 2 here). For instance, if there are three variables, then $\tau^2 \boldsymbol{x}_t{}^\top \boldsymbol{\Sigma} \boldsymbol{x}_t = x_{t1}^2 \tau_{11} + x_{t2}^2 \tau_{22} + x_{t3}^2 \tau_{33} + 2x_{t1}x_{t2}\tau_{12} + 2x_{t1}x_{t3}\tau_{13} + 2x_{t2}x_{t3}\tau_{23}$, where τ_{ij} are the elements of $\tau^2\boldsymbol{\Sigma}$. Therefore $\boldsymbol{z}_t$ consists of $x_{t1}^2, x_{t2}^2, x_{t3}^2, 2x_{t1}x_{t2}, 2x_{t1}x_{t3}, 2x_{t2}x_{t3}$, and $\boldsymbol{\alpha}^\top = [\tau_{11}, \tau_{22}, \tau_{33}, \tau_{12}, \tau_{13}, \tau_{23}]$. Then construct the matrix $\boldsymbol{Z}$ which has as its tth row the vector $\boldsymbol{z}_t^\top$; it follows $\mathcal{V}[\boldsymbol{\varepsilon}] = \operatorname{diag}(\boldsymbol{\gamma})$ where $\boldsymbol{\gamma} = \boldsymbol{Z}\boldsymbol{\alpha}$.

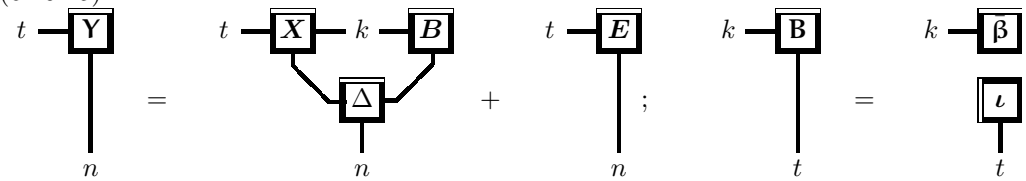
Using this notation and defining, as usual, $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$, and writing $\boldsymbol{m}_t$ for the tth column vector of $\boldsymbol{M}$, and furthermore writing $\boldsymbol{Q}$ for the matrix whose elements are the squares of the elements of $\boldsymbol{M}$, and writing $\boldsymbol{\delta}_t$ for the vector that has 1 in the tth place and 0 elsewhere, one can derive:

$$\begin{aligned}
\mathrm{E}[\hat{\varepsilon}_t^2] &= \mathrm{E}[(\boldsymbol{\delta}_t^\top \hat{\boldsymbol{\varepsilon}})^2] && (61.0.37)\\
&= \mathrm{E}[\hat{\boldsymbol{\varepsilon}}^\top \boldsymbol{\delta}_t \boldsymbol{\delta}_t^\top \hat{\boldsymbol{\varepsilon}}] && (61.0.38)\\
&= \mathrm{E}[\boldsymbol{\varepsilon}^\top \boldsymbol{M} \boldsymbol{\delta}_t \boldsymbol{\delta}_t^\top \boldsymbol{M} \boldsymbol{\varepsilon}] && (61.0.39)\\
&= \mathrm{E}[\operatorname{tr} \boldsymbol{M} \boldsymbol{\delta}_t \boldsymbol{\delta}_t^\top \boldsymbol{M} \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top] && (61.0.40)\\
&= \operatorname{tr} \boldsymbol{M} \boldsymbol{\delta}_t \boldsymbol{\delta}_t^\top \boldsymbol{M} \operatorname{diag}(\boldsymbol{\gamma}) = \operatorname{tr} \boldsymbol{m}_t \boldsymbol{m}^\top \operatorname{diag}(\boldsymbol{\gamma}) && (61.0.41)\\
&= \boldsymbol{m}_t^\top \operatorname{diag}(\boldsymbol{\gamma}) \boldsymbol{m}_t = m_{t1}\gamma_1 m_{1t} + \cdots + m_{tn}\gamma_n m_{nt} && (61.0.42)\\
&= \boldsymbol{q}_t^\top \boldsymbol{\gamma} = \boldsymbol{q}_t^\top \boldsymbol{Z}\boldsymbol{\alpha} && (61.0.43)\\
\mathrm{E}[\hat{\boldsymbol{\varepsilon}}^2] &= \boldsymbol{Q}\boldsymbol{Z}\boldsymbol{\alpha}, && (61.0.44)
\end{aligned}$$

where $\boldsymbol{\alpha}$ is as above. This allows one to get an estimate of $\boldsymbol{\alpha}$ by regressing the vector $[\hat{\varepsilon}_1^2, \ldots, \hat{\varepsilon}_n^2]^\top$ on $\boldsymbol{Q}\boldsymbol{Z}$, and then to use $\boldsymbol{Z}\boldsymbol{\alpha}$ to get an estimate of the variances $\tau^2 \boldsymbol{x}^\top \boldsymbol{\Sigma} \boldsymbol{x}$. Unfortunately, the estimated covariance matrix one gets in this way may not be nonnegative definite.

[**Gre97**, p. 669–674] brings this model in the following form:

(61.0.45)



PROBLEM 513. *Let $\mathbf{y}_i$ be the ith column of $\mathbf{Y}$. The random coefficients model as discussed in* [**Gre97**, p. 669–674] *specifies $\mathbf{y}_i = \boldsymbol{X}_i\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i$ with $\boldsymbol{\varepsilon}_i \sim (\boldsymbol{o}, \sigma_i^2\boldsymbol{I})$ and $\boldsymbol{\varepsilon}_i$ uncorrelated with $\boldsymbol{\varepsilon}_j$ for $i \neq j$. Furthermore also $\boldsymbol{\beta}_i$ is random, write it as $\boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{\nu}_i$, with $\boldsymbol{\nu}_i \sim (\boldsymbol{o}, \tau^2\boldsymbol{\Gamma})$ with a positive definite $\boldsymbol{\Gamma}$, and again $\boldsymbol{\nu}_i$ uncorrelated with $\boldsymbol{\nu}_j$ for $i \neq j$. Furthermore, all $\boldsymbol{\nu}_i$ are uncorrelated with all $\boldsymbol{\varepsilon}_j$.*

• a. *4 points In this model the disturbance term is really $\boldsymbol{w}_i = \boldsymbol{\varepsilon}_i + \boldsymbol{X}\boldsymbol{\nu}_i$, which has covariance matrix $\mathcal{V}[\boldsymbol{w}_i] = \sigma_i^2\boldsymbol{I} + \tau^2\boldsymbol{X}_i\boldsymbol{\Gamma}\boldsymbol{X}_i^\top$. As a preliminary calculation for the next part of the question show that*

$$\boldsymbol{X}_i^\top(\mathcal{V}[\boldsymbol{w}_i])^{-1} = \frac{1}{\tau^2}\boldsymbol{\Gamma}^{-1}(\boldsymbol{X}_i^\top\boldsymbol{X}_i + \kappa_i^2\boldsymbol{\Gamma}^{-1})^{-1}\boldsymbol{X}_i^\top \tag{61.0.46}$$

where $\kappa_i^2 = \sigma_i^2/\tau^2$. You are allowed to use, without proof, formula (A.8.13)*, which reads for inverses, not generalized inverses:*

$$\Big(\boldsymbol{A} + \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}\Big)^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{D} + \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}\boldsymbol{C}\boldsymbol{A}^{-1} \tag{61.0.47}$$

ANSWER. In (61.0.47) set $\boldsymbol{A} = \sigma_i^2\boldsymbol{I}$, $\boldsymbol{B} = \boldsymbol{X}_i$, $\boldsymbol{D}^{-1} = \tau^2\boldsymbol{\Gamma}$, and $\boldsymbol{C} = \boldsymbol{X}_i^\top$ to get

$$(\mathcal{V}[\boldsymbol{w}_i])^{-1} = \frac{1}{\sigma_i^2}(\boldsymbol{I} - \boldsymbol{X}_i(\boldsymbol{X}_i^\top\boldsymbol{X}_i + \kappa_i^2\boldsymbol{\Gamma}^{-1})^{-1}\boldsymbol{X}_i^\top) \tag{61.0.48}$$

Premultiply this by $\boldsymbol{X}_i^\top$ and add and subtract the same term:

$$\boldsymbol{X}_i^\top(\mathcal{V}[\boldsymbol{w}_i])^{-1} = \frac{1}{\sigma_i^2}\boldsymbol{X}_i^\top - \frac{1}{\sigma_i^2}(\boldsymbol{X}_i^\top\boldsymbol{X}_i + \kappa_i^2\boldsymbol{\Gamma}^{-1} - \kappa_i^2\boldsymbol{\Gamma}^{-1})(\boldsymbol{X}_i^\top\boldsymbol{X}_i + \kappa_i^2\boldsymbol{\Gamma}^{-1})^{-1}\boldsymbol{X}_i^\top \tag{61.0.49}$$

(61.0.50)

$$= \frac{1}{\sigma_i^2}\boldsymbol{X}_i^\top - \frac{1}{\sigma_i^2}\boldsymbol{X}_i^\top + \frac{1}{\sigma_i^2}\kappa_i^2\boldsymbol{\Gamma}^{-1}(\boldsymbol{X}_i^\top\boldsymbol{X}_i + \kappa_i^2\boldsymbol{\Gamma}^{-1})^{-1}\boldsymbol{X}_i^\top = \frac{1}{\tau^2}\boldsymbol{\Gamma}^{-1}(\boldsymbol{X}_i^\top\boldsymbol{X}_i + \kappa_i^2\boldsymbol{\Gamma}^{-1})^{-1}\boldsymbol{X}_i^\top.$$

□

• b. *2 points From* (61.0.46) *derive:*

$$\Big(\boldsymbol{X}_i^\top(\mathcal{V}[\boldsymbol{w}_i])^{-1}\boldsymbol{X}_i\Big)^{-1} = \sigma_i^2(\boldsymbol{X}_i^\top\boldsymbol{X}_i)^{-1} + \tau^2\boldsymbol{\Gamma} \tag{61.0.51}$$

ANSWER. From (61.0.46) follows

$$\boldsymbol{X}_i^\top(\mathcal{V}[\boldsymbol{w}_i])^{-1}\boldsymbol{X}_i = \frac{1}{\tau^2}\boldsymbol{\Gamma}^{-1}(\boldsymbol{X}_i^\top\boldsymbol{X}_i + \kappa_i^2\boldsymbol{\Gamma}^{-1})^{-1}\boldsymbol{X}_i^\top\boldsymbol{X}_i \tag{61.0.52}$$

This is the product of three matrices each of which has an inverse:

(61.0.53)

$$\Big(\boldsymbol{X}_i^\top(\mathcal{V}[\boldsymbol{w}_i])^{-1}\boldsymbol{X}_i\Big)^{-1} = \tau^2(\boldsymbol{X}_i^\top\boldsymbol{X}_i)^{-1}(\boldsymbol{X}_i^\top\boldsymbol{X}_i + \kappa_i^2\boldsymbol{\Gamma}^{-1})\boldsymbol{\Gamma} = (\tau^2\boldsymbol{I} + \sigma_i^2(\boldsymbol{X}_i^\top\boldsymbol{X}_i)^{-1}\boldsymbol{\Gamma}^{-1})\boldsymbol{\Gamma} = \tau^2\boldsymbol{\Gamma} + \sigma_i^2(\boldsymbol{X}_i^\top\boldsymbol{X}_i)^{-}$$

□

• c. *2 points Show that from* (61.0.46) *also follows that The GLS of each column of $\mathbf{Y}$ separately is the OLS $\hat{\boldsymbol{\beta}}_i = (\boldsymbol{X}_i^\top\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i^\top\mathbf{y}_i$.*

• d. *2 points Show that $\mathcal{V}[\hat{\boldsymbol{\beta}}_i] = \sigma_i^2(\boldsymbol{X}_i^\top\boldsymbol{X}_i)^{-1} + \tau^2\boldsymbol{\Gamma}$.*

ANSWER. Since $\mathcal{V}[\mathbf{y}_i] = \sigma_i^2\boldsymbol{I} + \tau^2\boldsymbol{X}_i\boldsymbol{\Gamma}\boldsymbol{X}_i^\top$, it follows $\mathcal{V}[\hat{\boldsymbol{\beta}}_i] = \sigma_i^2(\boldsymbol{X}_i^\top\boldsymbol{X}_i)^{-1} + \tau^2(\boldsymbol{X}_i^\top\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i^\top\boldsymbol{X}_i\boldsymbol{\Gamma}\boldsymbol{X}_i^\top\boldsymbol{X}$ as postulated. □

• e. *3 points* [**Gre97**, p. 670] *describes a procedure how to estimate the covariance matrices if they are unknown. Explain this procedure clearly in your own words, and spell out the conditions under which it is a consistent estimate.*

ANSWER. If $\boldsymbol{\Gamma}$ is unknown, it is possible to get it from the sample covariance matrix of the group-specific OLS estimates, as long as the σ_i^2 and the $\boldsymbol{X}_i$ are such that asymptotically $\frac{1}{n-1}\sum(\hat{\boldsymbol{\beta}}_i - \bar{\hat{\beta}})(\hat{\boldsymbol{\beta}}_i - \bar{\hat{\beta}})^\top$ is the same as $\frac{1}{n}\sum(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta})^\top$ which again is asymptotically the same as $\sum\frac{1}{n}\mathcal{V}[\hat{\boldsymbol{\beta}}_i]$. We also need that asymptotically $\sum\frac{1}{n}\mathrm{s}_i^2(\boldsymbol{X}_i^\top\boldsymbol{X}_i)^{-1} = \sum\frac{1}{n}\sigma_i^2(\boldsymbol{X}_i^\top\boldsymbol{X}_i)^{-1}$. If these substitutions can be made, then $\operatorname{plim}\frac{1}{n-1}\sum(\hat{\boldsymbol{\beta}}_i - \bar{\hat{\beta}})(\hat{\boldsymbol{\beta}}_i - \bar{\hat{\beta}})^\top - \sum\frac{1}{n}\sigma_i^2(\boldsymbol{X}_i^\top\boldsymbol{X}_i)^{-1} = \tau^2\boldsymbol{\Gamma}$, since $\sum\frac{1}{n}\mathcal{V}[\hat{\boldsymbol{\beta}}_i] = \tau^2\boldsymbol{\Gamma} + \sum\frac{1}{n}\sigma_i^2(\boldsymbol{X}_i^\top\boldsymbol{X}_i)^{-1}$. This is [**Gre97**, (15-29) on p. 670]. □

PROBLEM 514. *5 points Describe in words how the "Random Coefficient Model" differs from an ordinary regression model, how it can be estimated, and describe situations in which it may be appropriate. Use your own words instead of excerpting the notes, don't give unnecessary detail but give an overview which will allow one to decide whether this is a good model for a given situation.*

ANSWER. If $\boldsymbol{\Sigma}$ is known, estimation proceeds in two steps: first estimate $\bar{\boldsymbol{\beta}}$ by a heteroskedastic GLS model, and then predict, or better retrodict, the actual value taken by the β_t by the usual linear prediction formulas. But the most important aspect of the model is that it is possible to estimate $\boldsymbol{\Sigma}$ if it is not known! This is possible because each ν_t imposes a different but known pattern of heteroskedasticity on the error terms, it so to say leaves its footprints, and if one has enough observations, it is possible to reconstruct the covariance matrix from these footprints. □

PROBLEM 515. *4 points The specification is*

$$y_t = \alpha + \beta_t x_t + \gamma x_t^2 \tag{61.0.54}$$

(no separate disturbance term), where α and γ are constants, and β_t is the tth element of a random vector $\boldsymbol{\beta} \sim (\boldsymbol{\iota}\mu, \tau^2\boldsymbol{I})$. Explain how you would estimate α, γ, μ, and τ^2.

ANSWER. Set $\boldsymbol{\nu} = \boldsymbol{\beta} - \boldsymbol{\iota}\mu$; it is $\boldsymbol{\nu} \sim (\boldsymbol{o}, \tau^2\boldsymbol{I})$ and one gets

$$y_t = \alpha + \mu x_t + \gamma x_t^2 + \nu_t x_t \tag{61.0.55}$$

This is regression with a heteroskedastic disturbance term. Therefore one has to specify weights$=1/x_t^2$, if one does that, one gets

$$\frac{\mathfrak{y}_t}{x_t} = \frac{\alpha}{x_t} + \mu + \gamma x_t + \nu_t \tag{61.0.56}$$

the coefficient estimates are the obvious ones, and the variance estimate in this regression is an unbiased estimate of τ^2. □

CHAPTER 62

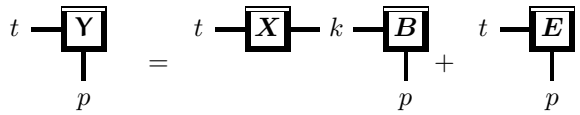
Multivariate Regression

62.1. Multivariate Econometric Models: A Classification

If the dependent variable $\mathbf{Y}$ is a matrix, then there are three basic models:

The simplest model is the multivariate regression model, in which all columns of $\mathbf{Y}$ have the same explanatory variables but different regression coefficients.

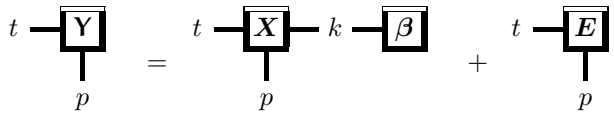
(62.1.1)



The most common application of these kinds of models are Vector Autoregressive Time Series models. If one adds the requirements that all coefficient vectors satisfy the same kind of linear constraint, one gets a model which is sometimes called a growth curve models. These models will be discussed in the remainder of this chapter.

In a second basic model, the explanatory variables are different, but the coefficient vector is the same. In tiles:

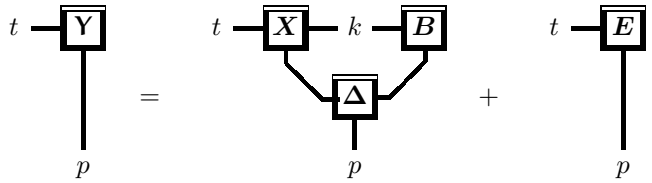
(62.1.2)



These models are used for pooling cross-sectional and timeseries data. They will be discussed in chapter 64.

In the third basic model, both explanatory variables and coefficient vectors are different.

(62.1.3)



These models are known under the name "seemingly unrelated" or "disturbance related" regression models. They will be discussed in chapter 65.

After this, chapter 66 will discuss "Simultaneous Equations Systems," in which the dependent variable in one equation may be the explanatory variable in another equation.

62.2. Multivariate Regression with Equal Regressors

The multivariate regression model with equal regressors reads

$$\mathbf{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E} \tag{62.2.1}$$

where we make the following assumptions: $\boldsymbol{X}$ is nonrandom and observed and $\mathbf{Y}$ is random and observed, $\boldsymbol{B}$ is nonrandom and not observed. $\boldsymbol{E}$ is random and not observed, but we know $\mathcal{E}[\boldsymbol{E}] = \boldsymbol{O}$ and the rows of $\boldsymbol{E}$ are independent drawings from the same $(\boldsymbol{o}, \boldsymbol{\Sigma})$ distribution with an unknown positive definite $\boldsymbol{\Sigma}$.

This model has applications for Vector Autoregressive Time Series and Multivariate Analysis of Variance (MANOVA).

The usual estimator of $\boldsymbol{B}$ in this model can be introduced by three properties: a least squares property, a BLUE property, and the maximum likelihood property (under the assumption of normality). In the univariate case, the least squares property is a scalar minimization, while the BLUE property involves matrix minimization. In the present case, the least squares property becomes a matrix minimization property, the BLUE property involves arrays of rank 4, and the maximum likelihood property is scalar maximization. In the univariate case, the scalar parameter σ^2 could be estimated alongside the linear estimator of $\boldsymbol{\beta}$, and now the whole covariance matrix $\boldsymbol{\Sigma}$ can.

62.2.1. Least Squares Property. The least squares principle can be applied here in the following form: given a matrix of observations $\mathbf{Y}$, estimate $\boldsymbol{B}$ by that value $\hat{\mathbf{B}}$ for which

$$\boldsymbol{B} = \hat{\mathbf{B}} \qquad \text{minimizes} \qquad (\mathbf{Y} - \boldsymbol{X}\boldsymbol{B})^\top(\mathbf{Y} - \boldsymbol{X}\boldsymbol{B}) \tag{62.2.2}$$

in the matrix sense, i.e., $(\mathbf{Y} - \boldsymbol{X}\hat{\mathbf{B}})^\top(\mathbf{Y} - \boldsymbol{X}\hat{\mathbf{B}})$ is by a nnd matrix smaller than any other $(\mathbf{Y} - \boldsymbol{X}\boldsymbol{B})^\top(\mathbf{Y} - \boldsymbol{X}\boldsymbol{B})$. And an unbiased estimator of $\boldsymbol{\Sigma}$ is $\hat{\boldsymbol{\Sigma}} = \frac{1}{n-k}(\mathbf{Y} - \boldsymbol{X}\hat{\mathbf{B}})^\top(\mathbf{Y} - \boldsymbol{X}\hat{\mathbf{B}})$.

Any $\hat{\mathbf{B}}$ which satisfies the normal equation

$$\boldsymbol{X}^\top\boldsymbol{X}\hat{\mathbf{B}} = \boldsymbol{X}^\top\mathbf{Y} \tag{62.2.3}$$

is a solution. There is always at least one such solution, and if $\boldsymbol{X}$ has full rank, then the solution is uniquely determined.

Proof: This is Problem 232. Due to the normal equations, the cross product disappears:

$$(\mathbf{Y}-\boldsymbol{X}\boldsymbol{B})^\top(\mathbf{Y}-\boldsymbol{X}\boldsymbol{B}) = (\mathbf{Y}-\boldsymbol{X}\hat{\mathbf{B}}+\boldsymbol{X}\hat{\mathbf{B}}-\boldsymbol{X}\boldsymbol{B})^\top(\mathbf{Y}-\boldsymbol{X}\hat{\mathbf{B}}+\boldsymbol{X}\hat{\mathbf{B}}-\boldsymbol{X}\boldsymbol{B})$$

$$= (\mathbf{Y}-\boldsymbol{X}\hat{\mathbf{B}})^\top(\mathbf{Y}-\boldsymbol{X}\hat{\mathbf{B}}) + (\boldsymbol{X}\hat{\mathbf{B}}-\boldsymbol{X}\boldsymbol{B})^\top(\boldsymbol{X}\hat{\mathbf{B}}-\boldsymbol{X}\boldsymbol{B}) \tag{62.2.4}$$

Note that the normal equation (62.2.3) simply reduces to the OLS normal equation for each column $\boldsymbol{\beta}_i$ of $\boldsymbol{B}$, with the corresponding column $\mathbf{y}_i$ of $\mathbf{Y}$ as dependent variable. In other words, for the estimation of $\boldsymbol{\beta}_i$, only the ith column $\mathbf{y}_i$ is used.

62.2.2. BLUE. To show that $\hat{\mathbf{B}}$ is the BLUE, write the equation $\mathbf{Y}=\boldsymbol{X}\boldsymbol{B}+\boldsymbol{E}$ in vectorized form, using (B.5.19), as

$$\operatorname{vec}(\mathbf{Y}) = (\boldsymbol{I}\otimes\boldsymbol{X})\operatorname{vec}(\boldsymbol{B})+\operatorname{vec}(\boldsymbol{E}) \tag{62.2.5}$$

Since $\mathcal{V}[\operatorname{vec}(\boldsymbol{E})]=\boldsymbol{\Sigma}\otimes\boldsymbol{I}$, the GLS estimate is, according to (26.0.2),

$$\operatorname{vec}(\hat{\mathbf{B}}) = \Big((\boldsymbol{I}\otimes\boldsymbol{X})^\top(\boldsymbol{\Sigma}\otimes\boldsymbol{I})^{-1}(\boldsymbol{I}\otimes\boldsymbol{X})\Big)^{-1}(\boldsymbol{I}\otimes\boldsymbol{X})^\top(\boldsymbol{\Sigma}\otimes\boldsymbol{I})^{-1}\operatorname{vec}(\mathbf{Y}) \tag{62.2.6}$$

$$= \Big((\boldsymbol{I}\otimes\boldsymbol{X}^\top)(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{I})(\boldsymbol{I}\otimes\boldsymbol{X})\Big)^{-1}(\boldsymbol{I}\otimes\boldsymbol{X}^\top)(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{I})\operatorname{vec}(\mathbf{Y}) \tag{62.2.7}$$

$$= \Big(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{X}^\top\boldsymbol{X}\Big)^{-1}(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{X}^\top)\operatorname{vec}(\mathbf{Y}) \tag{62.2.8}$$

$$= \Big(\boldsymbol{I}\otimes(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\Big)\operatorname{vec}(\mathbf{Y}) \tag{62.2.9}$$

and applying (B.5.19) again, this is equivalent to

$$\hat{\mathbf{B}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\mathbf{Y}. \tag{62.2.10}$$

From this vectorization one can also derive the dispersion matrix $\mathcal{V}[\operatorname{vec}(\hat{\mathbf{B}})]=\boldsymbol{\Sigma}\otimes(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$. In other words, $\mathcal{C}[\hat{\boldsymbol{\beta}}_i,\hat{\boldsymbol{\beta}}_j]=\sigma_{ij}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$, which can be estimated by $\hat{\sigma}_{ij}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$.

62.2.3. Maximum Likelihood. To derive the likelihood function, write the model in the row-partitioned form

$$\begin{bmatrix}\mathbf{y}_1^\top\\ \vdots\\ \mathbf{y}_n^\top\end{bmatrix} = \begin{bmatrix}\boldsymbol{x}_1^\top\\ \vdots\\ \boldsymbol{x}_n^\top\end{bmatrix}\boldsymbol{B} + \begin{bmatrix}\boldsymbol{\varepsilon}_1^\top\\ \vdots\\ \boldsymbol{\varepsilon}_n^\top\end{bmatrix} \tag{62.2.11}$$

Assuming normality, the ith row vector is $\mathbf{y}_i^\top\sim\mathsf{N}(\boldsymbol{x}_i^\top\boldsymbol{B},\boldsymbol{\Sigma})$, or $\mathbf{y}_i\sim\mathsf{N}(\boldsymbol{B}^\top\boldsymbol{x}_i,\boldsymbol{\Sigma})$. Since all rows are independent, the likelihood function is

$$f_{\mathbf{Y}}(\boldsymbol{Y}) = \prod_{i=1}^n\Big((2\pi)^{-r/2}(\det\boldsymbol{\Sigma})^{-1/2}\exp\big(-\frac{1}{2}(\boldsymbol{y}_i^\top-\boldsymbol{x}_i^\top\boldsymbol{B})\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_i-\boldsymbol{B}^\top\boldsymbol{x}_i)\big)\Big) \tag{62.2.12}$$

$$= (2\pi)^{-nr/2}(\det\boldsymbol{\Sigma})^{-n/2}\exp\big(-\frac{1}{2}\sum_i(\boldsymbol{y}_i^\top-\boldsymbol{x}_i^\top\boldsymbol{B})\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_i-\boldsymbol{B}^\top\boldsymbol{x}_i)\big). \tag{62.2.13}$$

The quadratic form in the exponent can be rewritten as follows:

$$\begin{aligned}\sum_{i=1}^n(\boldsymbol{y}_i^\top-\boldsymbol{x}_i^\top\boldsymbol{B})\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_i-\boldsymbol{B}^\top\boldsymbol{x}_i) &= \sum_{i=1}^n\operatorname{tr}(\boldsymbol{y}_i^\top-\boldsymbol{x}_i^\top\boldsymbol{B})\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_i-\boldsymbol{B}^\top\boldsymbol{x}_i)\\ &= \sum_{i=1}^n\operatorname{tr}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_i-\boldsymbol{B}^\top\boldsymbol{x}_i)(\boldsymbol{y}_i^\top-\boldsymbol{x}_i^\top\boldsymbol{B})\\ &= \operatorname{tr}\boldsymbol{\Sigma}^{-1}\sum_{i=1}^n(\boldsymbol{y}_i^\top-\boldsymbol{x}_i^\top\boldsymbol{B})^\top(\boldsymbol{y}_i^\top-\boldsymbol{x}_i^\top\boldsymbol{B})\\ &= \operatorname{tr}\boldsymbol{\Sigma}^{-1}\begin{bmatrix}\boldsymbol{y}_1^\top-\boldsymbol{x}_1^\top\boldsymbol{B}\\ \vdots\\ \boldsymbol{y}_n^\top-\boldsymbol{x}_n^\top\boldsymbol{B}\end{bmatrix}^\top\begin{bmatrix}\boldsymbol{y}_1^\top-\boldsymbol{x}_1^\top\boldsymbol{B}\\ \vdots\\ \boldsymbol{y}_n^\top-\boldsymbol{x}_n^\top\boldsymbol{B}\end{bmatrix}\\ &= \operatorname{tr}\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B})^\top(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B}).\end{aligned}$$

There are several different methods to maximize the likelihood function of a multivariate Normal Distribution. [**AO85**] gives a good survey: one can use matrix differentiation and matrix transformations, but also induction and inequalities.

The first step is obvious: using (62.2.4), the quadratic form in the exponent becomes:

$$\begin{aligned}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B})^\top(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B}) &= \operatorname{tr}\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{B}})^\top(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{B}})\\ &\quad+\operatorname{tr}(\boldsymbol{X}\hat{\boldsymbol{B}}-\boldsymbol{X}\boldsymbol{B})\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}\hat{\boldsymbol{B}}-\boldsymbol{X}\boldsymbol{B})^\top.\end{aligned}$$

The argument which minimizes this is $\boldsymbol{B}=\hat{\mathbf{B}}$, regardless of the value of $\boldsymbol{\Sigma}$. Therefore the concentrated likelihood function becomes, using the notation $\hat{\boldsymbol{E}}=(\mathbf{Y}-\boldsymbol{X}\hat{\mathbf{B}})$:

$$(2\pi)^{-nr/2}(\det\boldsymbol{\Sigma})^{-n/2}\exp\big(-\frac{1}{2}\operatorname{tr}\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{E}}^\top\hat{\boldsymbol{E}}\big). \tag{62.2.14}$$

In order to find the value of $\boldsymbol{\Sigma}$ which maximizes this we will use (A.8.21) in Theorem A.8.3 in the Mathematical Appendix. From (A.8.21) follows

$$(\det\boldsymbol{A})^{n/2}e^{-\frac{n}{2}\operatorname{tr}\boldsymbol{A}}\le e^{-rn/2}, \tag{62.2.15}$$

We want to apply (62.2.15). Set $\mathbf{A} = \frac{1}{n}(\hat{\boldsymbol{E}}^\top\hat{\boldsymbol{E}})^{1/2}\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{E}}^\top\hat{\boldsymbol{E}})^{1/2}$; then $\exp\bigl(-\frac{n}{2}\operatorname{tr}\mathbf{A}\bigr) = \exp\bigl(-\frac{1}{2}\operatorname{tr}\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{E}}^\top\hat{\boldsymbol{E}}\bigr)$, and $\det\mathbf{A} = \det(\frac{1}{n}\hat{\boldsymbol{E}}^\top\hat{\boldsymbol{E}})/\det\boldsymbol{\Sigma}$; therefore, using (62.2.15),

$$(2\pi)^{-nr/2}(\det\boldsymbol{\Sigma})^{-n/2}\exp\bigl(-\tfrac{1}{2}\operatorname{tr}\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{E}}^\top\hat{\boldsymbol{E}}\bigr) \le 2\pi e^{-nr/2}\det(\tfrac{1}{n}\hat{\boldsymbol{E}}^\top\hat{\boldsymbol{E}})^{-n/2},$$

with equality holding when $\mathbf{A} = \boldsymbol{I}$, i.e., for the value $\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\hat{\boldsymbol{E}}^\top\hat{\boldsymbol{E}}$.

(62.2.14) is the concentrated likelihood function even if one has prior knowledge about $\boldsymbol{\Sigma}$; in this case, the maximization is more difficult.

62.2.4. Distribution of the Least Squares Estimators. $\hat{\mathbf{B}}$ is normally distributed, with mean $\boldsymbol{B}$ and dispersion matrix $\mathcal{V}[\operatorname{vec}(\hat{\mathbf{B}})] = \boldsymbol{\Sigma}\otimes(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$. From the univariate result that $\operatorname{vec}(\hat{\boldsymbol{E}})$ and $\operatorname{vec}(\hat{\mathbf{B}})$ are uncorrelated, or from the univariate proof which goes through for the multivariate situation, follows in the Normal case that they are independent. Therefore $\hat{\mathbf{B}}$ is also independent of $\hat{\boldsymbol{\Sigma}}$. Since

$$\hat{\boldsymbol{E}}^\top\hat{\boldsymbol{E}} = \mathbf{Y}^\top(\boldsymbol{I}-\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top)\mathbf{Y} = \boldsymbol{E}^\top(\boldsymbol{I}-\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top)\boldsymbol{E} \tag{62.2.16}$$

and the matrix in the middle is idempotent, $\hat{\boldsymbol{E}}^\top\hat{\boldsymbol{E}}$ has a $\boldsymbol{W}(n-k,\boldsymbol{\Sigma})$ (Wishart) distribution, therefore

$$\text{the unbiased}\quad \hat{\boldsymbol{\Sigma}} = \frac{1}{n-k}\hat{\boldsymbol{E}}^\top\hat{\boldsymbol{E}} \sim \frac{1}{n-k}\boldsymbol{W}(n-k,\boldsymbol{\Sigma}) \tag{62.2.17}$$

and it is independent of $\hat{\mathbf{B}}$.

Let us look at the simplest example, in which $\boldsymbol{X} = \boldsymbol{\iota}$. Then $\boldsymbol{B}$ is a row vector, write it as $\boldsymbol{B} = \boldsymbol{\mu}^\top$, and the model reads

$$\mathbf{Y} = \boldsymbol{\iota}\boldsymbol{\mu}^\top + \boldsymbol{E}, \tag{62.2.18}$$

in other words, each row $\mathbf{y}$ of $\mathbf{Y}$ is an independent drawing from the same $\boldsymbol{\mu},\boldsymbol{\Sigma}$ distribution, and we want to estimate $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, and also the correlation coefficients. An elementary and detailed discussion of this model is given in chapter 63.

62.2.5. Testing. We will first look at tests of hypotheses of the form $\boldsymbol{RB} = \boldsymbol{U}$. This is a quite specialized hypothesis, meaning that each column of $\boldsymbol{B}$ is subject to the same linear constraint, although the values which these linear combinations take may differ from column to column. Remember in the univariate case we introduced several testing principles, the Wald test, the likelihood ratio test, and the Lagrange multiplier test, and showed that in the linear model they are equivalent. These principles can be directly transferred to the multivariate case. The Wald test consists in computing the unconstrained estimator $\hat{\mathbf{B}}$, and assessing, in terms of the (estimated, i.e., "studentized") Mahalanobis distance, how far $\boldsymbol{R}\hat{\mathbf{B}}$ is away from $\boldsymbol{U}$. The Likelihood ratio test (applied to the least squares objective function) consists in running both the constrained and the unconstrained multivarate regression, and then

determining how far the achieved values of the GLS objective function (which are matrices) are apart from each other.

Since the univariate t-test and its multivariate generalization, called Hotelling's T, is usually only applied in hypotheses where $\boldsymbol{R}$ is a row vector, we will limit our discussion to this case as well. The simplest example of such a test would be to test whether $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ in the above "simplest example" model with iid observations. The OLS estimate of $\boldsymbol{\mu}$ is $\bar{\mathbf{y}}$, which one gets by taking the column means of $\boldsymbol{Y}$. The dispersion matrix of this estimate is $\boldsymbol{\Sigma}/n$. The Mahalanobis distance of this estimate from $\boldsymbol{\mu}_0$ is therefore $n(\bar{\mathbf{y}}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{y}}-\boldsymbol{\mu})$, and replacing $\boldsymbol{\Sigma}$ by its unbiased $\boldsymbol{S} = \boldsymbol{W}/(n-1)$, one gets the following test statistic: $\mathsf{T}^2_{n-1} = n(\bar{\mathbf{y}}-\boldsymbol{\mu})^\top\boldsymbol{S}^{-1}(\bar{\mathbf{y}}-\boldsymbol{\mu})$.

Here use the following definition: if $\boldsymbol{z}\sim\mathsf{N}(\boldsymbol{o},\boldsymbol{\Sigma})$ is a r-vector, and $\boldsymbol{W}\sim\boldsymbol{W}(r,\boldsymbol{\Sigma})$ independent of $\boldsymbol{z}$ with the same $\boldsymbol{\Sigma}$, so that $\boldsymbol{S} = \boldsymbol{W}/r$ is an unbiased estimate of $\boldsymbol{\Sigma}$, then

$$\mathsf{T}^2_r = \boldsymbol{z}^\top\boldsymbol{S}^{-1}\boldsymbol{z} \tag{62.2.19}$$

is called a Hotelling $\mathsf{T}^2_{r,r}$ with r and r degrees of freedom.

One sees easily that the distribution of $\mathsf{T}^2_{r,r}$ is independent of $\boldsymbol{\Sigma}$. It can be written in the form

$$\mathsf{T}^2_r = \boldsymbol{z}^\top\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{S}\boldsymbol{\Sigma}^{-1/2})^{-1}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{z} \tag{62.2.20}$$

Here $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{z}\sim\mathsf{N}(\boldsymbol{o},\boldsymbol{I})$ and from $\boldsymbol{W} = \mathbf{Y}^\top\mathbf{Y}$ where each row of $\mathbf{Y}$ is a $\mathsf{N}(\boldsymbol{o},\boldsymbol{\Sigma})$, then $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{S}\boldsymbol{\Sigma}^{-1/2} = \mathbf{U}^\top\mathbf{U}$ where each row of $\mathbf{U}$ is a $\mathsf{N}(\boldsymbol{o},\boldsymbol{I})$.

From the interpretation of the Mahalanobis distance as the number of standard deviations the "worst" linear combination is away from its mean, Hotelling's T^2-test can again be interpreted as: make t-tests for all possible linear combinations of the components of $\boldsymbol{\mu}_0$ at an appropriately less stringent significance level, and reject the hypothesis if at least one of these t-tests rejects. This principle of constructing tests for multivariate hypotheses from those of simple hypotheses is called the "union-intersection principle" in multivariate statistics.

Since the usual F-statistic in univariate regression can also be considered the estimate of a Mahalanobis distance, it might be worth while to point out the difference. The difference is that in the case of the F-statistic, the dispersion matrix was known up to a factor σ^2, and only this factor had to be estimated. In the case of the Hotelling T^2, the whole dispersion matrix is unknown and all of it must be estimated (but one has also multivariate rather than univariate observations). Just as the distribution of the F statistic does not depend on the true value of σ^2, the distribution of Hotelling's T^2 does not depend on $\boldsymbol{\Sigma}$. Indeed, its distribution can be expressed in terms of the F-distribution. This is a deep result which we will not prove here:

If $\boldsymbol{\Sigma}$ is a $r \times r$ nonsingular matrix, then the distribution of Hotelling's $\mathsf{T}^2_{r,r}$ with r and r degrees of freedom can be expressed in terms of the F-distribution as follows:

$$\frac{r-r+1}{r}\frac{\mathsf{T}^2_{r,r}}{r} \sim \mathsf{F}_{r,r-r+1} \tag{62.2.21}$$

This apparatus with Hotelling's T^2 has been developed only for a very specific kind of hypothesis, namely, a hypothesis of the form $\boldsymbol{r}^\top \boldsymbol{B} = \boldsymbol{u}^\top$. Now let us turn to the more general hypothesis $\boldsymbol{RB} = \boldsymbol{U}$, where $\boldsymbol{R}$ has rank i, and apply the F-test principle. For this one runs the constrained and the unconstrained multivariate regression, calling the attained error sum of squares and products matrices $\hat{\boldsymbol{E}}_1^\top \hat{\boldsymbol{E}}_1$ (for the constrained) and $\hat{\boldsymbol{E}}^\top \hat{\boldsymbol{E}}$ (for the unconstrained model). Then one fills in the following table: Just as in the univariate case one shows that the S.P. matrices in

Source	D. F.	S. P. Matrix
Deviation from Hypothesis	$k-i$	$\hat{\boldsymbol{E}}_1^\top \hat{\boldsymbol{E}}_1 - \hat{\boldsymbol{E}}^\top \hat{\boldsymbol{E}}$
Error	$n-k$	$\hat{\boldsymbol{E}}^\top \hat{\boldsymbol{E}}$
(Restricted) Total	$n-i$	$\hat{\boldsymbol{E}}_1^\top \hat{\boldsymbol{E}}_1$

the first two rows are independent Wishart matrices, the first being central if the hypothesis is correct, and noncentral otherwise.

In the univariate case one has scalars instead of the S.P. matrices; then one divides each of these sum of squares by its degrees of freedom, and then takes the relation of the "Deviation from hypothesis" mean square error by the error mean square error. In this way one gets, for the error sum of squares an unbiased estimate of σ^2. If the hypothesis is true, the mean squared sum of errors explained by the hypothesis is an independent unbiased estimate of σ^2, otherwise it is biased upwards. The F-statistic is the ratio between those two estimates.

In the multivariate case, division by the degrees of freedom would give unbiased resp. upwardly biased estimates of the dispersion matrix $\boldsymbol{\Sigma}$. Then one faces the task of comparing these matrices. We will only discuss one criterion, Wilks's Lambda, which is at the same time the likelihood ratio test. It is a generalization of the F-test to the multivariate situation. Other criteria have been proposed (Roy's test).

This criterion based on the S.P. matrices does not divide these matrices by their degrees of freedom, but divides the determinant of the unconstrained error sum of squares and products matrix by the determinant of the total sum of squares and products matrix. This gives a statistic whose distribution is again independent of $\boldsymbol{\Sigma}$:

Definition of Wilks's Lambda: if $\boldsymbol{W}_1 \sim W_r(k_1, \boldsymbol{\Sigma})$ and $\boldsymbol{W}_2 \sim W_r(k_2, \boldsymbol{\Sigma})$ are independent (the subscript r indicating that $\boldsymbol{\Sigma}$ is $r \times r$), then

$$\frac{|\boldsymbol{W}_1|}{|\boldsymbol{W}_1 + \boldsymbol{W}_2|} \sim \Lambda(r, k_1, k_2) \tag{62.2.22}$$

is said to have Wilks's Lambda distribution.

If the matrix due to the hypothesis is of rank one, then this criterion is equivalent to Hotelling's T^2 criterion. To show this, assume you observe a $\boldsymbol{W} \sim \boldsymbol{W}_r(k, \boldsymbol{\Sigma})$ and an independent $\mathbf{d} \sim \mathsf{N}_r(\boldsymbol{o}, \boldsymbol{\Sigma})$. By theorem A.7.3, one obtains

$$1 + \mathbf{d}^\top \boldsymbol{W}^{-1} \mathbf{d} = \frac{\det(\boldsymbol{W} + \mathbf{d}\mathbf{d}^\top)}{\det(\boldsymbol{W})} \tag{62.2.23}$$

The righthand side is the inverse of a Wilks's Lambda with r, k and 1 degrees of freedom, the lefthand side is $1 + \mathsf{T}^2_{r,k}/k$.

We will show by one example that the Wilks's Lambda criterion is equivalent to the likelihood ratio criterion. Assume

$$\mathsf{Y} = \begin{bmatrix} \boldsymbol{X}_1 & \boldsymbol{X}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{B}_1 \\ \boldsymbol{B}_2 \end{bmatrix} + \boldsymbol{E} = \boldsymbol{X}_1 \boldsymbol{B}_1 + \boldsymbol{X}_2 \boldsymbol{B}_2 + \boldsymbol{E} \tag{62.2.24}$$

where Y is $n \times r$, $\boldsymbol{X}_1$ is $n \times q$, $\boldsymbol{X}_2$ is $n \times r$, $\boldsymbol{B}_1$ is $q \times r$, $\boldsymbol{B}_2$ is $r \times r$ and we want to test the hypothesis $\boldsymbol{B}_2 = \boldsymbol{O}$. Run the regressions on $\boldsymbol{X}_1$ alone to get $\hat{\boldsymbol{E}}_1$, and then on $\begin{bmatrix} \boldsymbol{X}_1 & \boldsymbol{X}_2 \end{bmatrix}$ to get $\hat{\boldsymbol{E}}$. The maximum values of the likelihood functions are then

$$2\pi e^{-nr/2} \det\Big(\frac{1}{n}\hat{\boldsymbol{E}}_1^\top \hat{\boldsymbol{E}}_1\Big)^{-n/2} \qquad \text{and} \qquad 2\pi e^{-nr/2} \det\Big(\frac{1}{n}\hat{\boldsymbol{E}}^\top \hat{\boldsymbol{E}}\Big)^{-n/2}. \tag{62.2.25}$$

The likelihood ratio, i.e., the constrained value divided by unconstrained value, is then

$$\Big(\frac{\det(\hat{\boldsymbol{E}}^\top \hat{\boldsymbol{E}})}{\det(\hat{\boldsymbol{E}}_1^\top \hat{\boldsymbol{E}}_1)}\Big)^{n/2}, \tag{62.2.26}$$

which is a power of Wilks's Lambda.

62.3. Growth Curve Models

One might wonder how to estimate the above model if there are linear restrictions on the rows of $\boldsymbol{B}$, for instance, they are all equal, or they all lie on a straight line, or on a qth order polynomial. This means, $\boldsymbol{B}$ can be written in the form $\boldsymbol{\Theta H}$ for some given $\boldsymbol{H}$ and some unrestricted parameter vector $\boldsymbol{\Theta}$. Models of this form are called growth curve models:

Problem 516. *6 points Assume*

$$\mathsf{Y} = \boldsymbol{X\Theta H} + \boldsymbol{E}; \quad \operatorname{vec}(\boldsymbol{E}) \sim \boldsymbol{o}, \sigma^2 \boldsymbol{\Psi} \otimes \boldsymbol{I} \tag{62.3.1}$$

$\boldsymbol{X}$, $\boldsymbol{H}$, and $\boldsymbol{\Psi}$ are known matrices of constants, and σ^2 and $\boldsymbol{\Theta}$ is the matrix of parameters to be estimated. Show that the BLUE in this model is

$$\hat{\mathbf{B}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\mathbf{Y}\boldsymbol{\Psi}^{-1}\boldsymbol{H}^\top(\boldsymbol{H}\boldsymbol{\Psi}^{-1}\boldsymbol{H}^\top)^{-1} \tag{62.3.2}$$

and that

$$\mathcal{V}[\operatorname{vec}(\hat{\mathbf{B}})] = \sigma^2(\boldsymbol{H}\boldsymbol{\Psi}^{-1}\boldsymbol{H}^\top)^{-1} \otimes (\boldsymbol{X}^\top\boldsymbol{X})^{-1}. \tag{62.3.3}$$

This is a so-called growth curve model, and the simplest proof uses (B.5.19)*and Problem 611.*

Answer. If $\boldsymbol{\Psi}$ is known, then the BLUE can be obtained as follows: Use (B.5.19) to write the equation $\mathbf{Y} = \boldsymbol{X}\boldsymbol{\Theta}\boldsymbol{H} + \boldsymbol{E}$ in vectorized form as

$$\operatorname{vec}(\mathbf{Y}) = (\boldsymbol{H}^\top \otimes \boldsymbol{X})\operatorname{vec}(\boldsymbol{\Theta}) + \operatorname{vec}(\boldsymbol{E}) \tag{62.3.4}$$

Since $\mathcal{V}[\operatorname{vec}(\boldsymbol{E})] = \sigma^2\boldsymbol{\Psi} \otimes \boldsymbol{I}$, the GLS estimate is

$$\operatorname{vec}(\hat{\mathbf{B}}) = \Big((\boldsymbol{H} \otimes \boldsymbol{X}^\top)(\boldsymbol{\Psi}^{-1} \otimes \boldsymbol{I})(\boldsymbol{H}^\top \otimes \boldsymbol{X})\Big)^{-1}(\boldsymbol{H} \otimes \boldsymbol{X}^\top)(\boldsymbol{\Psi}^{-1} \otimes \boldsymbol{I})\operatorname{vec}(\mathbf{Y}) \tag{62.3.5}$$

$$= \Big(\boldsymbol{H}\boldsymbol{\Psi}^{-1}\boldsymbol{H}^\top \otimes \boldsymbol{X}^\top\boldsymbol{X}\Big)^{-1}\Big(\boldsymbol{H}\boldsymbol{\Psi}^{-1} \otimes \boldsymbol{X}^\top\Big)\operatorname{vec}(\mathbf{Y}) \tag{62.3.6}$$

$$= \Big(\big((\boldsymbol{H}\boldsymbol{\Psi}^{-1}\boldsymbol{H}^\top)^{-1}\boldsymbol{H}\boldsymbol{\Psi}^{-1}\big) \otimes \big((\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\big)\Big)\operatorname{vec}(\mathbf{Y}) \tag{62.3.7}$$

and now apply (B.5.19) again to transform this back into matrix notation

$$\hat{\mathbf{B}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\mathbf{Y}\boldsymbol{\Psi}^{-1}\boldsymbol{H}^\top(\boldsymbol{H}\boldsymbol{\Psi}^{-1}\boldsymbol{H}^\top)^{-1} \tag{62.3.8}$$

□

Here is one scenario how such a model may arise: assume you have n plants, you group those plants into two different groups, the first group going from plant 1 until plant m, and the second from plant $m+1$ until plant n. These groups obtain different treatments. At r different time points you measure the same character on each of these plants. These measurements give the rows of your $\mathbf{Y}$-matrix. You assume the following: the dispersion matrix between these measurements are identical for all plants, call it $\boldsymbol{\Psi}$, and the expected values of these measurements evolves over time following two different quadratic polynomials, one for each treatment.

This can be expressed mathematically as follows (omitting the matrix of error terms):

(62.3.9)

$$\begin{bmatrix} \mathsf{y}_{11} & \mathsf{y}_{12} & \cdots & \mathsf{y}_{1r} \\ \vdots & \vdots & \ddots & \vdots \\ \mathsf{y}_{m1} & \mathsf{y}_{m2} & \cdots & \mathsf{y}_{mr} \\ \mathsf{y}_{m+1,1} & \mathsf{y}_{m+1,2} & \cdots & \mathsf{y}_{m+1,r} \\ \vdots & \vdots & \ddots & \vdots \\ \mathsf{y}_{n1} & \mathsf{y}_{n2} & \cdots & \mathsf{y}_{nr} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \theta_{10} & \theta_{11} & \theta_{12} \\ \theta_{20} & \theta_{21} & \theta_{22} \end{bmatrix} \begin{bmatrix} 1 & 1 & \cdots & 1 \\ t_1 & t_2 & \cdots & t_p \\ t_1^2 & t_2^2 & \cdots & t_p^2 \end{bmatrix}$$

This gives the desired result $\mathsf{y}_{11} = \theta_{10} + \theta_{11}t_1 + \theta_{12}t_1^2$ plus an error term, etc.

If one does not know $\boldsymbol{\Psi}$, then one has to estimate it.

CHAPTER 63

Independent Observations from the Same Multivariate Population

This Chapter discusses a model that is a special case of the model in Chapter 62.2, but it goes into more depth towards the end.

63.1. Notation and Basic Statistics

Notational conventions are not uniform among the different books about multivariate statistic. Johnson and Wichern arrange the data in a $r \times n$ matrix $\mathbf{X}$. Each column is a separate independent observation of a q vector with mean $\boldsymbol{\mu}$ and dispersion matrix $\boldsymbol{\Sigma}$. There are n observations.

We will choose an alternative notation, which is also found in the literature, and write the matrix as a $n \times r$ matrix $\mathbf{Y}$. As before, each column represents a variable, and each row a usually independent observation.

Decompose $\mathbf{Y}$ into its row vectors as follows:

$$\mathbf{Y} = \begin{bmatrix} \mathbf{y}_1^\top \\ \vdots \\ \mathbf{y}_n^\top \end{bmatrix}. \tag{63.1.1}$$

Each row (written as a column vector) $\mathbf{y}_i$ has mean $\boldsymbol{\mu}$ and dispersion matrix $\boldsymbol{\Sigma}$, and different rows are independent of each other. In other words, $\mathcal{E}[\mathbf{Y}] = \boldsymbol{\iota}\boldsymbol{\mu}^\top$. $\mathcal{V}[\mathbf{Y}]$ is an array of rank 4, not a matrix. In terms of Kronecker products one can write $\mathcal{V}[\operatorname{vec}\mathbf{Y}] = \boldsymbol{\Sigma} \otimes \boldsymbol{I}$.

One can form the following descriptive statistics: $\bar{\mathbf{y}} = \frac{1}{n}\mathbf{y}_i$ is the vector of sample means, $\boldsymbol{W} = \sum_i (\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})^\top$ is matrix of (corrected) squares and cross products, the sample covariance matrix is $\mathbf{S}^{(n)} = \frac{1}{n}\boldsymbol{W}$ with divisor n, and $\mathbf{R}$ is the matrix of sample correlation coefficients.

Notation: the ith sample variance is called s_{ii} (not s_i^2, as one might perhaps expect).

The sample means indicate location, the sample standard deviations dispersion, and the sample correlation coefficients linear relationship.

How do we get these descriptive statistics from the data $\mathbf{Y}$ through a matrix manipulation? $\bar{\mathbf{y}}^\top = \frac{1}{n}\boldsymbol{\iota}^\top\mathbf{Y}$; now $\mathbf{Y} - \boldsymbol{\iota}\bar{\mathbf{y}}^\top = (\boldsymbol{I} - \frac{\boldsymbol{\iota}\boldsymbol{\iota}^\top}{n})\mathbf{Y}$ is the matrix of observations with the appropriate sample mean taken out of each element, therefore

$$\begin{aligned} \boldsymbol{W} = \begin{bmatrix} \mathbf{y}_1 - \bar{\mathbf{y}} & \cdots & \mathbf{y}_n - \bar{\mathbf{y}} \end{bmatrix} \begin{bmatrix} (\mathbf{y}_1 - \bar{\mathbf{y}})^\top \\ \vdots \\ (\mathbf{y}_n - \bar{\mathbf{y}})^\top \end{bmatrix} = \\ = \mathbf{Y}^\top(\boldsymbol{I} - \frac{\boldsymbol{\iota}\boldsymbol{\iota}^\top}{n})^\top(\boldsymbol{I} - \frac{\boldsymbol{\iota}\boldsymbol{\iota}^\top}{n})\mathbf{Y} = \mathbf{Y}^\top(\boldsymbol{I} - \frac{\boldsymbol{\iota}\boldsymbol{\iota}^\top}{n})\mathbf{Y}. \end{aligned} \tag{63.1.2}$$

Then $\mathbf{S}^{(n)} = \frac{1}{n}\boldsymbol{W}$, and in order to get the sample correlation matrix $\mathbf{R}$, use

$$\mathbf{D}^{(n)} = \operatorname{diag}(\mathbf{S}^{(n)}) = \begin{bmatrix} \mathrm{s}_{11} & 0 & \cdots & 0 \\ 0 & \mathrm{s}_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathrm{s}_{nn} \end{bmatrix} \tag{63.1.3}$$

and then $\mathbf{R} = (\mathbf{D}^{(n)})^{-1/2}\mathbf{S}^{(n)}(\mathbf{D}^{(n)})^{-1/2}$.

In analogy to the formulas for variances and covariances of linear transformations of a vector, one has the following formula for sample variances and covariances of linear combinations $\mathbf{Y}\boldsymbol{a}$ and $\mathbf{Y}\boldsymbol{b}$: $\operatorname{est.cov}[\mathbf{Y}\boldsymbol{a}, \mathbf{Y}\boldsymbol{b}] = \boldsymbol{a}^\top\mathbf{S}^{(n)}\boldsymbol{b}$.

Problem 517. *Show that $\mathcal{E}[\bar{\mathbf{y}}] = \boldsymbol{\mu}$ and $\mathcal{V}[\bar{\mathbf{y}}] = \frac{1}{n}\boldsymbol{\Sigma}$. (The latter identity can be shown in two ways: once using the Kronecker product of matrices, and once by partitioning $\mathbf{Y}$ into its rows.)*

Answer. $\mathcal{E}[\bar{\mathbf{y}}] = \mathcal{E}[\frac{1}{n}\mathbf{Y}^\top\boldsymbol{\iota}] = \frac{1}{n}(\mathcal{E}[\mathbf{Y}])^\top\boldsymbol{\iota} = \frac{1}{n}\boldsymbol{\mu}\boldsymbol{\iota}^\top\boldsymbol{\iota} = \boldsymbol{\mu}$. Using Kronecker products, one obtains from $\bar{\mathbf{y}}^\top = \frac{1}{n}\boldsymbol{\iota}^\top\mathbf{Y}$ that

$$\bar{\mathbf{y}} = \operatorname{vec}(\bar{\mathbf{y}}^\top) = \frac{1}{n}(\boldsymbol{I} \otimes \boldsymbol{\iota}^\top)\operatorname{vec}\mathbf{Y}; \tag{63.1.4}$$

therefore

$$\mathcal{V}[\bar{\mathbf{y}}] = \frac{1}{n^2}(\boldsymbol{I} \otimes \boldsymbol{\iota}^\top)(\boldsymbol{\Sigma} \otimes \boldsymbol{I})(\boldsymbol{I} \otimes \boldsymbol{\iota}) = \frac{1}{n^2}(\boldsymbol{\Sigma} \otimes \boldsymbol{\iota}^\top\boldsymbol{\iota}) = \frac{1}{n}\boldsymbol{\Sigma} \tag{63.1.5}$$

The alternative way to do it is

$$\mathcal{V}[\bar{\mathbf{y}}] = \mathcal{E}[(\bar{\mathbf{y}} - \boldsymbol{\mu})(\bar{\mathbf{y}} - \boldsymbol{\mu})^\top] \tag{63.1.6}$$

$$= \mathcal{E}[\Big(\frac{1}{n}\sum_i(\mathbf{y}_i - \boldsymbol{\mu})\Big)\Big(\frac{1}{n}\sum_j(\mathbf{y}_j - \boldsymbol{\mu})\Big)^\top] \tag{63.1.7}$$

$$= \frac{1}{n^2}\sum_{i,j}\mathcal{E}[(\mathbf{y}_i - \boldsymbol{\mu})(\mathbf{y}_j - \boldsymbol{\mu})^\top] \tag{63.1.8}$$

$$= \frac{1}{n^2}\sum_{i}\mathcal{E}[(\mathbf{y}_i - \boldsymbol{\mu})(\mathbf{y}_i - \boldsymbol{\mu})^\top] \tag{63.1.9}$$

$$= \frac{n}{n^2}\,\mathcal{E}[(\mathbf{y}_i - \boldsymbol{\mu})(\mathbf{y}_i - \boldsymbol{\mu})^\top] = \frac{1}{n}\boldsymbol{\Sigma}. \tag{63.1.10}$$

□

PROBLEM 518. *Show that* $\mathcal{E}[\mathbf{S}^{(n)}] = \frac{n-1}{n}\boldsymbol{\Sigma}$, *therefore the unbiased* $\mathbf{S} = \frac{1}{n-1}\sum_i(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top$ *has* $\boldsymbol{\Sigma}$ *as its expected value.*

63.2. Two Geometries

One can distinguish two geometries, according to whether one takes the rows or the columns of $\mathbf{Y}$ as the points. Rows as points gives n points in r-dimensional space, the "scatterplot geometry." If $r = 2$, this is the scatter plot of the two variables against each other.

In this geometry, the sample mean is the center of balance or center of gravity. The dispersion of the observations around their mean defines a distance measure in this geometry.

The book introduces this distance by suggesting with its illustrations that the data are clustered in hyperellipsoids. The right way to introduce this distance would be to say: we are not only interested in the r coordinates separately but also in any linear combinations, then use our treatment of the Mahalanobis distance for a given population, and then transfer it to the empirical distribution given by the sample.

In the other geometry, all observations of a given random variable form one point, here called "vector." I.e., the basic entities are the columns of $\mathbf{Y}$. In this so-called "vector geometry," $\bar{\mathbf{x}}$ is the projection on the diagonal vector $\boldsymbol{\iota}$, and the correlation coefficient is the cosine of the angle between the deviation vectors.

Generalized sample variance is defined as determinant of $\mathbf{S}$. Its geometric intuition: in the scatter plot geometry it is proportional to the square of the volume of the hyperellipsoids, (see J&W, p. 103), and in the geometry in which the observations of each variabe form a vector it is

$$\det \mathbf{S} = (n-1)^{-r}(\text{volume})^2 \tag{63.2.1}$$

where the volume is that spanned by the deviation vectors.

63.3. Assumption of Normality

A more general version of this section is 62.2.3.

Assume that the $\mathbf{y}_i$, the row vectors of $\mathbf{Y}$, are independent, and each is $\sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ positive definite. Then the density function of $\mathbf{Y}$ is

$$f_{\mathbf{Y}}(\boldsymbol{Y}) = \prod_{j=1}^{n}\Big((2\pi)^{-r/2}(\det\boldsymbol{\Sigma})^{-1/2}\exp\big(-\frac{1}{2}(\boldsymbol{y}_j - \boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_j - \boldsymbol{\mu})\big)\Big) \tag{63.3.1}$$

$$= (2\pi)^{-nr/2}(\det\boldsymbol{\Sigma})^{-n/2}\exp\big(-\frac{1}{2}\sum_j(\boldsymbol{y}_j - \boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_j - \boldsymbol{\mu})\big). \tag{63.3.2}$$

The quadratic form in the exponent can be rewritten as follows:

$$\sum_{j=1}^{n}(\boldsymbol{y}_j - \boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_j - \boldsymbol{\mu}) = \sum_{j=1}^{n}(\boldsymbol{y}_j - \bar{\boldsymbol{y}} + \bar{\boldsymbol{y}} - \boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_j - \bar{\boldsymbol{y}} + \bar{\boldsymbol{y}} - \boldsymbol{\mu})$$

$$= \sum_{j=1}^{n}(\boldsymbol{y}_j - \bar{\boldsymbol{y}})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_j - \bar{\boldsymbol{y}}) + n(\bar{\boldsymbol{y}} - \boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{y}} - \boldsymbol{\mu}) \tag{63.3.3}$$

The first term can be simplified as follows:

$$\sum_j(\boldsymbol{y}_j - \bar{\boldsymbol{y}})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_j - \bar{\boldsymbol{y}}) = \sum_j \operatorname{tr}(\boldsymbol{y}_j - \bar{\boldsymbol{y}})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_j - \bar{\boldsymbol{y}})$$

$$= \sum_j \operatorname{tr}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_j - \bar{\boldsymbol{y}})(\boldsymbol{y}_j - \bar{\boldsymbol{y}})^\top$$

$$= \operatorname{tr}\boldsymbol{\Sigma}^{-1}\sum_j(\boldsymbol{y}_j - \bar{\boldsymbol{y}})(\boldsymbol{y}_j - \bar{\boldsymbol{y}})^\top = n\operatorname{tr}\boldsymbol{\Sigma}^{-1}\boldsymbol{S}^{(n)}$$

Using this one can write the density function as

$$f_{\mathbf{Y}}(\boldsymbol{Y}) = (2\pi)^{-nr/2}(\det\boldsymbol{\Sigma})^{-n/2}\exp\Big(-\frac{n}{2}\operatorname{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}^{(n)})\Big)\exp\Big(-\frac{n}{2}(\bar{\boldsymbol{y}} - \boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{y}} - \boldsymbol{\mu})\Big). \tag{63.3.4}$$

One sees, therefore, that the density function depends on the observation only through $\bar{\boldsymbol{y}}$ and $\boldsymbol{S}^{(n)}$, which means that $\bar{\mathbf{y}}$ and $\mathbf{S}^{(n)}$ are sufficient statistics.

Now we compute the maximum likelihood estimators: taking the maximum for $\boldsymbol{\mu}$ is simply $\hat{\boldsymbol{\mu}} = \bar{\mathbf{y}}$. This leaves the concentrated likelihood function

$$\max_{\boldsymbol{\mu}} f_{\mathbf{Y}}(\boldsymbol{Y}) = (2\pi)^{-nr/2}(\det\boldsymbol{\Sigma})^{-n/2}\exp\Big(-\frac{n}{2}\operatorname{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{S}^{(n)})\Big). \tag{63.3.5}$$

To obtain the maximum likelihood estimate of $\boldsymbol{\Sigma}$ one needs equation (A.8.21) in Theorem A.8.3 in the Appendix and (62.2.15).

If one sets $\boldsymbol{A} = \mathbf{S}^{(n)^{1/2}}\boldsymbol{\Sigma}^{-1}\mathbf{S}^{(n)^{1/2}}$, then $\operatorname{tr}\boldsymbol{A} = \operatorname{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{S}^{(n)})$ and $\det\boldsymbol{A} = (\det\boldsymbol{\Sigma})^{-1}\det\mathbf{S}^{(n)}$, in (62.2.15), therefore the concentrated likelihood function

$$(63.3.6)\qquad (2\pi)^{-nr/2}(\det\boldsymbol{\Sigma})^{-n/2}\exp\Big(-\frac{n}{2}\operatorname{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{S}^{(n)})\Big) \le (2\pi e)^{-rn/2}(\det\mathbf{S}^{(n)})^{-n/2}$$

with equality holding if $\hat{\boldsymbol{\Sigma}} = \mathbf{S}^{(n)}$. Note that the maximum value is a multiple of the estimated generalized variance.

63.4. EM-Algorithm for Missing Observations

The maximization of the likelihood function is far more difficult if some observations are missing. (Here assume they are missing randomly, i.e., the fact that they are missing is not related to the values of these entries. Otherwise one has sample selection bias!) In this case, a good iterative procedure to obtain the maximum likelihood estimate is the EM-algorithm (expectation-maximization algorithm). It is an iterative prediction and estimation.

Let's follow Johnson and Wichern's example on their p. 199. The matrix is

$$(63.4.1)\qquad \mathbf{Y} = \begin{bmatrix} - & 0 & 3 \\ 7 & 2 & 6 \\ 5 & 1 & 2 \\ - & - & 5 \end{bmatrix}$$

It is not so important how one gets the initial estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$: say $\tilde{\boldsymbol{\mu}}^\top = \begin{bmatrix}6 & 1 & 4\end{bmatrix}$, and to get $\tilde{\boldsymbol{\Sigma}}$ take deviations from the mean, putting zeros in for the missing values (which will of course underestimate the variances), and divide by the number of observations. (Since we are talking maximum likelihood, there is no adjustment for degrees of freedom.)

(63.4.2)

$$\tilde{\boldsymbol{\Sigma}} = \frac{1}{4}\underline{\mathbf{Y}}^\top\underline{\mathbf{Y}} \quad\text{where}\quad \underline{\mathbf{Y}} = \begin{bmatrix} 0 & -1 & -1 \\ 1 & 1 & 2 \\ -1 & 0 & -2 \\ 0 & 0 & 1 \end{bmatrix}, \quad\text{i.e.,}\quad \tilde{\boldsymbol{\Sigma}} = \begin{bmatrix} 1/2 & 1/4 & 1 \\ 1/4 & 1/2 & 3/4 \\ 1 & 3/4 & 5/2 \end{bmatrix}.$$

Given these estimates, the prediction step is next. The likelihood function depends on sample mean and sample dispersion matrix only. These, in turn, are simple functions of the vector of column sums $\mathbf{Y}^\top\boldsymbol{\iota}$ and the matrix of (uncentered) sums of squares and crossproducts $\mathbf{Y}^\top\mathbf{Y}$, which are complete sufficient statistics. To predict those we need predictions of the missing elements of $\mathbf{Y}$, of their squares, and of their products with each other and with the observed elements of $\mathbf{Y}$. Our method of predicting is to take conditional expectations, assuming $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\Sigma}}$ are the true mean and dispersion matrix.

For the prediction of the upper lefthand corner element of $\mathbf{Y}$, only the first row of $\mathbf{Y}$ is relevant. Partitioning this row into the observed and unobserved elements gives

(63.4.3)

$$\begin{bmatrix} \mathfrak{y}_{11} \\ 0 \\ 3 \end{bmatrix} \sim N\Big(\begin{bmatrix} 6 \\ 1 \\ 4 \end{bmatrix}, \begin{bmatrix} 1/2 & 1/4 & 1 \\ 1/4 & 1/2 & 3/4 \\ 1 & 3/4 & 5/2 \end{bmatrix}\Big) \quad\text{or}\quad \begin{bmatrix} \mathfrak{y}_1 \\ \mathfrak{y}_2 \end{bmatrix} \sim N\Big(\begin{bmatrix} \tilde{\boldsymbol{\mu}}_1 \\ \tilde{\boldsymbol{\mu}}_2 \end{bmatrix}, \begin{bmatrix} \tilde{\boldsymbol{\Sigma}}_{11} & \tilde{\boldsymbol{\Sigma}}_{12} \\ \tilde{\boldsymbol{\Sigma}}_{21} & \tilde{\boldsymbol{\Sigma}}_{22} \end{bmatrix}\Big).$$

The conditional mean of $\mathfrak{y}_1$ is the best linear predictor

$$(63.4.4)\qquad \mathcal{E}[\mathfrak{y}_1|\mathfrak{y}_2;\tilde{\boldsymbol{\mu}},\tilde{\boldsymbol{\Sigma}}] = \mathfrak{y}_1^* = \tilde{\boldsymbol{\mu}}_1 + \tilde{\boldsymbol{\Sigma}}_{12}\tilde{\boldsymbol{\Sigma}}_{22}^{-1}(\mathfrak{y}_2 - \tilde{\boldsymbol{\mu}}_2)$$

or in our numerical example

$$(63.4.5)\qquad \mathcal{E}[\mathfrak{y}_{11}|\cdots] = \mathfrak{y}_{11}^* = 6 + \begin{bmatrix}1/4 & 1\end{bmatrix}\begin{bmatrix} 1/2 & 3/4 \\ 3/4 & 5/2 \end{bmatrix}^{-1}\begin{bmatrix} 0-1 \\ 3-4 \end{bmatrix} = 5.73$$

Furthermore,

(63.4.6)

$$\mathcal{E}[(\mathfrak{y}_1-\mathfrak{y}_1^*)(\mathfrak{y}_1-\mathfrak{y}_1^*)^\top|\mathfrak{y}_2;\tilde{\boldsymbol{\mu}},\tilde{\boldsymbol{\Sigma}}] = \mathcal{E}[(\mathfrak{y}_1-\mathfrak{y}_1^*)(\mathfrak{y}_1-\mathfrak{y}_1^*)^\top] = \mathcal{MSE}[\mathfrak{y}_1^*;\mathfrak{y}_1] = \tilde{\boldsymbol{\Sigma}}_{11}-\tilde{\boldsymbol{\Sigma}}_{12}\tilde{\boldsymbol{\Sigma}}_{22}^{-1}\tilde{\boldsymbol{\Sigma}}_{21}.$$

These two data are sufficient to compute $\mathcal{E}[\mathfrak{y}_1\mathfrak{y}_1^\top|\mathfrak{y}_2;\tilde{\boldsymbol{\mu}},\tilde{\boldsymbol{\Sigma}}]$. From $\mathfrak{y}_1 = \mathfrak{y}_1 - \mathfrak{y}_1^* + \mathfrak{y}_1^*$ follows

$$(63.4.7)\quad \mathfrak{y}_1\mathfrak{y}_1^\top = (\mathfrak{y}_1-\mathfrak{y}_1^*)(\mathfrak{y}_1-\mathfrak{y}_1^*)^\top+(\mathfrak{y}_1-\mathfrak{y}_1^*)(\mathfrak{y}_1^*)^\top+(\mathfrak{y}_1^*)(\mathfrak{y}_1-\mathfrak{y}_1^*)^\top+(\mathfrak{y}_1^*)(\mathfrak{y}_1^*)^\top.$$

Now take conditional expectations:

$$(63.4.8)\qquad \mathcal{E}[\mathfrak{y}_1\mathfrak{y}_1^\top|\mathfrak{y}_2;\tilde{\boldsymbol{\mu}},\tilde{\boldsymbol{\Sigma}}] = \tilde{\boldsymbol{\Sigma}}_{11} - \tilde{\boldsymbol{\Sigma}}_{12}\tilde{\boldsymbol{\Sigma}}_{22}^{-1}\tilde{\boldsymbol{\Sigma}}_{21} + \boldsymbol{O} + \boldsymbol{O} + (\mathfrak{y}_1^*)(\mathfrak{y}_1^*)^\top$$

For the cross products with the observed values one can apply the linearity of the (conditional) expectations operator:

$$(63.4.9)\qquad \mathcal{E}[\mathfrak{y}_1\mathfrak{y}_2^\top|\mathfrak{y}_2;\tilde{\boldsymbol{\mu}},\tilde{\boldsymbol{\Sigma}}] = (\mathfrak{y}_1^*)\mathfrak{y}_2^\top$$

Therefore one obtains

$$(63.4.10)\quad \mathcal{E}[\begin{bmatrix} \mathfrak{y}_1\mathfrak{y}_1^\top & \mathfrak{y}_1\mathfrak{y}_2^\top \\ \mathfrak{y}_2\mathfrak{y}_1^\top & \mathfrak{y}_2\mathfrak{y}_2^\top \end{bmatrix}|\mathfrak{y}_2;\tilde{\boldsymbol{\mu}},\tilde{\boldsymbol{\Sigma}}] = \begin{bmatrix} \tilde{\boldsymbol{\Sigma}}_{11} - \tilde{\boldsymbol{\Sigma}}_{12}\tilde{\boldsymbol{\Sigma}}_{22}^{-1}\tilde{\boldsymbol{\Sigma}}_{21} + (\mathfrak{y}_1^*)(\mathfrak{y}_1^*)^\top & \mathfrak{y}_1^*\mathfrak{y}_2^\top \\ \mathfrak{y}_2(\mathfrak{y}_1^*)^\top & \mathfrak{y}_2\mathfrak{y}_2^\top \end{bmatrix}$$

In our numerical example this gives

$$(63.4.11)\qquad \mathrm{E}[\mathfrak{y}_{11}^2|\cdots] = 1/2 - \begin{bmatrix}1/4 & 1\end{bmatrix}\begin{bmatrix} 1/2 & 3/4 \\ 3/4 & 5/2 \end{bmatrix}^{-1}\begin{bmatrix} 1/4 \\ 1 \end{bmatrix} + (5.73)^2 = 32.99$$

(63.4.12)

$$\mathcal{E}[\begin{bmatrix}\mathfrak{y}_{11}\mathfrak{y}_{12} & \mathfrak{y}_{11}\mathfrak{y}_{13}\end{bmatrix}|\cdots] = 5.73\begin{bmatrix}0 & 3\end{bmatrix} = \begin{bmatrix}0 & 17.18\end{bmatrix}$$

PROBLEM 519. *Compute in the same way for the last row of* $\mathbf{Y}$*:*

$$\mathcal{E}[\begin{bmatrix}\mathsf{y}_{41} & \mathsf{y}_{42}\end{bmatrix}|\cdots] = \begin{bmatrix}\mathsf{y}^*_{41} & \mathsf{y}^*_{42}\end{bmatrix} = \begin{bmatrix}6.4 & 1.2\end{bmatrix} \tag{63.4.13}$$

$$\mathcal{E}[\begin{bmatrix}\mathsf{y}^2_{41} & \mathsf{y}_{41}\mathsf{y}_{42}\\ \mathsf{y}_{42}\mathsf{y}_{41} & \mathsf{y}^2_{42}\end{bmatrix}|\cdots] = \begin{bmatrix}41.06 & 8.27\\ 8.27 & 1.97\end{bmatrix} \tag{63.4.14}$$

$$\mathcal{E}[\begin{bmatrix}\mathsf{y}_{41}\mathsf{y}_{43}\\ \mathsf{y}_{42}\mathsf{y}_{43}\end{bmatrix}|\cdots] = \begin{bmatrix}32.0\\ 6.5\end{bmatrix} \tag{63.4.15}$$

ANSWER. This is in Johnson and Wichern, p. 200. □

Now switch back to the more usual notation, in which $\mathbf{y}_i$ is the ith row vector of $\mathbf{Y}$ and $\bar{\mathbf{y}}$ the vector of column means. Since $\mathbf{S}^{(n)} = \frac{1}{n}\sum \mathbf{y}_i\mathbf{y}_i^\top - \bar{\mathbf{y}}\bar{\mathbf{y}}^\top$, one can obtain from the above the value of

$$\mathcal{E}[\mathbf{S}^{(n)}|\text{ all observed values in }\mathbf{Y};\tilde{\boldsymbol{\mu}},\tilde{\boldsymbol{\Sigma}}]. \tag{63.4.16}$$

Of course, in a similar, much simpler fashion one obtains

$$\mathcal{E}[\bar{\mathbf{y}}|\text{ all observed values in }\mathbf{Y};\tilde{\boldsymbol{\mu}},\tilde{\boldsymbol{\Sigma}}]. \tag{63.4.17}$$

In our numerical example, therefore, we obtain

(63.4.18)

$$\mathcal{E}[\mathbf{Y}^\top\boldsymbol{\iota}|\cdots] = \mathcal{E}[\begin{bmatrix}\mathsf{y}_{11} & 7 & 5 & \mathsf{y}_{41}\\ 0 & 2 & 1 & \mathsf{y}_{42}\\ 3 & 6 & 2 & 5\end{bmatrix}\begin{bmatrix}1\\1\\1\\1\end{bmatrix}|\cdots] = \begin{bmatrix}5.73 & 7 & 5 & 6.4\\ 0 & 2 & 1 & 1.3\\ 3 & 6 & 2 & 5\end{bmatrix}\begin{bmatrix}1\\1\\1\\1\end{bmatrix} = \begin{bmatrix}24.13\\4.30\\16.00\end{bmatrix}$$

(63.4.19)

$$\mathcal{E}[\mathbf{Y}^\top\mathbf{Y}|\cdots] = \mathcal{E}[\begin{bmatrix}\mathsf{y}_{11} & 7 & 5 & \mathsf{y}_{41}\\ 0 & 2 & 1 & \mathsf{y}_{42}\\ 3 & 6 & 2 & 5\end{bmatrix}\begin{bmatrix}\mathsf{y}_{11} & 0 & 3\\ 7 & 2 & 6\\ 5 & 1 & 2\\ \mathsf{y}_{41} & \mathsf{y}_{42} & 5\end{bmatrix}] = \begin{bmatrix}148.05 & 27.27 & 101.18\\ 27.27 & 6.97 & 20.50\\ 101.18 & 20.50 & 74.00\end{bmatrix}$$

The next step is to plug those estimated values of $\mathbf{Y}^\top\boldsymbol{\iota}$ and $\mathbf{Y}^\top\mathbf{Y}$ into the likelihood function and get the maximum likelihood estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, in other words, set mean and dispersion matrix equal to the sample mean vector and sample dispersion matrices computed from these complete sufficient statistics:

$$\tilde{\tilde{\boldsymbol{\mu}}} = \frac{1}{n}\mathbf{Y}^\top\boldsymbol{\iota} = \begin{bmatrix}6.03\\1.08\\4.00\end{bmatrix} \tag{63.4.20}$$

$$\tilde{\tilde{\boldsymbol{\Sigma}}} = \frac{1}{n}\mathbf{Y}^\top(\boldsymbol{I} - \frac{1}{n}\boldsymbol{\iota}\boldsymbol{\iota}^\top)\mathbf{Y} = \frac{1}{n}\mathbf{Y}^\top\mathbf{Y} - \tilde{\tilde{\boldsymbol{\mu}}}\tilde{\tilde{\boldsymbol{\mu}}}^\top = \begin{bmatrix}.65 & .31 & 1.18\\ .31 & .58 & .81\\ 1.18 & .81 & 2.50\end{bmatrix}, \tag{63.4.21}$$

then predict the missing observations anew.

63.5. Wishart Distribution

The *Wishart distribution* is a multivariate generalization of the $\sigma^2\chi^2$. The noncentral Wishart is the distribution of $\mathbf{Y}^\top\mathbf{Y}$ if $\mathbf{Y}$ is normally distributed as above. But we will be mainly interested in the central Wishart distribution.

Let $\mathbf{Z} = \begin{bmatrix}\boldsymbol{z}_1^\top\\ \vdots\\ \boldsymbol{z}_r^\top\end{bmatrix}$ where $\boldsymbol{z}_j \sim \mathsf{NID}(\boldsymbol{o},\boldsymbol{\Sigma})$. Then the joint distribution of $\mathbf{Z}^\top\mathbf{Z} = \sum_{j=1}^r \boldsymbol{z}_j\boldsymbol{z}_j^\top$ is called a (central) Wishart distribution, notation $\mathbf{Z}^\top\mathbf{Z} \sim \boldsymbol{W}(r,\boldsymbol{\Sigma})$. r is the number of degrees of freedom. The following theorem is exactly parallel to theorem 10.4.3.

THEOREM 63.5.1. *Let* $\mathbf{Z} = \begin{bmatrix}\boldsymbol{z}_1^\top\\ \vdots\\ \boldsymbol{z}_n^\top\end{bmatrix}$ *where* $\boldsymbol{z}_j \sim \mathsf{NID}(\boldsymbol{o},\boldsymbol{\Sigma})$*, and let* $\boldsymbol{P}$ *be symmetric and of rank* r*. A necessary and sufficient condition for* $\mathbf{Z}^\top\boldsymbol{P}\mathbf{Z}$ *to have a Wishart distribution with covariance matrix* $\boldsymbol{\Sigma}$ *is* $\boldsymbol{P}^2 = \boldsymbol{P}$*. In this case, this Wishart distribution has* r *degrees of freedom.*

Proof of sufficiency: If $\boldsymbol{P}^2 = \boldsymbol{P}$ with rank r, a $r \times n$ matrix $\boldsymbol{T}$ exists with $\boldsymbol{P} = \boldsymbol{T}^\top\boldsymbol{T}$ and $\boldsymbol{T}\boldsymbol{T}^\top = \boldsymbol{I}$. Therefore $\mathbf{Z}^\top\boldsymbol{P}\mathbf{Z} = \mathbf{Z}^\top\boldsymbol{T}^\top\boldsymbol{T}\mathbf{Z}$. Define $\mathbf{X} = \boldsymbol{T}\mathbf{Z}$. Writing $\mathbf{x}_i$ for the *column* vectors of $\mathbf{X}$, we know $\mathcal{C}[\mathbf{x}_i,\mathbf{x}_j] = \sigma_{ij}\boldsymbol{T}\boldsymbol{T}^\top = \sigma_{ij}\boldsymbol{I}$. For the *rows* of $\mathbf{X}$ this means they are independent of each other and each of them $\sim \mathsf{N}(\boldsymbol{o},\boldsymbol{\Sigma})$. Since there are r rows, the result follows.

Necessity: Take a vector $\boldsymbol{c}$ with $\boldsymbol{c}^\top\boldsymbol{\Sigma}\boldsymbol{c} = 1$. Then $\boldsymbol{c}^\top\boldsymbol{z}_j \sim \mathsf{N}(0,1)$ for each j, and $\boldsymbol{c}^\top\boldsymbol{z}_j$ independent of $\boldsymbol{c}^\top\boldsymbol{z}_k$ for $j \neq k$. Therefore $\mathbf{Z}\boldsymbol{c} \sim \mathsf{N}(\boldsymbol{o},\boldsymbol{I})$. It follows also $\boldsymbol{T}\mathbf{Z}\boldsymbol{c} = \mathbf{X}\boldsymbol{c} \sim \mathsf{N}(\boldsymbol{o},\boldsymbol{I})$ (the first vector having n and the second r components). Therefore $\boldsymbol{c}^\top\mathbf{Z}^\top\boldsymbol{P}\mathbf{Z}\boldsymbol{c}$ is distributed as a χ^2, therefore we can use the necessity condition in theorem 10.4.3 to show that $\boldsymbol{P}$ is idempotent.

As an application it follows from (63.1.2) that $\mathbf{S}^{(n)} \sim \boldsymbol{W}(n-1,\boldsymbol{\Sigma})$.

One can also show the following generalization of Craig's theorem: If $\mathbf{Z}$ as above, then $\mathbf{Z}^\top\boldsymbol{P}\mathbf{Z}$ is independent of $\mathbf{Z}^\top\boldsymbol{Q}\mathbf{Z}$ if and only if $\boldsymbol{P}\boldsymbol{Q} = \boldsymbol{O}$.

63.6. Sample Correlation Coefficients

What is the distribution of the sample correlation coefficients, and also of the various multiple and partial correlation coefficients in the above model? Suffice it to remark at this point that this is a notoriously difficult question. We will only look at one special case, which also illustrates the use of random orthogonal transformations.

Look at the following scenario: our matrix $\mathbf{Y}$ hat two columns only, write it as

$$\mathbf{Y} = \begin{bmatrix}\mathbf{u} & \boldsymbol{v}\end{bmatrix} = \begin{bmatrix}\mathsf{u}_1 & \mathsf{v}_1 \\ \vdots & \vdots \\ \mathsf{u}_n & \mathsf{v}_n\end{bmatrix} \tag{63.6.1}$$

and we assume each row $\mathbf{y}_j = \begin{bmatrix}\mathsf{u}_j \\ \mathsf{v}_j\end{bmatrix}$ to be an independent sample of the same bivariate normal distribution, characterized by the means μ_{u}, μ_{v}, the variances σ_{uu}, σ_{vv}, and the correlation coefficient ρ (but none of these five paramters are known). The goal is to compute the distribution of the sample correlation coefficient

$$\mathsf{r} = \frac{\sum(\mathsf{u}_i - \bar{\mathsf{u}})(\mathsf{v}_i - \bar{\mathsf{v}})}{\sqrt{\sum(\mathsf{u}_i - \bar{\mathsf{u}})^2}\sqrt{\sum(\mathsf{v}_i - \bar{\mathsf{v}})^2}} \tag{63.6.2}$$

if the true ρ is zero.

We know that $\mathbf{u} \sim \mathsf{N}(\boldsymbol{o}, \sigma_{\mathsf{uu}}\boldsymbol{I})$. Under the null hypothesis, $\mathbf{u}$ is independent of $\boldsymbol{v}$, therefore its distribution conditionally on $\boldsymbol{v}$ is the same as its unconditional distribution. Furthermore look at the matrix consisting of random elements

$$\mathbf{P} = \begin{bmatrix} 1/\sqrt{n} & \cdots & 1/\sqrt{n} \\ (\mathsf{v}_1 - \bar{\mathsf{v}})/\sqrt{\mathsf{s}_{\mathsf{vv}}} & \cdots & (\mathsf{v}_n - \bar{\mathsf{v}})/\sqrt{\mathsf{s}_{\mathsf{vv}}}\end{bmatrix} \tag{63.6.3}$$

It satisfies $\mathbf{P}\mathbf{P}^\top = \boldsymbol{I}$, i.e., it is incomplete orthogonal. The use of random orthogonal transformations is an important trick which simplifies many proofs in multivariate statistics. Conditionally on $\boldsymbol{v}$, the matrix $\mathbf{P}$ is of course constant, and therefore, by theorem 10.4.2, conditionally on $\boldsymbol{v}$ the vector $\boldsymbol{w} = \mathbf{P}\mathbf{u}$ is standard normal with same variance σ_{uu}, and $\mathsf{q} = \mathbf{u}^\top\mathbf{u} - \boldsymbol{w}^\top\boldsymbol{w}$ is an independent $\sigma_{\mathsf{uu}}\chi^2_{n-2}$. In other words, conditionally on $\boldsymbol{v}$, the following three variables are mutually independent and have the following distributions:

$$w_1 = \sqrt{n}\bar{\mathsf{u}} \sim \mathsf{N}(0, \sigma_{\mathsf{uu}}) \tag{63.6.4}$$

$$w_2 = \frac{\sum(\mathsf{u}_i - \bar{\mathsf{u}})(\mathsf{v}_i - \bar{\mathsf{v}})}{\sqrt{\sum(\mathsf{v}_i - \bar{\mathsf{v}})^2}} = \mathsf{r}\sqrt{\mathsf{s}_{\mathsf{uu}}} \sim \mathsf{N}(0, \sigma_{\mathsf{uu}}) \tag{63.6.5}$$

$$\mathsf{q} = \sum \mathsf{u}_i^2 - n\bar{\mathsf{u}}^2 - w_2^2 = (1 - \mathsf{r}^2)\mathsf{s}_{\mathsf{uu}} \sim \sigma_{\mathsf{uu}}\chi^2_{n-2} \tag{63.6.6}$$

Since the values of $\boldsymbol{v}$ do not enter any of these distributions, these are also the unconditional distributions. Therefore we can form a simple function of r which has a t-distribution:

$$\frac{w_2}{\sqrt{\mathsf{q}/(n-2)}} = \frac{\mathsf{r}\sqrt{n-2}}{\sqrt{1-\mathsf{r}^2}} \sim t_{n-2} \tag{63.6.7}$$

This can be used to test whether $\rho = 0$.

CHAPTER 64

Pooling of Cross Section and Time Series Data

Given m cross-sectional units, each of which has been observed for t time periods. The dependent variable for cross sectional unit i at time s is y_{si}. There are also k independent variables, and the value of the jth independent variable for cross sectional unit i at time s is x_{sij}. I.e., instead of a vector, the dependent variable is a matrix, and instead of a matrix, the independent variables form a 3-way array. We will discuss three different models here which assign equal slope parameters to the different cross-sectional units but which differ in their treatment of the intercept.

64.1. OLS Model

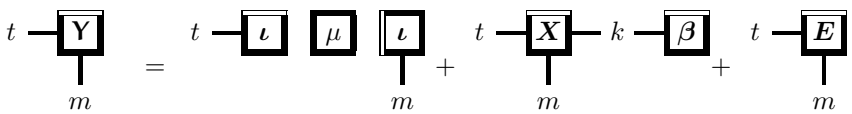
The most restrictive model of the three assumes that all cross-sectional units have the same intercept μ. I.e.,

$$\mathsf{y}_{si} = \mu + \sum_{j=1}^{k} x_{sij}\beta_j + \varepsilon_{si} \qquad s = 1, \ldots, t, \quad i = 1, \ldots, m, \tag{64.1.1}$$

where the error terms are uncorrelated and have equal variance σ_ε^2.

In tile notation:

(64.1.2)



In matrix notation this model can be written as

$$\boldsymbol{Y} = \boldsymbol{\iota}\mu\boldsymbol{\iota}^\top + \begin{bmatrix}\boldsymbol{X}_1\boldsymbol{\beta} & \cdots & \boldsymbol{X}_m\boldsymbol{\beta}\end{bmatrix} + \boldsymbol{E} \tag{64.1.3}$$

where $\boldsymbol{Y} = \begin{bmatrix}\mathsf{y}_1 & \cdots & \mathsf{y}_m\end{bmatrix}$ is $t \times m$, each of the $\boldsymbol{X}_i$ is $t \times k$, the first $\boldsymbol{\iota}$ is the t-vector of ones and the second $\boldsymbol{\iota}$ the m-vector of ones, μ is the intercept and $\boldsymbol{\beta}$ the k-vector of slope coefficients, and $\boldsymbol{E} = \begin{bmatrix}\varepsilon_1 & \cdots & \varepsilon_m\end{bmatrix}$ the matrix of disturbances. The notation $\begin{bmatrix}\boldsymbol{X}_1\boldsymbol{\beta} & \cdots & \boldsymbol{X}_m\boldsymbol{\beta}\end{bmatrix}$ represents a matrix obtained by the multiplication of a 3-way array with a vector. We assume $\operatorname{vec}\boldsymbol{E} \sim \boldsymbol{o}, \sigma^2\boldsymbol{I}$.

If one vectorizes this one gets

(64.1.4)

$$\operatorname{vec}(\boldsymbol{Y}) = \begin{bmatrix}\boldsymbol{\iota} & \boldsymbol{X}_1 \\ \boldsymbol{\iota} & \boldsymbol{X}_2 \\ \vdots & \vdots \\ \boldsymbol{\iota} & \boldsymbol{X}_m\end{bmatrix}\begin{bmatrix}\mu \\ \boldsymbol{\beta}\end{bmatrix} + \operatorname{vec}(\boldsymbol{E}) \quad \text{or} \quad \operatorname{vec}(\boldsymbol{Y}) = \begin{bmatrix}\boldsymbol{\iota} \\ \boldsymbol{\iota} \\ \vdots \\ \boldsymbol{\iota}\end{bmatrix}\mu + \begin{bmatrix}\boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_m\end{bmatrix}\boldsymbol{\beta} + \operatorname{vec}(\boldsymbol{E})$$

Using the abbreviation

$$\boldsymbol{Z} = \begin{bmatrix}\boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_m\end{bmatrix} \tag{64.1.5}$$

this can also be written

$$\operatorname{vec}(\boldsymbol{Y}) = \boldsymbol{\iota}\mu + \boldsymbol{Z}\boldsymbol{\beta} + \operatorname{vec}(\boldsymbol{E}) = \begin{bmatrix}\boldsymbol{\iota} & \boldsymbol{Z}\end{bmatrix} + \begin{bmatrix}\mu \\ \boldsymbol{\beta}\end{bmatrix}. \tag{64.1.6}$$

Problem 520. *1 point Show that* $\operatorname{vec}(\begin{bmatrix}\boldsymbol{X}_1\boldsymbol{\beta} & \cdots & \boldsymbol{X}_m\boldsymbol{\beta}\end{bmatrix}) = \boldsymbol{Z}\boldsymbol{\beta}$ *with* $\boldsymbol{Z}$ *as just defined.*

Answer.

$$\operatorname{vec}(\begin{bmatrix}\boldsymbol{X}_1\boldsymbol{\beta} & \cdots & \boldsymbol{X}_m\boldsymbol{\beta}\end{bmatrix}) = \begin{bmatrix}\boldsymbol{X}_1\boldsymbol{\beta} \\ \vdots \\ \boldsymbol{X}_m\boldsymbol{\beta}\end{bmatrix} = \begin{bmatrix}\boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_m\end{bmatrix}\boldsymbol{\beta} = \boldsymbol{Z}\boldsymbol{\beta} \tag{64.1.7}$$

□

One gets the paramater estimates by regressing running OLS on (64.1.4), i.e., regressing $\operatorname{vec}\boldsymbol{Y}$ on $\boldsymbol{Z}$ with an intercept.

64.2. The Between-Estimator

By premultiplying (64.1.3) by $\frac{1}{t}\boldsymbol{\iota}^\top$ one obtains the so-called "between"-regression. Defining $\bar{\mathsf{y}}^\top = \frac{1}{t}\boldsymbol{\iota}^\top\boldsymbol{Y}$, i.e., $\bar{\mathsf{y}}^\top$ is the row vector consisting of the column means, and in the same way $\bar{\boldsymbol{x}}_i^\top = \frac{1}{t}\boldsymbol{\iota}^\top\boldsymbol{X}_i$ and $\bar{\varepsilon}^\top = \frac{1}{t}\boldsymbol{\iota}^\top\boldsymbol{E}$, one obtains

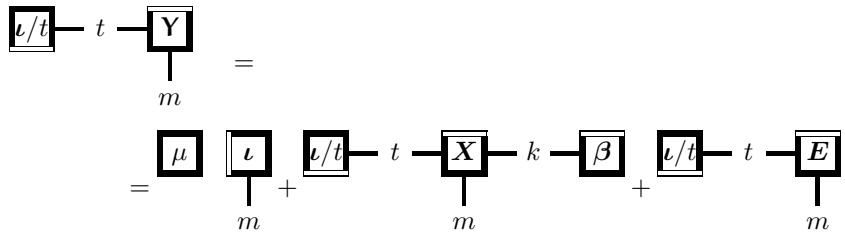
(64.2.1)

$$\bar{\mathsf{y}}^\top = \mu\boldsymbol{\iota}^\top + \begin{bmatrix}\bar{\boldsymbol{x}}_1^\top\boldsymbol{\beta} & \cdots & \bar{\boldsymbol{x}}_m^\top\boldsymbol{\beta}\end{bmatrix} + \bar{\varepsilon}^\top = \mu\boldsymbol{\iota}^\top + (\bar{\boldsymbol{X}}\boldsymbol{\beta})^\top + \bar{\varepsilon}^\top \quad \text{where} \quad \bar{\boldsymbol{X}} = \begin{bmatrix}\bar{\boldsymbol{x}}_1^\top \\ \vdots \\ \bar{\boldsymbol{x}}_m^\top\end{bmatrix}.$$

If one transposes this one obtains $\bar{\mathsf{y}} = \boldsymbol{\iota}\mu + \bar{\boldsymbol{X}}\boldsymbol{\beta} + \bar{\varepsilon}$.

In tiles, the between model is obtained from (64.1.2) by attaching $\boxed{\boldsymbol{\iota}/t}$ — t :

(64.2.2)



If one runs this regression one will get estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\beta}$ which are less efficient than those from the full regression. But these regressions are consistent even if the error terms in the same column are correlated (as they are in the Random Effects model).

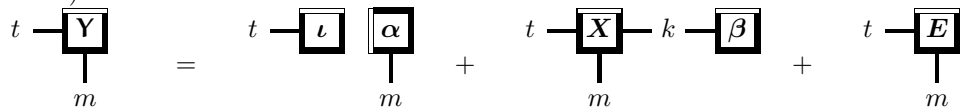
64.3. Dummy Variable Model (Fixed Effects)

While maintaining the assumption that the cross sectional units have the same slope parameters, we are now allowing a different intercept for each unit. I.e., the model is now

$$\mathsf{y}_{si} = \alpha_i + \sum_{j=1}^{k} x_{sij}\beta_j + \varepsilon_{si} \qquad s = 1, \ldots, t, \quad i = 1, \ldots, m, \tag{64.3.1}$$

where the error terms are uncorrelated and have equal variance σ_ε^2. In tile notation this is

(64.3.2)



One can write this model as the matrix equation

$$\mathsf{Y} = \boldsymbol{\iota}\boldsymbol{\alpha}^\top + \begin{bmatrix}\boldsymbol{X}_1\boldsymbol{\beta} & \cdots & \boldsymbol{X}_m\boldsymbol{\beta}\end{bmatrix} + \boldsymbol{E} \tag{64.3.3}$$

where $\mathsf{Y} = \begin{bmatrix}\mathsf{y}_1 & \cdots & \mathsf{y}_m\end{bmatrix}$ is $t \times m$, each of the $\boldsymbol{X}_i$ is $t \times k$, $\boldsymbol{\iota}$ is the t-vector of ones, $\boldsymbol{\alpha}$ is the m-vector collecting all the intercept terms, $\boldsymbol{\beta}$ the k-vector of slope coefficients, $\boldsymbol{E} = \begin{bmatrix}\boldsymbol{\varepsilon}_1 & \cdots & \boldsymbol{\varepsilon}_m\end{bmatrix}$ the matrix of disturbances. We assume $\operatorname{vec}\boldsymbol{E} \sim \boldsymbol{o}, \sigma^2\boldsymbol{I}$.

For estimation, it is convenient to vectorize (64.3.3) to get

$$\begin{bmatrix}\mathsf{y}_1\\ \vdots\\ \mathsf{y}_m\end{bmatrix} = \begin{bmatrix}\boldsymbol{\iota} & \boldsymbol{o} & \cdots & \boldsymbol{o}\\ \boldsymbol{o} & \boldsymbol{\iota} & \cdots & \boldsymbol{o}\\ \vdots & \vdots & \ddots & \vdots\\ \boldsymbol{o} & \boldsymbol{o} & \cdots & \boldsymbol{\iota}\end{bmatrix}\begin{bmatrix}\alpha_1\\ \vdots\\ \alpha_m\end{bmatrix} + \begin{bmatrix}\boldsymbol{X}_1\\ \vdots\\ \boldsymbol{X}_m\end{bmatrix}\boldsymbol{\beta} + \begin{bmatrix}\boldsymbol{\varepsilon}_1\\ \vdots\\ \boldsymbol{\varepsilon}_m\end{bmatrix} \tag{64.3.4}$$

Problem 521. *2 points Show that* $\operatorname{vec}(\boldsymbol{\iota}\boldsymbol{\alpha}^\top) = \boldsymbol{K}\boldsymbol{\alpha}$ *where* $\boldsymbol{K} = \boldsymbol{I} \otimes \boldsymbol{\iota}$ *is the matrix of dummies in* (64.3.4). *This is a special case of* (B.5.19), *but I would like you to prove it from scratch without using* (B.5.19).

Answer. $\boldsymbol{\iota}\boldsymbol{\alpha}^\top = \begin{bmatrix}\boldsymbol{\iota}\alpha_1 & \cdots & \boldsymbol{\iota}\alpha_m\end{bmatrix}$ and

$$\boldsymbol{K}\boldsymbol{\alpha} = \begin{bmatrix}\boldsymbol{\iota} & \boldsymbol{o} & \cdots & \boldsymbol{o}\\ \boldsymbol{o} & \boldsymbol{\iota} & \cdots & \boldsymbol{o}\\ \vdots & \vdots & \ddots & \vdots\\ \boldsymbol{o} & \boldsymbol{o} & \cdots & \boldsymbol{\iota}\end{bmatrix}\begin{bmatrix}\alpha_1\\ \vdots\\ \alpha_m\end{bmatrix} = \begin{bmatrix}\boldsymbol{\iota}\alpha_1\\ \vdots\\ \boldsymbol{\iota}\alpha_m\end{bmatrix}.$$

□

Using the $\boldsymbol{K}$ defined in Problem 521 and the $\boldsymbol{Z}$ defined in (64.1.5), (64.3.4) can also be written as

$$\operatorname{vec}(\mathsf{Y}) = \boldsymbol{K}\boldsymbol{\alpha} + \boldsymbol{Z}\boldsymbol{\beta} + \operatorname{vec}(\boldsymbol{E}) \tag{64.3.5}$$

[**JHG+88**] give a good example how such a model can arise: s is years, i is firms, y_{si} is costs, and there is only one x_{si} for every firm (i.e. $k = 1$), which is sales. These firms would have equal marginal costs but different fixed overhead charges.

In principle (64.3.4) presents no estimation problems, it is OLS with lots of dummy variables (if there are lots of cross-sectional units). But often it is advantageous to use the following sequential procedure: (1) in oder to get $\hat{\boldsymbol{\beta}}$ regress

$$\begin{bmatrix}\boldsymbol{D}\mathsf{y}_1\\ \vdots\\ \boldsymbol{D}\mathsf{y}_m\end{bmatrix} = \begin{bmatrix}\boldsymbol{D}\boldsymbol{X}_1\\ \vdots\\ \boldsymbol{D}\boldsymbol{X}_m\end{bmatrix}\hat{\boldsymbol{\beta}} + \text{residuals} \tag{64.3.6}$$

without a constant term (but if you leave the constant term in, this does not matter either, its coefficient will be exactly zero). Here $\boldsymbol{D}$ is the matrix which takes the mean out. I.e., take the mean out of every y individually and out of every $\boldsymbol{X}$ before running the regression. (2) Then you get each $\hat{\alpha}_i$ by the following equation:

$$\hat{\alpha}_i = \bar{\mathsf{y}}_i - \bar{\boldsymbol{x}}_i^\top\hat{\boldsymbol{\beta}} \tag{64.3.7}$$

where $\bar{\mathsf{y}}_i$ is the mean of y_i, and $\bar{\boldsymbol{x}}_i^\top$ is the row vector consisting of the column means of $\boldsymbol{X}_i$.

Problem 522. *Give a mathematical proof that this is the right procedure*

Answer. Equation (64.3.4) has the form of (30.0.1). Define $\boldsymbol{D} = \boldsymbol{I} - \boldsymbol{\iota}\boldsymbol{\iota}^\top/t$ and $\boldsymbol{W} = \boldsymbol{I} - \boldsymbol{K}(\boldsymbol{K}^\top\boldsymbol{K})^{-1}\boldsymbol{K}^\top = \boldsymbol{I} \otimes \boldsymbol{D}$. According to (30.0.3) and (30.0.4), $\hat{\boldsymbol{\beta}}$ and the vector of residuals can be obtained by regressing $\boldsymbol{W}\,\mathrm{vec}(\mathbf{Y})$ on $\boldsymbol{W}\boldsymbol{Z}$, and if one plugs this estimate $\hat{\boldsymbol{\beta}}$ back into the formula, then one obtains an estimate of $\boldsymbol{\alpha}$.

Without using the Kronecker product, this procedure can be described as follows: one gets the right $\hat{\boldsymbol{\beta}}$ if one estimates (64.3.3) premultiplied by $\boldsymbol{D}$. Since $\boldsymbol{D}\boldsymbol{\iota} = \boldsymbol{o}$, this premultiplication removes the first parameter vector $\boldsymbol{\alpha}$ from the regression, so that only

$$\boldsymbol{D}\mathbf{Y} = \begin{bmatrix}\boldsymbol{D}\boldsymbol{X}_1\boldsymbol{\beta} & \cdots & \boldsymbol{D}\boldsymbol{X}_m\boldsymbol{\beta}\end{bmatrix} + \boldsymbol{D}\boldsymbol{E} \tag{64.3.8}$$

remains—or, in vectorized form,

$$\begin{bmatrix}\boldsymbol{D}\mathbf{y}_1\\ \vdots \\ \boldsymbol{D}\mathbf{y}_m\end{bmatrix} = \begin{bmatrix}\boldsymbol{D}\boldsymbol{X}_1\\ \vdots \\ \boldsymbol{D}\boldsymbol{X}_m\end{bmatrix}\boldsymbol{\beta} + \begin{bmatrix}\boldsymbol{D}\boldsymbol{\varepsilon}_1\\ \vdots \\ \boldsymbol{D}\boldsymbol{\varepsilon}_m\end{bmatrix} \tag{64.3.9}$$

Although $\mathrm{vec}(\boldsymbol{D}\boldsymbol{E})$ is no longer spherically distributed, it can be shown that in the present situation the OLS of $\boldsymbol{\beta}$ is the BLUE.

After having obtained $\hat{\boldsymbol{\beta}}$, one obtains $\hat{\boldsymbol{\alpha}}$ by plugging this estimated $\hat{\boldsymbol{\beta}}$ into (64.3.3), which gives

$$\mathbf{Y} - \begin{bmatrix}\boldsymbol{X}_1\hat{\boldsymbol{\beta}} & \cdots & \boldsymbol{X}_m\hat{\boldsymbol{\beta}}\end{bmatrix} = \boldsymbol{\iota}\boldsymbol{\alpha}^\top + \boldsymbol{E} \tag{64.3.10}$$

Here each column of $\mathbf{Y}$ is independent of all the others, they no longer share common parameters, therefore one can run this regression column by column:

$$\mathbf{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}} = \boldsymbol{\iota}\alpha_i + \boldsymbol{\varepsilon}_i \qquad i = 1, \ldots, m \tag{64.3.11}$$

Since the regressor is the column of ones, one can write down the result immediately:

$$\hat{\alpha}_i = \bar{\mathbf{y}}_i - \bar{\boldsymbol{x}}_i^\top\hat{\boldsymbol{\beta}} \tag{64.3.12}$$

where $\bar{\mathbf{y}}_i$ is the mean of $\mathbf{y}_i$, and $\bar{\boldsymbol{x}}_i^\top$ is the row vector consisting of the column means of $\boldsymbol{X}_i$. □

To get the unbiased estimate of σ^2, one can almost take the s^2 from the regression (64.3.9), one only has to adjust it for the numbers of degrees of freedom.

Problem 523. *We are working in the dummy-variable model for pooled data, which can be written as*

$$\mathbf{Y} = \boldsymbol{\iota}\boldsymbol{\alpha}^\top + \begin{bmatrix}\boldsymbol{X}_1\boldsymbol{\beta} & \cdots & \boldsymbol{X}_m\boldsymbol{\beta}\end{bmatrix} + \boldsymbol{E} \tag{64.3.13}$$

where $\mathbf{Y} = \begin{bmatrix}\mathbf{y}_1 & \cdots & \mathbf{y}_m\end{bmatrix}$ *is* $t \times m$*, each of the* $\boldsymbol{X}_i$ *is* $t \times k$*,* $\boldsymbol{\iota}$ *is the* t*-vector of ones,* $\boldsymbol{E}$ *is a* $t \times m$ *matrix of identically distributed independent error terms with zero mean, and* $\boldsymbol{\alpha}$ *is a* m*-vector and* $\boldsymbol{\beta}$ *a* k*-vector of unknown nonrandom parameters.*

• a. *3 points Describe in words the characteristics of this model and how it can come about.*

Answer. Each of the m units has a different intercept, slope is the same. Equal marginal costs but different fixed costs. □

• b. *4 points Describe the issues in estimating this model and how it should be estimated.*

Answer. After vectorization OLS is fine, but design matrix very big. One can derive formulas that are easier to evaluate numerically because they involve smaller matrices, by exploiting the structure of the overall design matrix. First estimate the slope parameters by sweeping out the means, then the intercepts. □

• c. *3 points Set up an* F*-test testing whether the individual intercept parameters are indeed different, by running two separate regressions on the restricted and the unrestricted model and using the generic formula* (42.1.4) *for the* F*-test. Describe how you would run the restricted and how the unrestricted model. Give the number of constraints, the number of observations, and the number of coefficients in the unrestricted model in terms of* m*,* t*, and* k*.*

Answer. The unrestricted regression is the dummy variables regression which was described here: first form $\boldsymbol{D}\mathbf{Y}$ and all the $\boldsymbol{D}\boldsymbol{X}_i$, then run regression (64.3.9) without intercept, which is already enough to get the SSE_r.

Number of constraints is $m - 1$, number of observations is tm, and number of coefficients in the unrestricted model is $k + m$. The test statistic is given in [**JHG+88**, (11.4.25) on p. 475]:

$$\mathsf{F} = \frac{(\mathsf{SSE}_r - \mathsf{SSE}_u)/(m-1)}{\mathsf{SSE}_u/(mt - m - k)} \tag{64.3.14}$$

□

• d. *3 points An alternative model specification is the variance components model. Describe it as well as you can, and discuss situations when it would be more appropriate than the model above.*

Answer. If one believes that variances are similar, and if one is not interested in those particular firms in the sample, but in all firms. □

Problem 524. *3 points Enumerate as many commonalities and differences as you can between the dummy variable model for pooling cross sectional and time series data, and the seemingly unrelated regression model.*

Answer. Both models involve different cross-sectional units in overlapping time intervals. In the SUR model, the different equations are related through the disturbances only, while in the dummy variable model, no relationship at all is going through the disturbances, all the errors are independent! But in the dummy variable model, the equations are strongly related since all slope coefficients are equal in the different equations, only the intercepts may differ. In the SUR model, there is no relationship between the parameters in the different equations, the parameter vectors may even be of differant lengths. Unlike [**JHG+88**], I would not call the dummy variable model a special case of the SUR modiel, since I would no longer call it a SUR model if there are cross-equation restrictions. □

64.4. Relation between the three Models so far:

The Dummy-Variable model can also be called the "within" model and the estimation of the constrained model with grouped data is the "between"-model. [**Gre97**, 14.3.2] shows that the OLS estimator in the restricted model is a matrix-weighted

average of the between-group and the within-group estimator. If all intercepts are the same, then the between-model and the within-model are both inefficient, but in complementary ways.

64.5. Variance Components Model (Random Effects)

The model can still be written as

$$\mathfrak{y}_{si} = \alpha_i + \sum_{j=1}^{k} x_{sij}\beta_j + \varepsilon_{si} \qquad s = 1, \ldots, t, \quad i = 1, \ldots, m \tag{64.5.1}$$

or

$$\mathbf{Y} = \boldsymbol{\iota}\boldsymbol{\alpha}^\top + \begin{bmatrix} \boldsymbol{X}_1\boldsymbol{\beta} & \cdots & \boldsymbol{X}_m\boldsymbol{\beta} \end{bmatrix} + \boldsymbol{E} \tag{64.5.2}$$

with i.i.d. error terms ε whose variance is σ^2_ε, but this time the α_i are random too, they are elements of the vector $\boldsymbol{\alpha} \sim (\boldsymbol{\iota}\mu, \sigma^2_\alpha \boldsymbol{I})$ which is uncorrelated with $\boldsymbol{E}$. Besides $\boldsymbol{\beta}$, the two main parameters to be estimated are μ and σ^2_α, but sometimes one may also want to predict $\boldsymbol{\alpha}$.

In our example of firms, this specification would be appropriate if we are not interested in the fixed costs associated with the specific firms in our sample, but want to know the mean and variance of the fixed costs for all firms.

With the definition $\boldsymbol{\delta} = \boldsymbol{\alpha} - \boldsymbol{\iota}\mu$ for the random part of the intercept term, (64.5.2) becomes

$$\mathbf{Y} = \boldsymbol{\iota}\boldsymbol{\iota}^\top\mu + \begin{bmatrix} \boldsymbol{X}_1\boldsymbol{\beta} & \cdots & \boldsymbol{X}_m\boldsymbol{\beta} \end{bmatrix} + \boldsymbol{\iota}\boldsymbol{\delta}^\top + \boldsymbol{E} \tag{64.5.3}$$

The only difference between (64.1.3) and (64.5.3) is that the error term is now $\boldsymbol{\iota}\boldsymbol{\delta}^\top + \boldsymbol{E}$, which still has zero mean but no longer a spherical covariance matrix. The columns of $\boldsymbol{\iota}\boldsymbol{\delta}^\top + \boldsymbol{E}$ are independent of each other, and each of the columns has the same covariance matrix $\boldsymbol{V}$, which is "equicorrelated":

$$\boldsymbol{V} = \sigma^2_\alpha \boldsymbol{\iota}\boldsymbol{\iota}^\top + \sigma^2_\varepsilon \boldsymbol{I}_t = \begin{bmatrix} \sigma^2_\alpha + \sigma^2_\varepsilon & \sigma^2_\alpha & \cdots & \sigma^2_\alpha \\ \sigma^2_\alpha & \sigma^2_\alpha + \sigma^2_\varepsilon & \cdots & \sigma^2_\alpha \\ \vdots & \vdots & \ddots & \vdots \\ \sigma^2_\alpha & \sigma^2_\alpha & \cdots & \sigma^2_\alpha + \sigma^2_\varepsilon \end{bmatrix} \tag{64.5.4}$$

Problem 525. *3 points Using Problems 612 and 613 show that the covariance matrix of the error term in the random coefficients model (after the random part of the intercept has been added to it) is* $\mathcal{V}[\mathrm{vec}(\boldsymbol{\iota}\boldsymbol{\delta}^\top + \boldsymbol{E})] = \boldsymbol{I}_m \otimes \boldsymbol{V}$, *where* $\boldsymbol{V}$ *is defined in* (64.5.4).

Answer. $\mathcal{V}[\mathrm{vec}(\boldsymbol{\iota}\boldsymbol{\delta}^\top + \boldsymbol{E})] = \mathcal{V}[\mathrm{vec}(\boldsymbol{\iota}\boldsymbol{\delta}^\top) + \mathrm{vec}(\boldsymbol{E})] = \mathcal{V}[\boldsymbol{\delta} \otimes \boldsymbol{\iota} + \mathrm{vec}(\boldsymbol{E})]$. Now $\mathcal{V}[\boldsymbol{\delta} \otimes \boldsymbol{\iota}] = \mathcal{V}[\boldsymbol{\delta}] \otimes \boldsymbol{\iota}\boldsymbol{\iota}^\top = \sigma^2_\alpha \boldsymbol{I}_m \otimes \boldsymbol{\iota}\boldsymbol{\iota}^\top$. Furthermore, $\mathcal{V}[\mathrm{vec}(\boldsymbol{E})] = \sigma^2_\varepsilon \boldsymbol{I}_{tm} = \sigma^2_\varepsilon \boldsymbol{I}_m \otimes \boldsymbol{I}_t$. Since the two are uncorrelated, their covariance matrices add, therefore

$$\mathcal{V}[\mathrm{vec}(\boldsymbol{\iota}\boldsymbol{\delta}^\top + \boldsymbol{E})] = \sigma^2_\alpha \boldsymbol{I}_m \otimes \boldsymbol{\iota}\boldsymbol{\iota}^\top + \sigma^2_\varepsilon \boldsymbol{I}_m \otimes \boldsymbol{I}_t = \boldsymbol{I}_m \otimes (\sigma^2_\alpha \boldsymbol{\iota}\boldsymbol{\iota}^\top + \sigma^2_\varepsilon \boldsymbol{I}_t) = \boldsymbol{I}_m \otimes \boldsymbol{V} \tag{64.5.5}$$

□

Problem 526. *One has m timeseries regressions* $\mathfrak{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i$ *with the same coefficient vector* $\boldsymbol{\beta}$. $\mathcal{V}[\boldsymbol{\varepsilon}_i] = \tau^2\boldsymbol{\Sigma}_i$, *and for* $i \neq j$, $\boldsymbol{\varepsilon}_i$ *is independent of* $\boldsymbol{\varepsilon}_j$. *For the sake of the argument we assume here that all* $\boldsymbol{\Sigma}_i$ *are known.*

• a. *3 points Show that the BLUE in this model based on all observations is*

$$\hat{\hat{\boldsymbol{\beta}}} = \Big(\sum_i \boldsymbol{X}_i^\top \boldsymbol{\Sigma}_i^{-1} \boldsymbol{X}_i\Big)^{-1} \sum_i \boldsymbol{X}_i^\top \boldsymbol{\Sigma}_i^{-1} \mathfrak{y}_i \tag{64.5.6}$$

Answer. In vectorized form the model reads

$$\begin{bmatrix} \mathfrak{y}_1 \\ \vdots \\ \mathfrak{y}_m \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_m \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \vdots \\ \boldsymbol{\varepsilon}_m \end{bmatrix}; \qquad \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \vdots \\ \boldsymbol{\varepsilon}_m \end{bmatrix} \sim \begin{bmatrix} \boldsymbol{o} \\ \vdots \\ \boldsymbol{o} \end{bmatrix}, \tau^2 \begin{bmatrix} \boldsymbol{\Sigma}_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \boldsymbol{\Sigma}_m \end{bmatrix}. \tag{64.5.7}$$

Therefore the GLSE is

$$\hat{\hat{\boldsymbol{\beta}}} = \left(\begin{bmatrix} \boldsymbol{X}_1^\top & \cdots & \boldsymbol{X}_m^\top \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}_1 & \cdots & \boldsymbol{O} \\ \vdots & \ddots & \vdots \\ \boldsymbol{O} & \cdots & \boldsymbol{\Sigma}_m \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_m \end{bmatrix}\right)^{-1} \begin{bmatrix} \boldsymbol{X}_1^\top & \cdots & \boldsymbol{X}_m^\top \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}_1 & \cdots & \boldsymbol{O} \\ \vdots & \ddots & \vdots \\ \boldsymbol{O} & \cdots & \boldsymbol{\Sigma}_m \end{bmatrix}^{-1} \begin{bmatrix} \mathfrak{y}_1 \\ \vdots \\ \mathfrak{y}_m \end{bmatrix}. \tag{64.5.8}$$

One can take the inverses block by block, and gets the above. □

• b. *1 point Show that this is a matrix-weighted average of the BLUE's in the individual timeseries regressions, with the inverses of the covariance matrices of these BLUE's as the weighting matrices.*

Answer. Simple because

$$\hat{\hat{\boldsymbol{\beta}}} = \Big(\sum_i \boldsymbol{X}_i^\top \boldsymbol{\Sigma}_i^{-1} \boldsymbol{X}_i\Big)^{-1} \sum_i \boldsymbol{X}_i^\top \boldsymbol{\Sigma}_i^{-1} \boldsymbol{X}_i \Big(\boldsymbol{X}_i^\top \boldsymbol{\Sigma}_i^{-1} \boldsymbol{X}_i\Big)^{-1} \boldsymbol{X}_i^\top \boldsymbol{\Sigma}_i^{-1} \mathfrak{y}_i \tag{64.5.9}$$

□

Since the columns of $\boldsymbol{\iota}\boldsymbol{\delta}^\top + \boldsymbol{E}$ are independent and have equal covariance matrices, it is possible to transform $\boldsymbol{\iota}\boldsymbol{\delta}^\top + \boldsymbol{E}$ into a matrix of uncorrelated and homoskedastic error terms by simply premultiplying it by a suitable transformation matrix $\boldsymbol{P}$. The following amazing bit of matrix algebra helps us to compute $\boldsymbol{P}$:

Problem 527. *4 points Assume* $\boldsymbol{\Omega}$ *is a symmetric idempotent matrix. Show that for* $\nu \neq 0$ *and* $\omega \neq -\nu$, *the matrix* $\nu\boldsymbol{I} + \omega\boldsymbol{\Omega}$ *has the inverse* $\frac{1}{\nu}\boldsymbol{I} + (\frac{1}{\nu+\omega} - \frac{1}{\nu})\boldsymbol{\Omega}$, *and that the square root of this inverse is* $\frac{1}{\sqrt{\nu}}\boldsymbol{I} + (\frac{1}{\sqrt{\nu+\omega}} - \frac{1}{\sqrt{\nu}})\boldsymbol{\Omega}$.

The connection between the dummy variable model and the error components model becomes most apparent if we scale $\boldsymbol{P}$ such that $\boldsymbol{P}\boldsymbol{V}\boldsymbol{P}^\top = \sigma_\varepsilon^2\boldsymbol{I}$, i.e., $\boldsymbol{P}$ is the square root of the inverse of $\boldsymbol{V}/\sigma_\varepsilon^2$. In Problem 527 we must therefore set $\boldsymbol{\Omega} = \boldsymbol{\iota}\boldsymbol{\iota}^\top/t$, $\nu = 1$, and $\omega = t\sigma_\alpha^2/\sigma_\varepsilon^2$. The matrix which diagonalizes the error covariance matrix is therefore

$$\boldsymbol{P} = \boldsymbol{I} - \gamma\frac{\boldsymbol{\iota}\boldsymbol{\iota}^\top}{t} \qquad \text{where} \quad \gamma = 1 - \sqrt{\frac{\sigma_\varepsilon^2/t}{\sigma_\varepsilon^2/t + \sigma_\alpha^2}} \tag{64.5.10}$$

Problem 528. *4 points Using* (64.5.10) *double-check that every column of* $\boldsymbol{P}\boldsymbol{W}$ *is spherically distributed.*

Answer. Since $\boldsymbol{P}\boldsymbol{\iota} = \boldsymbol{\iota}(1-\gamma)$, $\boldsymbol{P}\boldsymbol{w}_i = \boldsymbol{\iota}\delta_i(1-\gamma) + \boldsymbol{P}\boldsymbol{\varepsilon}_i$ and $\mathcal{V}[\boldsymbol{P}\boldsymbol{w}_i] = \boldsymbol{\iota}\sigma_\alpha^2\frac{\sigma_\varepsilon^2/t}{\sigma_\varepsilon^2/t+\sigma_\alpha^2}\boldsymbol{\iota}^\top + \sigma_\varepsilon^2\boldsymbol{P}\boldsymbol{P}^\top$. Now $\boldsymbol{P}\boldsymbol{P}^\top = \boldsymbol{I} + \boldsymbol{\iota}\frac{\gamma^2-2\gamma}{t}\boldsymbol{\iota}^\top$, and

$$\gamma^2 - 2\gamma = (1-\gamma)^2 - 1 = \frac{-\sigma_\alpha^2}{\sigma_\varepsilon^2/t + \sigma_\alpha^2} \tag{64.5.11}$$

Therefore

$$\sigma_\varepsilon^2\boldsymbol{P}\boldsymbol{P}^\top = \sigma_\varepsilon^2\boldsymbol{I} + \boldsymbol{\iota}\sigma_\varepsilon^2/t(\gamma^2-2\gamma)\boldsymbol{\iota}^\top = \sigma_\varepsilon^2\boldsymbol{I} - \boldsymbol{\iota}\frac{\sigma_\alpha^2\sigma_\varepsilon^2/t}{\sigma_\varepsilon^2/t+\sigma_\alpha^2}\boldsymbol{\iota}^\top \tag{64.5.12}$$

and $\mathcal{V}[\boldsymbol{P}\boldsymbol{w}_i] = \sigma_\varepsilon^2$. □

Problem 529. *1 point Show that* $\boldsymbol{P}\boldsymbol{\iota} = \boldsymbol{\iota}(1-\gamma)$ *and that the other eigenvectors of* $\boldsymbol{P}$ *are exactly the vectors the elements of which sum to 0, with the eigenvalues 1. Derive from this the determinant of* $\boldsymbol{P}$.

Answer. $\boldsymbol{P}\boldsymbol{\iota} = (\boldsymbol{I} - \gamma\frac{\boldsymbol{\iota}\boldsymbol{\iota}^\top}{t})\boldsymbol{\iota} = \boldsymbol{\iota}(1-\gamma)$. Now if a vector $\boldsymbol{a}$ satisfies $\boldsymbol{\iota}^\top\boldsymbol{a} = 0$, then $\boldsymbol{P}\boldsymbol{a} = \boldsymbol{a}$. Since there are $t-1$ independent such vectors, this gives all eigenvectors. $\det(\boldsymbol{P}) = 1-\gamma$ (the product of all eigenvalues). □

Problem 530. *3 points Now write down this likelihood function, see* [**Gre97**, exercise 4 on p. 643].

Answer. Assuming normality, the ith column vector is $\mathbf{y}_i \sim \mathsf{N}(\boldsymbol{\iota}\mu + \boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{V})$ and different columns are independent. Since $\mathcal{V}[\boldsymbol{P}\boldsymbol{w}_i] = \boldsymbol{P}\boldsymbol{V}\boldsymbol{P}^\top = \sigma_\varepsilon^2\boldsymbol{I}$ it follows $\det(\boldsymbol{V}) = \sigma_\varepsilon^{2t}(\det\boldsymbol{P})^{-2}$. Therefore the density function is

$$f_{\mathbf{Y}}(\boldsymbol{Y}) = \prod_{i=1}^m\Big((2\pi)^{-t/2}(\det\boldsymbol{V})^{-1/2}\exp\Big(-\frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{\iota}\mu + \boldsymbol{X}_i\boldsymbol{\beta})^\top\boldsymbol{V}^{-1}(\boldsymbol{y}_i - \boldsymbol{\iota}\mu + \boldsymbol{X}_i\boldsymbol{\beta})\Big)\Big)$$

$$= (2\pi)^{-mt/2}(\sigma_\varepsilon^2)^{-mt/2}\,|1-\gamma|^m\exp\Big(-\frac{1}{2}\sum_{i=1}^m(\boldsymbol{y}_i - \boldsymbol{\iota}\mu + \boldsymbol{X}_i\boldsymbol{\beta})^\top\boldsymbol{V}^{-1}(\boldsymbol{y}_i - \boldsymbol{\iota}\mu + \boldsymbol{X}_i\boldsymbol{\beta})\Big). \tag{64.5.13}$$

□

Comparing this $\boldsymbol{P}$ with the $\boldsymbol{D}$ which we used to transform the dummy variable model, we see: instead of subtracting the mean from every column, we subtract γ times the mean from every column. This factor γ approaches 1 as t increases and as σ_α^2 increases. If one premultiplies (64.5.3) by $\boldsymbol{P}$ one gets

$$\boldsymbol{P}\boldsymbol{Y} = (1-\gamma)\boldsymbol{\iota}\boldsymbol{\iota}^\top\mu + \begin{bmatrix}\boldsymbol{P}\boldsymbol{X}_1\boldsymbol{\beta} & \cdots & \boldsymbol{P}\boldsymbol{X}_m\boldsymbol{\beta}\end{bmatrix} + (1-\gamma)\boldsymbol{\iota}\boldsymbol{\delta}^\top + \boldsymbol{P}\boldsymbol{E} \tag{64.5.14}$$

and after vectorization this reads

$$\begin{bmatrix}\boldsymbol{P}\mathbf{y}_1\\ \vdots\\ \boldsymbol{P}\mathbf{y}_t\end{bmatrix} = (1-\gamma)\begin{bmatrix}\boldsymbol{\iota}\\ \vdots\\ \boldsymbol{\iota}\end{bmatrix}\mu + \begin{bmatrix}\boldsymbol{P}\boldsymbol{X}_1\\ \vdots\\ \boldsymbol{P}\boldsymbol{X}_m\end{bmatrix}\boldsymbol{\beta} + \text{spherical disturbances.} \tag{64.5.15}$$

For estimation it is advantageous to write it as

$$\begin{bmatrix}\boldsymbol{P}\mathbf{y}_1\\ \vdots\\ \boldsymbol{P}\mathbf{y}_t\end{bmatrix} = \begin{bmatrix}\boldsymbol{\iota} & \boldsymbol{P}\boldsymbol{X}_1\\ \vdots & \vdots\\ \boldsymbol{\iota} & \boldsymbol{P}\boldsymbol{X}_m\end{bmatrix}\begin{bmatrix}(1-\gamma)\mu\\ \boldsymbol{\beta}\end{bmatrix} + \text{spherical disturbances,} \tag{64.5.16}$$

To sum up, if one knows γ, one can construct $\boldsymbol{P}$ and has to apply $\boldsymbol{P}$ to $\mathbf{Y}$ and all $\boldsymbol{X}_i$ and then run a regression with an intercept. The estimate of μ is this estimated intercept divided by $1-\gamma$.

How can we estimate the variances? There is a rich literature about estimation of the variances in variance component models. ITPE gives a very primitive but intuitive estimator. An estimate of σ_ε^2 can be obtained from the dummy variable model, since the projection operator in (64.3.9) removes $\boldsymbol{\alpha}$ together with its error term.

Information about σ_α^2 can be obtained from the variance from the "between"-regression which one gets by premultiplying (64.5.3) by $\frac{1}{t}\boldsymbol{\iota}^\top$. Defining $\bar{\mathbf{y}}^\top = \frac{1}{t}\boldsymbol{\iota}^\top\mathbf{Y}$, i.e., $\bar{\mathbf{y}}^\top$ is the row vector consisting of the column means, and in the same way $\bar{\boldsymbol{x}}_i^\top = \frac{1}{t}\boldsymbol{\iota}^\top\boldsymbol{X}_i$ and $\bar{\boldsymbol{\varepsilon}}^\top = \frac{1}{t}\boldsymbol{\iota}^\top\boldsymbol{E}$, one obtains

$$\bar{\mathbf{y}}^\top = \mu\boldsymbol{\iota}^\top + \begin{bmatrix}\bar{\boldsymbol{x}}_1^\top\boldsymbol{\beta} & \cdots & \bar{\boldsymbol{x}}_m^\top\boldsymbol{\beta}\end{bmatrix} + \boldsymbol{\delta}^\top + \bar{\boldsymbol{\varepsilon}}^\top = \mu\boldsymbol{\iota}^\top + (\bar{\boldsymbol{X}}\boldsymbol{\beta})^\top + \boldsymbol{\delta}^\top + \bar{\boldsymbol{\varepsilon}}^\top \quad \text{where} \quad \bar{\boldsymbol{X}} = \begin{bmatrix}\bar{\boldsymbol{x}}_1^\top\\ \vdots\\ \bar{\boldsymbol{x}}_m^\top\end{bmatrix}. \tag{64.5.17}$$

If one transposes this one obtains $\bar{\mathbf{y}} = \boldsymbol{\iota}\mu + \bar{\boldsymbol{X}}\boldsymbol{\beta} + \boldsymbol{\delta} + \bar{\boldsymbol{\varepsilon}}$.

Here the error term $\boldsymbol{\delta} + \bar{\boldsymbol{\varepsilon}}$ is the sum of two spherically distributed independent error terms, thefore it is still spherically distributed, and its variance, call it σ_w^2, is $\sigma_w^2 = \sigma_\alpha^2 + \sigma_\varepsilon^2/t$. In terms of σ_ε^2 and σ_w^2, the factor γ can be written as $\gamma = 1 - \sqrt{\sigma_\varepsilon^2/(t\sigma_w^2)}$. Unfortunately it is possible that the estimate $\mathrm{s}_w^2 - \mathrm{s}_\varepsilon^2/t < 0$, which implies a negative estimate for σ_α^2.

PROBLEM 531. *4 points Several times we have encountered the problem that estimated variances may be negative, or that an estimated covariance matrix may not be nonnegative definite. In which cases is this true, and how can it come about?*

ANSWER. One can get guarantee for nnd covariance matrix estimates only if the covariance matrix is estimated as a whole, for instance in the SUR model with equal numbers of observations in each equation, in which the covariance matrix is estimated as the sample covariance matrix of the residuals. If the covariance matrix is not estimated as one, but piece by piece, as for instance in the SUR model when the number of observations varies, or in the random coefficients model, then nnd is no longer guaranteed. Individual variances can obtain negative estimates when the formula for the variance contains several parameters which are estimated separately. In the variance components model, the variance is estimated as the difference between two other variances, which are estimated separately so that there is no guarantee that their difference is nonnegative. If the estimated ρ in an AR1-process comes out to be greater than 1, then the estimated covariance matrix is no longer nnd, and the formula $\operatorname{var}[\varepsilon] = \operatorname{var}[\nu]/(1-\rho^2)$ yields negative variances. □

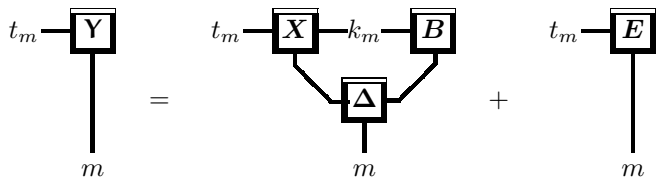
64.5.1. Testing. The variance component model relies on one assumption which is often not satisfied: the errors in $\boldsymbol{\alpha}$ and the errors in $\boldsymbol{E}$ must be uncorrelated. If this is not the case, then the variance components estimator suffers from omitted variables bias. Hausman used this as a basis for a test: if the errors are uncorrelated, then the GLS is the BLUE, and the Dummy variable estimator is consistent, but not efficient. If the errors are correlated, then the Dummy variables estimator is still consistent, but the GLS is no longer BLUE. I.e., one should expect that the difference between these estimators is much greater when the error terms are correlated. And under the null hypothesis that the error terms are orthogonal, there is an easy way to get the covariance matris of the estimators: since the GLS is BLUE and the other estimator is unbiased, the dispersion matrix of the difference of the estimators is the difference of their dispersion matrices. For more detail see [**Gre97**, 14.4.4].

CHAPTER 65

Disturbance Related (Seemingly Unrelated) Regressions

One has m timeseries regression equations $\mathfrak{y}_i = \boldsymbol{X}_i\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i$. Everything is different: the dependent variables, the explanatory variables, the coefficient vectors. Even the numbers of the observations may be different, The ith regression has k_i explanatory variables and t_i observations. They may be time series covering different but partly overlapping time periods. This is why they are called "seemingly unrelated" regressions. The only connection between the regressions is that for those observations which overlap in time the disturbances for different regressions are contemperaneously correlated, and these correlations are assumed to be constant over time. In tiles, this model is

(65.0.18)



65.1. The Supermatrix Representation

One can combine all these regressions into one big "supermatrix" as follows:

$$\begin{bmatrix} \mathfrak{y}_1 \\ \mathfrak{y}_2 \\ \vdots \\ \mathfrak{y}_m \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_1 & \boldsymbol{O} & \cdots & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{X}_2 & \cdots & \boldsymbol{O} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{O} & \boldsymbol{O} & \cdots & \boldsymbol{X}_m \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_m \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_m \end{bmatrix} \tag{65.1.1}$$

The covariance matrix of the disturbance term in (65.1.1) has the following "striped" form:

$$\mathcal{V}\left[\begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_m \end{bmatrix}\right] = \begin{bmatrix} \sigma_{11}\boldsymbol{I}_{11} & \sigma_{12}\boldsymbol{I}_{12} & \cdots & \sigma_{1m}\boldsymbol{I}_{1m} \\ \sigma_{21}\boldsymbol{I}_{21} & \sigma_{22}\boldsymbol{I}_{22} & \cdots & \sigma_{2m}\boldsymbol{I}_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{m1}\boldsymbol{I}_{m1} & \sigma_{m2}\boldsymbol{I}_{m2} & \cdots & \sigma_{mm}\boldsymbol{I}_{mm} \end{bmatrix} \tag{65.1.2}$$

Here $\boldsymbol{I}_{ij}$ is the $t_i \times t_j$ matrix which has zeros everywhere except at the intersections of rows and columns denoting the same time period.

In the special case that all time periods are identical, i.e., all $t_i = t$, one can define the matrices $\mathbf{Y} = \begin{bmatrix} \mathfrak{y}_1 & \cdots & \mathfrak{y}_m \end{bmatrix}$ and $\boldsymbol{E} = \begin{bmatrix} \boldsymbol{\varepsilon}_1 & \cdots & \boldsymbol{\varepsilon}_m \end{bmatrix}$, and write the equations in matrix form as follows:

$$\mathbf{Y} = \begin{bmatrix} \boldsymbol{X}_1\boldsymbol{\beta}_1 & \ldots & \boldsymbol{X}_m\boldsymbol{\beta}_m \end{bmatrix} + \boldsymbol{E} = \boldsymbol{H}(\boldsymbol{B}) + \boldsymbol{E} \tag{65.1.3}$$

The vector of dependent variables and the vector of disturbances in the supermatrix representation (65.1.1) can in this special case be written in terms of the vectorization operator as $\operatorname{vec}\mathbf{Y}$ and $\operatorname{vec}\boldsymbol{E}$. And the covariance matrix can be written as a Kronecker product: $\mathcal{V}[\operatorname{vec}\boldsymbol{E}] = \boldsymbol{\Sigma} \otimes \boldsymbol{I}$, since all $\boldsymbol{I}_{ij}$ in (65.1.2) are $t \times t$ identity matrices. If $t = 5$ and $m = 3$, the covariance matrix would be

$$\begin{bmatrix}
\sigma_{11} & 0 & 0 & 0 & 0 & \sigma_{12} & 0 & 0 & 0 & 0 & \sigma_{13} & 0 & 0 & 0 & 0 \\
0 & \sigma_{11} & 0 & 0 & 0 & 0 & \sigma_{12} & 0 & 0 & 0 & 0 & \sigma_{13} & 0 & 0 & 0 \\
0 & 0 & \sigma_{11} & 0 & 0 & 0 & 0 & \sigma_{12} & 0 & 0 & 0 & 0 & \sigma_{13} & 0 & 0 \\
0 & 0 & 0 & \sigma_{11} & 0 & 0 & 0 & 0 & \sigma_{12} & 0 & 0 & 0 & 0 & \sigma_{13} & 0 \\
0 & 0 & 0 & 0 & \sigma_{11} & 0 & 0 & 0 & 0 & \sigma_{12} & 0 & 0 & 0 & 0 & \sigma_{13} \\
\sigma_{21} & 0 & 0 & 0 & 0 & \sigma_{22} & 0 & 0 & 0 & 0 & \sigma_{23} & 0 & 0 & 0 & 0 \\
0 & \sigma_{21} & 0 & 0 & 0 & 0 & \sigma_{22} & 0 & 0 & 0 & 0 & \sigma_{23} & 0 & 0 & 0 \\
0 & 0 & \sigma_{21} & 0 & 0 & 0 & 0 & \sigma_{22} & 0 & 0 & 0 & 0 & \sigma_{23} & 0 & 0 \\
0 & 0 & 0 & \sigma_{21} & 0 & 0 & 0 & 0 & \sigma_{22} & 0 & 0 & 0 & 0 & \sigma_{23} & 0 \\
0 & 0 & 0 & 0 & \sigma_{21} & 0 & 0 & 0 & 0 & \sigma_{22} & 0 & 0 & 0 & 0 & \sigma_{23} \\
\sigma_{31} & 0 & 0 & 0 & 0 & \sigma_{32} & 0 & 0 & 0 & 0 & \sigma_{33} & 0 & 0 & 0 & 0 \\
0 & \sigma_{31} & 0 & 0 & 0 & 0 & \sigma_{32} & 0 & 0 & 0 & 0 & \sigma_{33} & 0 & 0 & 0 \\
0 & 0 & \sigma_{31} & 0 & 0 & 0 & 0 & \sigma_{32} & 0 & 0 & 0 & 0 & \sigma_{33} & 0 & 0 \\
0 & 0 & 0 & \sigma_{31} & 0 & 0 & 0 & 0 & \sigma_{32} & 0 & 0 & 0 & 0 & \sigma_{33} & 0 \\
0 & 0 & 0 & 0 & \sigma_{31} & 0 & 0 & 0 & 0 & \sigma_{32} & 0 & 0 & 0 & 0 & \sigma_{33}
\end{bmatrix}$$

If in addition all regressions have the same number of regressors, one can combine the coefficients into a matrix $\boldsymbol{B}$ and can write the system as

$$\operatorname{vec}\mathbf{Y} = \boldsymbol{Z}\operatorname{vec}\boldsymbol{B} + \operatorname{vec}\boldsymbol{E} \qquad \operatorname{vec}\boldsymbol{E} \sim (\boldsymbol{o}, \boldsymbol{\Sigma} \otimes \boldsymbol{I}), \tag{65.1.4}$$

where $\boldsymbol{Z}$ contains the regressors arranged in a block-diagonal "supermatrix."

If one knows $\boldsymbol{\Sigma}$ up to a multiplicative factor, and if all regressions cover the same time period, then one can apply (26.0.2) to (65.1.4) to get the following formula for the GLS estimator and at the same time maximum likelihood estimator:

$$\text{(65.1.5)} \qquad \operatorname{vec}(\hat{\mathbf{B}}) = \left(\boldsymbol{Z}^\top(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{I})\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^\top(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{I})\operatorname{vec}(\mathbf{Y}).$$

To evaluate this, note first that $\boldsymbol{Z}^\top(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{I}) =$

$$\begin{bmatrix} \boldsymbol{X}_1^\top & \boldsymbol{O} & \cdots & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{X}_2^\top & \cdots & \boldsymbol{O} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{O} & \boldsymbol{O} & \cdots & \boldsymbol{X}_m^\top \end{bmatrix} \begin{bmatrix} \sigma^{11}\boldsymbol{I} & \sigma^{12}\boldsymbol{I} & \cdots & \sigma^{1m}\boldsymbol{I} \\ \sigma^{21}\boldsymbol{I} & \sigma^{22}\boldsymbol{I} & \cdots & \sigma^{2m}\boldsymbol{I} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma^{m1}\boldsymbol{I} & \sigma^{m2}\boldsymbol{I} & \cdots & \sigma^{mm}\boldsymbol{I} \end{bmatrix} = \begin{bmatrix} \sigma^{11}\boldsymbol{X}_1^\top & \cdots & \sigma^{1m}\boldsymbol{X}_1^\top \\ \vdots & \ddots & \vdots \\ \sigma^{m1}\boldsymbol{X}_m^\top & \cdots & \sigma^{mm}\boldsymbol{X}_m^\top \end{bmatrix}$$

where σ^{ij} are the elements of the inverse of $\boldsymbol{\Sigma}$, therefore

$$\text{(65.1.6)} \qquad \begin{bmatrix} \hat{\boldsymbol{\beta}}_1 \\ \vdots \\ \hat{\boldsymbol{\beta}}_m \end{bmatrix} = \begin{bmatrix} \sigma^{11}\boldsymbol{X}_1^\top\boldsymbol{X}_1 & \cdots & \sigma^{1m}\boldsymbol{X}_1^\top\boldsymbol{X}_m \\ \vdots & \ddots & \vdots \\ \sigma^{m1}\boldsymbol{X}_m^\top\boldsymbol{X}_1 & \cdots & \sigma^{mm}\boldsymbol{X}_m^\top\boldsymbol{X}_m \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{X}_1^\top\sum_{i=1}^m \sigma^{1i}\mathbf{y}_i \\ \vdots \\ \boldsymbol{X}_m^\top\sum_{i=1}^m \sigma^{mi}\mathbf{y}_i \end{bmatrix}.$$

In the seemingly unrelated regression model, OLS on each equation singly is therefore less efficient than an approach which estimates all the equations simultaneously. If the numbers of observations in the different regressions are unequal, then the formula for the GLSE is no longer so simple. It is given in [**JHG+88**, (11.2.59) on p. 464].

65.2. The Likelihood Function

We know therefore what to do in the hypothetical case that $\boldsymbol{\Sigma}$ is known. What if it is not known? We will derive here the maximum likelihood estimator. For the exponent of the likelihood function we need the following mathematical tool:

Problem 532. *Show that* $\sum_{s=1}^t \boldsymbol{a}_s^\top\boldsymbol{\Omega}\boldsymbol{a}_s = \operatorname{tr}\boldsymbol{A}^\top\boldsymbol{\Omega}\boldsymbol{A}$ *where* $\boldsymbol{A} = \begin{bmatrix}\boldsymbol{a}_1 & \dots & \boldsymbol{a}_t\end{bmatrix}$.

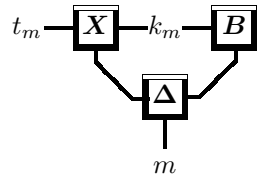
Answer.

$$\boldsymbol{A}^\top\boldsymbol{\Omega}\boldsymbol{A} = \begin{bmatrix}\boldsymbol{a}_1^\top \\ \vdots \\ \boldsymbol{a}_t^\top\end{bmatrix}\boldsymbol{\Omega}\begin{bmatrix}\boldsymbol{a}_1 & \dots & \boldsymbol{a}_t\end{bmatrix} = \begin{bmatrix} \boldsymbol{a}_1^\top\boldsymbol{\Omega}\boldsymbol{a}_1 & \boldsymbol{a}_1^\top\boldsymbol{\Omega}\boldsymbol{a}_2 & \cdots & \boldsymbol{a}_1^\top\boldsymbol{\Omega}\boldsymbol{a}_t \\ \boldsymbol{a}_2^\top\boldsymbol{\Omega}\boldsymbol{a}_1 & \boldsymbol{a}_2^\top\boldsymbol{\Omega}\boldsymbol{a}_2 & \cdots & \boldsymbol{a}_2^\top\boldsymbol{\Omega}\boldsymbol{a}_t \\ \boldsymbol{a}_t^\top\boldsymbol{\Omega}\boldsymbol{a}_1 & \boldsymbol{a}_t^\top\boldsymbol{\Omega}\boldsymbol{a}_2 & \cdots & \boldsymbol{a}_t^\top\boldsymbol{\Omega}\boldsymbol{a}_t \end{bmatrix}$$

Now take the trace of this. □

To derive the likelihood function, define the matrix function $\boldsymbol{H}(\boldsymbol{B})$ as follows: $\boldsymbol{H}(\boldsymbol{B})$ is a $t\times m$ matrix the ith column of which is $\boldsymbol{X}_i\boldsymbol{\beta}_i$, i.e., $\boldsymbol{H}(\boldsymbol{B})$ as a column-partitioned matrix is $\boldsymbol{H}(\boldsymbol{B}) = \begin{bmatrix}\boldsymbol{X}_1\boldsymbol{\beta}_1 & \cdots & \boldsymbol{X}_m\boldsymbol{\beta}_m\end{bmatrix}$. In tiles,

(65.2.1) $\boldsymbol{H}(\boldsymbol{B}) =$



The above notation follows [**DM93**, 315–318]. [**Gre97**, p. 683 top] writes this same $\boldsymbol{H}$ as the matrix product

$$\text{(65.2.2)} \qquad \boldsymbol{H}(\boldsymbol{B}) = \boldsymbol{Z}\boldsymbol{\Pi}(\boldsymbol{B})$$

where $\boldsymbol{Z}$ has all the different regressors in the different regressions as columns (it is $\boldsymbol{Z} = \begin{bmatrix}\boldsymbol{X}_1 & \cdots & \boldsymbol{X}_n\end{bmatrix}$ with duplicate columns deleted), and the ith column of $\boldsymbol{\Pi}$ has zeros for those regressors which are not in the ith equation, and elements of $\boldsymbol{B}$ for those regressors which are in the ith equation.

Using $\boldsymbol{H}$, the model is simply, as in (65.0.18),

$$\text{(65.2.3)} \qquad \mathbf{Y} = \boldsymbol{H}(\boldsymbol{B}) + \boldsymbol{E}, \quad \operatorname{vec}(\boldsymbol{E}) \sim \mathsf{N}(\boldsymbol{o}, \boldsymbol{\Sigma}\otimes\boldsymbol{I})$$

This is a matrix generalization of (56.0.21).

The likelihood function which we are going to derive now is valid not only for this particular $\boldsymbol{H}$ but for more general, possibly nonlinear $\boldsymbol{H}$. Define $\boldsymbol{\eta}_s(\boldsymbol{B})$ to be the sth *row* of $\boldsymbol{H}$, written as a column vector, i.e., as a row-partitioned matrix we have $\boldsymbol{H}(\boldsymbol{B}) = \begin{bmatrix}\boldsymbol{\eta}_1^\top(\boldsymbol{B}) \\ \vdots \\ \boldsymbol{\eta}_t^\top(\boldsymbol{B})\end{bmatrix}$. Then (65.2.3) in row-partitioned form reads

$$\text{(65.2.4)} \qquad \begin{bmatrix}\mathbf{y}_1^\top \\ \vdots \\ \mathbf{y}_t^\top\end{bmatrix} = \begin{bmatrix}\boldsymbol{\eta}_1^\top(\boldsymbol{B}) \\ \vdots \\ \boldsymbol{\eta}_t^\top(\boldsymbol{B})\end{bmatrix} + \begin{bmatrix}\boldsymbol{\varepsilon}_1^\top \\ \vdots \\ \boldsymbol{\varepsilon}_t^\top\end{bmatrix}$$

We assume Normality, the sth row vector is $\mathbf{y}_s^\top \sim \mathsf{N}(\boldsymbol{\eta}_s^\top(\boldsymbol{B}), \boldsymbol{\Sigma})$, or $\mathbf{y}_s \sim \mathsf{N}(\boldsymbol{\eta}_s(\boldsymbol{B}), \boldsymbol{\Sigma})$, and we assume that different rows are independent. Therefore the density function

is

$$
\begin{aligned}
f_{\boldsymbol{Y}}(\boldsymbol{Y}) &= \prod_{s=1}^{t}\Big((2\pi)^{-m/2}(\det\boldsymbol{\Sigma})^{-1/2}\exp\big(-\frac{1}{2}(\boldsymbol{y}_s-\boldsymbol{\eta}_s(\boldsymbol{B}))^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_s-\boldsymbol{\eta}_s(\boldsymbol{B}))\big)\Big)\\
&= (2\pi)^{-mt/2}(\det\boldsymbol{\Sigma})^{-t/2}\exp\big(-\frac{1}{2}\sum_s(\boldsymbol{y}_s-\boldsymbol{\eta}_s(\boldsymbol{B}))^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_s-\boldsymbol{\eta}_s(\boldsymbol{B}))\big)\\
&= (2\pi)^{-mt/2}(\det\boldsymbol{\Sigma})^{-t/2}\exp\big(-\frac{1}{2}\operatorname{tr}(\boldsymbol{Y}-\boldsymbol{H}(\boldsymbol{B}))\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}-\boldsymbol{H}(\boldsymbol{B}))^\top\big)\\
&= (2\pi)^{-mt/2}(\det\boldsymbol{\Sigma})^{-t/2}\exp\big(-\frac{1}{2}\operatorname{tr}(\boldsymbol{Y}-\boldsymbol{H}(\boldsymbol{B}))^\top(\boldsymbol{Y}-\boldsymbol{H}(\boldsymbol{B}))\boldsymbol{\Sigma}^{-1}\big). \qquad (65.2.5)
\end{aligned}
$$

PROBLEM 533. *Expain exactly the step in the derivation of* (65.2.5) *in which the trace enters.*

ANSWER. Write the quadratic form in the exponent as follows:

(65.2.6)

$$
\sum_{s=1}^{t}(\boldsymbol{y}_s-\boldsymbol{\eta}_s(\boldsymbol{B}))^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_s-\boldsymbol{\eta}_s(\boldsymbol{B})) = \sum_{s=1}^{t}\operatorname{tr}(\boldsymbol{y}_s-\boldsymbol{\eta}_s(\boldsymbol{B}))^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_s-\boldsymbol{\eta}_s(\boldsymbol{B}))
$$

$$
= \sum_{s=1}^{t}\operatorname{tr}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_s-\boldsymbol{\eta}_s(\boldsymbol{B}))(\boldsymbol{y}_s-\boldsymbol{\eta}_s(\boldsymbol{B}))^\top \qquad (65.2.7)
$$

$$
= \operatorname{tr}\boldsymbol{\Sigma}^{-1}\sum_{s=1}^{t}(\boldsymbol{y}_s-\boldsymbol{\eta}_s(\boldsymbol{B}))(\boldsymbol{y}_s-\boldsymbol{\eta}_s(\boldsymbol{B}))^\top \qquad (65.2.8)
$$

$$
= \operatorname{tr}\boldsymbol{\Sigma}^{-1}\begin{bmatrix}(\boldsymbol{y}_1-\boldsymbol{\eta}_1(\boldsymbol{B})) & \cdots & (\boldsymbol{y}_t-\boldsymbol{\eta}_t(\boldsymbol{B}))\end{bmatrix}\begin{bmatrix}(\boldsymbol{y}_1-\boldsymbol{\eta}_1(\boldsymbol{B}))^\top\\ \vdots\\ (\boldsymbol{y}_t-\boldsymbol{\eta}_t(\boldsymbol{B}))^\top\end{bmatrix} \qquad (65.2.9)
$$

$$
= \operatorname{tr}\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}-\boldsymbol{H}(\boldsymbol{B}))^\top(\boldsymbol{Y}-\boldsymbol{H}(\boldsymbol{B})) \qquad (65.2.10)
$$

□

The log likelihood function $\ell(\boldsymbol{Y};\boldsymbol{B},\boldsymbol{\Sigma})$ is therefore

$$
\ell = -\frac{mt}{2}\log 2\pi - \frac{t}{2}\log\det\boldsymbol{\Sigma} - \frac{1}{2}\operatorname{tr}(\boldsymbol{Y}-\boldsymbol{H}(\boldsymbol{B}))^\top(\boldsymbol{Y}-\boldsymbol{H}(\boldsymbol{B}))\boldsymbol{\Sigma}^{-1}. \qquad (65.2.11)
$$

In order to concentrate out $\boldsymbol{\Sigma}$ it is simpler to take the partial derivatives with respect to $\boldsymbol{\Sigma}^{-1}$ than those with respect to $\boldsymbol{\Sigma}$ itself. Using the matrix differentiation rules (C.1.24) and (C.1.16) and noting that $-t/2\log\det\boldsymbol{\Sigma} = t/2\log\det\boldsymbol{\Sigma}^{-1}$ one gets:

$$
\frac{\partial\ell}{\partial\boldsymbol{\Sigma}^{-1}} = \frac{t}{2}\boldsymbol{\Sigma} - \frac{1}{2}(\boldsymbol{Y}-\boldsymbol{H}(\boldsymbol{B}))^\top(\boldsymbol{Y}-\boldsymbol{H}(\boldsymbol{B})), \qquad (65.2.12)
$$

and if we set this zero we get

$$
\hat{\boldsymbol{\Sigma}}(\boldsymbol{B}) = \frac{1}{t}(\boldsymbol{Y}-\boldsymbol{H}(\boldsymbol{B}))^\top(\boldsymbol{Y}-\boldsymbol{H}(\boldsymbol{B})). \qquad (65.2.13)
$$

Written row vector by row vector this is

$$
\hat{\boldsymbol{\Sigma}} = \frac{1}{t}\sum_{s=1}^{t}(\mathfrak{y}_s-\boldsymbol{\eta}_s(\boldsymbol{B}))(\mathfrak{y}_s-\boldsymbol{\eta}_s(\boldsymbol{B}))^\top \qquad (65.2.14)
$$

The maximum likelihood estimator of $\boldsymbol{\Sigma}$ is therefore simply the sample covariance matrix of the residuals taken with the maximum likelihood estimates of $\boldsymbol{B}$.

We know therefore what the maximum likelihood estimator of $\boldsymbol{\Sigma}$ is if $\boldsymbol{B}$ is known: it is the sample covariance matrix of the residuals. And we know what the maximum likelihood estimator of $\boldsymbol{B}$ is if $\boldsymbol{\Sigma}$ is known: it is given by equation (65.1.6). In such a situation, one good numerical method is to iterate: start with an initial estimate of $\boldsymbol{\Sigma}$ (perhaps from the OLS residuals), get from this an estimate of $\boldsymbol{B}$, then use this to get a second estimate of $\boldsymbol{\Sigma}$, etc., until it converges. This iterative scheme is called *iterated Zellner* or *iterated SUR*. See [**Ruu00**, p. 706], the original article is [**Zel62**].

65.3. Concentrating out the Covariance Matrix (Incomplete)

One can rewrite (65.2.11) using (65.2.13) as a definition:

$$
\ell = -\frac{mt}{2}\log 2\pi - \frac{t}{2}\log\det\boldsymbol{\Sigma} - \frac{t}{2}\operatorname{tr}\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}} \qquad (65.3.1)
$$

and therefore the concentrated log likelihood function is, compare [**Gre97**, 15-53 on p. 685]:

$$
\begin{aligned}
\ell_c &= -\frac{mt}{2}\log 2\pi - \frac{t}{2}\log\det\hat{\boldsymbol{\Sigma}} - \frac{t}{2}\operatorname{tr}\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\Sigma}}\\
&= -\frac{mt}{2}(1+\log 2\pi) - \frac{t}{2}\log\det\hat{\boldsymbol{\Sigma}}(\boldsymbol{B}). \qquad (65.3.2)
\end{aligned}
$$

This is an important formula which is valid for all the different models, including nonlinear models, which can be written in the form (65.2.3).

As a next step we will write, following [**Gre97**, p. 683], $\boldsymbol{H}(\boldsymbol{B}) = \boldsymbol{Z}\boldsymbol{\Pi}(\boldsymbol{B})$ and derive the following formula from [**Gre97**, p. 685]:

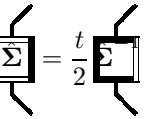
$$
\frac{\partial\ell_c}{\partial\boldsymbol{\Pi}^\top} = \hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{Y}-\boldsymbol{Z}\boldsymbol{\Pi})^\top\boldsymbol{Z} \qquad (65.3.3)
$$

Here is a derivation of this using tile notation. We use the notation $\hat{\boldsymbol{E}} = \boldsymbol{Y}-\boldsymbol{H}(\boldsymbol{B})$ for the matrix of residuals, and apply the chain rule to get the derivatives:

$$
\frac{\partial\ell_c}{\partial\boldsymbol{\Pi}^\top} = \frac{\partial\ell_c}{\partial\hat{\boldsymbol{\Sigma}}^\top}\frac{\partial\hat{\boldsymbol{\Sigma}}}{\partial\hat{\boldsymbol{E}}^\top}\frac{\partial\hat{\boldsymbol{E}}}{\partial\boldsymbol{\Pi}^\top} \qquad (65.3.4)
$$

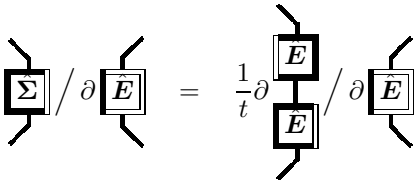
The product here is not a matrix product but the concatenation of a matrix with three arrays of rank 4. In tile notation, the first term in this product is

(65.3.5) $$\frac{\partial \ell_c}{\partial \hat{\boldsymbol{\Sigma}}^\top} = \partial \boxed{\ell_c} \Big/ \partial \boxed{\boldsymbol{\Sigma}} = \frac{t}{2}$$

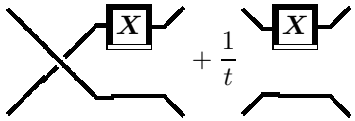


This is an array of rank 2, i.e., a matrix, but the other factors are arrays of rank 4: Using (C.1.22) we get

$$\frac{\partial \hat{\boldsymbol{\Sigma}}}{\partial \hat{\boldsymbol{E}}^\top} = \partial \boxed{\hat{\boldsymbol{\Sigma}}} \Big/ \partial \boxed{\hat{\boldsymbol{E}}} = \frac{1}{t} \partial \boxed{\boldsymbol{E}}\boxed{\boldsymbol{E}} \Big/ \partial \boxed{\hat{\boldsymbol{E}}} =$$



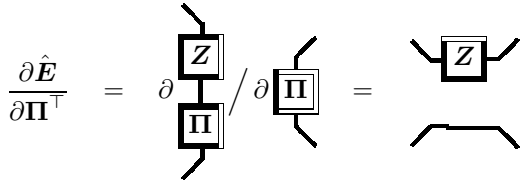
$$= \frac{1}{t}$$



$$+ \frac{1}{t}$$

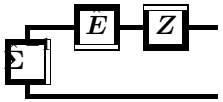
Finally, by (C.1.18),

$$\frac{\partial \hat{\boldsymbol{E}}}{\partial \boldsymbol{\Pi}^\top} = \partial \boxed{\boldsymbol{Z}}\boxed{\boldsymbol{\Pi}} \Big/ \partial \boxed{\boldsymbol{\Pi}} =$$



Putting it all together, using the symmetry of the first term (65.3.5) (which has the effect that the term with the crossing arms is the same as the straight one), gives

$$\frac{\partial \ell_c}{\partial \boldsymbol{\Pi}^\top} = \partial \boxed{\ell_c} \Big/ \partial \boxed{\boldsymbol{\Pi}} =$$



which is exactly (65.3.3).

65.4. Situations in which OLS is Best

One of the most amazing results regarding seemingly unrelated regressions is: if the $\boldsymbol{X}$ matrices are identical, then it is not necessary to do GLS, because OLS on each equation separately gives exactly the same result. Question 534 gives three different proofs of this:

Problem 534. *Given a set of disturbance related regression equations*

(65.4.1) $$\mathbf{y}_i = \boldsymbol{X}\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i \qquad i = 1, \dots m$$

in which all $\boldsymbol{X}_i$ are equal to $\boldsymbol{X}$, note that equation (65.4.1) *has no subscript at the matrices of explanatory variables.*

• a. *1 point Defining* $\mathbf{Y} = \begin{bmatrix}\mathbf{y}_1 & \cdots & \mathbf{y}_m\end{bmatrix}$, $\mathbf{B} = \begin{bmatrix}\boldsymbol{\beta}_1 & \cdots & \boldsymbol{\beta}_m\end{bmatrix}$ *and* $\boldsymbol{E} = \begin{bmatrix}\boldsymbol{\varepsilon}_1 & \cdots & \boldsymbol{\varepsilon}_m\end{bmatrix}$, *show that the m equations* (65.4.1) *can be combined into the single matrix equation*

(65.4.2) $$\mathbf{Y} = \boldsymbol{X}\mathbf{B} + \boldsymbol{E}.$$

Answer. The only step needed to show this is that $\boldsymbol{X}\mathbf{B}$, column by column, can be written $\boldsymbol{X}\mathbf{B} = \begin{bmatrix}\boldsymbol{X}\boldsymbol{\beta}_1 & \dots & \boldsymbol{X}\boldsymbol{\beta}_m\end{bmatrix}$. □

• b. *1 point The contemporaneous correlations of the disturbances can now be written* $\operatorname{vec}(\boldsymbol{E}) \sim (\boldsymbol{o}, \boldsymbol{\Sigma} \otimes \boldsymbol{I})$.

• c. *4 points For this part of the Question you will need the following properties of* vec *and* $\otimes$*:* $(\boldsymbol{A}\otimes\boldsymbol{B})^\top = \boldsymbol{A}^\top \otimes \boldsymbol{B}^\top$, $(\boldsymbol{A}\otimes\boldsymbol{B})(\boldsymbol{C}\otimes\boldsymbol{D}) = (\boldsymbol{A}\boldsymbol{C})\otimes(\boldsymbol{B}\boldsymbol{D})$, $(\boldsymbol{A}\otimes\boldsymbol{B})^{-1} = \boldsymbol{A}^{-1} \otimes \boldsymbol{B}^{-1}$, $\operatorname{vec}(\boldsymbol{A}+\boldsymbol{B}) = \operatorname{vec}(\boldsymbol{A}) + \operatorname{vec}(\boldsymbol{B})$, *and finally the important identity*

((B.5.19)) $$\operatorname{vec}(\boldsymbol{A}\boldsymbol{B}\boldsymbol{C}) = (\boldsymbol{C}^\top \otimes \boldsymbol{A}) \operatorname{vec}(\boldsymbol{B}).$$

By applying the vec *operator to* (65.4.2) *show that the BLUE of the matrix* $\mathbf{B}$ *is* $\hat{\mathbf{B}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\mathbf{Y}$, *i.e., show that, despite the fact that the dispersion matrix is not spherical, one simply has to apply OLS to every equation separately.*

Answer. Use (B.5.19) to write (65.4.2) in vectorized form as

$$\operatorname{vec}(\mathbf{Y}) = (\boldsymbol{I} \otimes \boldsymbol{X}) \operatorname{vec}(\mathbf{B}) + \operatorname{vec}(\boldsymbol{E})$$

Since $\mathcal{V}[\operatorname{vec}(\boldsymbol{E})] = \boldsymbol{\Sigma} \otimes \boldsymbol{I}$, the GLS estimate is

$$\begin{aligned}
\operatorname{vec}(\hat{\mathbf{B}}) &= \Big((\boldsymbol{I} \otimes \boldsymbol{X})^\top (\boldsymbol{\Sigma} \otimes \boldsymbol{I})^{-1} (\boldsymbol{I} \otimes \boldsymbol{X})\Big)^{-1} (\boldsymbol{I} \otimes \boldsymbol{X})^\top (\boldsymbol{\Sigma} \otimes \boldsymbol{I})^{-1} \operatorname{vec}(\mathbf{Y}) \\
&= \Big((\boldsymbol{I} \otimes \boldsymbol{X}^\top)(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{I})(\boldsymbol{I} \otimes \boldsymbol{X})\Big)^{-1} (\boldsymbol{I} \otimes \boldsymbol{X}^\top)(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{I}) \operatorname{vec}(\mathbf{Y}) \\
&= \Big(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{X}^\top\boldsymbol{X}\Big)^{-1} (\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{X}^\top) \operatorname{vec}(\mathbf{Y}) \\
&= \Big(\boldsymbol{I} \otimes (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\Big) \operatorname{vec}(\mathbf{Y})
\end{aligned}$$

and applying (B.5.19) again, this is equivalent to

(65.4.3) $$\hat{\mathbf{B}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\mathbf{Y}.$$

□

• d. *3 points* [**DM93**, p. 313] *appeals to Kruskal's theorem, which is Question 499, to prove this. Supply the details of this proof.*

Answer. Look at the derivation of (65.4.3) again. The $\boldsymbol{\Sigma}^{-1}$ in numerator and denominator cancel out since they commute with $\boldsymbol{Z}$. defining $\boldsymbol{\Omega} = \boldsymbol{\Sigma} \otimes \boldsymbol{I}$, this "commuting" is the formula $\boldsymbol{\Omega Z} = \boldsymbol{ZK}$ for some $\boldsymbol{K}$, i.e.,

$$\begin{bmatrix} \sigma_{11}\boldsymbol{I} & \dots & \sigma_{1m}\boldsymbol{I} \\ \vdots & \ddots & \vdots \\ \sigma_{m1}\boldsymbol{I} & \dots & \sigma_{mm}\boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{X} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \boldsymbol{X} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \boldsymbol{X} \end{bmatrix} \begin{bmatrix} \sigma_{11}\boldsymbol{I} & \dots & \sigma_{1m}\boldsymbol{I} \\ \vdots & \ddots & \vdots \\ \sigma_{m1}\boldsymbol{I} & \dots & \sigma_{mm}\boldsymbol{I} \end{bmatrix}. \tag{65.4.4}$$

Note that the $\boldsymbol{I}$ on the lefthand side are $m \times m$, and those on the right are $k \times k$. This "commuting" allows us to apply Kruskal's theorem. □

• e. *4 points Theil* [**The71**, pp. 500–502] *gives a different proof: he maximizes the likelihood function of* $\mathbf{Y}$ *with respect to* $\mathbf{B}$ *for the given* $\boldsymbol{\Sigma}$*, using the fact that the matrix of OLS estimates* $\hat{\boldsymbol{B}}$ *has the property that* $(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{B}})^\top(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{B}})$ *is by a nnd matrix smaller than any other* $(\boldsymbol{Y} - \boldsymbol{XB})^\top(\boldsymbol{Y} - \boldsymbol{XB})$*. Carry out this proof in detail.*

Answer. Let $\boldsymbol{B} = \hat{\boldsymbol{B}} + \boldsymbol{A}$; then $(\boldsymbol{Y} - \boldsymbol{XB})^\top(\boldsymbol{Y} - \boldsymbol{XB}) = (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{B}})^\top(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{B}}) + \boldsymbol{A}^\top\boldsymbol{X}^\top\boldsymbol{XA}$ because the cross product terms $\boldsymbol{A}^\top\boldsymbol{X}^\top(Y - \boldsymbol{X}\hat{\mathbf{B}}) = \boldsymbol{O}$ since $\hat{\mathbf{B}}$ satisfies the normal equation $\boldsymbol{X}^\top(Y - \boldsymbol{X}\hat{\mathbf{B}}) = \boldsymbol{O}$.

Instead of maximizing the likelihood function with respect to $\mathbf{B}$ and $\boldsymbol{\Sigma}$ simultaneously, Theil in [**The71**, p. 500–502] only maximizes it with respect to $\mathbf{B}$ for the given $\boldsymbol{\Sigma}$ and finds a solution which is independent of $\boldsymbol{\Sigma}$. The likelihood function of $\mathbf{Y}$ is (65.2.5) with $\boldsymbol{H}(\boldsymbol{B}) = \boldsymbol{XB}$, i.e.,

$$f_{\mathbf{Y}}(\boldsymbol{Y}) = (2\pi)^{-tm/2}(\det\boldsymbol{\Sigma})^{-t/2}\exp\Big(-\frac{1}{2}\operatorname{tr}\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{XB}))^\top(\boldsymbol{Y} - \boldsymbol{XB})\Big) \tag{65.4.5}$$

The trace in the exponent can be split up into $\operatorname{tr}(\boldsymbol{\Sigma}^{-1}(Y - \boldsymbol{X}\hat{\mathbf{B}})^\top(Y - \boldsymbol{X}\hat{\mathbf{B}}) + \operatorname{tr}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}^\top\boldsymbol{X}^\top\boldsymbol{XA}$; but this last term is equal to $\operatorname{tr}\boldsymbol{XA\Sigma}^{-1}\boldsymbol{A}^\top\boldsymbol{X}^\top$, which is ≥ 0. □

Joint estimation has therefore the greatest efficiency gains over OLS if the correlations between the errors are high and the correlations between the explanatory variables are *low*.

Problem 535. *Are following statements true or false?*

• a. *1 point In a seemingly unrelated regression framework, joint estimation of the whole model is much better than estimation of each equation singly if the errors are highly correlated. True or false?*

Answer. True □

• b. *1 point In a seemingly unrelated regression framework, joint estimation of the whole model is much better than estimation of each equation singly if the independent variables in the different regressions are highly correlated. True or false?*

Answer. False. □

Assume I have two equations whose disturbances are correlated, and the second has *all* variables that the first has, *plus* some additional ones. Then the inclusion of the second equation does not give additional information for the first; however, including the first gives additional information for the second!

What is the rationale for this? Since the first equation has fewer variables than the second, I know the disturbances better. For instance, if the equation would not have any variables, then I would know the disturbances exactly. But if I know these disturbances, and know that they are correlated with the disturbances of the second equation, then I can also say something about the disturbances of the second equation, and therefore estimate the parameters of the second equation better.

Problem 536. *You have two disturbance-related equations*

$$\mathbf{y}_1 = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}_1, \quad \mathbf{y}_2 = \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}_2, \qquad \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \end{bmatrix} \sim \begin{bmatrix} \boldsymbol{o} \\ \boldsymbol{o} \end{bmatrix}, \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} \otimes \boldsymbol{I} \tag{65.4.6}$$

where all σ_{ij} *are known, and the set of explanatory variables in* $\boldsymbol{X}_1$ *is a subset of those in* $\boldsymbol{X}_2$*. One of the following two statements is correct, the other is false. Which is correct? (a) in order to estimate* $\boldsymbol{\beta}_1$*, OLS on the first equation singly is as good as SUR. (b) in order to estimate* $\boldsymbol{\beta}_2$*, OLS on the second equation singly is as good as SUR. Which of these two is true?*

Answer. The first is true. One cannot obtain a more efficient estimator of $\boldsymbol{\beta}_1$ by considering the whole system. This is [**JGH+85**, p. 469]. □

65.5. Unknown Covariance Matrix

What to do when we don't know $\boldsymbol{\Sigma}$? Two main possibilities: One is "feasible GLS", which uses the OLS residuals to estimate $\boldsymbol{\Sigma}$, and then uses the GLS formula with the estimated elements of $\boldsymbol{\Sigma}$. This is the most obvious method; unfortunately if the numbers of observations are unequal, then this may no longer give a nonnegative definite matrix. The other is the maximum likelihood estimation of $\mathbf{B}$ and $\boldsymbol{\Sigma}$ simultaneously. If one iterates the "feasible GLS" method, i.e., uses the residuals of the feasible GLS equation to get new estimates of $\boldsymbol{\Sigma}$, then does feasible GLS with the new $\boldsymbol{\Sigma}$, etc., then one will get the maximum likelihood estimator.

Problem 537. *4 points Explain how to do iterated EGLS (i.e., GLS with an estimated covariance matrix) in a model with first-order autoregression, and in a seemingly unrelated regression model. Will you end up with the (normal) maximum likelihood estimator if you iterate until convergence?*

Answer. You will only get the Maximum Likelihood estimator in the SUR case, not in the AR1 case, because the determinant term will never come in by iteration, and in the AR1 case, EGLS is known to underestimate the ρ. Of course, iterated EGLS is in both situations asymtotically as good as Maximum Likelihood, but the question was whether it is in small samples already equal to the ML. You can have asymptotically equivalent estimates which differ greatly in small samples. □

Asymptotically, feasible GLS is as good as Maximum likelihood. This is really nothing new and nothing exciting. The two estimators may have quite different properties before the asymptotic limit is reached! But there is another, much stronger result: already for finite sample size, *iterated* feasible GLS is equal to the maximum likelihood estimator.

PROBLEM 538. *5 points Define "seemingly unrelated equations" and discuss the estimation issues involved.*

CHAPTER 66

Simultaneous Equations Systems

This was a central part of econometrics in the fifties and sixties.

66.1. Examples

[**JHG+88**, 14.1 Introduction] gives examples. The first example is clearly not identified, indeed it has no exogenous variables. But the idea of a simultaneous equations system is not dependent on this:

$$\mathbf{y}_d = \boldsymbol{\iota}\alpha + \mathbf{p}\beta + \boldsymbol{\varepsilon}_1 \tag{66.1.1}$$

$$\mathbf{y}_s = \boldsymbol{\iota}\gamma + \mathbf{p}\delta + \boldsymbol{\varepsilon}_2 \tag{66.1.2}$$

$$\mathbf{y}_d = \mathbf{y}_s \tag{66.1.3}$$

$\mathbf{y}_d$, $\mathbf{y}_s$, and $\mathbf{p}$ are the jointly determined endogenous variables. The first equation describes the behavior of the consumers, the second the behavior of producers.

Problem 539. [**Gre97**, p. 709 ff]. *Here is a demand and supply curve with* $\mathbf{q}$ *quantity,* $\mathbf{p}$ *price,* $\mathbf{y}$ *income, and* $\boldsymbol{\iota}$ *is the vector of ones. All vectors are t-vectors.*

$$\mathbf{q} = \alpha_0\boldsymbol{\iota} + \alpha_1\mathbf{p} + \alpha_2\mathbf{y} + \boldsymbol{\varepsilon}_d \qquad \boldsymbol{\varepsilon}_d \sim (\boldsymbol{o}, \sigma_d^2\boldsymbol{I}) \qquad \textit{(demand)} \tag{66.1.4}$$

$$\mathbf{q} = \beta_0\boldsymbol{\iota} + \beta_1\mathbf{p} + \boldsymbol{\varepsilon}_s \qquad \boldsymbol{\varepsilon}_s \sim (\boldsymbol{o}, \sigma_s^2\boldsymbol{I}) \qquad \textit{(supply)} \tag{66.1.5}$$

$\boldsymbol{\varepsilon}_d$ *and* $\boldsymbol{\varepsilon}_s$ *are independent of* $\mathbf{y}$*, but amongst each other they are contemporaneously correlated, with their covariance constant over time:*

$$\operatorname{cov}[\varepsilon_{dt}, \varepsilon_{su}] = \begin{cases} 0 & \textit{if } t \neq u \\ \sigma_{ds} & \textit{if } t = u \end{cases} \tag{66.1.6}$$

• a. *1 point Which variables are exogenous and which are endogenous?*

Answer. $\mathbf{p}$ and $\mathbf{q}$ are called *jointly dependent* or *endogenous*. $\mathbf{y}$ is determined outside the system or *exogenous*. □

• b. *2 points Assuming* $\alpha_1 \neq \beta_1$*, verify that the reduced-form equations for* $\mathbf{p}$ *and* $\mathbf{q}$ *are as follows:*

$$\mathbf{p} = \frac{\alpha_0 - \beta_0}{\beta_1 - \alpha_1}\boldsymbol{\iota} + \frac{\alpha_2}{\beta_1 - \alpha_1}\mathbf{y} + \frac{\boldsymbol{\varepsilon}_d - \boldsymbol{\varepsilon}_s}{\beta_1 - \alpha_1} \tag{66.1.7}$$

$$\mathbf{q} = \frac{\beta_1\alpha_0 - \beta_0\alpha_1}{\beta_1 - \alpha_1}\boldsymbol{\iota} + \frac{\beta_1\alpha_2}{\beta_1 - \alpha_1}\mathbf{y} + \frac{\beta_1\boldsymbol{\varepsilon}_d - \alpha_1\boldsymbol{\varepsilon}_s}{\beta_1 - \alpha_1} \tag{66.1.8}$$

Answer. One gets the reduced form equation for $\mathbf{p}$ by simply setting the righthand sides equal:

$$\begin{aligned} \beta_0\boldsymbol{\iota} + \beta_1\mathbf{p} + \boldsymbol{\varepsilon}_s &= \alpha_0\boldsymbol{\iota} + \alpha_1\mathbf{p} + \alpha_2\mathbf{y} + \boldsymbol{\varepsilon}_d \\ (\beta_1 - \alpha_1)\mathbf{p} &= (\alpha_0 - \beta_0)\boldsymbol{\iota} + \alpha_2\mathbf{y} + \boldsymbol{\varepsilon}_d - \boldsymbol{\varepsilon}_s, \end{aligned}$$

hence (66.1.7). To get the reduced form equation for $\mathbf{q}$, plug that for $\mathbf{p}$ into the supply function (one might also plug it into the demand function but the math would be more complicated):

$$\mathbf{q} = \beta_0\boldsymbol{\iota} + \beta_1\mathbf{p} + \boldsymbol{\varepsilon}_s = \beta_0\boldsymbol{\iota} + \frac{\beta_1(\alpha_0 - \beta_0)}{\beta_1 - \alpha_1}\boldsymbol{\iota} + \frac{\beta_1\alpha_2}{\beta_1 - \alpha_1}\mathbf{y} + \frac{\beta_1(\boldsymbol{\varepsilon}_d - \boldsymbol{\varepsilon}_s)}{\beta_1 - \alpha_1} + \boldsymbol{\varepsilon}_s$$

Combining the first two and the last two terms gives (66.1.8). □

• c. *2 points Show that one will in general not get consistent estimates of the supply equation parameters if one regresses* $\mathbf{q}$ *on* $\mathbf{p}$ *(with an intercept).*

Answer. By (66.1.7) (the reduced form equation for $\mathbf{p}$), $\operatorname{cov}[\varepsilon_{st}, \mathrm{p}_t] = \operatorname{cov}[\varepsilon_{st}, \frac{\varepsilon_{dt} - \varepsilon_{st}}{\beta_1 - \alpha_1}] = \frac{\sigma_{sd} - \sigma_s^2}{\beta_1 - \alpha_1}$. This is generally $\neq 0$, therefore inconsistency. □

• d. *2 points If one estimates the supply function by instrumental variables, using* $\mathbf{y}$ *as an instrument for* $\mathbf{p}$ *and* $\boldsymbol{\iota}$ *as instrument for itself, write down the formula for the resulting estimator* $\tilde{\beta}_1$ *of* β_1 *and show that it is consistent. You are allowed to use, without proof, equation* (52.0.12).

Answer. $\tilde{\beta}_1 = \frac{\frac{1}{n}\sum(\mathrm{y}_i - \bar{\mathrm{y}})(\mathrm{q}_i - \bar{\mathrm{q}})}{\frac{1}{n}\sum(\mathrm{y}_i - \bar{\mathrm{y}})(\mathrm{p}_i - \bar{\mathrm{p}})}$. Its plim is $\frac{\operatorname{cov}[\mathbf{y},\mathbf{q}]}{\operatorname{cov}[\mathbf{y},\mathbf{p}]} = \frac{\beta_1\alpha_2 \operatorname{var}[\mathbf{y}]/(\beta_1 - \alpha_1)}{\alpha_2 \operatorname{var}[\mathbf{y}]/(\beta_1 - \alpha_1)} = \beta_1$. These covariances were derived from (66.1.7) and (66.1.8). □

• e. *2 points Show that the Indirect Least Squares estimator of* β_1 *is identical to the instrumental variables estimator.*

Answer. For indirect least squares one estimates the two reduced form equations by OLS:

the slope parameter in (66.1.7), $\dfrac{\alpha_2}{\beta_1 - \alpha_1}$, estimated by $\dfrac{\sum(\mathrm{y}_i - \bar{\mathrm{y}})(\mathrm{p}_i - \bar{\mathrm{p}})}{\sum(\mathrm{y}_i - \bar{\mathrm{y}})^2}$;

the slope parameter in (66.1.8), $\dfrac{\beta_1\alpha_2}{\beta_1 - \alpha_1}$, estimated by $\dfrac{\sum(\mathrm{y}_i - \bar{\mathrm{y}})(\mathrm{q}_i - \bar{\mathrm{q}})}{\sum(\mathrm{y}_i - \bar{\mathrm{y}})^2}$

Divide to get

$$\beta_1 \text{ estimated by } \frac{\sum(\mathsf{y}_i-\bar{\mathsf{y}})(\mathsf{q}_i-\bar{\mathsf{q}})}{\sum(\mathsf{y}_i-\bar{\mathsf{y}})(\mathsf{p}_i-\bar{\mathsf{p}})}$$

which is the same $\tilde{\beta}_1$ as in part d. □

• f. *1 point Since the error terms in the reduced form equations are contemporaneously correlated, wouldn't one get more precise estimates if one estimates the reduced form equations as a seemingly unrelated system, instead of OLS?*

Answer. Not as long as one does not impose any constraints on the reduced form equations, since all regressors are the same. □

• g. *2 points We have shown above that the regression of* $\mathbf{q}$ *on* $\mathbf{p}$ *does not give a consistent estimator of* β_1*. However one does get a consistent estimator of* β_1 *if one regresses* $\mathbf{q}$ *on the predicted values of* $\mathbf{p}$ *from the reduced form equation. (This is 2SLS.) Show that this estimator is also the same as above.*

Answer. This gives $\tilde{\beta}_1 = \frac{\sum(\mathsf{q}_i-\bar{\mathsf{q}})(\hat{\mathsf{p}}_i-\bar{\mathsf{p}})}{\sum(\hat{\mathsf{p}}_i-\bar{\mathsf{p}})^2}$. Now use $\hat{\mathsf{p}}_i-\bar{\mathsf{p}} = \hat{\pi}_1(\mathsf{y}_i-\bar{\mathsf{y}})$ where $\hat{\pi}_1 = \sum\frac{(\mathsf{p}_i-\bar{\mathsf{p}})(\mathsf{y}_i-\bar{\mathsf{y}})}{\sum(\mathsf{y}_i-\bar{\mathsf{y}})^2}$. Therefore $\tilde{\beta} = \hat{\pi}_1\frac{\sum(\mathsf{q}_i-\bar{\mathsf{q}})(\mathsf{y}_i-\bar{\mathsf{y}})}{\hat{\pi}_1^2\sum(\mathsf{y}_i-\bar{\mathsf{y}})^2} = \frac{\sum(\mathsf{y}_i-\bar{\mathsf{y}})(\mathsf{q}_i-\bar{\mathsf{q}})}{\sum(\mathsf{y}_i-\bar{\mathsf{y}})(\mathsf{p}_i-\bar{\mathsf{p}})}$ again. □

• h. *1 point So far we have only discussed estimators of the parameters in the supply function. How would you estimate the demand function?*

Answer. You can't. The supply function can be estimated because it stays put while the demand function shifts around, therefore the observed intersection points lie on the same supply function but different demand functions. The demand function itself cannot be estimated, it is underidentified in this system. □

Here is an example from [**WW79**, 257–266]. Take a simple Keynesian income-expenditure model of a consumption function with investment i exogenous:

$$\mathsf{c} = \alpha + \beta\mathsf{y} + \varepsilon \tag{66.1.9}$$

$$\mathsf{y} = \mathsf{c} + \mathsf{i} \tag{66.1.10}$$

Exogenous means: determined outside the system. By definition this always means: it is independent of all the disturbance terms in the equations (here there is just one disturbance term). Then the first claim is: y is correlated with ε, because y and c are determined simultaneously once i and ε is given, and both depend on i and ε. Let us do that in more detail and write the *reduced form equation* for y. That means, let us express y in terms of the exogenous variable and the disturbances only. Plug $\mathsf{c} = \mathsf{y} - \mathsf{i}$ into (66.1.9) to get

$$\mathsf{y} - \mathsf{i} = \alpha + \beta\mathsf{y} + \varepsilon \tag{66.1.11}$$

$$\text{or} \qquad \mathsf{y}(1-\beta) = \alpha + \mathsf{i} + \varepsilon \tag{66.1.12}$$

This gives the reduced form equations

$$\mathsf{y} = \frac{\alpha}{1-\beta} + \frac{1}{1-\beta}\mathsf{i} + \frac{1}{1-\beta}\varepsilon \tag{66.1.13}$$

$$\text{and} \qquad \mathsf{c} = \mathsf{y} - \mathsf{i} = \frac{\alpha}{1-\beta} + \frac{\beta}{1-\beta}\mathsf{i} + \frac{1}{1-\beta}\varepsilon \tag{66.1.14}$$

From this one can see

$$\operatorname{cov}(\mathsf{y},\varepsilon) = 0 + 0 + \frac{1}{1-\beta}\operatorname{cov}(\varepsilon,\varepsilon) = \frac{\sigma^2}{1-\beta} \tag{66.1.15}$$

Therefore OLS applied to equation (66.1.9) gives inconsistent results.

Problem 540. *4 points Show that OLS applied to equation* (66.1.9) *gives an estimate which is in the* plim *larger than the true* β*.*

Answer.

$$\operatorname{plim}\hat{\beta} = \frac{\operatorname{cov}[\mathsf{y},\mathsf{c}]}{\operatorname{var}[\mathsf{y}]} = \frac{\frac{\beta}{(1-\beta)^2}\operatorname{var}[\mathsf{i}] + \frac{1}{(1-\beta)^2}\operatorname{var}[\varepsilon]}{\frac{1}{(1-\beta)^2}\operatorname{var}[\mathsf{i}] + \frac{1}{(1-\beta)^2}\operatorname{var}[\varepsilon]} = \tag{66.1.16}$$

$$= \frac{\beta\operatorname{var}[\mathsf{i}] + \operatorname{var}[\varepsilon]}{\operatorname{var}[\mathsf{i}] + \operatorname{var}[\varepsilon]} = \beta + \frac{(1-\beta)\operatorname{var}[\varepsilon]}{\operatorname{var}[\mathsf{i}] + \operatorname{var}[\varepsilon]} > \beta$$

□

One way out is to estimate with instrumental variables. The model itself provides an instrument for $\mathbf{y}$, namely, the exogenous variable $\mathbf{i}$.

As an alternative one might also estimate the reduced form equation (66.1.13) and then get the structural parameters from that. I.e., let $\hat{a}$ and $\hat{b}$ be the regression coefficients of (66.1.13). Then one can set, for the slope parameter β,

$$\hat{b} = \frac{1}{1-\hat{\beta}} \qquad \text{or} \qquad \hat{\beta} = \frac{\hat{b}-1}{\hat{b}}. \tag{66.1.17}$$

This estimation method is called ILS, indirect least squares, because the estimates were obtained indirectly, by estimating the reduced form equations.

Which of these two estimation methods is better? It turns out that they are exactly the same. Proof: from $\hat{b} = \widehat{\operatorname{cov}}(\mathbf{y},\mathbf{i})/\widehat{\operatorname{var}}(\mathbf{i})$ follows

$$\hat{\beta} = \frac{\hat{b}-1}{\hat{b}} = \frac{\widehat{\operatorname{cov}}(\mathbf{y},\mathbf{i}) - \widehat{\operatorname{var}}(\mathbf{i})}{\widehat{\operatorname{cov}}(\mathbf{y},\mathbf{i})} = \frac{\widehat{\operatorname{cov}}(\mathbf{c},\mathbf{i})}{\widehat{\operatorname{cov}}(\mathbf{y},\mathbf{i})}. \tag{66.1.18}$$

66.2. General Mathematical Form

Definition of Important Terms:

- endogenous
- exogenous
- Lagged endogenous
- Exogenous and lagged endogenous together are called predetermined.
- Nonobservable random errors, uncorrelated with exogenous variables and contemperaneously uncorrelated with predetermined variables.

What to consider when building an economic model:

- classification of economic variables (whether endogenous or exogenous)
- which variables enter which equation
- possible lags involved
- nonsample information about a single parameter or combination of parameters
- how many equations (structural equations) there should be and how the system should be "closed"
- algebraic form of the equations, also the question in which scales (logarithmic scale, prices or inverse prices, etc.) the variables are to be measured.
- distribution of the random errors

A general mathematical form for a simultaneous equations system is

$$\mathbf{Y}\boldsymbol{\Gamma} = \mathbf{X}\boldsymbol{B} + \boldsymbol{E} \tag{66.2.1}$$

If one splits $\mathbf{Y}$, $\mathbf{X}$, and $\boldsymbol{E}$ into their columns one gets

$$\begin{bmatrix} \mathbf{y}_1 & \dots & \mathbf{y}_m \end{bmatrix} \begin{bmatrix} \gamma_{11} & \dots & \gamma_{1m} \\ \vdots & \ddots & \vdots \\ \gamma_{M1} & \dots & \gamma_{mm} \end{bmatrix} = \\ = \begin{bmatrix} \mathbf{x}_1 & \dots & \mathbf{x}_k \end{bmatrix} \begin{bmatrix} \beta_{11} & \dots & \beta_{1m} \\ \vdots & \ddots & \vdots \\ \beta_{K1} & \dots & \beta_{km} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 & \dots & \boldsymbol{\varepsilon}_m \end{bmatrix}$$

The standard assumptions are that $\mathcal{E}[\boldsymbol{E}|\mathbf{X}] = \boldsymbol{O}$ and $\mathcal{V}[\operatorname{vec}\boldsymbol{E}|\mathbf{X}] = \boldsymbol{\Sigma} \otimes \boldsymbol{I}$ with an unknown nonsingular $\boldsymbol{\Sigma}$. $\boldsymbol{\Gamma}$ is assumed nonsingular as well. Furthermore it is assumed that $\operatorname{plim} \frac{1}{t}\mathbf{X}^\top\mathbf{X}$ exists and is nonsingular, and that $\operatorname{plim} \frac{1}{n}\mathbf{X}^\top\boldsymbol{\varepsilon} = \boldsymbol{o}$.

Problem 541. *1 point If $\mathcal{V}[\operatorname{vec}\boldsymbol{E}] = \boldsymbol{\Sigma} \otimes \boldsymbol{I}$, this means (check the true answer or answers) that*

- *different rows of $\boldsymbol{E}$ are uncorrelated, and every row has the same covariance matrix, or*

- *different columns of $\boldsymbol{E}$ are uncorrelated, and every column has the same covariance matrix, or*
- *all ε_{ij} are uncorrelated.*

Answer. The first answer is right. □

Now the reduced form equations: postmultiplying by $\boldsymbol{\Gamma}^{-1}$ and setting $\boldsymbol{\Pi} = \boldsymbol{B}\boldsymbol{\Gamma}^{-1}$ and $\mathbf{V} = \boldsymbol{E}\boldsymbol{\Gamma}^{-1}$ one obtains $\mathbf{Y} = \mathbf{X}\boldsymbol{\Pi} + \mathbf{V}$.

Problem 542. *If $\mathcal{V}[\operatorname{vec}\boldsymbol{E}] = \boldsymbol{\Sigma} \otimes \boldsymbol{I}$ and $\mathbf{V} = \boldsymbol{E}\boldsymbol{\Gamma}^{-1}$, show that $\mathcal{V}[\operatorname{vec}\mathbf{V}] = (\boldsymbol{\Gamma}^{-1})^\top\boldsymbol{\Sigma}\boldsymbol{\Gamma}^{-1} \otimes \boldsymbol{I}$.*

Answer. First use (B.5.19) to develop $\operatorname{vec}\mathbf{V} = -\operatorname{vec}(\boldsymbol{I}\boldsymbol{E}\boldsymbol{\Gamma}^{-1}) = -\Big((\boldsymbol{\Gamma}^{-1})^\top \otimes \boldsymbol{I}\Big)\operatorname{vec}\boldsymbol{E}$, therefore

$$\mathcal{V}[\operatorname{vec}\mathbf{V}] = \Big((\boldsymbol{\Gamma}^{-1})^\top \otimes \boldsymbol{I}\Big)\Big(\mathcal{V}[\operatorname{vec}\boldsymbol{E}]\Big)\Big(\boldsymbol{\Gamma}^{-1} \otimes \boldsymbol{I}\Big) = (\boldsymbol{\Gamma}^{-1})^\top\boldsymbol{\Sigma}\boldsymbol{\Gamma}^{-1} \otimes \boldsymbol{I}. \tag{66.2.2}$$

□

Here is an example, inspired by, but not exactly identical to, [**JHG+88**, pp. 607–9]. The structural equations are:

$$\mathbf{y}_1 = -\mathbf{y}_2\gamma_{21} + \mathbf{x}_2\beta_{21} + \boldsymbol{\varepsilon}_1 \tag{66.2.3}$$

$$\mathbf{y}_2 = -\mathbf{y}_1\gamma_{12} + \mathbf{x}_1\beta_{12} + \mathbf{x}_3\beta_{32} + \boldsymbol{\varepsilon}_2 \tag{66.2.4}$$

This is the form in which structural equations usually arise naturally: one of the endogenous variables is on the left of each of the structural equations. There are as many structural equations as there are endogenous variables. The notation for the unknown parameters and minus sign in front of γ_{12} and γ_{21} come from the fact that these parameters are elements of the matrices $\boldsymbol{\Gamma}$ and $\boldsymbol{B}$.

In matrix notation, this system of structural equations becomes

$$\begin{bmatrix} \mathbf{y}_1 & \mathbf{y}_2 \end{bmatrix} \begin{bmatrix} 1 & \gamma_{12} \\ \gamma_{21} & 1 \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \mathbf{x}_3 \end{bmatrix} \begin{bmatrix} 0 & \beta_{12} \\ \beta_{21} & 0 \\ 0 & \beta_{32} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 & \boldsymbol{\varepsilon}_2 \end{bmatrix} \tag{66.2.5}$$

Note that normalization conventions and exclusion restrictions are built directly into $\boldsymbol{\Gamma}$ and $\boldsymbol{B}$. In general it is not necessary that each structural equation has a different endogenous variable on the left. Often the same endogenous variable may be on the lefthand side of more than one structural equation. In this case, $\boldsymbol{\Gamma}$ in (66.2.5) does not have 1 in the diagonal but has a 1 somewhere in every column.

In the present hypothetical exercise we are playing God and therefore know the true parameter values $-\gamma_{21} = 1$, $-\gamma_{12} = 2$, $\beta_{21} = 2$, $\beta_{12} = 3$, and $\beta_{32} = 1$. And

while the earthly researcher only knows that the following two matrices exist and are nonsingular, we know their precise values:

$$\operatorname{plim}\frac{1}{n}\begin{bmatrix}\boldsymbol{\varepsilon}_1^\top\\ \boldsymbol{\varepsilon}_2^\top\end{bmatrix}\begin{bmatrix}\boldsymbol{\varepsilon}_1 & \boldsymbol{\varepsilon}_2\end{bmatrix}=\operatorname{plim}\frac{1}{n}\begin{bmatrix}\boldsymbol{\varepsilon}_1^\top\boldsymbol{\varepsilon}_1 & \boldsymbol{\varepsilon}_1^\top\boldsymbol{\varepsilon}_2\\ \boldsymbol{\varepsilon}_2^\top\boldsymbol{\varepsilon}_1 & \boldsymbol{\varepsilon}_2^\top\boldsymbol{\varepsilon}_2\end{bmatrix}=\begin{bmatrix}5 & 1\\ 1 & 1\end{bmatrix}$$

$$\operatorname{plim}\frac{1}{n}\begin{bmatrix}\mathbf{x}_1^\top\\ \mathbf{x}_2^\top\\ \mathbf{x}_3^\top\end{bmatrix}\begin{bmatrix}\mathbf{x}_1 & \mathbf{x}_2 & \mathbf{x}_3\end{bmatrix}=\operatorname{plim}\frac{1}{n}\begin{bmatrix}\mathbf{x}_1^\top\mathbf{x}_1 & \mathbf{x}_1^\top\mathbf{x}_2 & \mathbf{x}_1^\top\mathbf{x}_3\\ \mathbf{x}_2^\top\mathbf{x}_1 & \mathbf{x}_2^\top\mathbf{x}_2 & \mathbf{x}_2^\top\mathbf{x}_3\\ \mathbf{x}_3^\top\mathbf{x}_1 & \mathbf{x}_3^\top\mathbf{x}_2 & \mathbf{x}_3^\top\mathbf{x}_3\end{bmatrix}=\begin{bmatrix}1 & 1 & 0\\ 1 & 2 & 0\\ 0 & 0 & 1\end{bmatrix}$$

Besides, the assumption is always (and this is known to the earthly researcher too):

$$\operatorname{plim}\frac{1}{n}\begin{bmatrix}\mathbf{x}_1^\top\\ \mathbf{x}_2^\top\\ \mathbf{x}_3^\top\end{bmatrix}\begin{bmatrix}\boldsymbol{\varepsilon}_1 & \boldsymbol{\varepsilon}_2\end{bmatrix}=\operatorname{plim}\frac{1}{n}\begin{bmatrix}\mathbf{x}_1^\top\boldsymbol{\varepsilon}_1 & \mathbf{x}_1^\top\boldsymbol{\varepsilon}_2\\ \mathbf{x}_2^\top\boldsymbol{\varepsilon}_1 & \mathbf{x}_2^\top\boldsymbol{\varepsilon}_2\\ \mathbf{x}_3^\top\boldsymbol{\varepsilon}_1 & \mathbf{x}_3^\top\boldsymbol{\varepsilon}_2\end{bmatrix}=\begin{bmatrix}0 & 0\\ 0 & 0\\ 0 & 0\end{bmatrix}$$

First let us compute the true values of the reduced form parameters. Insert the known parameter values into (66.2.5):

$$\begin{bmatrix}\mathbf{y}_1 & \mathbf{y}_2\end{bmatrix}\begin{bmatrix}1 & -2\\ -1 & 1\end{bmatrix}=\begin{bmatrix}\mathbf{x}_1 & \mathbf{x}_2 & \mathbf{x}_3\end{bmatrix}\begin{bmatrix}0 & 3\\ 2 & 0\\ 0 & 1\end{bmatrix}+\begin{bmatrix}\boldsymbol{\varepsilon}_1 & \boldsymbol{\varepsilon}_2\end{bmatrix}\tag{66.2.6}$$

Using

$$\begin{bmatrix}1 & -2\\ -1 & 1\end{bmatrix}^{-1}=-\begin{bmatrix}1 & 2\\ 1 & 1\end{bmatrix}\quad\text{and}\quad\begin{bmatrix}0 & 3\\ 2 & 0\\ 0 & 1\end{bmatrix}\begin{bmatrix}1 & 2\\ 1 & 1\end{bmatrix}=\begin{bmatrix}3 & 3\\ 2 & 4\\ 1 & 1\end{bmatrix},$$

we can solve as follows:

$$\begin{bmatrix}\mathbf{y}_1 & \mathbf{y}_2\end{bmatrix}=-\begin{bmatrix}\mathbf{x}_1 & \mathbf{x}_2 & \mathbf{x}_3\end{bmatrix}\begin{bmatrix}3 & 3\\ 2 & 4\\ 1 & 1\end{bmatrix}-\begin{bmatrix}\boldsymbol{\varepsilon}_1 & \boldsymbol{\varepsilon}_2\end{bmatrix}\begin{bmatrix}1 & 2\\ 1 & 1\end{bmatrix}\tag{66.2.7}$$

I.e., the true parameter matrix $\mathbf{\Pi}$ in the reduced form equation $\mathbf{Y}=\mathbf{X}\mathbf{\Pi}+\boldsymbol{E}\mathbf{\Gamma}^{-1}$ is $\mathbf{\Pi}=-\begin{bmatrix}3 & 3\\ 2 & 4\\ 1 & 1\end{bmatrix}$. If we postmultiply (66.2.7) by $\left[\begin{smallmatrix}1\\ 0\end{smallmatrix}\right]$ and $\left[\begin{smallmatrix}0\\ 1\end{smallmatrix}\right]$ we get the reduced-form equation written column by column:

$$\mathbf{y}_1=-\begin{bmatrix}\mathbf{x}_1 & \mathbf{x}_2 & \mathbf{x}_3\end{bmatrix}\begin{bmatrix}3\\ 2\\ 1\end{bmatrix}-\begin{bmatrix}\boldsymbol{\varepsilon}_1 & \boldsymbol{\varepsilon}_2\end{bmatrix}\begin{bmatrix}1\\ 1\end{bmatrix}=-3\mathbf{x}_1-2\mathbf{x}_2-\mathbf{x}_3-\boldsymbol{\varepsilon}_1-\boldsymbol{\varepsilon}_2$$

$$\mathbf{y}_2=-\begin{bmatrix}\mathbf{x}_1 & \mathbf{x}_2 & \mathbf{x}_3\end{bmatrix}\begin{bmatrix}3\\ 4\\ 1\end{bmatrix}-\begin{bmatrix}\boldsymbol{\varepsilon}_1 & \boldsymbol{\varepsilon}_2\end{bmatrix}\begin{bmatrix}2\\ 1\end{bmatrix}=-3\mathbf{x}_1-4\mathbf{x}_2-\mathbf{x}_3-2\boldsymbol{\varepsilon}_1-\boldsymbol{\varepsilon}_2$$

Problem 543. *Show that the* plim*s of the OLS estimates in equation* (66.2.3) *are* $\operatorname{plim}\hat{\gamma}_{21;OLS}=-0.6393\neq-1$ *and* $\operatorname{plim}\hat{\beta}_{21;OLS}=0.0164\neq 2$*, i.e., OLS is inconsistent. Do these* plim*s depend on the covariance matrix of the disturbances?*

Answer. The first structural equation is

$$\mathbf{y}_1=\begin{bmatrix}\mathbf{y}_2 & \mathbf{x}_2\end{bmatrix}\begin{bmatrix}-\gamma_{21}\\ \beta_{21}\end{bmatrix}+\boldsymbol{\varepsilon}_1=\mathbf{Z}_1\boldsymbol{\delta}_1+\boldsymbol{\varepsilon}_1\tag{66.2.8}$$

The OLS estimators are $\hat{\boldsymbol{\delta}}=(\mathbf{Z}_1^\top\mathbf{Z}_1)^{-1}\mathbf{Z}_1^\top\mathbf{y}_1$ or, with factors $1/n$ so that we can take plims,

$$\begin{bmatrix}-\hat{\gamma}_{21;OLS}\\ \hat{\beta}_{21;OLS}\end{bmatrix}=\left(\frac{1}{n}\begin{bmatrix}\mathbf{y}_2^\top\mathbf{y}_2 & \mathbf{y}_2^\top\mathbf{x}_2\\ \mathbf{x}_2^\top\mathbf{y}_2 & \mathbf{x}_2^\top\mathbf{x}_2\end{bmatrix}\right)^{-1}\frac{1}{n}\begin{bmatrix}\mathbf{y}_2^\top\mathbf{y}_1\\ \mathbf{x}_2^\top\mathbf{y}_1\end{bmatrix}\tag{66.2.9}$$

The plims of the squares and cross products of the $\mathbf{x}_i$ and $\mathbf{y}_i$ can be computed from those of the $\mathbf{x}_i$ and $\boldsymbol{\varepsilon}_i$ which we know since we are playing God. Here are those relevant for running OLS on the first equation: Since

$$\begin{aligned}\mathbf{y}_2^\top\mathbf{y}_2=&\begin{bmatrix}3 & 4 & 1\end{bmatrix}\begin{bmatrix}\mathbf{x}_1^\top\\ \mathbf{x}_2^\top\\ \mathbf{x}_3^\top\end{bmatrix}\begin{bmatrix}\mathbf{x}_1 & \mathbf{x}_2 & \mathbf{x}_3\end{bmatrix}\begin{bmatrix}3\\ 4\\ 1\end{bmatrix}\\ &+2\cdot\begin{bmatrix}3 & 4 & 1\end{bmatrix}\begin{bmatrix}\mathbf{x}_1^\top\\ \mathbf{x}_2^\top\\ \mathbf{x}_3^\top\end{bmatrix}\begin{bmatrix}\boldsymbol{\varepsilon}_1 & \boldsymbol{\varepsilon}_2\end{bmatrix}\begin{bmatrix}2\\ 1\end{bmatrix}\\ &+\begin{bmatrix}2 & 1\end{bmatrix}\begin{bmatrix}\boldsymbol{\varepsilon}_1^\top\\ \boldsymbol{\varepsilon}_2^\top\end{bmatrix}\begin{bmatrix}\boldsymbol{\varepsilon}_1 & \boldsymbol{\varepsilon}_2\end{bmatrix}\begin{bmatrix}2\\ 1\end{bmatrix},\end{aligned}$$

it follows, after taking plims

$$\operatorname{plim}\frac{1}{n}\mathbf{y}_2^\top\mathbf{y}_2=\begin{bmatrix}3 & 4 & 1\end{bmatrix}\begin{bmatrix}1 & 1 & 0\\ 1 & 2 & 0\\ 0 & 0 & 1\end{bmatrix}\begin{bmatrix}3\\ 4\\ 1\end{bmatrix}+\begin{bmatrix}2 & 1\end{bmatrix}\begin{bmatrix}5 & 1\\ 1 & 1\end{bmatrix}\begin{bmatrix}2\\ 1\end{bmatrix}=66+25=91$$

Equally

$$\mathbf{x}_2^\top\mathbf{y}_2=-\begin{bmatrix}\mathbf{x}_2^\top\mathbf{x}_1 & \mathbf{x}_2^\top\mathbf{x}_2 & \mathbf{x}_2^\top\mathbf{x}_3\end{bmatrix}\begin{bmatrix}3\\ 4\\ 1\end{bmatrix}-\begin{bmatrix}\mathbf{x}_2^\top\boldsymbol{\varepsilon}_1 & \mathbf{x}_2^\top\boldsymbol{\varepsilon}_2\end{bmatrix}\begin{bmatrix}2\\ 1\end{bmatrix}$$

$$\operatorname{plim}\frac{1}{n}\mathbf{x}_2^\top\mathbf{y}_2=-\begin{bmatrix}1 & 2 & 0\end{bmatrix}\begin{bmatrix}3\\ 4\\ 1\end{bmatrix}=-11$$

Also

$$\mathbf{y}_2^\top \mathbf{y}_1 = \begin{bmatrix} 3 & 4 & 1 \end{bmatrix} \begin{bmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \mathbf{x}_3^\top \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \mathbf{x}_3 \end{bmatrix} \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix} + \begin{bmatrix} 3 & 4 & 1 \end{bmatrix} \begin{bmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \mathbf{x}_3^\top \end{bmatrix} \begin{bmatrix} \varepsilon_1 & \varepsilon_2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} +$$

$$+ \begin{bmatrix} 2 & 1 \end{bmatrix} \begin{bmatrix} \varepsilon_1^\top \\ \varepsilon_2^\top \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \mathbf{x}_3 \end{bmatrix} \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix} + \begin{bmatrix} 2 & 1 \end{bmatrix} \begin{bmatrix} \varepsilon_1^\top \\ \varepsilon_2^\top \end{bmatrix} \begin{bmatrix} \varepsilon_1 & \varepsilon_2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$\operatorname{plim} \frac{1}{n} \mathbf{y}_2^\top \mathbf{y}_1 = \begin{bmatrix} 3 & 4 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix} + \begin{bmatrix} 2 & 1 \end{bmatrix} \begin{bmatrix} 5 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = 44 + 14 = 58$$

Finally

$$\mathbf{x}_2^\top \mathbf{y}_1 = -\begin{bmatrix} \mathbf{x}_2^\top \mathbf{x}_1 & \mathbf{x}_2^\top \mathbf{x}_2 & \mathbf{x}_2^\top \mathbf{x}_3 \end{bmatrix} \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix} - \begin{bmatrix} \mathbf{x}_2^\top \varepsilon_1 & \mathbf{x}_2^\top \varepsilon_2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$\operatorname{plim} \frac{1}{n} \mathbf{x}_2^\top \mathbf{y}_1 = -\begin{bmatrix} 1 & 2 & 0 \end{bmatrix} \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix} = -7$$

One sees that the covariance matrix of the disturbance terms enters some of these results. Putting it all together gives

$$\operatorname{plim} \begin{bmatrix} -\hat{\gamma}_{21} \\ \hat{\beta}_{21} \end{bmatrix} = \begin{bmatrix} 91 & -11 \\ -11 & 2 \end{bmatrix}^{-1} \begin{bmatrix} 58 \\ -7 \end{bmatrix} = \frac{1}{61} \begin{bmatrix} 2 & 11 \\ 11 & 91 \end{bmatrix} \begin{bmatrix} 58 \\ -7 \end{bmatrix} = \frac{1}{61} \begin{bmatrix} 39 \\ 1 \end{bmatrix} = \begin{bmatrix} 0.6393 \\ 0.0164 \end{bmatrix} \neq \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

Therefore OLS is inconsistent. □

66.3. Indirect Least Squares

If OLS is inconsistent, what estimation methods can be used instead? An obvious choice of a consistent estimator is Indirect Least Squares (ILS), i.e., run OLS on the reduced form equations, and then compute the structural parameters from the reduced form parameters. (The reduced form parameters themselves are usually not much of interest, since they represent a mixture of the different effects which are separated out in the structural equations.) In this estimation of the structural equations, one uses OLS on every equation individually, because one considers it a SUR system with equal $\mathbf{X}$-matrices.

However if one applies this in the present system, one has 6 parameters to estimate in the reduced form equation (66.2.7), and only 5 in the structural equations (66.2.5). To understand how this discrepancy arises, look at the relationship between the structural and reduced form parameters, which can most simply be written as $\mathbf{\Pi\Gamma} = \boldsymbol{B}$:

$$\begin{bmatrix} \pi_{11} & \pi_{12} \\ \pi_{21} & \pi_{22} \\ \pi_{31} & \pi_{32} \end{bmatrix} \begin{bmatrix} 1 & -\gamma_{12} \\ -\gamma_{21} & 1 \end{bmatrix} = \begin{bmatrix} 0 & \beta_{12} \\ \beta_{21} & 0 \\ 0 & \beta_{32} \end{bmatrix} \tag{66.3.1}$$

The coefficients of the first structural equation are in the first columns of $\mathbf{\Gamma}$ and $\boldsymbol{B}$. Let us write these first columns separately:

$$\begin{bmatrix} \pi_{11} & \pi_{12} \\ \pi_{21} & \pi_{22} \\ \pi_{31} & \pi_{32} \end{bmatrix} \begin{bmatrix} 1 \\ -\gamma_{21} \end{bmatrix} = \begin{bmatrix} 0 \\ \beta_{21} \\ 0 \end{bmatrix} \quad \text{or} \quad (66.3.2) \qquad \begin{aligned} \pi_{11} - \pi_{12}\gamma_{21} &= 0 \\ \pi_{21} - \pi_{22}\gamma_{21} &= \beta_{21} \\ \pi_{31} - \pi_{32}\gamma_{21} &= 0 \end{aligned}$$

One sees that there are two ways to get γ_{21} from the elements of $\mathbf{\Pi}$: $\gamma_{21} = \pi_{11}/\pi_{12}$ or $\gamma_{21} = \pi_{31}/\pi_{32}$. The ILS principle gives us therefore two different consistent estimates of γ_{21}, but no obvious way to combine them. This is called: the first structural equation is "overidentified." If one looks at the true values one sees that indeed $\pi_{11}/\pi_{12} = \pi_{31}/\pi_{32}$. The estimation of the reduced form equations does not take advantage of all the information given in the structural equations: they should have been estimated as a constrained estimate, not with a linear constraint but a bilinear constraint of the form $\pi_{11}\pi_{32} = \pi_{31}\pi_{12}$. ILS is therefore not the most efficient estimation method for the first structural equation.

How about the second structural equation?

$$\begin{bmatrix} \pi_{11} & \pi_{12} \\ \pi_{21} & \pi_{22} \\ \pi_{31} & \pi_{32} \end{bmatrix} \begin{bmatrix} -\gamma_{12} \\ 1 \end{bmatrix} = \begin{bmatrix} \beta_{12} \\ 0 \\ \beta_{32} \end{bmatrix} \quad \text{or} \quad (66.3.3) \qquad \begin{aligned} -\pi_{11}\gamma_{12} + \pi_{12} &= \beta_{12} \\ -\pi_{21}\gamma_{12} + \pi_{22} &= 0 \\ -\pi_{31}\gamma_{12} + \pi_{32} &= \beta_{32} \end{aligned}$$

This can be solved uniquely: $\gamma_{12} = \pi_{22}/\pi_{21}$, $\beta_{12} = \pi_{12} - \pi_{11}\pi_{22}/\pi_{21}$, $\beta_{32} = \pi_{32} - \pi_{31}\pi_{22}/\pi_{21}$. Therefore one says that the second equaton is exactly identified.

It is also possible that an equation is not identified. This identification status is not a property of ILS, but a property of the model.

66.4. Instrumental Variables (2SLS)

A somewhat more sophisticated approach to estimation in a simultaneous equations system would be: use those exogenous variables which are not included in the ith structural equation as instruments for the endogenous variables on the righthand side of the ith structural equation.

I.e., for the first structural equation in our example, (66.2.3), we can use $\mathbf{x}_1$ and $\mathbf{x}_3$ as instruments for $\mathbf{y}_2$ (while $\mathbf{x}_2$ is its own instrument), and in the second structural equation we can use $\mathbf{x}_2$ as instrument for $\mathbf{y}_1$ (while $\mathbf{x}_1$ and $\mathbf{x}_3$ are their own instruments).

In order to show how this is connected with ILS, I will prove that ILS is identical to instrumental variables in the special case that there are as many instruments as regressors. I will show by the example of our second structural equation that ILS is exactly equal to instrumental variables. This is the proof given in [**JHG+88**, Section 15.1.1], not in general but in our example.

Remember how we got the ILS estimates $\tilde{\boldsymbol{\Gamma}}$ and $\tilde{\mathbf{B}}$: First we ran the regression on the unrestricted reduced form to get $\hat{\boldsymbol{\Pi}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$, and then we solved the equation $\hat{\boldsymbol{\Pi}}\tilde{\boldsymbol{\Gamma}} = \tilde{\mathbf{B}}$ where $\tilde{\boldsymbol{\Gamma}}$ and $\tilde{\mathbf{B}}$ have the zeros and the normalization ones inserted at the right places, see (66.3.3).

In the case of the 2nd equation this becomes

$$\hat{\boldsymbol{\Pi}}\tilde{\boldsymbol{\gamma}}_2 = \tilde{\boldsymbol{\beta}}_2 \tag{66.4.1}$$

Now premultiply (66.4.1) by $\mathbf{X}^\top\mathbf{X}$ to get

$$\mathbf{X}^\top\mathbf{Y}\hat{\boldsymbol{\gamma}}_2 = \mathbf{X}^\top\mathbf{X}\hat{\boldsymbol{\beta}}_2 \tag{66.4.2}$$

or

$$\begin{bmatrix}\mathbf{x}_1^\top\\ \mathbf{x}_2^\top\\ \mathbf{x}_3^\top\end{bmatrix}\begin{bmatrix}\mathbf{y}_1 & \mathbf{y}_2\end{bmatrix}\begin{bmatrix}\tilde{\gamma}_{12}\\ 1\end{bmatrix} = \begin{bmatrix}\mathbf{x}_1^\top\\ \mathbf{x}_2^\top\\ \mathbf{x}_3^\top\end{bmatrix}\begin{bmatrix}\mathbf{x}_1 & \mathbf{x}_2 & \mathbf{x}_3\end{bmatrix}\begin{bmatrix}\tilde{\beta}_{12}\\ 0\\ \tilde{\beta}_{32}\end{bmatrix} \tag{66.4.3}$$

This simplifies

$$\begin{bmatrix}\mathbf{x}_1^\top\\ \mathbf{x}_2^\top\\ \mathbf{x}_3^\top\end{bmatrix}(\mathbf{y}_1\tilde{\gamma}_{12} + \mathbf{y}_2) = \begin{bmatrix}\mathbf{x}_1^\top\\ \mathbf{x}_2^\top\\ \mathbf{x}_3^\top\end{bmatrix}(\mathbf{x}_1\tilde{\beta}_{12} + \mathbf{x}_3\tilde{\beta}_{32}) \tag{66.4.4}$$

Now rearrange

$$\begin{bmatrix}\mathbf{x}_1^\top\\ \mathbf{x}_2^\top\\ \mathbf{x}_3^\top\end{bmatrix}\mathbf{y}_2 = \begin{bmatrix}\mathbf{x}_1^\top\\ \mathbf{x}_2^\top\\ \mathbf{x}_3^\top\end{bmatrix}(\mathbf{x}_1\tilde{\beta}_{12} - \mathbf{y}_1\tilde{\gamma}_{12} + \mathbf{x}_3\tilde{\beta}_{32}) = \begin{bmatrix}\mathbf{x}_1^\top\\ \mathbf{x}_2^\top\\ \mathbf{x}_3^\top\end{bmatrix}\begin{bmatrix}\mathbf{x}_1 & \mathbf{y}_1 & \mathbf{x}_3\end{bmatrix}\begin{bmatrix}\tilde{\beta}_{12}\\ -\tilde{\gamma}_{12}\\ \tilde{\beta}_{32}\end{bmatrix} \tag{66.4.5}$$

I will show that (66.4.5) is exactly the normal equation for the IV estimator. Write the second structural equation as

$$\mathbf{y}_2 = \begin{bmatrix}\mathbf{x}_1 & \mathbf{y}_1 & \mathbf{x}_3\end{bmatrix}\begin{bmatrix}\beta_{12}\\ -\gamma_{12}\\ \beta_{32}\end{bmatrix} + \boldsymbol{\varepsilon}_2 \tag{66.4.6}$$

The matrix of instruments is $\mathbf{W} = \begin{bmatrix}\mathbf{x}_1 & \mathbf{x}_2 & \mathbf{x}_3\end{bmatrix}$, i.e., $\mathbf{x}_1$ and $\mathbf{x}_3$ are instruments for themselves, and $\mathbf{x}_2$ is an instrument for $\mathbf{y}_1$. Now remember the IV normal equation in this simplified case: instead of $\mathbf{X}^\top\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^\top\mathbf{y}$ one has $\mathbf{W}^\top\mathbf{X}\tilde{\boldsymbol{\beta}} = \mathbf{W}^\top\mathbf{y}$. In our situation this gives

$$\begin{bmatrix}\mathbf{x}_1^\top\\ \mathbf{x}_2^\top\\ \mathbf{x}_3^\top\end{bmatrix}\mathbf{y}_2 = \begin{bmatrix}\mathbf{x}_1^\top\\ \mathbf{x}_2^\top\\ \mathbf{x}_3^\top\end{bmatrix}\begin{bmatrix}\mathbf{x}_1 & \mathbf{y}_1 & \mathbf{x}_3\end{bmatrix}\begin{bmatrix}\tilde{\beta}_{12}\\ -\tilde{\gamma}_{12}\\ \tilde{\beta}_{32}\end{bmatrix} \tag{66.4.7}$$

which is, as claimed, the same as (66.4.5).

Now in the overidentified case, ILS does not have a good method to offer. There are more than one ways to get the reduced form estimates from the structural estimates, and the ILS principle says that one could use either one, but there is no easy way to combine them. The estimation approach by Instrumental Variables, on the other hand, has an obvious way to take advantage of overidentification: one will do Instrumental Variables in the generalized case in which there are "too many" instruments. This is exactly 2SLS.

Problem 544. *1 point Describe the two "stages" in the two stages least squares estimation of a structural equation which is part of a simultaneous equations system.*

66.5. Identification

How can one tell by looking at the structural equations whether the equation is exactly identified or underidentified or overidentified? If one just has one system, solving the reduced form equations by hand is legitimate.

The so-called "order condition" is not sufficient but necessary for identification. One possible formulation of it is: each equation must have at least $m-1$ exclusions. One can also say, and this is the formulation which I prefer: for each endogenous variable on the righthand side of the structural equation, at least one exogenous variable must be excluded from this equation.

Problem 545. *This example is adapted from* [**JHG+88**, (14.5.8) on p. 617]*:*

• a. *2 points Use the order condition to decide which of the following equations are exactly identified, overidentified, not identified.*

$$\mathbf{y}_1 = -\mathbf{y}_2\gamma_{21} - \mathbf{y}_4\gamma_{41} + \mathbf{x}_1\beta_{11} + \mathbf{x}_4\beta_{41} + \boldsymbol{\varepsilon}_1 \tag{66.5.1}$$

$$\mathbf{y}_2 = -\mathbf{y}_1\gamma_{12} + \mathbf{x}_1\beta_{12} + \mathbf{x}_2\beta_{22} + \boldsymbol{\varepsilon}_2 \tag{66.5.2}$$

$$\mathbf{y}_1 = -\mathbf{y}_2\gamma_{23} - \mathbf{y}_3\gamma_{33} - \mathbf{y}_4\gamma_{43} + \mathbf{x}_1\beta_{13} + \mathbf{x}_4\beta_{43} + \boldsymbol{\varepsilon}_3 \tag{66.5.3}$$

$$\mathbf{y}_4 = \mathbf{x}_1\beta_{14} + \mathbf{x}_2\beta_{24} + \mathbf{x}_3\beta_{34} + \mathbf{x}_4\beta_{44} + \boldsymbol{\varepsilon}_4 \tag{66.5.4}$$

Answer. (66.5.4) is exactly identified since there are no endogenous variable on the right hand side, but all exogenous variables are on the right hand side. (66.5.3) is not identified, it has 3 $\mathbf{y}$'s on the right hand side but only excludes two $\mathbf{x}$'s. (66.5.2) overfulfils the order condition, overidentified. (66.5.1) is exactly identified. □

• b. *1 point Write down the matrices* $\boldsymbol{\Gamma}$ *and* $\boldsymbol{B}$ *(indicating where there are zeros and ones) in the matrix representation of this system, which has the form*

$$\begin{bmatrix}\mathbf{y}_1 & \mathbf{y}_2 & \mathbf{y}_3 & \mathbf{y}_4\end{bmatrix}\begin{bmatrix}\gamma_{11} & \gamma_{12} & \gamma_{13} & \gamma_{14}\\ \gamma_{21} & \gamma_{22} & \gamma_{23} & \gamma_{24}\\ \gamma_{31} & \gamma_{32} & \gamma_{33} & \gamma_{34}\\ \gamma_{41} & \gamma_{42} & \gamma_{43} & \gamma_{44}\end{bmatrix} = \tag{66.5.5}$$

$$= \begin{bmatrix}\mathbf{x}_1 & \mathbf{x}_2 & \mathbf{x}_3 & \mathbf{x}_4\end{bmatrix}\begin{bmatrix}\beta_{11} & \beta_{12} & \beta_{13} & \beta_{14}\\ \beta_{21} & \beta_{22} & \beta_{23} & \beta_{24}\\ \beta_{31} & \beta_{32} & \beta_{33} & \beta_{34}\\ \beta_{41} & \beta_{42} & \beta_{43} & \beta_{44}\end{bmatrix} + \begin{bmatrix}\boldsymbol{\varepsilon}_1 & \boldsymbol{\varepsilon}_2 & \boldsymbol{\varepsilon}_3 & \boldsymbol{\varepsilon}_4\end{bmatrix}$$

Answer.

$$\boldsymbol{\Gamma} = \begin{bmatrix}1 & \gamma_{12} & 1 & 0\\ \gamma_{21} & 1 & \gamma_{23} & 0\\ 0 & 0 & \gamma_{33} & 0\\ \gamma_{41} & 0 & \gamma_{43} & 1\end{bmatrix} \qquad \boldsymbol{B} = \begin{bmatrix}\beta_{11} & \beta_{12} & \beta_{13} & \beta_{14}\\ 0 & \beta_{22} & 0 & \beta_{24}\\ 0 & 0 & 0 & \beta_{34}\\ \beta_{41} & 0 & \beta_{43} & \beta_{44}\end{bmatrix}$$

□

Criteria which are necessary *and* sufficient for identification are called "rank conditions." There are various equivalent forms for it. We will pick out here one of these equivalent formulations, that which is preferred by ITPE, and give a recipe how to apply it. We will give no proofs.

First of all, define the matrix

$$\boldsymbol{\Delta} = \begin{bmatrix}\boldsymbol{\Gamma}\\ \boldsymbol{B}\end{bmatrix} \tag{66.5.6}$$

$\boldsymbol{\Delta}$ contains in its ith column the coefficients of the ith structural equation. In our example if is

$$\boldsymbol{\Delta} = \begin{bmatrix}\gamma_{11} & \gamma_{12} & \gamma_{13} & 0\\ \gamma_{21} & \gamma_{22} & \gamma_{23} & 0\\ 0 & 0 & \gamma_{33} & 0\\ \gamma_{41} & 0 & \gamma_{43} & \gamma_{44}\\ \beta_{11} & \beta_{12} & \beta_{13} & \beta_{14}\\ 0 & \beta_{22} & 0 & \beta_{24}\\ 0 & 0 & 0 & \beta_{34}\\ \beta_{41} & 0 & \beta_{43} & \beta_{44}\end{bmatrix} \tag{66.5.7}$$

Each column of $\boldsymbol{\Delta}$ is subject to a different set of exclusion restrictions, say the ith column of $\boldsymbol{\Delta}$ is $\boldsymbol{\delta}_i$ and it satisfies $\boldsymbol{R}_i\boldsymbol{\delta}_i = \boldsymbol{o}$. For instance in the first equation (66.2.3)

$\gamma_{31} = 0$, $\beta_{21} = 0$, and $\beta_{31} = 0$, therefore

$$\boldsymbol{R}_1 = \begin{bmatrix}0 & 0 & 1 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0\end{bmatrix} \tag{66.5.8}$$

Now instead of looking at $\boldsymbol{R}_i\boldsymbol{\delta}_i$ look at $\boldsymbol{R}_i\boldsymbol{\Delta}$. Since $\boldsymbol{\Delta}$ has m columns, and the ith column is annulled by $\boldsymbol{R}_i$, the rank of $\boldsymbol{R}_i\boldsymbol{\Delta}$ can at most be $m-1$. The rank condition for identification says that this rank *must* be $m-1$ for the ith equation to be identified. In our example,

$$\boldsymbol{R}_1\boldsymbol{\Delta} = \begin{bmatrix}0 & 0 & \gamma_{33} & 0\\ 0 & \beta_{22} & 0 & \beta_{24}\\ 0 & 0 & 0 & \beta_{34}\end{bmatrix} \tag{66.5.9}$$

All columns except the first must be linearly independent.

Problem 546. *1 point Show that the columns of the matrix*

$$\begin{bmatrix}0 & \gamma_{33} & 0\\ \beta_{22} & 0 & \beta_{24}\\ 0 & 0 & \beta_{34}\end{bmatrix} \tag{66.5.10}$$

are linearly independent if γ_{33}*,* β_{22}*, and* β_{34} *are nonzero.*

Answer.

$$\begin{bmatrix}0\\ \beta_{22}\\ 0\end{bmatrix}\alpha_1 + \begin{bmatrix}\gamma_{33}\\ 0\\ 0\end{bmatrix}\alpha_2 + \begin{bmatrix}0\\ \beta_{24}\\ \beta_{34}\end{bmatrix}\alpha_3 = \begin{bmatrix}0\\ 0\\ 0\end{bmatrix} \tag{66.5.11}$$

$\gamma_{33}\alpha_2 = 0$, therefore $\alpha_2 = 0$. It also implies $\gamma_{34}\alpha_3 = 0$, therefore also $\alpha_3 = 0$. It remains $\beta_{22}\alpha_1 + \beta_{24}\alpha_3 = 0$, but since we already know $\alpha_3 = 0$ this means that also $\alpha_1 = 0$. □

To understand why this rank condition is necessary for identification, assume the jth equation has exactly the same endogenous and exogenous variables as the ith which you are interested in. Then these two effects cannot be distinguished, i.e., neither equation is identified.

Here is more detail about the rank conditions, taken from [**JHG+88**, p. 624]:

If rank $\boldsymbol{R}_i\boldsymbol{\Delta} < m-1$ then the ith equation is not identified.

If rank $\boldsymbol{R}_i\boldsymbol{\Delta} = m-1$ and rank $\boldsymbol{R}_i = m-1$ then the ith equation is exactly identified.

If rank $\boldsymbol{R}_i\boldsymbol{\Delta} = m-1$ and rank $\boldsymbol{R}_i > m-1$ then the ith equation is overidentified.

66.6. Other Estimation Methods

To get an overview over the other estimation methods, we have to distinguish between single-equation methods and system methods, and between maximum likelihood estimators (based on the assumption that the error terms are multivariate

normal, which however also have good properties if this is not the case; see here [**DM93**, p. 641]) and estimators based on instrumental variables.

Single-equation estimators are simpler to compute and they are also more robust: if only one of the equations is mis-specified, then a systems estimator is inconsistent, but single-equations estimators of the other equations may still be consistent. Systems estimators are more efficient: they exploit the correlation between the residuals of the different equations, they allow exclusion restrictions in one equation to benefit another equation, and they also allow cross-equation restrictions on the parameters which cannot be handled by single-equations systems.

Maximum likelihood estimation of the whole model (FIML) requires numerical methods and is a demanding task. We assume $\boldsymbol{X}$ nonrandom, or we condition on $\mathbf{X} = \boldsymbol{X}$, therefore we write

$$\mathbf{Y}\boldsymbol{\Gamma} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E} \tag{66.6.1}$$

In (**??**) we split $\mathbf{Y}$, $\boldsymbol{X}$, and $\boldsymbol{E}$ into their columns; now we will split them into their rows:

$$\begin{bmatrix} \mathbf{y}_1^\top \\ \vdots \\ \mathbf{y}_t^\top \end{bmatrix} \boldsymbol{\Gamma} = \begin{bmatrix} \boldsymbol{x}_1^\top \\ \vdots \\ \boldsymbol{x}_t^\top \end{bmatrix} \boldsymbol{B} + \begin{bmatrix} \boldsymbol{\varepsilon}_1^\top \\ \vdots \\ \boldsymbol{\varepsilon}_t^\top \end{bmatrix} \tag{66.6.2}$$

Here $\boldsymbol{\varepsilon}_s \sim \mathsf{N}(\boldsymbol{o}, \boldsymbol{\Sigma})$ and $\boldsymbol{\varepsilon}_r$ independent of $\boldsymbol{\varepsilon}_s$ for $r \neq s$. Density of each $\boldsymbol{\varepsilon}$ is

$$f_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon}_s) = (2\pi)^{-m/2} (\det \boldsymbol{\Sigma})^{-1/2} \exp(-\frac{1}{2} \boldsymbol{\varepsilon}_s^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\varepsilon}_s) \tag{66.6.3}$$

According to the transformation rules for densities, we have to express the random variable whose density we want to know in terms of the random variable whose density we know: $\boldsymbol{\varepsilon}_s^\top = \mathbf{y}_s^\top \boldsymbol{\Gamma} - \boldsymbol{x}_s^\top \boldsymbol{B}$ or $\boldsymbol{\varepsilon}_s = \boldsymbol{\Gamma}^\top \mathbf{y}_s - \boldsymbol{B}^\top \boldsymbol{x}_s$. The Jacobian matrix is the derivative of this, it is $\partial \boldsymbol{\varepsilon}_s / \partial \mathbf{y}_s^\top = \boldsymbol{\Gamma}^\top$. Therefore the density of each $\mathbf{y}$ is

$$f_{\mathbf{y}}(\boldsymbol{y}_s) = (2\pi)^{-m/2} \left|\det \boldsymbol{\Gamma}\right| (\det \boldsymbol{\Sigma})^{-1/2} \exp(-\frac{1}{2} (\boldsymbol{y}_s^\top \boldsymbol{\Gamma} - \boldsymbol{x}_s^\top \boldsymbol{B}) \boldsymbol{\Sigma}^{-1} (\boldsymbol{\Gamma}^\top \boldsymbol{y}_s - \boldsymbol{B}^\top \boldsymbol{x}_s)).$$

Multiply this for all the rows and remember Problem 532 to get $f_{\mathbf{Y}}(\boldsymbol{Y}) =$

$$= (2\pi)^{-mt/2} \left|\det \boldsymbol{\Gamma}\right|^t (\det \boldsymbol{\Sigma})^{-t/2} \exp\bigl(-\frac{1}{2} \sum_{x=1}^{t} (\boldsymbol{y}_s^\top \boldsymbol{\Gamma} - \boldsymbol{x}_s^\top \boldsymbol{B}) \boldsymbol{\Sigma}^{-1} (\boldsymbol{\Gamma}^\top \boldsymbol{y}_s - \boldsymbol{B}^\top \boldsymbol{x}_s)\bigr)$$

$$= (2\pi)^{-mt/2} \left|\det \boldsymbol{\Gamma}\right|^t (\det \boldsymbol{\Sigma})^{-t/2} \exp\bigl(-\frac{1}{2} \operatorname{tr} (\boldsymbol{Y}\boldsymbol{\Gamma} - \boldsymbol{X}\boldsymbol{B})^\top (\boldsymbol{Y}\boldsymbol{\Gamma} - \boldsymbol{X}\boldsymbol{B}) \boldsymbol{\Sigma}^{-1}\bigr)$$

Therefore the log likelihood function $\ell = \log f_{\mathbf{Y}}(\mathbf{Y}) =$

$$= -\frac{mt}{2} \log(2\pi) + t \log \left|\det \boldsymbol{\Gamma}\right| - \frac{t}{2} \log(\det \boldsymbol{\Sigma}) - \frac{1}{2} \operatorname{tr} (\mathbf{Y}\boldsymbol{\Gamma} - \boldsymbol{X}\boldsymbol{B})^\top (\mathbf{Y}\boldsymbol{\Gamma} - \boldsymbol{X}\boldsymbol{B}) \boldsymbol{\Sigma}^{-1}.$$

If one compares this with the log likelihood function (65.2.5) for simultenous equations systems, one sees many similarities: the last item is a function of the residuals, $\boldsymbol{\Sigma}$ enters in exactly the same way, the only difference is the term $t \log \left|\det \boldsymbol{\Gamma}\right|$. Therefore the next steps here are parallel to our development in Chapter 65. However one can see already now the following shortcut if the system is a recursive system, i.e., if $\boldsymbol{\Gamma}$ is lower diagonal with 1s in the diagonal. Then $\det \boldsymbol{\Gamma} = 1$, and in this case one can just use the formalism developed for seemingy unrelated systems, simply ignoring the fact that some of the explanatory variables are endogenous, i.e., treating them in the same way as the exogonous variables.

But now let us go in with the general casae. In order to concentrate out $\boldsymbol{\Sigma}$ it is simpler to take the partial derivatives with respect to $\boldsymbol{\Sigma}^{-1}$ than those with respect to $\boldsymbol{\Sigma}$ itself. Using the matrix differentiation rules (C.1.24) and (C.1.16) and noting that $-t/2 \log \det \boldsymbol{\Sigma} = t/2 \log \det \boldsymbol{\Sigma}^{-1}$ one gets:

$$\partial \ell / \partial \boldsymbol{\Sigma}^{-1} = \frac{t}{2} \boldsymbol{\Sigma} - \frac{1}{2} (\mathbf{Y}\boldsymbol{\Gamma} - \boldsymbol{X}\boldsymbol{B})^\top (\mathbf{Y}\boldsymbol{\Gamma} - \boldsymbol{X}\boldsymbol{B}) \tag{66.6.4}$$

and if one sets this zero one gets $\hat{\boldsymbol{\Sigma}} = \frac{1}{t} (\mathbf{Y}\boldsymbol{\Gamma} - \boldsymbol{X}\boldsymbol{B})^\top (\mathbf{Y}\boldsymbol{\Gamma} - \boldsymbol{X}\boldsymbol{B})$. Plugging this in gives the concentrated log likelihood function $\log f(\mathbf{Y})_c =$

$$= -\frac{mt}{2} \log(2\pi) + t \log \left|\det \boldsymbol{\Gamma}\right| - \frac{t}{2} \log\bigl(\det \frac{1}{t} (\mathbf{Y}\boldsymbol{\Gamma} - \boldsymbol{X}\boldsymbol{B})^\top (\mathbf{Y}\boldsymbol{\Gamma} - \boldsymbol{X}\boldsymbol{B})\bigr) - \frac{mt}{2}.$$

This is not just a minimization of the SSE because of the $t \log \left|\det \boldsymbol{\Gamma}\right|$ term. This term makes things very complicated, since the information matrix is no longer block diagonal, see [**Ruu00**, p. 724] for more detail. One sees here that Simultaneous Equations is the SUR system of reduced form equations with nonlinear restrictions. Must be maximized subject to exclusion restrictions; difficult but can be done. References in [**DM93**, 640]. Since maximization routine will usually not cross the loci with $\det \boldsymbol{\Gamma} = 0$, careful selection of the starting value is important.

Here is an alternative derivation of the same result, using (65.3.2):

$$\begin{bmatrix} \mathbf{y}_1^\top \\ \vdots \\ \mathbf{y}_t^\top \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1^\top \\ \vdots \\ \boldsymbol{x}_t^\top \end{bmatrix} \boldsymbol{B}\boldsymbol{\Gamma}^{-1} + \begin{bmatrix} \boldsymbol{w}_1^\top \\ \vdots \\ \boldsymbol{w}_t^\top \end{bmatrix} \tag{66.6.5}$$

where $\boldsymbol{w}_s^\top = \boldsymbol{\varepsilon}_s^\top \boldsymbol{\Gamma}^{-1}$ or $\boldsymbol{w}_s = (\boldsymbol{\Gamma}^{-1})^\top \boldsymbol{\varepsilon}_s$, therefore $\boldsymbol{w}_s \sim \mathsf{N}(\boldsymbol{o}, (\boldsymbol{\Gamma}^{-1})^\top \boldsymbol{\Sigma} \boldsymbol{\Gamma}^{-1})$. According to (65.3.2) the concentrated likelihood function is

$$\begin{aligned}\ell_c &= -\frac{mt}{2}(1+\log 2\pi) - \frac{t}{2}\log\det(\mathbf{Y} - \boldsymbol{XB}\boldsymbol{\Gamma}^{-1})^\top(\mathbf{Y} - \boldsymbol{XB}\boldsymbol{\Gamma}^{-1}) \\ &= -\frac{mt}{2}(1+\log 2\pi) - \frac{t}{2}\log\det((\boldsymbol{\Gamma}^{-1})^\top(\mathbf{Y}\boldsymbol{\Gamma} - \boldsymbol{XB})^\top(\mathbf{Y}\boldsymbol{\Gamma} - \boldsymbol{XB})\boldsymbol{\Gamma}^{-1}) \\ &= -\frac{mt}{2}(1+\log 2\pi) + t\log|\det\boldsymbol{\Gamma}| - \frac{t}{2}\log\det(\mathbf{Y}\boldsymbol{\Gamma} - \boldsymbol{XB})^\top(\mathbf{Y}\boldsymbol{\Gamma} - \boldsymbol{XB})\end{aligned}$$

This must be maximized subject to the bilinear constraints imposed by the overidentifying restrictions.

Since FIML is so difficult and expensive, researchers often omit specification tests. [**DM93**] recommend to make these tests with the unrestricted reduced form. This is based on the assumption that most of these mis-specifications already show up on the unrestricted reduced form: serial correlation or heteroskedasticity of the error terms, test whether parameters change over the sample period.

Another specification test is also a test of the overidentifying restrictions: a LR test comparing the attained level of the likelihood function of the FIML estimator with that of the unrestricted reduced form estimator. Twice the difference between the restricted and unrestricted value of the log likelihood function $\sim \chi^2$ where the number of degrees of freedom is the number of the overidentifying restrictions.

LIML (limited information maximum likelihood) is a single-equation method based on maximum likelihood. In the model $\mathbf{Y}\boldsymbol{\Gamma} = \mathbf{X}\boldsymbol{B} + \boldsymbol{E}$ the ith equation is

$$\begin{bmatrix}\mathbf{y}_1 & \dots & \mathbf{y}_m\end{bmatrix}\begin{bmatrix}\gamma_{1i}\\ \vdots \\ \gamma_{mi}\end{bmatrix} = \begin{bmatrix}\mathbf{x}_1 & \dots & \mathbf{x}_K\end{bmatrix}\begin{bmatrix}\beta_{1i}\\ \vdots \\ \beta_{Ki}\end{bmatrix} + \boldsymbol{\varepsilon}_i \tag{66.6.6}$$

Some of the γ_{gi} and β_{hi} must be zero, and one of the γ_{gi} is 1. Rearrange the columns of $\mathbf{Y}$ and $\mathbf{X}$ such that $\gamma_{1i} = 1$, and that the zero coefficients come last:

$$\begin{bmatrix}\mathbf{y}_1 & \mathbf{Y}_2 & \mathbf{Y}_3\end{bmatrix}\begin{bmatrix}1\\ \boldsymbol{\gamma} \\ \boldsymbol{o}\end{bmatrix} = \begin{bmatrix}\mathbf{X}_1 & \mathbf{X}_2\end{bmatrix}\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{o}\end{bmatrix} + \boldsymbol{\varepsilon}_i \tag{66.6.7}$$

Now write the reduced form equations conformably:

$$\begin{bmatrix}\mathbf{y}_1 & \mathbf{Y}_2 & \mathbf{Y}_3\end{bmatrix} = \begin{bmatrix}\mathbf{X}_1 & \mathbf{X}_2\end{bmatrix}\begin{bmatrix}\boldsymbol{\pi}_{11} & \boldsymbol{\Pi}_{12} & \boldsymbol{\Pi}_{13}\\ \boldsymbol{\pi}_{21} & \boldsymbol{\Pi}_{22} & \boldsymbol{\Pi}_{23}\end{bmatrix} + \begin{bmatrix}\boldsymbol{v}_1 & \boldsymbol{V}_2 & \boldsymbol{V}_3\end{bmatrix} \tag{66.6.8}$$

Then LIML for the ith equation is maximum likelihood on the following system:

$$\mathbf{y}_1 + \mathbf{Y}_2\boldsymbol{\gamma} = \mathbf{X}_1\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i \tag{66.6.9}$$

$$\mathbf{Y}_2 = \mathbf{X}_1\boldsymbol{\Pi}_{12} + \mathbf{X}_2\boldsymbol{\Pi}_{22} + \boldsymbol{V}_2 \tag{66.6.10}$$

I.e., it includes the ith structural equation and the unrestricted reduced form equations for all the endogenous variables on the righthand side of the ith structural equation. Written as one partitioned matrix equation:

$$\begin{bmatrix}\mathbf{y}_1 & \mathbf{Y}_2\end{bmatrix}\begin{bmatrix}1 & \boldsymbol{o}^\top\\ \boldsymbol{\gamma} & \boldsymbol{I}\end{bmatrix} = \begin{bmatrix}\mathbf{X}_1 & \mathbf{X}_2\end{bmatrix}\begin{bmatrix}\boldsymbol{\beta} & \boldsymbol{\Pi}_{21}\\ \boldsymbol{o} & \boldsymbol{\Pi}_{22}\end{bmatrix} + \begin{bmatrix}\boldsymbol{\varepsilon}_i & \boldsymbol{V}_2\end{bmatrix} \tag{66.6.11}$$

Since $\boldsymbol{\Gamma} = \begin{bmatrix}1 & \boldsymbol{o}^\top\\ \boldsymbol{\gamma} & \boldsymbol{I}\end{bmatrix}$ is lower triangular, its determinant is the product of the diagonal elements, i.e., it is $= 1$. Therefore the Jacobian term in the likelihood function is $= 1$, and therefore the likelihood function is the same as that of a seemingly unrelated regression model. One can therefore compute the LIML estimator from (66.6.11) using the software for seemingly unrelated regressions, disregarding the difference between endogenous and exogenous variables. But there are other ways to compute this estimator which are simpler. They will not be discussed here. They either amount to (1) an eigenvalue problem. or (2) a "least variance ratio" estimator, or (3) a "k-class" estimator. See [**DM93**, pp. 645–647]. Although LIML is used less often than 2SLS, it has certain advantages: (1) it is invariant under reparametrization, and (2) 2SLS can be severly biased in small samples.

For 3SLS write the equations as

$$\begin{bmatrix}\mathbf{y}_1\\ \mathbf{y}_2\\ \vdots\\ \mathbf{y}_m\end{bmatrix} = \begin{bmatrix}\mathbf{Z}_1 & \boldsymbol{O} & \cdots & \boldsymbol{O}\\ \boldsymbol{O} & \mathbf{Z}_2 & \cdots & \boldsymbol{O}\\ \vdots & \vdots & \ddots & \vdots\\ \boldsymbol{O} & \boldsymbol{O} & \cdots & \mathbf{Z}_m\end{bmatrix}\begin{bmatrix}\boldsymbol{\delta}_1\\ \boldsymbol{\delta}_2\\ \vdots\\ \boldsymbol{\delta}_n\end{bmatrix} + \begin{bmatrix}\boldsymbol{\varepsilon}_1\\ \boldsymbol{\varepsilon}_2\\ \vdots\\ \boldsymbol{\varepsilon}_m\end{bmatrix} \qquad \mathcal{V}[\operatorname{vec}\boldsymbol{E}] = \boldsymbol{\Sigma}\otimes\boldsymbol{I} \tag{66.6.12}$$

Here the $\mathbf{Z}_i$ contain endogenous and exogenous variables, therefore OLS is inconsistent. But if we do 2SLS, i.e., if we take $\hat{\mathbf{Z}}_i = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Z}_i$ as regressors, we get consistent estimates:

(66.6.13)

$$\begin{bmatrix}\mathbf{y}_1\\ \mathbf{y}_2\\ \vdots\\ \mathbf{y}_m\end{bmatrix} = \begin{bmatrix}\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Z}_1 & \boldsymbol{O} & \cdots & \boldsymbol{O}\\ \boldsymbol{O} & \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Z}_2 & \cdots & \boldsymbol{O}\\ \vdots & \vdots & \ddots & \vdots\\ \boldsymbol{O} & \boldsymbol{O} & \cdots & \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Z}_m\end{bmatrix}\begin{bmatrix}\boldsymbol{\delta}_1\\ \boldsymbol{\delta}_2\\ \vdots\\ \boldsymbol{\delta}_m\end{bmatrix} + \begin{bmatrix}\boldsymbol{\varepsilon}_1\\ \boldsymbol{\varepsilon}_2\\ \vdots\\ \boldsymbol{\varepsilon}_m\end{bmatrix}$$

In the case of SUR we know that OLS singly is not efficient, but GLS is. We use this same method here: (1) estimate σ_{ij}—not from the residuals in (66.6.13) but as $\hat{\sigma}_{ij} = \frac{1}{t}\hat{\boldsymbol{\varepsilon}}_i\hat{\boldsymbol{\varepsilon}}_j$ where $\hat{\boldsymbol{\varepsilon}}_i = \mathbf{y}_i - \mathbf{Z}_i\hat{\boldsymbol{\delta}}_{i;2SLS}$. With this estimated covariance matrix do GLS

$$\operatorname{vec}(\hat{\mathbf{B}})_{3SLS} = \left(\hat{\mathbf{Z}}^\top(\hat{\boldsymbol{\Sigma}}\otimes\boldsymbol{I})^{-1}\hat{\mathbf{Z}}\right)^{-1}\hat{\mathbf{Z}}^\top(\hat{\boldsymbol{\Sigma}}\otimes\boldsymbol{I})^{-1}\operatorname{vec}(\mathbf{Y}), \tag{66.6.14}$$

which can also be written as

$$\operatorname{vec}(\hat{\mathbf{B}})_{3SLS} = \left(\hat{\mathbf{Z}}^\top(\hat{\boldsymbol{\Sigma}}^{-1} \otimes \boldsymbol{I})\mathbf{Z}\right)^{-1}\hat{\mathbf{Z}}^\top(\hat{\boldsymbol{\Sigma}}^{-1} \otimes \boldsymbol{I})\operatorname{vec}(\mathbf{Y}). \tag{66.6.15}$$

It is instrumental variables with a nonspherical covariance matrix (and can be derived as a GMM estimator). This is much easier to estimate than FIML, but it is nevertheless asymptotically as good as FIML.

Problem 547.

• a. *6 points Give an overview over the main issues in the estimation of a simultaneous equations system, and discuss the estimation principles involved.*

• b. *2 points How would you test whether a simultaneous equations system is correctly specified?*

CHAPTER 67

Timeseries Analysis

A time series y with typical element y_s is a (finite or infinite) sequence of random variables. Usually, the subscript s goes from 1 to ∞, i.e., the time series is written $\mathsf{y}_1, \mathsf{y}_2, \ldots$, but it may have different (finite or inifinite) starting or ending values.

67.1. Covariance Stationary Timeseries

A time series is *covariance-stationary* if and only if:

$$\mathrm{E}[\mathsf{y}_s] = \mu \quad \text{for all } s \tag{67.1.1}$$

$$\mathrm{var}[\mathsf{y}_s] < \infty \quad \text{for all } s \tag{67.1.2}$$

$$\mathrm{cov}[\mathsf{y}_s, \mathsf{y}_{s+k}] = \gamma_k \quad \text{for all } s \text{ and } k \tag{67.1.3}$$

I.e., the means do not depend on s, and the covariances only depend on the distances and not on s. A covariance stationary time series is characterized by the expected value of each observation μ, the variance of each observation σ^2, and the "autocorrelation function" ρ_k for $k \geq 1$ or, alternatively, by μ and the "autocovariance function" γ_k for $k \geq 0$. The autocovariance and autocorrelation functions are vectors containing the unique elements of the covariance and correlation matrices.

The simplest time series has all $\mathsf{y}_t \sim \mathrm{IID}(\mu, \sigma^2)$, i.e., all covariances between different elements are zero. If $\mu = 0$ this is called "white noise."

A covariance-stationary process y_t $(t = 1, \ldots, n)$ with expected value $\mu = \mathrm{E}[\mathsf{y}_i]$ is said to be *ergodic for the mean* if

$$\operatorname*{plim}_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \mathsf{y}_t = \mu. \tag{67.1.4}$$

We will usually require ergodicity along with stationarity.

Problem 548. [**Ham94**, pp. 46/7] *Give a simple example for a stationary time series process which is not ergodic for the mean.*

Answer. White noise plus a mean which is drawn once and for all from a $\mathrm{N}(0, \tau^2)$ independent of the white noise. □

67.1.1. Moving Average Processes. The following is based on [**Gra89**, pp. 63–91] and on [**End95**].

We just said that the simplest stationary process is a constant plus "white noise" (all autocorrelations zero). The next simplest process is a moving average process of order 1, also called a MA(1) process:

$$\mathsf{y}_t = \mu + \varepsilon_t + \beta\varepsilon_{t-1} \qquad \varepsilon_t \sim \mathrm{IID}(0, \sigma^2) \tag{67.1.5}$$

where the first y, say it is y_1, depends on the pre-sample ε_0.

Problem 549. *Compute the autocovariance and autocorrelation function of the time series defined in* (67.1.5)*, and show that the following process*

$$\mathsf{y}_t = \mu + \eta_t + \frac{1}{\beta}\eta_{t-1} \qquad \eta_t \sim \mathrm{IID}(0, \beta^2\sigma^2) \tag{67.1.6}$$

generates a timeseries with equal statistical properties as (67.1.5)*.*

Answer. (67.1.5): $\mathrm{var}[\mathsf{y}_t] = \sigma^2(1+\beta^2)$, $\mathrm{cov}[\mathsf{y}_t, \mathsf{y}_{t-1}] = \beta\sigma^2$, and $\mathrm{cov}[\mathsf{y}_t, \mathsf{y}_{t-h}] = 0$ for $h > 1$. $\mathrm{corr}[\mathsf{y}_t, \mathsf{y}_{t-1}] = \beta/(1-\beta^2)$. (67.1.6) gives the same variance $\beta^2\sigma^2(1+1/\beta^2) = \sigma^2(1+\beta^2)$ and the same correlation $(1/\beta)/(1+1/\beta^2) = \beta/(1+\beta^2)$ □

The moving-average representation of a timeseries is therefore not unique. It is not possible to tell from observation of the time series alone whether the process generating it was (67.1.5) or (67.1.6). One can say in general that unless $\beta = 1$ every MA(1) process could have been generated by a process in which $|\beta| < 1$. This process is called the invertible form or the fundamental representation of the time series.

Problem 550. *What are the implications for estimation of the fact that a MA-process can have different data-generating processes?*

Answer. Besides looking how the timeseries fits the data, the econometrician should also look whether the disturbances are plausible values in light of the actual history of the process, in order to ascertain that one is using the right representation. □

The fundamental representation of the time series is needed for forecasting. Let us first look at the simplest situation: the time series at hand is generated by the process (67.1.5) with $|\beta| < 1$, the parameters μ and β are known, and one wants to forecast y_{t+1} on the basis of all past and present observations. Clearly, the past and present has no information about ε_{t+1}, therefore the best we can hope to do is to forecast y_{t+1} by $\mu + \beta\varepsilon_t$.

But do we know ε_t? If a time series is generated by an invertible process, then someone who knows μ, β, and the current and all past values of y can use this to

reconstruct the value of the current disturbance. One sees this as follows:

$$\mathfrak{y}_t = \mu + \varepsilon_t + \beta\varepsilon_{t-1} \tag{67.1.7}$$

$$\varepsilon_t = \mathfrak{y}_t - \mu - \beta\varepsilon_{t-1} \tag{67.1.8}$$

$$\varepsilon_{t-1} = \mathfrak{y}_{t-1} - \mu - \beta\varepsilon_{t-2} \tag{67.1.9}$$

$$\varepsilon_t = \mathfrak{y}_t - \mu - \beta(\mathfrak{y}_{t-1} - \mu - \beta\varepsilon_{t-2}) \tag{67.1.10}$$

$$= -\mu(1-\beta) + \mathfrak{y}_t - \beta\mathfrak{y}_{t-1} + \beta^2\varepsilon_{t-2} \tag{67.1.11}$$

after the next step

$$\varepsilon_t = -\mu(1-\beta+\beta^2) + \mathfrak{y}_t - \beta\mathfrak{y}_{t-1} + \beta^2\mathfrak{y}_{t-2} - \beta^3\varepsilon_{t-3} \tag{67.1.12}$$

and after t steps

$$\varepsilon_t = -\mu\big(1-\beta+\beta^2-\cdots+(-\beta)^{t-1}\big) \tag{67.1.13}$$

$$+\mathfrak{y}_t - \beta\mathfrak{y}_{t-1} + \beta^2\mathfrak{y}_{t-2} - \cdots + (-\beta)^{t-1}\mathfrak{y}_1 + (-\beta)^t\varepsilon_0 \tag{67.1.14}$$

$$= -\mu\frac{1+(-\beta)^t}{1+\beta} + \sum_{i=0}^{t-1}(-\beta)^i\mathfrak{y}_{t-i} + (-\beta)^t\varepsilon_0 \tag{67.1.15}$$

If $|\beta| < 1$, the last term of the right hand side, which depends on the unobservable ε_0, becomes less and less important. Therefore, if μ and β are known, and all past values of $\mathfrak{y}_t$ are known, this is enough information to compute the value of the present disturbance ε_t. Equation (67.1.15) can be considered the "inversion" of the MA1-process, i.e., its representation as an infinite autoregressive process.

The disturbance in the invertible process is called the "fundamental innovation" because every $\mathfrak{y}_t$ is composed of a part which is determined by the history $\mathfrak{y}_{t-1}, \mathfrak{y}_{t-2}, \ldots$ plus ε_t which is new to the present period.

The invertible representation can therefore be used for forecasting: the best predictor of $\mathfrak{y}_{t+1}$ is $\mu + \beta\varepsilon_t$.

Even if a time series was actually generated by a non-invertible process, the formula based on the invertible process is still the best formula for prediction, but now it must be given a different interpretation.

All this can be generalized for higher order MA processes. [**Ham94**, pp. 64–68] says: for any noninvertible MA process (which is not borderline in the sense that $|\beta| = 1$) there is an invertible MA process which has same means, variances, and autocorrelations. It is called the "fundamental representation" of this process.

The fundamental representation of a process is the one which leads to very simple equations for forecasting. It used to be a matter of course to assume at the same time that also the true process which generated the timeseries must be an invertible process, although the reasons given to justify this assumption were usually

vague. The classic monograph [**BJ76**, p. 51] says, for instance: "The requirement of invertibility is needed if we are interested in associating present events with *past* happenings in a sensible manner." [**Dea92**, p. 85] justifies the requirement of invertibility as follows: "Without [invertibility] the consumer would have no way of calculating the innovation from current and past values of income."

But recently it has been discovered that certain economic models naturally lead to non-invertible data generating processes, see problem 552. This is a process in which the economic agents observe and act upon information which the econometrician cannot observe.

If one goes over to infinite MA processes, then one gets all indeterministic stationary processes. According to the so-called Wold decomposition, every stationary process can be represented as a (possibly infinite) moving average process plus a "linearly deterministic" term, i.e., a term which can be linearly predicted without error from its past. There is consensus that economic time series do not contain such linearly deterministic terms.

The errors in the infinite Moving Average representation also have to do with prediction: can be considered the errors in the best one-step ahead linear prediction based on the infinite past [**Rei93**, p. 7].

A stationary process without a linear deterministic term has therfore the form

$$\mathfrak{y}_t = \mu + \sum_{j=0}^{\infty}\psi_j\varepsilon_{t-j} \tag{67.1.16}$$

or, in vector notation

$$\mathfrak{y} = \boldsymbol{\iota}\mu + \sum_{j=0}^{\infty}\psi_j\boldsymbol{B}^j\boldsymbol{\varepsilon} \tag{67.1.17}$$

where the timeseries ε_s is white noise, and $\boldsymbol{B}$ is the backshift operator satisfying $\boldsymbol{e}_t^\top\boldsymbol{B} = \boldsymbol{e}_{t-1}^\top$ (here $\boldsymbol{e}_t$ is the tth unit vector which picks out the tth element of the time series).

The coefficients satiisfy $\sum\psi_i^2 < \infty$, and if they satisfy the stronger condition $\sum|\psi_i| < \infty$, then the process is called *causal*.

Problem 551. *Show that without loss of generality* $\psi_0 = 1$ *in* (67.1.16).

Answer. If say ψ_k is the first nonzero ψ, then simply write $\boldsymbol{\eta}_j = \psi_k\boldsymbol{\varepsilon}_{j+k}$ □

Dually, one can also represent each fully indeterministic stationary processs as an infinite AR-process $\mathfrak{y}_t - \mu + \sum_{j=1}^{p}\phi_i(\mathfrak{y}_{t-i} - \mu) = \varepsilon_t$. This representation is called *invertible* if it satisfies $\sum|\theta_i| < \infty$.

67.1.2. The Box Jenkins Approach. Now assume that the operator $\boldsymbol{\Psi}(\boldsymbol{B}) = \sum_{j=0}^{\infty} \psi_j \boldsymbol{B}^j$ can be written as the product $\boldsymbol{\Psi} = \boldsymbol{\Phi}^{-1}\boldsymbol{\Theta}$ where each $\boldsymbol{\Phi}$ and $\boldsymbol{\Theta}$ are finite polynomials in $\boldsymbol{B}$. Again, without loss of generality, the leading coefficients in $\boldsymbol{\Psi}$ and $\boldsymbol{\Theta}$ can be assumed to be $= 1$. Then the time series can be written

$$\mathsf{y}_t - \mu + \sum_{j=1}^{p} \phi_i(\mathsf{y}_{t-i} - \mu) = \varepsilon_t + \sum_{j=1}^{\infty} \theta_j \varepsilon_{t-j} \tag{67.1.18}$$

A process is an ARMA-process if it satisfies this relation, regardless of whether the process y_t is stationary or not. See [**Rei93**, p. 8]. Again, there may be more than one such representation for a given process.

The Box-Jenkins approach is based on the assumption that empirically occurring stationary timeseries can be modeled as low-order ARMA processes. This would for instance be the case if the time series is built up recursively from its own past, with innovations which extend over more than one period.

If this general assumption is satisfied, this has the following implications for methodology:

- Some simple procedures have been developed how to recognize which of these time series one is dealing with.
- In the case of autoregressive time series, estimation is extremely simple and can be done using the regression framework.

67.1.3. Moving Average Processes. In order to see what order a finite moving average process is, one should look at the correlation coefficients. If the order is j, then the *theoretical* correlation coefficients are zero for all values $> j$, and therefore the estimates of these correlation coefficients, which have the form

$$r_k = \frac{\sum_{t=k+1}^{n}(\mathsf{y}_t - \bar{\mathsf{y}})(\mathsf{y}_{t-k} - \bar{\mathsf{y}})}{\sum_{t=1}^{n}(\mathsf{y}_t - \bar{\mathsf{y}})^2} \tag{67.1.19}$$

must be insignificant.

For estimation the preferred estimate is the maximum likelihood estimate. It can not be represented in closed form, therefore we have to rely on numerical maximization procedures.

67.1.4. Autoregressive Processes. The common wisdom in econometrics is that economic time series are often built up recursively from their own past. Example of an AR(1) process is

$$\mathsf{y}_t = \alpha \mathsf{y}_{t-1} + \varepsilon_t \tag{67.1.20}$$

where the first observation, say it is y_1, depends on the pre-sample y_0. (67.1.20) is called a difference equation.

This process generates a stationary timeseries only if $|\alpha| < 1$. Proof: $\operatorname{var}[\mathsf{y}_t] = \operatorname{var}[\mathsf{y}_{t-1}]$ means $\operatorname{var}[\mathsf{y}_t] = \alpha^2 \operatorname{var}[\mathsf{y}_t] + \sigma^2$ and therefore $\operatorname{var}[\mathsf{y}_t](1 - \alpha^2) = \sigma^2$, and since $\sigma^2 > 0$ by assumption, it follows that $1 - \alpha^2 > 0$.

Solution (i.e., Wold representation as a MA process) is

$$\mathsf{y}_t = \mathsf{y}_0 \alpha^t + (\varepsilon_t + \alpha\varepsilon_{t-1} + \cdots + \alpha^{t-1}\varepsilon_1) \tag{67.1.21}$$

As proof that this is a solution, write down $\alpha \mathsf{y}_{t-1}$ and check that it is equal to $\mathsf{y}_t - \varepsilon_t$.

67.1.5. Difference Equations. Let's make here a digression about nth order linear difference equations with constant coefficients. Definition from [**End95**, p. 8]:

$$\mathsf{y}_t = \alpha_0 + \sum_{i=1}^{n} \alpha_i \mathsf{y}_{t-i} + \mathsf{x}_t \tag{67.1.22}$$

here x_t is called the "forcing process." A solution of this difference equation is an expression of y_t in terms of present and past values of x and of t and of initial values of y_t. Difference equations usually have more than one solution, this is why these initial values are needed to identify the solution.

In order to solve this, the following 4 steps are needed (this is [**End95**, p. 17]):

(1) Form the homogeneous equation and find all n homogeneous solutions.

(2) Find a particular solution.

(3) Then the general solution is the sum of the particular solution and an arbitrary linear combination of all homogeneous solutions.

(4) Eliminate the arbitrary constant(s) by imposing the initial condition(s) on the general solution.

Let us apply this to $\mathsf{y}_t = \alpha \mathsf{y}_{t-1} + \varepsilon_t$. The homogeneous equation is $\mathsf{y}_t = \alpha \mathsf{y}_{t-1}$ and this has the general solution $\mathsf{y}_t = \beta\alpha^t$ where β is an arbitrary constant. If the timeseries goes back to $-\infty$, the particular solution is $\mathsf{y}_t = \sum_{i=0}^{\infty} \alpha^i \varepsilon_{t-i}$, but if the timeseries only exists for $t \geq 1$ the particular solution is $\mathsf{y}_t = \sum_{i=0}^{t-1} \alpha^i \varepsilon_{t-i}$. This gives solution (67.1.21).

Now let us look at a second order process: $\mathsf{y}_t = \alpha_1 \mathsf{y}_{t-1} + \alpha_2 \mathsf{y}_{t-2} + \mathsf{x}_t$. In order to get solutions of the homogeneous equation $\mathsf{y}_t = \alpha_1 \mathsf{y}_{t-1} + \alpha_2 \mathsf{y}_{t-2}$ try $\mathsf{y}_t = \beta\gamma^t$. This gives the following condition for γ: $\gamma^t = \alpha_1\gamma^{t-1} + \alpha_2\gamma^{t-2}$ or $\gamma^2 - \alpha_1\gamma + \alpha_2 = 0$. The solution of this quadratic equation is

$$\gamma = \frac{\alpha_1 \pm \sqrt{\alpha_1^2 + 4\alpha_2}}{2} \tag{67.1.23}$$

If this equation has two real roots, then everything is fine. If it has only one real root, i.e., if $\alpha_2 = -\alpha_1^2/4$, then $\gamma = \alpha_1/2$, i.e., $\mathsf{y}_t = \beta_1(\alpha_1/2)^t$ is one solution. But there is also a second solution, which is not obvious: $\mathsf{y}_t = \beta_2 t(\alpha_1/2)^t$ is a solution as well. One sees this by checking:

$$t(\alpha_1/2)^t = \alpha_1(t-1)(\alpha_1/2)^{t-1} + \alpha_2(t-2)(\alpha_1/2)^{t-2} \tag{67.1.24}$$

Simplify this and you will see that it holds.

If the roots of the characteristic equation are complex, one needs linear combinations of these complex roots, which are trigonometric functions. Here the homogeneous solution can be written in the form

$$\mathsf{y}_t = \beta_1 r^t \cos(\theta t + \beta_2) \tag{67.1.25}$$

where $r = \sqrt{-\alpha_2}$ and θ is defined by $\cos(\theta) = \alpha_1/2r$. This formula is from [**End95**, p. 29], and more explanations can be found there.

But in all these cases the roots of the characteristic equations determine the character of the homogeneous solution. They also determine whether the difference equation is stable, i.e., whether the homogeneous solutions die out over time or not. For stability, all roots must lie in the unit circle.

In terms of the coefficients themselves, these stability conditions are much more complicated. See [**End95**, pp. 31–33].

These stability conditions are also important for stochastic difference equations: in order to have stationary solutions, it must be stable.

It is easy to estimate AR processes: simply regress the time series on its lags. But before one can do this estimation one has to know the order of the autoregressive process. A useful tool for this are the partial autocorrelation coefficients.

We discussed partial correlation coefficients in chapter 19. The kth partial autocorrelation coefficient is the correlation between y_t and y_{t-k} with the influence of the invervening lags partialled out. The kth *sample* partial autocorrelation coefficient is the last coefficient in the regression of the timeseries on its first k lags. It is the effect which the kth lag has which cannot be explained by earlier lagged values. In an autoregressive process of order k, the "theoretical" partial autocorrelations are zero for lags greater than k, therefore the *estimated* partial autocorrelation coefficients should be insignificant for those lags. The asymptotic distribution of these estimates is normal with zero mean and variance $1/T$, therefore one often finds lines at $2/\sqrt{T}$ and $-2/\sqrt{T}$ in the plot of the estimated partial autocorrelation coefficients, which give an indication which values are significant at the 95% level and which are not.

67.1.6. ARMA(p,q) Processes. Sometimes it is appropriate to estimate a stationary process as having both autoregressive and moving average components (ARMA) or, if they are not stationary, they may be autoregressive or moving average after differencing them one or several times (ARIMA).

An $ARMA(p,q)$ process is the solution of a pth order difference equation with a $MA(q)$ as driving process.

These models have been very successful. On the one hand, there is reason to believe on theoretical grounds that many economic timeseries are $ARMA(p,q)$. [**Gra89**, p. 64] cites an interesting theorem which also contributes to the usefulness of

$ARMA$ processes: the sum of two independent series, one of which is $ARMA(p_1,q_1)$ and the other $ARMA(p_2,q_2)$, is $ARMA\big(p_1+p_2, \max(p_1+q_2, p_2+q_1)\big)$.

Box and Jenkins recommend to use the autocorrelations and partial autocorrelations for determining the order of the autoregressive or moving average parts, although this more difficult for an ARMA process than for an MA or AR process.

The last step after what in the time series context is called "*identification*" (a more generally used term might be "specification" or "model selection") and *estimation* is *diagnostic checking*, i.e., a check whether the results bear out the assumptions made by the model. Such diagnostic checks are necessary because mis-specification is possible if one follows this procedure. One way would be to see whether the residuals resemble a white noise process, by looking at the autocorrelation coefficients of the residuals. The so-called portmanteau test statistics test whether a given series is white noise: there is either the Box-Pierce statistic which is the sum of the squared sample autocorrelations

$$Q = T\sum_{k=1}^{p} r_k^2 \tag{67.1.26}$$

or the Ljung-Box statistic

$$Q = T(T-2)\sum_{k=1}^{p} \frac{r_k^2}{T-k} \tag{67.1.27}$$

which is asymptotically the same as the Box-Pierce statistic but seems to have better small-sample properties.

A second way to check the model is to overfit the model and see if the additional coefficients are zero. A third way would be to use the model for forecasting and to see whether important features of the original timeseries are captured (whether it can forecast turning points, etc.)

[**Gre97**, 839–841] gives an example. Eyeballing the timeseries does not give the impression that it is a stationary process, but the statistics seem to suggest an AR-2 process.

67.2. Vector Autoregressive Processes

[**JHG+88**, Chapter 18.1] start with an example in which an economic timeseries is not adequately modelled by a function of its own past plus some present innovations, but where two timeseries are jointly determined by their past plus some innovation: consumption function

$$\mathsf{c}_t = \eta_1 + \mathsf{y}_t\alpha + \mathsf{c}_{t-1}\beta + \varepsilon_{1t} \tag{67.2.1}$$

and then there is also a lagged effect of consumption on income

$$\mathsf{y}_t = \eta_2 + \mathsf{c}_{t-1}\gamma + \mathsf{y}_{t-1}\delta + \varepsilon_{2t} \tag{67.2.2}$$

This is the structural form of a dynamic simultaneous equations system. Identification status: first equation has y_t on righthand side and does not have y_{t-1}, therefore exactly identified. Second equation has no endogenous variables on the righthand side, which makes it also exactly identified. One can see it also by solving the reduced form equation for the structural coefficients. Therefore lets look at its reduced form. The second equation is already in reduced form, since it only has lagged values of c and y on the righthand side. The first becomes

$$\mathsf{c}_t = (\eta_1 + \alpha\eta_2) + \mathsf{c}_{t-1}(\beta + \alpha\gamma) + \mathsf{y}_{t-1}\alpha\delta + (\varepsilon_{1t} + \alpha\varepsilon_{2t}) \tag{67.2.3}$$

This reduced form is an unconstrained VAR(1) process:

$$\begin{bmatrix}\mathsf{c}_t & \mathsf{y}_t\end{bmatrix} = \begin{bmatrix}\eta_1 & \eta_2\end{bmatrix} + \begin{bmatrix}\mathsf{c}_{t-1} & \mathsf{y}_{t-1}\end{bmatrix}\begin{bmatrix}\theta_1 & \psi_1\\ \gamma_1 & \delta_1\end{bmatrix} + \begin{bmatrix}\varepsilon_{1t} & \varepsilon_{2t}\end{bmatrix} \tag{67.2.4}$$

Disturbances have same properties as errors in simultaneous equations systems (it is called vector white noise). VAR processes are special cases of multivariate time series. Therefore we will first take a look at multivariate time series in general. A good source here is [**Rei93**].

Covariance stationarity of multivariate time series is the obvious extension of the univariate definition (67.1.1)–(67.1.2):

$$\mathcal{E}[\mathbf{y}_t] = \boldsymbol{\mu} \tag{67.2.5}$$

$$\operatorname{var}[\mathsf{y}_{mt}] < \infty \tag{67.2.6}$$

$$\mathcal{C}[\mathbf{y}_t, \mathbf{y}_{t-h}] \quad \text{only depends on } h. \tag{67.2.7}$$

One can write a VAR(j) process as

$$\mathbf{y}_t^\top = \boldsymbol{\mu}^\top + (\mathbf{y}_{t-1} - \boldsymbol{\mu})^\top\boldsymbol{\Theta}_1 + \cdots + (\mathbf{y}_{t-n} - \boldsymbol{\mu})^\top\boldsymbol{\Theta}_n + \boldsymbol{\varepsilon}_t^\top \tag{67.2.8}$$

or equivalently

$$(\mathbf{y}_t - \boldsymbol{\mu})^\top - \sum_{j=1}^{n}(\mathbf{y}_{t-j} - \boldsymbol{\mu})^\top\boldsymbol{\Theta}_j = \boldsymbol{\varepsilon}_t^\top \tag{67.2.9}$$

In this notation the contemperaneous dependencies are in the covariance matrix of the disturbances. But there are other ways to write it too: but one can also write it

$$\sum_{j=0}^{n}(\mathbf{y}_{t-j} - \boldsymbol{\mu})^\top\boldsymbol{\Theta}_j = \boldsymbol{\varepsilon}_t^\top \tag{67.2.10}$$

where $\boldsymbol{\Theta}_0$ is lower diagonal and the covariance matrix of the disturbances is diagonal. For each permutation of the variables there is a unique lower diagonal $\boldsymbol{\Theta}_0$

which makes the covariance matrix of the disturbances the identity matrix, here prior knowledge about the order in which the variables depend on each other is necessary. But if one has a representation like this, one can build an impulse response function.

Condition for a VAR(n) process to be stationary is, using (67.2.9):

$$\det[\boldsymbol{I} - \boldsymbol{\Theta}_1 z - \boldsymbol{\Theta}_2 z^2 - \cdots - \boldsymbol{\Theta}_n z^n] \tag{67.2.11}$$

has all its roots outside the unit circle. These are the same conditions as the stability conditions.

Under general conditions, all stationary vector time series are $VAR(P)$ of a possibly infinite degree.

Estimation: the reduced form is like a disturbance-related equation system with all explanatory variables the same: therefore OLS is consistent, efficient, and asymptotically normal. But OLS is insensitive, since one has so many parameters to estimate. Therefore one may introduce restrictions, not all lagged variables appear in all equations, or one can use Bayesian methods (Minnesota prior, see [**BLR99**, pp. 269–72]).

Instead of using theory and prior knowledge to determine the number of lags, we use statistical criteria. Minimize an adaptation of Akaike's AIC criterion

$$AIC(n) = \log\det(\tilde{\boldsymbol{\Sigma}}_n) + \frac{2M^2 n}{T} \tag{67.2.12}$$

$$SC(n) = \log\det(\tilde{\boldsymbol{\Sigma}}_n) + \frac{M^2 n \log T}{T} \tag{67.2.13}$$

where M = number of variables in the system, T = sample size, n = number of lags included, and $\tilde{\boldsymbol{\Sigma}}$ has elements $\tilde{\sigma}_{ij} = \frac{\hat{\boldsymbol{\varepsilon}}_i^\top \hat{\boldsymbol{\varepsilon}}_j}{T}$

Again, diagnostic checks necessary because mis-specification is possible.

What to do with the estimation once it is finished? (1) forecasting really easy, the AR-framework gives natural forecasts. One-step ahead forecasts by simply using present and past values of the timeseries and setting the future innovations zero, and in order to get forecasts more than one step ahead, use the one-step etc. forecasts for those date which have not yet been observed.

67.2.1. Granger Causality. Granger causality tests are tests whether certain autoregressive coefficients are zero. It makes more sense to speak of Granger-noncausality: the time series $\mathbf{x}$ *fails* to Granger-cause $\mathbf{y}$ if $\mathbf{y}$ can be predicted as well from its own past as from the past of $\mathbf{x}$ and $\mathbf{y}$. An equivalent expression is: in a regression of y_t on its own lagged values $\mathsf{y}_{t-1}, \mathsf{y}_{t-1}, \ldots$ and the lagged values $\mathsf{x}_{t-1}, \mathsf{x}_{t-2}, \ldots$, the coefficients of $\mathsf{x}_{t-1}, \mathsf{x}_{t-2}, \ldots$ are not significantly different from zero.

Alternative test proposed by Sims: $\mathbf{x}$ fails to Granger-cause $\mathbf{y}$ if in a regression of y_t on lagged, current, and future x_q, the coefficients of the future x_q are zero.

I have this from [**Mad88**, 329/30]. Leamer says that this should be called precedence, not causality, because all we are testing is precedence. I disagree; these tests do have implications on whether the researcher would want to draw causal inferences from his or her data, and the discussion of causality should be included in statistics textbooks.

Innovation accounting or impulse response functions: make a moving average representation, and then you can pick the timepath of the innovations: perhaps a 1-period shock, or a stepped increase, whatever is of economic interest. Then you can see how these shocks are propagated through the system.

Caveats:

(1) do not make these experiments too dissimilar to what actually transpired in the data from which the parameters were estimated.

(2) Innovations are correlated, and if you increase one without increasing another which is highly correlated with it then you may get misleading results.

Way out would be: transform the innovations in such a way that their estimated covariance matrix is diagonal, and only experiment with these diagonalized innovations. But there are more than one way to do this.

If one has the variables ordered in a halfways sensible way, then one could use the Cholesky decomposition, which diagonalizes this ordering of the variables.

Other approaches: forecast error (MSE) can be decomposed into a sum of contributions coming from the different innovations: but this decomposition is not unique!

Then the MA-representation is the answer to: how can one make policy recommendations with such a framework.

Here is an example how an economic model can lead to a non-invertible VARMA process. It is from [**AG97**, p. 119], originally in [**Qua90**] and [**BQ89**]. Income at time t is the sum of a permanent and a transitory component $\mathsf{y}_t = \mathsf{y}^{\mathsf{p}}{}_t + \mathsf{y}^{\mathsf{t}}{}_t$; the permanent follows a random walk $\mathsf{y}^{\mathsf{p}}{}_t = \mathsf{y}^{\mathsf{p}}{}_{t-1} + \delta_t$ while the transitory income is white noise, i.e., $\mathsf{y}^{\mathsf{t}}{}_t = \varepsilon_t$. $\operatorname{var}[\varepsilon_t] = \operatorname{var}[\delta_t] = \sigma^2$, and all disturbances are mutually independent. Consumers know which part of their income is transitory and which part is permanent; they have this information because they know their own particular circumstances, but this kind of information is not directly available to the econometrician. Consumers act on their privileged information: their increase in consumption is all of their increase in permanent income plus fraction $\beta < 1$ of their transitory income $\mathsf{c}_t - \mathsf{c}_{t-1} = \delta_t + \beta\varepsilon_t$. One can combine all this into

$$\mathsf{y}_t - \mathsf{y}_{t-1} = \delta_t + \varepsilon_t - \varepsilon_{t-1} \qquad \delta_i \sim (0, \sigma^2) \tag{67.2.14}$$

$$\mathsf{c}_t - \mathsf{c}_{t-1} = \delta_t + \beta\varepsilon_t \qquad \varepsilon_i \sim (0, \sigma^2) \tag{67.2.15}$$

This is a vector-moving-average process for the first differences

$$\begin{bmatrix} \mathsf{y}_t - \mathsf{y}_{t-1} \\ \mathsf{c}_t - \mathsf{c}_{t-1} \end{bmatrix} = \begin{bmatrix} 1 & 1-L \\ 1 & \beta \end{bmatrix} \begin{bmatrix} \delta_t \\ \varepsilon_t \end{bmatrix} \tag{67.2.16}$$

but it is not invertible. In other words, the econometrician cannot consistently estimate the values of the present disturances from the past of this timeseries. who only sees the timepaths of income and consumption, cannot reconstruct from this these data the information which the agents themselves used to make their consumption decision.

There is an invertible data generating process too, but it has the coefficients

$$\begin{bmatrix} \mathsf{y}_t - \mathsf{y}_{t-1} \\ \mathsf{c}_t - \mathsf{c}_{t-1} \end{bmatrix} = \frac{1}{\sqrt{1+\beta^2}} \begin{bmatrix} 1-(1-\beta)L & 1+\beta-\beta L \\ 0 & 1+\beta^2 \end{bmatrix} \begin{bmatrix} \xi_t \\ \zeta_t \end{bmatrix} \tag{67.2.17}$$

If the econometrician uses an estimation method which automatically generates the invertible representation, he will get the wrong answer. He will think that the shocks which have a permanent impact on y also have a delayed effect in the opposite direction on next year's income, but have no effect on consumption; and that the shocks affecting consumption this period also have an effect on this period's income and an opposite effect on next period's income. This is a quite different scenario, and in many respects the opposite scenario, than that in equation (67.2.16).

Problem 552. *It is the purpose of this question to show that the following two vector moving averages are empirically indistinguishable:*

$$\begin{bmatrix} \mathsf{u}_t \\ \mathsf{v}_t \end{bmatrix} = \begin{bmatrix} 1 & 1-L \\ 1 & \beta \end{bmatrix} \begin{bmatrix} \delta_t \\ \varepsilon_t \end{bmatrix} \tag{67.2.18}$$

and

$$\begin{bmatrix} \mathsf{u}_t \\ \mathsf{v}_t \end{bmatrix} = \frac{1}{\sqrt{1+\beta^2}} \begin{bmatrix} 1-(1-\beta)L & 1+\beta-\beta L \\ 0 & 1+\beta^2 \end{bmatrix} \begin{bmatrix} \xi_t \\ \zeta_t \end{bmatrix} \tag{67.2.19}$$

where all error terms δ, ε, ξ, and ζ are independent with equal variances σ^2.

- a. *Show that in both situations*

$$\mathcal{V}\Big[\begin{bmatrix} \mathsf{u}_t \\ \mathsf{v}_t \end{bmatrix}\Big] = \sigma^2 \begin{bmatrix} 3 & 1+\beta \\ 1+\beta & 1+\beta^2 \end{bmatrix}, \qquad \mathcal{C}\Big[\begin{bmatrix} \mathsf{u}_t \\ \mathsf{v}_t \end{bmatrix}, \begin{bmatrix} \mathsf{u}_{t-1} \\ \mathsf{v}_{t-1} \end{bmatrix}\Big] = \sigma^2 \begin{bmatrix} -1 & -\beta \\ 0 & 0 \end{bmatrix} \tag{67.2.20}$$

and that the higher lags have zero covariances.

Answer. First scenario: $u_t = \delta_t + \varepsilon_t - \varepsilon_{t-1}$ and $v_t = \delta_t + \beta\varepsilon_t$. Therefore $\operatorname{var}[u_t] = 3\sigma^2$; $\operatorname{cov}[u_t, v_t] = \sigma^2 + \beta\sigma^2$, $\operatorname{var}[v_t] = \sigma^2 + \beta^2\sigma^2$; $\operatorname{cov}[u_t, u_{t-1}] = -\sigma^2$; $\operatorname{cov}[u_t, v_{t-1}] = -\beta\sigma^2$, $\operatorname{cov}[v_t, u_{t-1}] = \operatorname{cov}[v_t, v_{t-1}] = 0$.

Second scenario: leaving out the factor $\frac{1}{\sqrt{1+\beta^2}}$, we have $u_t = \xi_t - (1-\beta)\xi_{t-1} + (1+\beta)\zeta_t - \beta\zeta_{t-1}$ and $v_t = (1+\beta^2)\zeta_t$ Therefore $\operatorname{var}[u_t] = 3\sigma^2$; $\operatorname{cov}[u_t, v_t] = \sigma^2 + \beta\sigma^2$, $\operatorname{var}[v_t] = \sigma^2 + \beta^2\sigma^2$; $\operatorname{cov}[u_t, u_{t-1}] = -\sigma^2$; $\operatorname{cov}[u_t, v_{t-1}] = -\beta\sigma^2$, $\operatorname{cov}[v_t, u_{t-1}] = \operatorname{cov}[v_t, v_{t-1}] = 0$. □

• b. *Show also that the first representation has characteristic root* $1-\beta$*, and the second has characteristic root* $\frac{1}{1-\beta}$*. I.e., with* $\beta < 1$*, the first is not invertible but the second is.*

Answer. Replace the Lag operator L by the complex variable z, and compute the determinant:

$$\det \begin{bmatrix} 1 & 1-z \\ 1 & \beta \end{bmatrix} = \beta - (1-z) \tag{67.2.21}$$

setting this determinant zero gives $z = 1-\beta$, i.e., the first representation has a root within the unit circle, therefore it is not invertible. For the second representation we get

$$\det \begin{bmatrix} 1-(1-\beta)z & 1+\beta-\beta z \\ 0 & 1+\beta^2 \end{bmatrix} = (1-(1-\beta)z)(1+\beta^2) \tag{67.2.22}$$

Setting this zero gives $1-(1-\beta)z = 0$ or $z = \frac{1}{1-\beta}$, which is outside the unit circle. Therefore this representation is invertible. □

[**AG97**, p. 119] writes: "When the agents' information set and the econometricians' information set do coincide, then the MA representation is fundamental." Non-fundamental representations for the observed variables are called-for when the theoretical framework postulates that agents observe variables that the econometrician cannot observe.

67.3. Nonstationary Processes

Here are some stylized facts about timeseries in economics, taken from [**End95**, p. 136–7]:

- Most series contain a clear trend.
- Some series seem to meander.
- Any shock to a series displays a high degree of persistence.
- Volatility of many series is not constant over time. (ARCH)
- Some series share comovements with other series (cointegration).
- Many series are seasonal.

What to do about trend/meandering? One can fit for instance a linear or polynomial time trend:

$$\mathfrak{y}_t = a_0 + a_1 t + a_2 t^2 + \cdots + a_n t^n + \varepsilon_t \tag{67.3.1}$$

But it may also be the case that there is a stochastic trend. To study this look at the random walk:

$$\mathfrak{y}_t = \mathfrak{y}_{t-1} + \varepsilon_t \tag{67.3.2}$$

I.e., the effects of the disturbances do not die out but they are permanent. The MA-representation of this series is

$$\mathfrak{y}_t = \mathfrak{y}_0 + \sum_{i=1}^{t} \varepsilon_i \tag{67.3.3}$$

n-step-ahead forecasts at time t are $\mathfrak{y}_t$.

Problem 553. *Show that in a random walk process* (67.3.3) *(with* $\mathfrak{y}_0$ *non-stochastic)* $\operatorname{var}[\mathfrak{y}_t] = t\sigma^2$ *(i.e., it is nonstationary),* $\operatorname{cov}[\mathfrak{y}_t, \mathfrak{y}_{t-h}] = \sigma^2(t-h)$*, and* $\operatorname{corr}[\mathfrak{y}_t, \mathfrak{y}_{t-h}] = \sqrt{(t-h)/t}$*.*

Answer. [**End95**, p. 168]: $\operatorname{cov}[\mathfrak{y}_t, \mathfrak{y}_{t-h}] = \operatorname{cov}[e_1 + \cdots + e_t, e_1 + \cdots + e_{t-h}$. $\operatorname{corr}[\mathfrak{y}_t, \mathfrak{y}_{t-h}] = \frac{\sigma^2(t-h)}{\sqrt{\sigma^2 t}\sqrt{\sigma^2(t-h)}}$. □

The significance of this last formula is: the autocorrelation functions of a nonstationary random walk look similar to those of an autoregressive stationary process.

Then Enders discusses some variations: random walk plus drift $\mathfrak{y}_t = \mathfrak{y}_{t-1} + \mu + \varepsilon_t$ which is Enders's (3.36), random walk plus noise $\mathfrak{y}_t = \mu_t + \eta_t$ where $\mu_t = \mu_{t-1} + \varepsilon_t$ with η_t and ε_t independent white noise processes is Enders's (3.38–39). Both can be combined in (3.41), and the so-called local linear trend model (3.45).

How to remove the trend? Random walk (with or without drift) is ARIMA(0,1,0), i.e., its first difference is a constant plus white noise.

The random walk with noise (or with drift and noise) is ARIMA(0,1,1):

Problem 554. *3 points Show: If you difference a random walk with noise process, you get a MA(1) process with a correlation that is between 0 and* $-1/2$*.*

Answer. Let $\mathfrak{y}_t$ be a random walk with noise, i.e., $\mathfrak{y}_t = \mu_t + \eta_t$ where $\mu_t = \mu_{t-1} + \varepsilon_t$ with η_t and ε_t independent white noise processes. Since $\Delta\mu_t = \varepsilon_t$, it follows $\Delta\mathfrak{y}_t = \varepsilon_t + \eta_t - \eta_{t-1}$. Stationary. $\operatorname{var}[\Delta\mathfrak{y}_t] = \sigma_\varepsilon^2 + 2\sigma_\eta^2$. $\operatorname{cov}[\Delta\mathfrak{y}_t, \Delta\mathfrak{y}_{t-1}] = \operatorname{cov}[\varepsilon_t + \eta_t - \eta_{t-1}, \varepsilon_{t-1} + \eta_{t-1} - \eta_{t-2}] = -\sigma_\eta^2$. $\operatorname{corr}[\Delta\mathfrak{y}_t, \Delta\mathfrak{y}_{t-1}] = -\sigma_\eta^2/(\sigma_\varepsilon^2 + 2\sigma_\eta^2)$ between $-1/2$ and 0. Higher covariances are zero. □

The local linear trend model is an example of a model which leads to a stationary process after differencing twice: it is an ARIMA(0,2,2) model.

I.e., certain time series are such that differencing is the right thing to do. But if a time series is the sum of a deterministic trend and white noise then differencing is not called for: From $\mathfrak{y}_t = \mathfrak{y}_0 + \alpha t + \varepsilon_t$ follows $\Delta\mathfrak{y}_t = \alpha + \varepsilon_t - \varepsilon_{t-1}$. This is not an invertible process. The appropriate method of detrending here is to regress the timeseries on t and take the residuals.

Difference stationary (DS) models can be made stationary by differencing, and trend stationary models (TS) can be made stationary by removing deterministic time trend.

Nelson and Plosser (1982) found evidence that, contrary to common wisdom, many macroeconomic timeseries are DS instead of TS. Therefore the question arises: how can we test for this? The obvious way would be to regress $\Delta\mathfrak{y}$ on $\mathfrak{y}_{t-1}$ and to see whether the coefficient is zero. Usually, one of the following three regressions is run:

$$\Delta\mathfrak{y}_1 = (\alpha - 1)\mathfrak{y}_{t-1} + u_t \tag{67.3.4}$$

$$\Delta\mathfrak{y}_1 = \beta_0 + (\alpha - 1)\mathfrak{y}_{t-1} + u_t \tag{67.3.5}$$

$$\Delta\mathfrak{y}_1 = \beta_0 + \beta_1 t + (\alpha - 1)\mathfrak{y}_{t-1} + u_t \tag{67.3.6}$$

and one tests whether $\alpha - 1 = 0$. But one cannot simply make an ordinary t-test because in the case of a random walk the t-statistic has different asymptotic properties.

The t-statistic in such a spurious regression rejects the null hypothesis of no relationship far too often. And if one increases the sample size, one gets even more spurious rejections. Asymptotically, this $\mathfrak{t}$ statistic will always be rejected.

In order to get an idea why this is so, look at a situation in which two independent random walks are regressed on each other. [**End95**, p. 218] shows the scatter diagram, which suggests a significant relationship (but the plot of the residuals shows nonstationarity).

There are two ways around this, both connected with the names Dickey and Fuller (DF-tests): Either one maintains the usual formula for the $\mathfrak{t}$-statistic but different significance points which were obtained by Monte-Carlo experiments. These are the so-called τ-tests. Based on the three above regressions [**DM93**, p. 703] calls them τ_{nc}, τ_c, and τ_{ct} (like: no constant, constant, and constant and trend). Or one uses a different test statistic in which one divides by T instead of $\sqrt{T}$ (and it turns out that one does not have to divide by the estimated standard deviation). This is the so-called z-statistic.

67.4. Cointegration

Two timeseries $\mathfrak{y}_0$ and $\mathfrak{y}_1$ which are $I(1)$ are called co-integrated if there is a linear combination of them which is $I(0)$. What this means is especially obvious if this linear combination is their difference, see the graphs in [**CD97**, pp, 123/4]. Usually in economic applications this linear combination also depends on exogenous variables; then the definition is that $\eta_1\mathfrak{y}_1 + \eta_2\mathfrak{y}_2 = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ whith a stationary $\boldsymbol{\varepsilon}$. These coefficients are determined only up to a multiplicative constant, therefore one can normalize them by setting say $\eta_1 = 1$.

How to estimate cointegration? The simplest way is to regress $\mathfrak{y}_1$ on $\mathfrak{y}_2$ and $\boldsymbol{X}$ (if the normalization is such that $\eta_1 = 1$). If there is no cointegration, this gives a spurious regression with nonstationary residuals–from which follows that one tests

for cointegration by stationarity of the residuals. Now if there is cointegration, the cointegrating relationship is stronger than the spurious regression effect.

Since cointegrated variables are usually jointly determined, there will be correlation between the error term and the regressor $\mathfrak{y}_2$ in the above regression. However the coefficient estimate itself is super-consistent, i.e., instead of approaching the true value at a rate of $t^{-1/2}$ it approaches them at a rate of t^{-1}. Therefore the correlation of the error terms with the price, which is only of the order $t^{-1/2}$, cannot make this extimator inconsistent.

Problem 555. *Assume the model (with time series data) can be written in the form $\mathfrak{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$, but the data were dynamically generated in one of the following ways. Discuss what you would do in these cases.*

• a. *2 points $\mathfrak{y}$ depends on various lags of the explanatory variable,*

Answer. Distributed lags, first estimate lag length, then due to multicollinearity perhaps polynomial distributed lags, estimate degree of polynomial. □

• b. *2 points $\mathfrak{y}$ depends on its own lagged values and some other explanatory variables.*

Answer. Do OLS, only a problem if the errors are also autocorrelated. □

• c. *2 points The error terms of different time periods are correlated (but, for simplicity, they are assumed to form a stationary process).*

• d. *2 points What should be considered when more than one of the above situations occur?*

CHAPTER 68

Seasonal Adjustment

Causes of seasonal movement:

- Weather: temperature, precipitation, hours of sunshine
- Calendar events: timing of religious or secular festivals (Christmas, July 4th)
- Timing decisions: school vacations, tax years, dates for dividend payments.

Seasonal adjustment is necessary because economic data are subject to different kinds of influences: seasonal factors (which are assumed to be beyond the policymaker's reach) and economic factors. Seasonally adjusted data are counterfactual data: they are an attempt to reconstruct what the economy would have been in the absence of the seasonal influences.

Seasonal adjustment has been criticized in [**Hyl92**, p. 231] on the grounds that it cannot be explained what the adjusted series is measuring. Signal extraction in electrical engineering has the goal to restore the original signal which actually existed before it was degraded by noise. But there is no actually existing "original signal" which the de-seasonalized economic timeseries tries to measure. Someone concluded from this, I have to find the quote and the exact wording again, "These adjusted timeseries must be considered uncomplicated aids in decision making, without a real counterpart." Look at [**Hyl92**, p. 102]. Here it is necessary to make the depth-realist distinction between the real and the actual. It is true that seasonally adjusted data have no *actual* counterpart; they are counterfactual, but they do have a *real* basis, namely, the underlying economic mechanisms which would also have been active in the absence of the seasonal factors.

Natural scientists can investigate their subject under controlled experimental conditions, shielded from non-essential influences. Economists cannot do this; they cannot run the economy inside a building in which all seasonal variations of weather and scheduling are eliminated, in order to see how the economy evolves in the absence of these disturbances and therefore to understand the character of economic mechanisms better.

Seasonal adjustment of the data is an imperfect substitute for this. It exploits the fact that phenomena which are generated by seasonal factors have a different

empirical footprint than those generated by other factors, namely, their periodicity, which is a very obvious feature of most economic timeseries. The removal of the periodicity from the data is their attempt to infer what the economy would have been like in the absence of the seasonal influences.

This is in principle no different than some of the methods by which we try to eliminate other non-economic influences from the data: many statistical methods make the the assumption that *fast* variations in the data are the result of random, i.e., non-economic influences, because the economy does not move this fast.

These limitations of seasonal adjustment point to a basic methodological flaw of this research method. The attempt to take data generated by an economy which is subject to seasonal influences and submit them to a mathematical procedure in order to see how the economy would have evolved in the absence of these influences really commits the "fallacy of misplaced concreteness" [**Col89**, pp. 27?, 52?]: if two different mechanism are at work, this does not mean that the events generated by them can be divided into two groups, or that the data generated by these mechanisms can be decomposed into two components: that generated by the first mechanism, and that generated by the second. This is why it is recommended so often that the seasonality should be incorporated in the *model* instead of adjusting the data. (In some simple cases, as in Problem 559, these two procedures are equivalent, but usually these two methods give different results.)

Seasonal adjustment as a scientific method encounters its limits whenever there are more than negligible interactions between seasonal and economic mechanisms:

- Leakages from seasonality to economics: The hot summers in the 1970s in Denmark caused a lot of investment into irrigation systems which then greatly changed agricultural technology. [**Hyl92**, which page?]
- Seasonality altering economic interactions: The building boom in Denmark in the 1970s caused seasonal labor shortages which gave rise to a change in construction technology. [**Hyl92**, which page?]

Furthermore, [**Hyl92**, chapter 6, need reference by author and title] shows by a theoretical model that the empirical expressions of seasonal influences do not necessarily move in synch with the seasons: optimal adjustment to seasonal demand leads to a seasonally-induced component of economic activity which has its power not restricted to the seasonal frequencies

Miron [**Mir96**, pp, 57–66] does not look at the frequency but at the amplitude of the seasonal variations. He argues that the seasonal variations observed in the economy are much stronger than the magnitude of the above external influences might justify. He concludes that there is a "seasonal business cycle" which shares many characteristics with the usual business cycle.

Despite these flaws, seasonal adjustment can have its uses.

Since seasonal adjustment has to do with the interaction between noneconomic and certain economic mechanisms, there can be no a priori theory about what is the right way to do seasonal adjustments; this question must be decided by experience. But from the success of some and the failure of other methods one should be able to make second-order inferences about the general character of the economy.

[**Mir96**, p. 10–13] did some work in this direction: he extracted the seasonal component by one of the statistical methods, and then looked whether this seasonal component was stable. Results: seasonal effects are not different in cyclical downturns than upturns, they are also not different when production is high than when production is low, and the turning points also do not have a statistically different pattern. There is a difference between the first and the second half of the post-World-War II period, but it is small compared with the seasonal effects themselves. Perhaps the changed character of ther seasonal rhythm of the economy is an indicator of other important changes in the structure of the economy?

[**Hyl92**, p. 105] cites several articles which say that one can get better forecasts from unadjusted data. In the face of this "loss of information" due to seasonal adjustment it is argued that one should *not* seasonally adjust the data. My answer: prediction is not the purpose of seasonal adjustment. One can predict quite well without knowing the underlying mechanisms. There is a difference between prediction and the exploration of underlying mechanisms.

[**BF91**] apply various adjustment mechanisms to real (and one simulated) time-series and ask whether the results have desirable properties. For instance, the X11-procedure was apparently adopted because it gave good results on many different time series. This shows that seasonal adjustment methods are selected not so much on the basis of prior theory, but on the basis of what works.

68.1. Methods of Seasonal Adjustment

The following is only a very small sampling of the methods in use:

Fixed additive method according to [**BF91**, p. 40/1]:

$$y_{ij} = g_{ij} + s_i + u_{ij} \tag{68.1.1}$$

(trend-cycle, seasonal, and irregular components). Here the trend-cycle component g_{ij} is estimated by the centered 12-month moving average of y_{ij}, i.e., 11 months have a weight of 1/12 and the two months at both ends of the relevant period have a weight of 1/24. For each month the mean of the difference between the trend-cycle component thus estimated and the original series $y_{ij} - g_{ij}$ is determined. This yields the preliminary estimate of the seasonal component s'_i. In order to let the seasonal components sum to zero over the year, they are derived from these preliminary

components as follows:

$$s_i = s'_i - \frac{1}{12}\sum_{j-1}^{1} 2s'_j \tag{68.1.2}$$

The irregular component is then the residual.

Problem 556. *How would you modify this method at the ends of the sampling period?*

[**ESS97**, p. 39 and 47–49] give the following prescription for a time-varying seasonal adjustment: $x_t = g_t + s_t + u_t$; here g_t is trend plus cyclical component; s_t is the seasonal component, and u_t the residual. We want g_t to be smooth in the sense of small third differences: $v_t = g_t - 3g_{t-1} + 3g_{t-2} - g_{t-3}$, and we want s_t to sum up to zero over the year (which is split into m periods) i.e., $w_t = \sum_{i=0}^{m-1} s_{t-i}$ must be minimized, and the squared sum of the errors must be minimized. i.e., one mimimizes

$$\sum_{t=1}^{n} u_t^2 + \alpha_1 \sum_{t=4}^{n} v_t^2 + \alpha_2 \sum_{t=m}^{n} w_t^2 \tag{68.1.3}$$

Here a large α_1 gives a smoother series, a small α_1 gives smaller residuals. A large α_2 gives a series whose seasonal behavior is fixed over time, while a small α_2 gives a more flexible seasonal pattern.

Problem 557. *Show that v_t is indeed the third difference.*

Answer. Define the first differences $d_t = g_t - g_{t-1}$, the second differences $e_t = d_t - d_{t-1}$, and the third difference $f_t = e_t - e_{t-1}$. Then you will see that $f_t = v_t$. Let's go through this: $e_t = g_t - g_{t-1} - (g_{t-1} - g_{t-2}) = g_t - 2g_{t-1} + g_{t-2}$. Therefore $f_t = g_t - 2g_{t-1} + g_{t-2} - (g_{t-1} - 2g_{t-2} + g_{t-3}) = g_t - 3g_{t-1} + 3g_{t-2} - g_{t-3} = v_t$. □

BTW the smooth component g_t is not being produced in the USA, but in Europe [**ESS97**, p. 98], this book has a few articles arguing that the smooth component is good for policy etc.

68.2. Seasonal Dummies in a Regression

The following is a more technical discussion, using the math of linear regression:

Problem 558. *Regression models incorporate seasonality often by the assumption that the intercept of the regression is different in every season, while the slopes remain the same. Assuming $\boldsymbol{X}$ contains quarterly data (but the constant term is not*

incorporated in $\boldsymbol{X}$), this can be achieved in several different ways: You may write your model as

$$\text{(68.2.1)} \qquad \mathbf{y} = \boldsymbol{X\beta} + \boldsymbol{C\delta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{C} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}.$$

Alternatively, you may write your model in the form

$$\text{(68.2.2)} \qquad \mathbf{y} = \boldsymbol{\iota}\alpha + \boldsymbol{X\beta} + \boldsymbol{K\delta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\iota} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ \vdots \end{bmatrix} \qquad \boldsymbol{K} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \end{bmatrix}$$

In R *this is the default method to generate dummy variables from a seasonal factor variable. (`Splus` has a different default.) This is also the procedure shown in* [**Gre97**, p. 383]. *But the following third alternative is often preferrable:*

$$\text{(68.2.3)} \qquad \mathbf{y} = \boldsymbol{\iota}\alpha + \boldsymbol{X\beta} + \boldsymbol{K\delta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\iota} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ \vdots \end{bmatrix} \qquad \boldsymbol{K} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & -1 & -1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & -1 & -1 \\ \vdots & \vdots & \vdots \end{bmatrix}$$

In R *one gets these dummy variables from a seasonal factor variable if one specifies* `contrast="contr.sum"`.

3 points What is the meaning of the seasonal dummies δ_1, δ_2, δ_3, and of the constant term α or the fourth seasonal dummy δ_4, in models (68.2.1), (68.2.2), *and* (68.2.3)*?*

Answer. Clearly, in model (68.2.1), δ_i is the intercept in the ith season. For (68.2.2) and (68.2.3), it is best to write the regression equation for each season separately, filling in the values the dummies take for these seasons, in order to see the meaning of these dummies. Assuming $\boldsymbol{X}$ consists of one column only, (68.2.2) becomes

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ \vdots \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ \vdots \end{bmatrix} \alpha + \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \\ x_8 \\ \vdots \end{bmatrix} \beta + \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \end{bmatrix} \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \vdots \end{bmatrix}$$

or, written element by element

$$\begin{aligned} y_1 &= 1 \cdot \alpha + x_1 \cdot \beta + 0 \cdot \delta_1 + 0 \cdot \delta_2 + 0 \cdot \delta_3 + \varepsilon_1 && \text{winter} \\ y_2 &= 1 \cdot \alpha + x_2 \cdot \beta + 1 \cdot \delta_1 + 0 \cdot \delta_2 + 0 \cdot \delta_3 + \varepsilon_2 && \text{spring} \\ y_3 &= 1 \cdot \alpha + x_3 \cdot \beta + 0 \cdot \delta_1 + 1 \cdot \delta_2 + 0 \cdot \delta_3 + \varepsilon_3 && \text{summer} \\ y_4 &= 1 \cdot \alpha + x_4 \cdot \beta + 0 \cdot \delta_1 + 0 \cdot \delta_2 + 1 \cdot \delta_3 + \varepsilon_4 && \text{autumn} \end{aligned}$$

therefore the overall intercept α is the intercept of the first quarter (winter); δ_1 is the difference between the spring intercept and the winter intercept, etc.

(68.2.3) becomes

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ \vdots \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ \vdots \end{bmatrix} \alpha + \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \\ x_8 \\ \vdots \end{bmatrix} \beta + \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & -1 & -1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & -1 & -1 \\ \vdots & \vdots & \vdots \end{bmatrix} \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \vdots \end{bmatrix}$$

or, written element by element

$$\begin{aligned} y_1 &= 1 \cdot \alpha + x_1 \cdot \beta + 1 \cdot \delta_1 + 0 \cdot \delta_2 + 0 \cdot \delta_3 + \varepsilon_1 && \text{winter} \\ y_2 &= 1 \cdot \alpha + x_2 \cdot \beta + 0 \cdot \delta_1 + 1 \cdot \delta_2 + 0 \cdot \delta_3 + \varepsilon_2 && \text{spring} \\ y_3 &= 1 \cdot \alpha + x_3 \cdot \beta + 0 \cdot \delta_1 + 0 \cdot \delta_2 + 1 \cdot \delta_3 + \varepsilon_3 && \text{summer} \\ y_4 &= 1 \cdot \alpha + x_4 \cdot \beta - 1 \cdot \delta_1 - 1 \cdot \delta_2 - 1 \cdot \delta_3 + \varepsilon_4 && \text{autumn} \end{aligned}$$

Here the winter intercept is $\alpha + \delta_1$, the spring intercept $\alpha + \delta_2$, summer $\alpha + \delta_3$, and autmumn $\alpha - \delta_1 - \delta_2 - \delta_3$. Summing this and dividing by 4 shows that the constant term α is the arithmetic

mean of all intercepts, therefore δ_1 is the difference between the winter intercept and the arithmetic mean of all intercepts, etc. □

PROBLEM 559. [**DM93**, pp. 23/4], [**JGH+85**, p. 260]. *Your dependent variable* $\mathbf{y}$ *and the explanatory variables* $\boldsymbol{X}$ *are quarterly timeseries data. Your regression includes a constant term (not included in* $\boldsymbol{X}$*). We also assume that your data set spans* m *full years, i.e., the number of observations is* $4m$*. The purpose of this exercise is to show that the following two procedures are equivalent:*

• a. *1 point You create a "seasonally adjusted" version of your data set, call them* $\underline{\mathbf{y}}$ *and* $\underline{\boldsymbol{X}}$*, by taking the seasonal mean out of every variable and adding the overall mean back, and you regress* $\underline{\mathbf{y}}$ *on* $\underline{\boldsymbol{X}}$ *with a constant term. (The underlining does not denote taking out of the mean, but the taking out of the seasonal means and adding back of the overall mean.) In the simple example where* $\mathbf{y} = \begin{bmatrix}1 & 3 & 8 & 4 & 5 & 3 & 2 & 6\end{bmatrix}^\top$*, compute* $\underline{\mathbf{y}}$*. Hint: the solution vector contains the numbers 7,3,6,4 in sequence.*

ANSWER. Subtract the seasonal means, and add back the overall mean to get:

$$\begin{bmatrix}1\\3\\8\\4\\5\\3\\2\\6\end{bmatrix} - \begin{bmatrix}3\\3\\5\\5\\3\\3\\5\\5\end{bmatrix} + \begin{bmatrix}4\\4\\4\\4\\4\\4\\4\\4\end{bmatrix} = \begin{bmatrix}2\\4\\7\\3\\6\\4\\1\\5\end{bmatrix} \tag{68.2.4}$$

□

• b. *2 points The alternative equivalent procedure is: You use the original data* $\mathbf{y}$ *and* $\boldsymbol{X}$ *but you add three seasonal dummies to your regression, i.e., you write your model in the form*

$$\mathbf{y} = \boldsymbol{\iota}\alpha + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{K}\boldsymbol{\delta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{K} = \begin{bmatrix}1 & 0 & 0\\0 & 1 & 0\\0 & 0 & 1\\-1 & -1 & -1\\1 & 0 & 0\\0 & 1 & 0\\0 & 0 & 1\\-1 & -1 & -1\\\vdots & \vdots & \vdots\end{bmatrix} \tag{68.2.5}$$

Explain the meaning of the three dummy coefficients in this regression.

ANSWER. See the answer to question 558 □

• c. *2 points Show that*

$$(\boldsymbol{K}^\top\boldsymbol{K})^{-1} = \frac{1}{4m}\begin{bmatrix}3 & -1 & -1\\-1 & 3 & -1\\-1 & -1 & 3\end{bmatrix} = \frac{1}{m}(\boldsymbol{I} - \frac{1}{4}\boldsymbol{\iota}\boldsymbol{\iota}^\top) \tag{68.2.6}$$

where $\boldsymbol{I}$ *is the* 3×3 *identity matrix, and* $\boldsymbol{\iota}$ *is a 3-vector of ones. Hint: this can easily be computed element by element, but the most elegant way is to write* $\boldsymbol{K}^\top = \begin{bmatrix}\boldsymbol{I} & -\boldsymbol{\iota} & \dots & \boldsymbol{I} & -\boldsymbol{\iota}\end{bmatrix}$ *where the* $\begin{bmatrix}\boldsymbol{I} & -\boldsymbol{\iota}\end{bmatrix}$ *group is repeated* m *times.*

ANSWER. $\boldsymbol{K}^\top\boldsymbol{K}$ before taking the inverses is

$$\boldsymbol{K}^\top\boldsymbol{K} = m\begin{bmatrix}2 & 1 & 1\\1 & 2 & 1\\1 & 1 & 2\end{bmatrix} = m(\boldsymbol{I} + \boldsymbol{\iota}\boldsymbol{\iota}^\top). \tag{68.2.7}$$

If it is written in the second way one can apply formula (31.2.6) to get the inverse. Of course, in the present case, the inverse is already given, therefore one can simply multiply the matrix with its inverse to verify that it is indeed the inverse. □

• d. *3 points Show that*

$$\boldsymbol{K}(\boldsymbol{K}^\top\boldsymbol{K})^{-1}\boldsymbol{K}^\top = \frac{1}{m}\begin{bmatrix}1&0&0&0&1&0&0&0\\0&1&0&0&0&1&0&0\\0&0&1&0&0&0&1&0\\0&0&0&1&0&0&0&1\\1&0&0&0&1&0&0&0\\0&1&0&0&0&1&0&0\\0&0&1&0&0&0&1&0\\0&0&0&1&0&0&0&1\end{bmatrix} - \frac{1}{4m}\begin{bmatrix}1&1&1&1&1&1&1&1\\1&1&1&1&1&1&1&1\\1&1&1&1&1&1&1&1\\1&1&1&1&1&1&1&1\\1&1&1&1&1&1&1&1\\1&1&1&1&1&1&1&1\\1&1&1&1&1&1&1&1\\1&1&1&1&1&1&1&1\end{bmatrix} \tag{68.2.8}$$

or, in a different notation

$$\boldsymbol{K}(\boldsymbol{K}^\top\boldsymbol{K})^{-1}\boldsymbol{K}^\top = \frac{1}{m}\begin{bmatrix}\boldsymbol{I} & \boldsymbol{o} & \boldsymbol{I} & \boldsymbol{o}\\\boldsymbol{o}^\top & 1 & \boldsymbol{o}^\top & 1\\\boldsymbol{I} & \boldsymbol{o} & \boldsymbol{I} & \boldsymbol{o}\\\boldsymbol{o}^\top & 1 & \boldsymbol{o}^\top & 1\end{bmatrix} - \frac{1}{4m}\begin{bmatrix}\boldsymbol{\iota}\boldsymbol{\iota}^\top & \boldsymbol{\iota} & \boldsymbol{\iota}\boldsymbol{\iota}^\top & \boldsymbol{\iota}\\\boldsymbol{\iota}^\top & 1 & \boldsymbol{\iota}^\top & 1\\\boldsymbol{\iota}\boldsymbol{\iota}^\top & \boldsymbol{\iota} & \boldsymbol{\iota}\boldsymbol{\iota}^\top & \boldsymbol{\iota}\\\boldsymbol{\iota}^\top & 1 & \boldsymbol{\iota}^\top & 1\end{bmatrix} \tag{68.2.9}$$

ANSWER. $\boldsymbol{K}(\boldsymbol{K}^\top\boldsymbol{K})^{-1}\boldsymbol{K}^\top = \frac{1}{m}(\boldsymbol{K}\boldsymbol{K}^\top - \frac{1}{4}\boldsymbol{K}\boldsymbol{\iota}\boldsymbol{\iota}^\top\boldsymbol{K}^\top)$. Since $\boldsymbol{K}$ is peridodic with period 2, we only need the 4 upper left partitions.

$$m\boldsymbol{K}(\boldsymbol{K}^\top\boldsymbol{K})^{-1}\boldsymbol{K}^\top = \begin{bmatrix}\boldsymbol{I} & -\boldsymbol{\iota}\\\boldsymbol{\iota}^\top & 3\end{bmatrix} - \frac{1}{4}\begin{bmatrix}\boldsymbol{\iota}\boldsymbol{\iota}^\top & -3\boldsymbol{\iota}\\-3\boldsymbol{\iota}^\top & 9\end{bmatrix} = \begin{bmatrix}\boldsymbol{I} - \frac{1}{4}\boldsymbol{\iota}\boldsymbol{\iota}^\top & -\frac{1}{4}\boldsymbol{\iota}\\-\frac{1}{4}\boldsymbol{\iota}^\top & \frac{3}{4}\end{bmatrix} = \boldsymbol{I}_4 - \frac{1}{4}\boldsymbol{\iota}_4\boldsymbol{\iota}_4^\top \tag{68.2.10}$$

□

• e. *2 points Using the above equations, show that the OLS estimate* $\hat{\boldsymbol{\beta}}$ *in this model is exactly the same as the OLS estimate in the regression of the seasonally*

adjusted data $\underline{\mathfrak{y}}$ *on* $\underline{\boldsymbol{X}}$. *Hint: All you have to show is that* $\boldsymbol{M}_1\mathfrak{y} = \underline{\mathfrak{y}}$, *and* $\boldsymbol{M}_1\boldsymbol{X} = \underline{\boldsymbol{X}}$, *where* $\boldsymbol{M}_1 = \boldsymbol{I} - \boldsymbol{K}(\boldsymbol{K}^\top\boldsymbol{K})^{-1}\boldsymbol{K}^\top$.

Answer. This gives

$$\boldsymbol{I} - \boldsymbol{K}(\boldsymbol{K}^\top\boldsymbol{K})^{-1}\boldsymbol{K}^\top = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} - \tag{68.2.11}$$

$$- \frac{1}{m}\begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix} + \frac{1}{4m}\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

The $\boldsymbol{I}_4$-striping takes the seasonal means out, and the $\frac{1}{4m}\boldsymbol{\iota}\boldsymbol{\iota}^\top$ adds the overall mean back.

□

CHAPTER 69

Binary Choice Models

69.1. Fisher's Scoring and Iteratively Reweighted Least Squares

This section draws on chapter 55 about Numerical Minimization. Another important "natural" choice for the positive definite matrix $\boldsymbol{R}_i$ in the gradient method is available if one maximizes a likelihood function: then $\boldsymbol{R}_i$ can be the inverse of the information matrix for the parameter values $\boldsymbol{\beta}_i$. This is called Fisher's Scoring method. It is closely related to the Newton-Raphson method. The Newton-Raphson method uses the Hessian matrix, and the information matrix is minus the expected value of the Hessian. Apparently Fisher first used the information matrix as a computational simplification in the Newton-Raphson method. Today IRLS is used in the GLIM program for generalized linear models.

As in chapter 56 discussing nonlinear least squares, $\boldsymbol{\beta}$ is the vector of parameters of interest, and we will work with an intermediate vector $\boldsymbol{\eta}(\boldsymbol{\beta})$ of predictors whose dimension is comparable to that of the observations. Therefore the likelihood function has the form $\mathsf{L} = L\big(\mathbf{y}, \boldsymbol{\eta}(\boldsymbol{\beta})\big)$. By the chain rule (C.1.23) one can write the Jacobian of the likelihood function as $\frac{\partial \mathsf{L}}{\partial \boldsymbol{\beta}^\top}(\boldsymbol{\beta}) = \mathbf{u}^\top \boldsymbol{X}$, where $\mathbf{u}^\top = \frac{\partial \mathsf{L}}{\partial \boldsymbol{\eta}^\top}(\boldsymbol{\eta}(\boldsymbol{\beta}))$ is the Jacobian of L as a function of $\boldsymbol{\eta}$, evaluated at $\boldsymbol{\eta}(\boldsymbol{\beta})$, and $\boldsymbol{X} = \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\beta}^\top}(\boldsymbol{\beta})$ is the Jacobian of $\boldsymbol{\eta}$. This is the same notation as in the discussion of the Gauss-Newton regression.

Define $\boldsymbol{A} = \mathcal{E}[\mathbf{u}\mathbf{u}^\top]$. Since $\boldsymbol{X}$ does not depend on the random variables, the information matrix of $\mathbf{y}$ with respect to $\boldsymbol{\beta}$ is then $\mathcal{E}[\boldsymbol{X}^\top \mathbf{u}\mathbf{u}^\top \boldsymbol{X}] = \boldsymbol{X}^\top \boldsymbol{A} \boldsymbol{X}$. If one uses the inverse of this information matrix as the $\boldsymbol{R}$-matrix in the gradient algorithm, one gets

$$\boldsymbol{\beta}_{i+1} = \boldsymbol{\beta}_i + \alpha_i \big(\boldsymbol{X}^\top \boldsymbol{A} \boldsymbol{X}\big)^{-1} \boldsymbol{X}^\top \mathbf{u} \tag{69.1.1}$$

The Iterated Reweighted Least Squares interpretation of this comes from rewriting (69.1.1) as

$$\boldsymbol{\beta}_{i+1} = \boldsymbol{\beta}_i + \big(\boldsymbol{X}^\top \boldsymbol{A} \boldsymbol{X}\big)^{-1} \boldsymbol{X}^\top \boldsymbol{A} \boldsymbol{A}^{-1} \mathbf{u}, \tag{69.1.2}$$

i.e., one obtains the step by regressing $\boldsymbol{A}^{-1}\mathbf{u}$ on $\boldsymbol{X}$ with weighting matrix $\boldsymbol{A}$.

Justifications of IRLS are: the information matrix is usually analytically simpler than the Hessian of the likelihood function, therefore it is a convenient approximation, and one needs the information matrix anyway at the end for the covariance matrix of the M.L. estimators.

69.2. Binary Dependent Variable

Assume each individual in the sample makes an independent random choice between two alternatives, which can conveniently be coded as $\mathsf{y}_i = 0$ or 1. The probability distribution of y_i is fully determined by the probability $\pi_i = \Pr[\mathsf{y}_i = 1]$ of the event which has y_i as its indicator function. Then $\mathrm{E}[\mathsf{y}_i] = \pi_i$ and $\mathrm{var}[\mathsf{y}_i] = \mathrm{E}[\mathsf{y}_i^2] - \big(\mathrm{E}[\mathsf{y}_i]\big)^2 = \mathrm{E}[\mathsf{y}_i] - \big(\mathrm{E}[\mathsf{y}_i]\big)^2 = \pi_i(1-\pi_i)$.

It is usually assumed that the individual choices are stochastically independent of each other, i.e., the distribution of the data is fully characterized by the π_i. Each π_i is assumed to depend on a vector of explanatory variables $\boldsymbol{x}_i$. There are different approaches to modelling this dependence.

The regression model $\mathsf{y}_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i$ with $\mathrm{E}[\varepsilon_i] = 0$ is inappropriate because $\boldsymbol{x}_i^\top \boldsymbol{\beta}$ can take any value, whereas $0 \le \mathrm{E}[\mathsf{y}_i] \le 1$. Nevertheless, people have been tinkering with it. The obvious first tinker is based on the observation that the ε_i are no longer homoskedastic, but their variance, which is a function of π_i, can be estimated, therefore one can correct for this heteroskedasticity. But things get complicated very quickly and then the main appeal of OLS, its simplicity, is lost. This is a wrong-headed approach, and any smart ideas which one may get when going down this road are simply wasted.

The right way to do this is to set $\pi_i = \mathrm{E}[\mathsf{y}_i] = \Pr[\mathsf{y}_i = 1] = h(\boldsymbol{x}_i^\top \boldsymbol{\beta})$ where h is some (necessarily nonlinear) function with values between 0 and 1.

69.2.1. Logit Specification (Logistic Regression). The logit or logistic specification is $\pi_i = e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}/(1 + e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}})$. Invert to get $\log(\pi_i/(1-\pi_i)) = \boldsymbol{x}_i^\top \boldsymbol{\beta}$. I.e., the logarithm of the odds depends linearly on the predictors. The log odds are a natural re-scaling of probabilities to a scale which goes from $-\infty$ to $+\infty$, and which is symmetric in that the log odds of the complement of an event is just the negative of the log odds of the event itself. (See my remarks about the odds ratio in Question 222.)

Problem 560. *1 point If $y = \log \frac{p}{1-p}$ (logit function), show that $p = \frac{\exp y}{1+\exp y}$ (logistic function).*

Answer. $\exp y = \frac{p}{1-p}$, now multiply by $1-p$ to get $\exp y - p \exp y = p$, collect terms $\exp y = p(1+\exp y)$, now divide by $1+\exp y$. □

PROBLEM 561. *Sometimes one finds the following alternative specification of the logit model:* $\pi_i = 1/(1+e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}})$. *What is the difference between it and our formulation of the logit model? Are these two formulations equivalent?*

ANSWER. It is simply a different parametrization. They get this because they come from index number problem. □

The logit function is also the canonical link function for the binomial distribution, see Problem 113.

69.2.2. Probit Model. An important class of functions with values between 0 and 1 is the class of cumulative probability distribution functions. If h is a cumulative distribution function, then one can give this specification an interesting interpretation in terms of an unobserved "index variable."

The index variable model specifies: there is a variable z_i with the property that $\mathsf{y}_i = 1$ if and only if $\mathsf{z}_i > 0$. For instance, the decision y_i whether or not individual i moves to a different location can be modeled by the calculation whether the net benefit of moving, i.e., the wage differential minus the cost of relocation and finding a new job, is positive or not. This moving example is worked out, with references, in [**Gre93**, pp. 642/3].

The value of the variable z_i is not observed, one only observes y_i, i.e., the only thing one knows about the value of z_i is whether it is positive or not. But it is assumed that z_i is the sum of a deterministic part which is specific to the individual and a random part which has the same distribution for all individuals and is stochastically independent between different individuals. The deterministic part specific to the individual is assumed to depend linearly on individual i's values of the covariates, with coefficients which are common to all individuals. In other words, $\mathsf{z}_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i$, where the ε_i are i.i.d. with cumulative distribution function F_ε. Then it follows $\pi_i = \Pr[\mathsf{y}_i = 1] = \Pr[\mathsf{z}_i > 0] = \Pr[\varepsilon_i > -\boldsymbol{x}_i^\top \boldsymbol{\beta}] = 1 - \Pr[\varepsilon_i \leq -\boldsymbol{x}_i^\top \boldsymbol{\beta}] = 1 - F_\varepsilon(-\boldsymbol{x}_i^\top \boldsymbol{\beta})$. I.e., in this case, $h(\eta) = 1 - F_\varepsilon(-\eta)$. If the distribution of ε_i is symmetric and has a density, then one gets the simpler formula $h(\eta) = F_\varepsilon(\eta)$.

Which cumulative distribution function should be chosen?

- In practice, the probit model, in which z_i is normal, is the only one used.
- The linear model, in which h is the line segment from $(a, 0)$ to $(b, 1)$, can also be considered generated by an in index function z_i which is here uniformly distributed.
- An alternative possible specification with the Cauchy distribution is proposed in [**DM93**, p. 516]. They say that curiously only logit and probit are being used.

In practice, the probit model is very similar to the logit model, once one has rescaled the variables to make the variances equal, but the logit model is easier to handle mathematically.

69.2.3. Replicated Data. Before discussing estimation methods I want to briefly address the issue whether or not to write the data in replicated form [**MN89**, p. 99–101]. If there are several observations for every individual, or if there are several individuals for the same values of the covariates (which can happen if all covariates are categorical), then one can write the data more compactly if one groups the data into so-called "covariate classes," i.e., groups of observations which share the same values of $\boldsymbol{x}_i$, and defines y_i to be the number of times the decision came out positive in this group. Then one needs a second variable, m_i, which is assumed nonrandom, indicating how many individual decisions are combined in the respective group. This is an equivalent formulation of the data, the only thing one loses is the order in which the observations were made (which may be relevant if there are training or warm-up effects). The original representation of the data is a special case of the grouped form: in the non-grouped form, all $m_i = 1$. We will from now on write our formulas for the grouped form.

69.2.4. Estimation. Maximum likelihood is the preferred estimation method. The likelihood function has the form $\mathcal{L} = \prod \pi_i^{\mathsf{y}_i}(1-\pi_i)^{(m_i - \mathsf{y}_i)}$. This likelihood function is not derived from a density, but from a probability mass function. For instance, in the case with non-replicated data, all $m_i = 1$, if you have n binary measurements, then you can have only 2^n different outcomes, and the probability of the sequence $y_1, \dots y_n = 0, 1, 0, 0, \dots, 1$ is as given above.

This is a highly nonlinear maximization and must be done numerically. Let us go through the method of scoring in the example of a logit distribution.

$$L = \sum_i \Big(\mathsf{y}_i \log \pi_i + (m_i - \mathsf{y}_i)\log(1-\pi_i)\Big) \tag{69.2.1}$$

$$\frac{\partial L}{\partial \pi_i} = \Big(\frac{\mathsf{y}_i}{\pi_i} - \frac{m_i - \mathsf{y}_i}{1-\pi_i}\Big) \tag{69.2.2}$$

$$\frac{\partial^2 L}{\partial \pi_i^2} = -\Big(\frac{\mathsf{y}_i}{\pi_i^2} + \frac{m_i - \mathsf{y}_i}{(1-\pi_i)^2}\Big) \tag{69.2.3}$$

Defining $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$, the logit specification can be written as $\pi_i = e^{\eta_i}/(1+e^{\eta_i})$. Differentiation gives $\frac{\partial \pi_i}{\partial \eta_i} = \pi_i(1-\pi_i)$. Combine this with (69.2.2) to get

$$u_i = \frac{\partial L}{\partial \eta_i} = \Big(\frac{\mathsf{y}_i}{\pi_i} - \frac{m_i - \mathsf{y}_i}{1-\pi_i}\Big)\pi_i(1-\pi_i) = \mathsf{y}_i - m_i\pi_i. \tag{69.2.4}$$

These are the elements of $\boldsymbol{u}$ in (69.1.1), and they have a very simple meaning: it is just the observations minus their expected values. Therefore one obtains immediately $\boldsymbol{A} = \mathcal{E}[\boldsymbol{u}\boldsymbol{u}^\top]$ is a diagonal matrix with $m_i\pi_i(1-\pi_i)$ in the diagonal.

PROBLEM 562. *6 points Show that for the maximization of the likelihood function of the logit model, Fisher's scoring method is equivalent to the Newton-Raphson algorithm.*

PROBLEM 563. *Show that in the logistic model,* $\sum m_i\hat{\pi}_i = \sum \mathtt{y}_i$.

69.3. The Generalized Linear Model

The binary choice models show how the linear model can be generalized. [**MN89**, p. 27–32] develop a unified theory of many different interesting models, called the "generalized linear model." The following few paragraphs are indebted to the elaborate and useful web site about Generalized Linear Models maintained by Gordon K. Smyth at `www.maths.uq.oz.au/~gks/research/glm`

In which cases is it necessary to go beyond linear models? The most important and common situation is one in which $\mathtt{y}_i$ and $\mu_i = \mathrm{E}[\mathtt{y}_i]$ are bounded:

- If $\mathtt{y}$ represents the amount of some physical substance then we may have $\mathtt{y} \geq 0$ and $\mu \geq 0$.
- If $\mathtt{y}$ is binary, i.e., $\mathtt{y} = 1$ if an animal survives and $\mathtt{y} = 0$ if it does not, then $0 \leq \mu \leq 1$.

The linear model is inadequate here because complicated and unnatural constraints on $\boldsymbol{\beta}$ would be required to make sure that μ stays in the feasible range. Generalized linear models instead assume a link linear relationship

$$g(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta} \tag{69.3.1}$$

where $g()$ is some known monotonic function which acts pointwise on $\boldsymbol{\mu}$. Typically $g()$ is used to transform the μ_i to a scale on which they are unconstrained. For example we might use $g(\mu) = \log(\mu)$ if $\mu_i > 0$ or $g(\mu) = \log\big(\mu/(1-\mu)\big)$ if $0 < \mu_i < 1$.

The same reasons which force us to abandon the linear model also force us to abandon the assumption of normality. If $\mathtt{y}$ is bounded then the variance of $\mathtt{y}$ must depend on its mean. Specifically if μ is close to a boundary for $\mathtt{y}$ then $\mathrm{var}(\mathtt{y})$ must be small. For example, if $\mathtt{y} > 0$, then we must have $\mathrm{var}(\mathtt{y}) \to 0$ as $\mu \to 0$. For this reason strictly positive data almost always shows increasing variability with increased size. If $0 < \mathtt{y} < 1$, then $\mathrm{var}(\mathtt{y}) \to 0$ as $\mu \to 0$ or $\mu \to 1$. For this reason, generalized linear models assume that

$$\mathrm{var}(\mathtt{y}_i) = \phi \cdot V(\mu_i) \tag{69.3.2}$$

where ϕ is an unknown scale factor and $V()$ is some known variance function appropriate for the data at hand.

We therefore estimate the nonlinear regression equation (69.3.1) weighting the observations inversely according to the variance functions $V(\mu_i)$. This weighting procedure turns out to be exactly equivalent to maximum likelihood estimation when the observations actually come from an exponential family distribution.

PROBLEM 564. *Describe estimation situations in which a linear model and Normal distribution are not appropriate.*

The generalized linear model has the following components:

- Random component: Instead of being *normally* distributed, the components of $\mathbf{y}$ have a distribution in the *exponential family.*
- . Introduce a new symbol $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$.
- A monotonic univariate link function g so that $\boldsymbol{\eta}_i = g(\mu_i)$ where $\boldsymbol{\mu} = \mathcal{E}[\mathbf{y}]$.

The generalized linear model allows for a nonlinear *link function* g specifying that transformation of the expected value of the response variable which depends linearly on the predictors:

$$g(\mathrm{E}[\mathtt{y}_i]) = \boldsymbol{x}_i^\top\boldsymbol{\beta}, \tag{69.3.3}$$

Its random specification is such that $\mathrm{var}[\mathtt{y}]$ depends on $\mathrm{E}[\mathtt{y}]$ through a *variance function* $\phi \cdot V$ (where ϕ is a constant taking the place of σ^2 in the regression model:)

$$\mathrm{var}[\mathtt{y}] = \phi \cdot V(\mathrm{E}[\mathtt{y}]) \tag{69.3.4}$$

We have seen earlier that these mean- and variance functions are not an artificial construct, but that the distributions from the "exponential dispersion family," see Section 6.2, naturally give rise to such mean and variance functions. But just as much of the theory of the linear model can be derived without the assumption that the residuals are normally distributed, many of the results about generalized linear models do not require us to specify the whole distribution but can be derived on the basis of the mean and variance functions alone.

CHAPTER 70

Multiple Choice Models

Discrete choice between three or more alternatives; came from choice of transportation.

The outcomes of these choices should no longer be represented by a vector $\mathbf{y}$, but one needs a matrix $\mathbf{Y}$ with $\mathsf{y}_{ij} = 1$ if the ith individual chooses the jth alternative, and 0 otherwise. Consider only three alternatives $j = 1, 2, 3$, and define $\Pr(\mathsf{y}_{ij} = 1) = \pi_{ij}$.

Conditional Logit model is a model which makes all π_{ij} dependent on $\boldsymbol{x}_i$. It is very simple extension of binary choice. In binary choice we had $\log \frac{\pi_i}{1-\pi_i} = \boldsymbol{x}_i^\top \boldsymbol{\beta}$, log of *odds ratio*. Here this is generalized to $\log \frac{\pi_{i2}}{\pi_{i1}} = \boldsymbol{x}_i^\top \boldsymbol{\beta}_2$, and $\log \frac{\pi_{i3}}{\pi_{i1}} = \boldsymbol{x}_i^\top \boldsymbol{\beta}_3$. From this we obtain

$$\pi_{i1} = 1 - \pi_{i2} - \pi_{i3} = 1 - \pi_{i1} e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}_2} - \pi_{i1} e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}_3}, \tag{70.0.5}$$

or

$$\pi_{i1} = \frac{1}{1 + e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}_2} + e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}_3}}, \tag{70.0.6}$$

$$\pi_{i2} = \frac{e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}_2}}{1 + e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}_2} + e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}_3}}, \tag{70.0.7}$$

$$\pi_{i3} = \frac{e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}_3}}{1 + e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}_2} + e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}_3}}. \tag{70.0.8}$$

One can write this as $\pi_{ij} = \frac{e^{\alpha_j + \beta_j X_i}}{\sum e^{\alpha_k + \beta_k X_i}}$ if one defines $\alpha_1 = \beta_1 = 0$. The only estimation method used is MLE.

$$\mathcal{L} = \prod \pi_{i1}^{\mathsf{y}_{i1}} \pi_{i2}^{\mathsf{y}_{i2}} \Pi_{i3}^{\mathsf{y}_{i3}} = \prod \frac{(e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}_2})^{\mathsf{y}_{i2}} (e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}_3})^{\mathsf{y}_{i3}}}{1 + e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}_2} + e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}_3}}. \tag{70.0.9}$$

Note: the odds are independent of all other alternatives. Therefore the alternatives must be chosen such that this independence is a good assumption. The choice between walking, car, red buses, and blue buses does not satisfy this. See [**Cra91**, p. 47] for the best explanation of this which I found till now.

APPENDIX A

Matrix Formulas

In this Appendix, efforts are made to give some of the familiar matrix lemmas in their most general form. The reader should be warned: the concept of a deficiency matrix and the notation which uses a thick fraction line multiplication with a scalar g-inverse are my own.

A.1. A Fundamental Matrix Decomposition

THEOREM A.1.1. *Every matrix $\boldsymbol{B}$ which is not the null matrix can be written as a product of two matrices $\boldsymbol{B} = \boldsymbol{C}\boldsymbol{D}$, where $\boldsymbol{C}$ has a left inverse $\boldsymbol{L}$ and $\boldsymbol{D}$ a right inverse $\boldsymbol{R}$, i.e., $\boldsymbol{L}\boldsymbol{C} = \boldsymbol{D}\boldsymbol{R} = \boldsymbol{I}$. This identity matrix is $r \times r$, where r is the rank of $\boldsymbol{B}$.*

A proof is in [**Rao73**, p. 19]. This is the fundamental theorem of algebra, that every homomorphism can be written as a product of epimorphism and monomorphism, together with the fact that all epimorphisms and monomorphisms split, i.e., have one-sided inverses.

One such factorization is given by the singular value theorem: If $\boldsymbol{B} = \boldsymbol{P}^\top \boldsymbol{\Lambda} \boldsymbol{Q}$ is the svd as in Theorem A.9.2, then one might set e.g. $\boldsymbol{C} = \boldsymbol{P}^\top \boldsymbol{\Lambda}$ and $\boldsymbol{D} = \boldsymbol{Q}$, consequently $\boldsymbol{L} = \boldsymbol{\Lambda}^{-1}\boldsymbol{P}$ and $\boldsymbol{R} = \boldsymbol{Q}^\top$. In this decomposition, the first row/column carries the largest weight and gives the best approximation in a least squares sense, etc.

The trace of a square matrix is defined as the sum of its diagonal elements. The rank of a matrix is defined as the number of its linearly independent rows, which is equal to the number of its linearly independent columns (row rank = column rank).

THEOREM A.1.2. $\operatorname{tr} \boldsymbol{B}\boldsymbol{C} = \operatorname{tr} \boldsymbol{C}\boldsymbol{B}$.

PROBLEM 565. *Prove theorem A.1.2.*

PROBLEM 566. *Use theorem A.1.1 to prove that if $\boldsymbol{B}\boldsymbol{B} = \boldsymbol{B}$, then* $\operatorname{rank} \boldsymbol{B} = \operatorname{tr} \boldsymbol{B}$.

ANSWER. Premultiply the equation $\boldsymbol{C}\boldsymbol{D} = \boldsymbol{C}\boldsymbol{D}\boldsymbol{C}\boldsymbol{D}$ by $\boldsymbol{L}$ and postmultiply it by $\boldsymbol{R}$ to get $\boldsymbol{D}\boldsymbol{C} = \boldsymbol{I}_r$. This is useful for the trace: $\operatorname{tr} \boldsymbol{B} = \operatorname{tr} \boldsymbol{C}\boldsymbol{D} = \operatorname{tr} \boldsymbol{D}\boldsymbol{C} = \operatorname{tr} \boldsymbol{I}_r = r$. I have this proof from [**Rao73**, p. 28]. □

THEOREM A.1.3. $\boldsymbol{B} = \boldsymbol{O}$ *if and only if* $\boldsymbol{B}^\top \boldsymbol{B} = \boldsymbol{O}$.

A.2. The Spectral Norm of a Matrix

The spectral norm of a matrix extends the Euclidean norm $\|\boldsymbol{z}\|$ from vectors to matrices. Its definition is $\|\boldsymbol{A}\| = \max_{\|\boldsymbol{z}\|=1} \|\boldsymbol{A}\boldsymbol{z}\|$. This spectral norm is the maximum singular value μ_{max}, and if $\boldsymbol{A}$ is square, then $\left\|\boldsymbol{A}^{-1}\right\| = 1/\mu_{min}$. It is a true norm, i.e., $\|\boldsymbol{A}\| = 0$ if and only if $\boldsymbol{A} = \boldsymbol{O}$, furthermore $\|\lambda \boldsymbol{A}\| = |\lambda| \cdot \|\boldsymbol{A}\|$, and the triangle inequality $\|\boldsymbol{A} + \boldsymbol{B}\| \leq \|\boldsymbol{A}\| + \|\boldsymbol{B}\|$. In addition, it obeys $\|\boldsymbol{A}\boldsymbol{B}\| \leq \|\boldsymbol{A}\| \cdot \|\boldsymbol{B}\|$.

PROBLEM 567. *Show that the spectral norm is the maximum singular value.*

ANSWER. Use the definition

$$\|\boldsymbol{A}\|^2 = \max \frac{\boldsymbol{z}^\top \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{z}}{\boldsymbol{z}^\top \boldsymbol{z}}. \tag{A.2.1}$$

Write $\boldsymbol{A} = \boldsymbol{P}^\top \boldsymbol{\Lambda} \boldsymbol{Q}$ as in (A.9.1), Then $\boldsymbol{z}^\top \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{z} = \boldsymbol{z}^\top \boldsymbol{Q}^\top \boldsymbol{\Lambda}^2 \boldsymbol{Q}\boldsymbol{z}$. Therefore we can first show: there is a $\boldsymbol{z}$ in the form $\boldsymbol{z} = \boldsymbol{Q}^\top \boldsymbol{x}$ which attains this maximum. Proof: for every $\boldsymbol{z}$ which has a nonzero value in the numerator of (A.2.1), set $\boldsymbol{x} = \boldsymbol{Q}\boldsymbol{z}$. Then $\boldsymbol{x} \neq \boldsymbol{o}$, and $\boldsymbol{Q}^\top \boldsymbol{x}$ attains the same value as $\boldsymbol{z}$ in the numerator of (A.2.1), and a smaller or equal value in the denominator. Therefore one can restrict the search for the maximum argument to vectors of the form $\boldsymbol{Q}^\top \boldsymbol{x}$. But for them the objective function becomes $\frac{\boldsymbol{x}^\top \boldsymbol{\Lambda}^2 \boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}}$, which is maximized by $\boldsymbol{x} = \boldsymbol{i}_1$, the first unit vector (or column vector of the unit matrix). Therefore the squared spectral norm is λ_{ii}^2, and therefore the spectral norm itself is λ_{ii}. □

A.3. Inverses and g-Inverses of Matrices

A g-inverse of a matrix $\boldsymbol{A}$ is any matrix $\boldsymbol{A}^-$ satisfying

$$\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^-\boldsymbol{A}. \tag{A.3.1}$$

It always exists but is not always unique. If $\boldsymbol{A}$ is square and nonsingular, then $\boldsymbol{A}^{-1}$ is its only g-inverse.

PROBLEM 568. *Show that a symmetric matrix $\boldsymbol{\Omega}$ has a g-inverse which is also symmetric.*

ANSWER. Use $\boldsymbol{\Omega}^- \boldsymbol{\Omega} {\boldsymbol{\Omega}^-}^\top$. □

The definition of a g-inverse is apparently due to [**Rao62**]. It is sometimes called the "conditional inverse" [**Gra83**, p. 129]. This g-inverse, and not the Moore-Penrose generalized inverse or pseudoinverse $\boldsymbol{A}^+$, is needed for the linear model, The Moore-Penrose generalized inverse is a g-inverse that in addition satisfies $\boldsymbol{A}^+\boldsymbol{A}\boldsymbol{A}^+ = \boldsymbol{A}^+$, and $\boldsymbol{A}\boldsymbol{A}^+$ as well as $\boldsymbol{A}^+\boldsymbol{A}$ symmetric. It always exists and is also unique, but the additional requirements are burdensome ballast. [**Gre97**, pp. 44-5] also advocates the Moore-Penrose inverse, but he does not really use it. If he were to try to use it, he would probably soon discover that it is not appropriate. The book [**Alb72**] does

the linear model with the Moore-Penrose inverse. It is a good demonstration of how complicated everything gets if one uses an inappropriate mathematical tool.

PROBLEM 569. *Use theorem A.1.1 to prove that every matrix has a g-inverse.*

ANSWER. Simple: a null matrix has its transpose as g-inverse, and if $\boldsymbol{A} \neq \boldsymbol{O}$ then $\boldsymbol{RL}$ is such a g-inverse. □

The g-inverse of a number is its inverse if the number is nonzero, and is arbitrary otherwise. Scalar expressions written as fractions are in many cases the multiplication by a g-inverse. We will use a fraction with a thick horizontal rule to indicate where this is the case. In other words, by definition,

$$\frac{a}{b} = b^{-}a. \qquad \text{Compare that with the ordinary fraction} \quad \frac{a}{b}. \tag{A.3.2}$$

This idiosyncratic notation allows to write certain theorems in a more concise form, but it requires more work in the proofs, because one has to consider the additional case that the denominator is zero. Theorems A.5.8 and A.8.2 are examples.

THEOREM A.3.1. *If $\boldsymbol{B} = \boldsymbol{AA}^{-}\boldsymbol{B}$ holds for one g-inverse $\boldsymbol{A}^{-}$ of $\boldsymbol{A}$, then it holds for all g-inverses. If $\boldsymbol{A}$ is symmetric and $\boldsymbol{B} = \boldsymbol{AA}^{-}\boldsymbol{B}$, then also $\boldsymbol{B}^{\top} = \boldsymbol{B}^{\top}\boldsymbol{A}^{-}\boldsymbol{A}$. If $\boldsymbol{B} = \boldsymbol{BA}^{-}\boldsymbol{A}$ and $\boldsymbol{C} = \boldsymbol{AA}^{-}\boldsymbol{C}$ then $\boldsymbol{BA}^{-}\boldsymbol{C}$ is independent of the choice of g-inverses.*

PROOF. Assume the identity $\boldsymbol{B} = \boldsymbol{AA}^{+}\boldsymbol{B}$ holds for some fixed g-inverse $\boldsymbol{A}^{+}$ (which may be, as the notation suggests, the Moore Penrose g-inverse, but this is not necessary), and let $\boldsymbol{A}^{-}$ be an different g-inverse. Then $\boldsymbol{AA}^{-}\boldsymbol{B} = \boldsymbol{AA}^{-}\boldsymbol{AA}^{+}\boldsymbol{B} = \boldsymbol{AA}^{+}\boldsymbol{B} = \boldsymbol{B}$. For the second statement one merely has to take transposes and note that a matrix is a g-inverse of a symmetric $\boldsymbol{A}$ if and only if its transpose is. For the third statement: $\boldsymbol{BA}^{+}\boldsymbol{C} = \boldsymbol{BA}^{-}\boldsymbol{AA}^{+}\boldsymbol{AA}^{-}\boldsymbol{C} = \boldsymbol{BA}^{-}\boldsymbol{AA}^{-}\boldsymbol{C} = \boldsymbol{BA}^{-}\boldsymbol{C}$. Here $^{+}$ signifies a different g-inverse; again, it is not necessarily the Moore-Penrose one. □

PROBLEM 570. *Show that $\boldsymbol{x}$ satisfies $\boldsymbol{x} = \boldsymbol{Ba}$ for some $\boldsymbol{a}$ if and only if $\boldsymbol{x} = \boldsymbol{BB}^{-}\boldsymbol{x}$.*

THEOREM A.3.2. *Both $\boldsymbol{A}^{\top}(\boldsymbol{AA}^{\top})^{-}$ and $(\boldsymbol{A}^{\top}\boldsymbol{A})^{-}\boldsymbol{A}$ are g-inverses of $\boldsymbol{A}$.*

PROOF. We have to show

$$\boldsymbol{A} = \boldsymbol{AA}^{\top}(\boldsymbol{AA}^{\top})^{-}\boldsymbol{A} \tag{A.3.3}$$

which is [**Rao73**, (1b.5.5) on p. 26]. Define $\boldsymbol{D} = \boldsymbol{A} - \boldsymbol{AA}^{\top}(\boldsymbol{AA}^{\top})^{-}\boldsymbol{A}$ and show, by multiplying out, that $\boldsymbol{DD}^{\top} = \boldsymbol{O}$. □

A.4. Deficiency Matrices

Here is again some idiosyncratic terminology and notation. It gives an explicit algebraic formulation for something that is often done implicitly or in a geometric paradigm. A matrix $\boldsymbol{G}$ will be called a "left deficiency matrix" of $\boldsymbol{S}$, in symbols, $\boldsymbol{G} \perp \boldsymbol{S}$, if $\boldsymbol{GS} = \boldsymbol{O}$, and for all $\boldsymbol{Q}$ with $\boldsymbol{QS} = \boldsymbol{O}$ there is an $\boldsymbol{X}$ with $\boldsymbol{Q} = \boldsymbol{XG}$. This factorization property is an algebraic formulation of the geometric concept of a null space. It is symmetric in the sense that $\boldsymbol{G} \perp \boldsymbol{S}$ is also equivalent with: $\boldsymbol{GS} = \boldsymbol{O}$, and for all $\boldsymbol{R}$ with $\boldsymbol{GR} = \boldsymbol{O}$ there is a $\boldsymbol{Y}$ with $\boldsymbol{R} = \boldsymbol{SY}$. In other words, $\boldsymbol{G} \perp \boldsymbol{S}$ and $\boldsymbol{S}^{\top} \perp \boldsymbol{G}^{\top}$ are equivalent.

This symmetry follows from the following characterization of a deficiency matrix which is symmetric:

THEOREM A.4.1. *$\boldsymbol{T} \perp \boldsymbol{U}$ iff $\boldsymbol{TU} = \boldsymbol{O}$ and $\boldsymbol{T}^{\top}\boldsymbol{T} + \boldsymbol{UU}^{\top}$ nonsingular.*

PROOF. This proof here seems terribly complicated. There must be a simpler way. Proof of "⇒": Assume $\boldsymbol{T} \perp \boldsymbol{U}$. Take any $\boldsymbol{\gamma}$ with $\boldsymbol{\gamma}^{\top}\boldsymbol{T}^{\top}\boldsymbol{T}\boldsymbol{\gamma} + \boldsymbol{\gamma}^{\top}\boldsymbol{UU}^{\top}\boldsymbol{\gamma} = 0$, i.e., $\boldsymbol{T}\boldsymbol{\gamma} = \boldsymbol{o}$ and $\boldsymbol{\gamma}^{\top}\boldsymbol{U} = \boldsymbol{o}^{\top}$. From this one can show that $\boldsymbol{\gamma} = \boldsymbol{o}$: since $\boldsymbol{T}\boldsymbol{\gamma} = \boldsymbol{o}$, there is a $\boldsymbol{\xi}$ with $\boldsymbol{\gamma} = \boldsymbol{U}\boldsymbol{\xi}$, therefore $\boldsymbol{\gamma}^{\top}\boldsymbol{\gamma} = \boldsymbol{\gamma}^{\top}\boldsymbol{U}\boldsymbol{\xi} = 0$. To prove "⇐" assume $\boldsymbol{TU} = \boldsymbol{O}$ and $\boldsymbol{T}^{\top}\boldsymbol{T} + \boldsymbol{UU}^{\top}$ is nonsingular. To show that $\boldsymbol{T} \perp \boldsymbol{U}$ take any $\boldsymbol{B}$ with $\boldsymbol{BU} = \boldsymbol{O}$. Then $\boldsymbol{B} = \boldsymbol{B}(\boldsymbol{T}^{\top}\boldsymbol{T} + \boldsymbol{UU}^{\top})(\boldsymbol{T}^{\top}\boldsymbol{T} + \boldsymbol{UU}^{\top})^{-1} = \boldsymbol{BT}^{\top}\boldsymbol{T}(\boldsymbol{T}^{\top}\boldsymbol{T} + \boldsymbol{UU}^{\top})^{-1}$. In the same way one gets $\boldsymbol{T} = \boldsymbol{TT}^{\top}\boldsymbol{T}(\boldsymbol{T}^{\top}\boldsymbol{T} + \boldsymbol{UU}^{\top})^{-1}$. Premultiply this last equation by $\boldsymbol{T}^{\top}\boldsymbol{T}(\boldsymbol{T}^{\top}\boldsymbol{TT}^{\top}\boldsymbol{T})^{-}\boldsymbol{T}^{\top}$ and use theorem A.3.2 to get $\boldsymbol{T}^{\top}\boldsymbol{T}(\boldsymbol{T}^{\top}\boldsymbol{TT}^{\top}\boldsymbol{T})^{-}\boldsymbol{T}^{\top}\boldsymbol{T} = \boldsymbol{T}^{\top}\boldsymbol{T}(\boldsymbol{T}^{\top}\boldsymbol{T} + \boldsymbol{UU}^{\top})^{-1}$. Inserting this into the equation for $\boldsymbol{B}$ gives $\boldsymbol{B} = \boldsymbol{BT}^{\top}\boldsymbol{T}(\boldsymbol{T}^{\top}\boldsymbol{TT}^{\top}\boldsymbol{T})^{-}\boldsymbol{T}^{\top}\boldsymbol{T}$, i.e., $\boldsymbol{B}$ factors over $\boldsymbol{T}$. □

The R/`Splus`-function `Null` gives the transpose of a deficiency matrix.

THEOREM A.4.2. *If for all $\boldsymbol{Y}$, $\boldsymbol{BY} = \boldsymbol{O}$ implies $\boldsymbol{AY} = \boldsymbol{O}$, then a $\boldsymbol{X}$ exists with $\boldsymbol{A} = \boldsymbol{XB}$.*

PROBLEM 571. *Prove theorem A.4.2.*

ANSWER. Let $\boldsymbol{B} \perp \boldsymbol{C}$. Choosing $\boldsymbol{Y} = \boldsymbol{B}$ follows $\boldsymbol{AB} = \boldsymbol{O}$, hence $\boldsymbol{X}$ exists. □

PROBLEM 572. *Show that $\boldsymbol{I} - \boldsymbol{SS}^{-} \perp \boldsymbol{S}$.*

ANSWER. Clearly, $(\boldsymbol{I} - \boldsymbol{SS}^{-})\boldsymbol{S} = \boldsymbol{O}$. Now if $\boldsymbol{QS} = \boldsymbol{O}$, then $\boldsymbol{Q} = \boldsymbol{Q}(\boldsymbol{I} - \boldsymbol{SS}^{-})$, i.e., the $\boldsymbol{X}$ whose existence is postulated in the definition of a deficiency matrix is $\boldsymbol{Q}$ itself. □

PROBLEM 573. *Show that $\boldsymbol{S} \perp \boldsymbol{U}$ if and only if $\boldsymbol{S}$ is a matrix with maximal rank which satisfies $\boldsymbol{SU} = \boldsymbol{O}$. In other words, one cannot add linearly independent rows to $\boldsymbol{S}$ in such a way that the new matrix still satisfies $\boldsymbol{TU} = \boldsymbol{O}$.*

Answer. First assume $\boldsymbol{S} \perp \boldsymbol{U}$ and take any additional row $\boldsymbol{t}^\top$ so that $\begin{bmatrix} \boldsymbol{S} \\ \boldsymbol{t}^\top \end{bmatrix} \boldsymbol{U} = \begin{bmatrix} \boldsymbol{O} \\ \boldsymbol{o}^\top \end{bmatrix}$. Then exists a $\begin{bmatrix} \boldsymbol{Q} \\ \boldsymbol{r} \end{bmatrix}$ such that $\begin{bmatrix} \boldsymbol{S} \\ \boldsymbol{t}^\top \end{bmatrix} = \begin{bmatrix} \boldsymbol{Q} \\ \boldsymbol{r} \end{bmatrix} \boldsymbol{S}$, i.e., $\boldsymbol{SQ} = \boldsymbol{S}$, and $\boldsymbol{t}^\top = \boldsymbol{r}^\top \boldsymbol{S}$. But this last equation means that $\boldsymbol{t}^\top$ is a linear combination of the rows of $\boldsymbol{S}$ with the r_i as coefficients. Now conversely, assume $\boldsymbol{S}$ is such that one cannot add a linearly independent row $\boldsymbol{t}^\top$ such that $\begin{bmatrix} \boldsymbol{S} \\ \boldsymbol{t}^\top \end{bmatrix} \boldsymbol{U} = \begin{bmatrix} \boldsymbol{O} \\ \boldsymbol{o}^\top \end{bmatrix}$, and let $\boldsymbol{PU} = \boldsymbol{O}$. Then all rows of $\boldsymbol{P}$ must be linear combinations of rows of $\boldsymbol{S}$ (otherwise one could add such a row to $\boldsymbol{S}$ and get the result which was just ruled out), therefore $\boldsymbol{P} = \boldsymbol{SS}$ where $\boldsymbol{A}$ is the matrix of coefficients of these linear combinations. □

The deficiency matrix is not unique, but we will use the concept of a deficiency matrix in a formula only then when this formula remains correct for every deficiency matrix. One can make deficiency matrices unique if one requires them to be projection matrices.

Problem 574. *Given $\boldsymbol{X}$ and a symmetric nonnegative definite $\boldsymbol{\Omega}$ such that $\boldsymbol{X} = \boldsymbol{\Omega W}$ for some $\boldsymbol{W}$. Show that $\boldsymbol{X} \perp \boldsymbol{U}$ if and only if $\boldsymbol{X}^\top \boldsymbol{\Omega}^- \boldsymbol{X} \perp \boldsymbol{U}$.*

Answer. One has to show that $\boldsymbol{XY} = \boldsymbol{O}$ is equivalent to $\boldsymbol{X}^\top \boldsymbol{\Omega}^- \boldsymbol{XY} = \boldsymbol{O}$. $\Rightarrow$ clear; for $\Leftarrow$ note that $\boldsymbol{X}^\top \boldsymbol{\Omega}^- \boldsymbol{X} = \boldsymbol{W}^\top \boldsymbol{\Omega W}$, therefore $\boldsymbol{XY} = \boldsymbol{\Omega WY} = \boldsymbol{\Omega W}(\boldsymbol{W}^\top \boldsymbol{\Omega W})^- \boldsymbol{W}^\top \boldsymbol{\Omega WY} = \boldsymbol{\Omega W}(\boldsymbol{W}^\top \boldsymbol{\Omega W})^- \boldsymbol{X}^\top \boldsymbol{\Omega}^- \boldsymbol{XY} = \boldsymbol{O}$.

□

A matrix is said to have full column rank if all its columns are linearly independent, and full row rank if its rows are linearly independent. The deficiency matrix provides a "holistic" definition for which it is not necessary to look at single rows and columns. $\boldsymbol{X}$ has full column rank if and only if $\boldsymbol{X} \perp \boldsymbol{O}$, and full row rank if and only if $\boldsymbol{O} \perp \boldsymbol{X}$.

Problem 575. *Show that the following three statements are equivalent: (1) $\boldsymbol{X}$ has full column rank, (2) $\boldsymbol{X}^\top \boldsymbol{X}$ is nonsingular, and (3) $\boldsymbol{X}$ has a left inverse.*

Answer. Here use $\boldsymbol{X} \perp \boldsymbol{O}$ as the definition of "full column rank." Then (1) $\Leftrightarrow$ (2) is theorem A.4.1. Now (1) $\Rightarrow$ (3): Since $\boldsymbol{IO} = \boldsymbol{O}$, a $\boldsymbol{P}$ exists with $\boldsymbol{I} = \boldsymbol{PX}$. And (3) $\Rightarrow$ (1): if a $\boldsymbol{P}$ exists with $\boldsymbol{I} = \boldsymbol{PX}$, then any $\boldsymbol{Q}$ with $\boldsymbol{QO} = \boldsymbol{O}$ can be factored over $\boldsymbol{X}$, simply say $\boldsymbol{Q} = \boldsymbol{QPX}$. □

Note that the usual solution of linear matrix equations with g-inverses involves a deficiency matrix:

Theorem A.4.3. *The solution of the consistent matrix equation $\boldsymbol{TX} = \boldsymbol{A}$ is*

$$\boldsymbol{X} = \boldsymbol{T}^- \boldsymbol{A} + \boldsymbol{UW} \tag{A.4.1}$$

where $\boldsymbol{T} \perp \boldsymbol{U}$ and $\boldsymbol{W}$ is arbitrary.

Proof. Given consistency, i.e., the existence of at least one $\boldsymbol{Z}$ with $\boldsymbol{TZ} = \boldsymbol{A}$, (A.4.1) defines indeed a solution, since $\boldsymbol{TX} = \boldsymbol{TT}^- \boldsymbol{TZ}$. Conversely, if $\boldsymbol{Y}$ satisfies $\boldsymbol{TY} = \boldsymbol{A}$, then $\boldsymbol{T}(\boldsymbol{Y} - \boldsymbol{T}^- \boldsymbol{A}) = \boldsymbol{O}$, therefore $\boldsymbol{Y} - \boldsymbol{T}^- \boldsymbol{A} = \boldsymbol{UW}$ for some $\boldsymbol{W}$. □

Theorem A.4.4. *Let $\boldsymbol{L} \perp \boldsymbol{T} \perp \boldsymbol{U}$ and $\boldsymbol{J} \perp \boldsymbol{HU} \perp \boldsymbol{R}$; then*

$$\begin{bmatrix} \boldsymbol{L} & \boldsymbol{O} \\ -\boldsymbol{JHT}^- & \boldsymbol{J} \end{bmatrix} \perp \begin{bmatrix} \boldsymbol{T} \\ \boldsymbol{H} \end{bmatrix} \perp \boldsymbol{UR}.$$

Proof. First deficiency relation: Since $\boldsymbol{I} - \boldsymbol{TT}^- = \boldsymbol{UW}$ for some $\boldsymbol{W}$, $-\boldsymbol{JHT}^- \boldsymbol{T} + \boldsymbol{JH} = \boldsymbol{O}$, therefore the matrix product is zero. Now assume $\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \end{bmatrix} \begin{bmatrix} \boldsymbol{T} \\ \boldsymbol{H} \end{bmatrix} = \boldsymbol{O}$. Then $\boldsymbol{BHU} = \boldsymbol{O}$, i.e., $\boldsymbol{B} = \boldsymbol{DJ}$ for some $\boldsymbol{D}$. Then $\boldsymbol{AT} = -\boldsymbol{DJH}$, which has as general solution $\boldsymbol{A} = -\boldsymbol{DJHT}^- + \boldsymbol{CL}$ for some $\boldsymbol{C}$. This together gives $\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \end{bmatrix} = \begin{bmatrix} \boldsymbol{C} & \boldsymbol{D} \end{bmatrix} \begin{bmatrix} \boldsymbol{L} & \boldsymbol{O} \\ -\boldsymbol{JHT}^- & \boldsymbol{J} \end{bmatrix}$. Now the second deficiency relation: clearly, the product of the matrices is zero. If $\boldsymbol{M}$ satisfies $\boldsymbol{TM} = \boldsymbol{O}$, then $\boldsymbol{M} = \boldsymbol{UN}$ for some $\boldsymbol{N}$. If $\boldsymbol{M}$ furthermore satisfies $\boldsymbol{HM} = \boldsymbol{O}$, then $\boldsymbol{HUN} = \boldsymbol{O}$, therefore $\boldsymbol{N} = \boldsymbol{RP}$ for some $\boldsymbol{P}$, therefore $\boldsymbol{M} = \boldsymbol{URP}$. □

Theorem A.4.5. *Assume $\boldsymbol{\Omega}$ is nonnegative definite symmetric and $\boldsymbol{K}$ is such that $\boldsymbol{K\Omega}$ is defined. Then the matrix*

$$\boldsymbol{\Xi} = \boldsymbol{\Omega} - \boldsymbol{\Omega K}^\top (\boldsymbol{K\Omega K}^\top)^- \boldsymbol{K\Omega} \tag{A.4.2}$$

has the following properties:

(1) $\boldsymbol{\Xi}$ does not depend on the choice of g-inverse of $\boldsymbol{K\Omega K}^\top$ used in (A.4.2).

(2) Any g-inverse of $\boldsymbol{\Omega}$ is also a g-inverse of $\boldsymbol{\Xi}$, i.e. $\boldsymbol{\Xi\Omega}^- \boldsymbol{\Xi} = \boldsymbol{\Xi}$.

(3) $\boldsymbol{\Xi}$ is nonnegative definite and symmetric.

(4) For every $\boldsymbol{P} \perp \boldsymbol{\Omega}$ follows $\begin{bmatrix} \boldsymbol{K} \\ \boldsymbol{P} \end{bmatrix} \perp \boldsymbol{\Xi}$

(5) If $\boldsymbol{T}$ is any other right deficiency matrix of $\begin{bmatrix} \boldsymbol{K} \\ \boldsymbol{P} \end{bmatrix}$, i.e., if $\begin{bmatrix} \boldsymbol{K} \\ \boldsymbol{P} \end{bmatrix} \perp \boldsymbol{T}$, then

$$\boldsymbol{\Xi} = \boldsymbol{T}(\boldsymbol{T}^\top \boldsymbol{\Omega}^- \boldsymbol{T})^- \boldsymbol{T}^\top. \tag{A.4.3}$$

Hint: show that any $\boldsymbol{D}$ satisfying $\boldsymbol{\Xi} = \boldsymbol{TDT}^\top$ is a g-inverse of $\boldsymbol{T}^\top \boldsymbol{\Omega}^- \boldsymbol{T}$.

In order to apply (A.4.3) *show that the matrix $\boldsymbol{T} = \boldsymbol{SK}$ where $\boldsymbol{K} \perp \boldsymbol{S}$ and $\boldsymbol{PS} \perp \boldsymbol{K}$ is a right deficiency matrix of $\begin{bmatrix} \boldsymbol{K} \\ \boldsymbol{P} \end{bmatrix}$.*

Proof of theorem A.4.5: Independence of choice of g-inverse follows from theorem A.5.10. That $\boldsymbol{\Omega}^-$ is a g-inverse is also an immediate consequence of theorem A.5.10. From the factorization $\boldsymbol{\Xi} = \boldsymbol{\Xi\Omega}^- \boldsymbol{\Xi}$ follows also that $\boldsymbol{\Xi}$ is nnd symmetric (since every nnd symmetric $\boldsymbol{\Omega}$ also has a symmetric nnd g-inverse). (4) Deficiency property:

From $\begin{bmatrix}\boldsymbol{K}\\ \boldsymbol{P}\end{bmatrix}\boldsymbol{Q}=\boldsymbol{O}$ follows $\boldsymbol{KQ}=\boldsymbol{O}$ and $\boldsymbol{PQ}=\boldsymbol{O}$. From this second equation and $\boldsymbol{P}\perp\boldsymbol{\Omega}$ follows $\boldsymbol{Q}=\boldsymbol{\Omega R}$ for some $\boldsymbol{R}$. Since $\boldsymbol{K\Omega R}=\boldsymbol{KQ}=\boldsymbol{O}$, it follows $\boldsymbol{Q}=\boldsymbol{\Omega R}=(\boldsymbol{\Omega}-\boldsymbol{\Omega K}^\top(\boldsymbol{K\Omega K}^\top)^-\boldsymbol{K\Omega})\boldsymbol{R}$.

Proof of (5): Since $\begin{bmatrix}\boldsymbol{K}\\ \boldsymbol{P}\end{bmatrix}\boldsymbol{\Xi}=\boldsymbol{O}$ it follows $\boldsymbol{\Xi}=\boldsymbol{TA}$ for some $\boldsymbol{A}$, and therefore $\boldsymbol{\Xi}=\boldsymbol{\Xi\Omega}^-\boldsymbol{\Xi}=\boldsymbol{TA\Omega}^-\boldsymbol{A}^\top\boldsymbol{T}^\top=\boldsymbol{TDT}^\top$ where $\boldsymbol{D}=\boldsymbol{A\Omega}^-\boldsymbol{A}^\top$.

Before going on we need a lemma. Since $(\boldsymbol{I}-\boldsymbol{\Omega\Omega}^-)\boldsymbol{\Omega}=\boldsymbol{O}$, there exists a $\boldsymbol{N}$ with $\boldsymbol{I}-\boldsymbol{\Omega\Omega}^-=\boldsymbol{NP}$, therefore $\boldsymbol{T}-\boldsymbol{\Omega\Omega}^-\boldsymbol{T}=\boldsymbol{NPT}=\boldsymbol{O}$ or

$$\boldsymbol{T}=\boldsymbol{\Omega\Omega}^-\boldsymbol{T} \tag{A.4.4}$$

Using (A.4.4) one can show the hint: that any $\boldsymbol{D}$ satisfying $\boldsymbol{\Xi}=\boldsymbol{TDT}^\top$ is a g-inverse of $\boldsymbol{T}^\top\boldsymbol{\Omega}^-\boldsymbol{T}$:

$$\boldsymbol{T}^\top\boldsymbol{\Omega}^-\boldsymbol{TDT}^\top\boldsymbol{\Omega}^-\boldsymbol{T}\equiv\boldsymbol{T}^\top\boldsymbol{\Omega}^-(\boldsymbol{\Omega}-\boldsymbol{\Omega K}^\top(\boldsymbol{K\Omega K}^\top)^-\boldsymbol{K\Omega})\boldsymbol{\Omega}^-\boldsymbol{T}=\boldsymbol{T}^\top\boldsymbol{\Omega}^-\boldsymbol{T}. \tag{A.4.5}$$

To complete the proof of (5) we have to show that the expression $\boldsymbol{T}(\boldsymbol{T}^\top\boldsymbol{\Omega}^-\boldsymbol{T})^-\boldsymbol{T}^\top$ does not depend on the choice of the g-inverse of $\boldsymbol{T}^\top\boldsymbol{\Omega}^-\boldsymbol{T}$. This follows from $\boldsymbol{T}(\boldsymbol{T}^\top\boldsymbol{\Omega}^-\boldsymbol{T})^-\boldsymbol{T}^\top=\boldsymbol{\Omega\Omega}^-\boldsymbol{T}(\boldsymbol{T}^\top\boldsymbol{\Omega}^-\boldsymbol{T})^-\boldsymbol{T}^\top\boldsymbol{\Omega}^-\boldsymbol{\Omega}$ and theorem A.5.10.

THEOREM A.4.6. *Given two matrices $\boldsymbol{T}$ and $\boldsymbol{U}$. Then $\boldsymbol{T}\perp\boldsymbol{U}$ if and only if for any $\boldsymbol{D}$ the following two statements are equivalent:*

$$\boldsymbol{TD}=\boldsymbol{O} \tag{A.4.6}$$

and

(A.4.7) *For all $\boldsymbol{C}$ which satisfy $\boldsymbol{CU}=\boldsymbol{O}$ follows $\boldsymbol{CD}=\boldsymbol{O}$.*

A.5. Nonnegative Definite Symmetric Matrices

By definition, a symmetric matrix $\boldsymbol{\Omega}$ is nonnegative definite if $\boldsymbol{a}^\top\boldsymbol{\Omega a}\geq 0$ for all vectors $\boldsymbol{a}$. It is positive definite if $\boldsymbol{a}^\top\boldsymbol{\Omega a}>0$ for all vectors $\boldsymbol{a}\neq\boldsymbol{o}$.

THEOREM A.5.1. *$\boldsymbol{\Omega}$ nonnegative definite symmetric if and only if it can be written in the form $\boldsymbol{\Omega}=\boldsymbol{A}^\top\boldsymbol{A}$ for some $\boldsymbol{A}$.*

THEOREM A.5.2. *If $\boldsymbol{\Omega}$ is nonnegative definite, and $\boldsymbol{a}^\top\boldsymbol{\Omega a}=0$, then already $\boldsymbol{\Omega a}=\boldsymbol{o}$.*

THEOREM A.5.3. *$\boldsymbol{A}$ is positive definite if and only it is nonnegative definite and nonsingular.*

THEOREM A.5.4. *If the symmetric matrix $\boldsymbol{A}$ has a nnd g-inverse then $\boldsymbol{A}$ itself is also nnd.*

THEOREM A.5.5. *If $\boldsymbol{\Omega}$ and $\boldsymbol{\Sigma}$ are positive definite, then $\boldsymbol{\Omega}-\boldsymbol{\Sigma}$ is positive (nonnegative) definite if and only if $\boldsymbol{\Sigma}^{-1}-\boldsymbol{\Omega}^{-1}$ is.*

THEOREM A.5.6. *If $\boldsymbol{\Omega}$ and $\boldsymbol{\Sigma}$ are nonnegative definite, then $\operatorname{tr}(\boldsymbol{\Omega\Sigma})\geq 0$.*

PROBLEM 576. *Prove theorem A.5.6.*

ANSWER. Find any factorization $\boldsymbol{\Sigma}=\boldsymbol{PP}^\top$. Then $\operatorname{tr}(\boldsymbol{\Omega\Sigma})=\operatorname{tr}(\boldsymbol{P}^\top\boldsymbol{\Omega P})\geq 0$. □

THEOREM A.5.7. *If $\boldsymbol{\Omega}$ is nonnegative definite symmetric, then*

$$(\boldsymbol{g}^\top\boldsymbol{\Omega a})^2\leq\boldsymbol{g}^\top\boldsymbol{\Omega g}\,\boldsymbol{a}^\top\boldsymbol{\Omega a}, \tag{A.5.1}$$

for arbitrary vectors $\boldsymbol{a}$ and $\boldsymbol{g}$. Equality holds if and only if $\boldsymbol{\Omega g}$ and $\boldsymbol{\Omega a}$ are linearly dependent, i.e., α and β exist, not both zero, such that $\boldsymbol{\Omega g}\alpha+\boldsymbol{\Omega a}\beta=\boldsymbol{o}$.

Proof: First we will show that the condition for equality is sufficient. Therefore assume $\boldsymbol{\Omega g}\alpha+\boldsymbol{\Omega a}\beta=0$ for a certain α and β, which are not both zero. Without loss of generality we can assume $\alpha\neq 0$. Then we can solve $\boldsymbol{a}^\top\boldsymbol{\Omega g}\alpha+\boldsymbol{a}^\top\boldsymbol{\Omega a}\beta=0$ to get $\boldsymbol{a}^\top\boldsymbol{\Omega g}=-(\beta/\alpha)\boldsymbol{a}^\top\boldsymbol{\Omega a}$, therefore the lefthand side of (A.5.1) is $(\beta/\alpha)^2(\boldsymbol{a}^\top\boldsymbol{\Omega a})^2$. Furthermore we can solve $\boldsymbol{g}^\top\boldsymbol{\Omega g}\alpha+\boldsymbol{g}^\top\boldsymbol{\Omega a}\beta=0$ to get $\boldsymbol{g}^\top\boldsymbol{\Omega g}=-(\beta/\alpha)\boldsymbol{g}^\top\boldsymbol{\Omega a}=(\beta/\alpha)^2\boldsymbol{a}^\top\boldsymbol{\Omega a}$, therefore the righthand side of (A.5.1) is $(\beta/\alpha)^2(\boldsymbol{a}^\top\boldsymbol{\Omega a})^2$ as well—i.e., (A.5.1) holds with equality.

Secondly we will show that (A.5.1) holds in the general case and that, if it holds with equality, $\boldsymbol{\Omega g}$ and $\boldsymbol{\Omega a}$ are linearly dependent. We will split this second half of the proof into two substeps. First verify that (A.5.1) holds if $\boldsymbol{g}^\top\boldsymbol{\Omega g}=0$. If this is the case, then already $\boldsymbol{\Omega g}=\boldsymbol{o}$, therefore the $\boldsymbol{\Omega g}$ and $\boldsymbol{\Omega a}$ are linearly dependent and, by the first part of the proof, (A.5.1) holds with equality.

The second substep is the main part of the proof. Assume $\boldsymbol{g}^\top\boldsymbol{\Omega g}\neq 0$. Since $\boldsymbol{\Omega}$ is nonnegative definite, it follows

(A.5.2)
$$0\leq\Big(\boldsymbol{a}-\boldsymbol{g}\frac{\boldsymbol{g}^\top\boldsymbol{\Omega a}}{\boldsymbol{g}^\top\boldsymbol{\Omega g}}\Big)^\top\boldsymbol{\Omega}\Big(\boldsymbol{a}-\boldsymbol{g}\frac{\boldsymbol{g}^\top\boldsymbol{\Omega a}}{\boldsymbol{g}^\top\boldsymbol{\Omega g}}\Big)=\boldsymbol{a}^\top\boldsymbol{\Omega a}-2\frac{(\boldsymbol{g}^\top\boldsymbol{\Omega a})^2}{\boldsymbol{g}^\top\boldsymbol{\Omega g}}+\frac{(\boldsymbol{g}^\top\boldsymbol{\Omega a})^2}{\boldsymbol{g}^\top\boldsymbol{\Omega g}}=\boldsymbol{a}^\top\boldsymbol{\Omega a}-\frac{(\boldsymbol{g}^\top\boldsymbol{\Omega a})^2}{\boldsymbol{g}^\top\boldsymbol{\Omega g}}.$$

From this follows (A.5.1). If (A.5.2) is an equality, then already $\boldsymbol{\Omega}\Big(\boldsymbol{a}-\boldsymbol{g}\frac{\boldsymbol{g}^\top\boldsymbol{\Omega a}}{\boldsymbol{g}^\top\boldsymbol{\Omega g}}\Big)=\boldsymbol{o}$, which means that $\boldsymbol{\Omega g}$ and $\boldsymbol{\Omega a}$ are linearly dependent.

THEOREM A.5.8. *In the situation of theorem A.5.7, one can take g-inverses as follows without disturbing the inequality*

$$\frac{(\boldsymbol{g}^\top\boldsymbol{\Omega a})^2}{\boldsymbol{g}^\top\boldsymbol{\Omega g}}\leq\boldsymbol{a}^\top\boldsymbol{\Omega a}. \tag{A.5.3}$$

Equality holds if and only if a $\gamma\neq 0$ exists with $\boldsymbol{\Omega g}=\boldsymbol{\Omega a}\gamma$.

PROBLEM 577. *Show that if $\boldsymbol{\Omega}$ is nonnegative definite, then its elements satisfy*

$$\omega_{ij}^2 \leq \omega_{ii}\omega_{jj} \tag{A.5.4}$$

ANSWER. Let $\boldsymbol{a}$ and $\boldsymbol{b}$ be the ith and jth unit vector. Then

$$\frac{(\boldsymbol{b}^\top\boldsymbol{\Omega}\boldsymbol{a})^2}{\boldsymbol{b}^\top\boldsymbol{\Omega}\boldsymbol{b}} \leq \max_{\boldsymbol{g}} \frac{(\boldsymbol{g}^\top\boldsymbol{\Omega}\boldsymbol{a})^2}{\boldsymbol{g}^\top\boldsymbol{\Omega}\boldsymbol{g}} = \boldsymbol{a}^\top\boldsymbol{\Omega}\boldsymbol{a}. \tag{A.5.5}$$

□

PROBLEM 578. *Assume $\boldsymbol{\Omega}$ nonnegative definite symmetric. If $\boldsymbol{x}$ satisfies $\boldsymbol{x} = \boldsymbol{\Omega}\boldsymbol{a}$ for some $\boldsymbol{a}$, show that*

$$\max_{\boldsymbol{g}} \frac{(\boldsymbol{g}^\top\boldsymbol{x})^2}{\boldsymbol{g}^\top\boldsymbol{\Omega}\boldsymbol{g}} = \boldsymbol{x}^\top\boldsymbol{\Omega}^-\boldsymbol{x}. \tag{A.5.6}$$

Furthermore show that equality holds if and only if $\boldsymbol{\Omega}\boldsymbol{g} = \boldsymbol{x}\gamma$ for some $\gamma \neq 0$.

ANSWER. From $\boldsymbol{x} = \boldsymbol{\Omega}\boldsymbol{a}$ follows $\boldsymbol{g}^\top\boldsymbol{x} = \boldsymbol{g}^\top\boldsymbol{\Omega}\boldsymbol{a}$ and $\boldsymbol{x}^\top\boldsymbol{\Omega}^-\boldsymbol{x} = \boldsymbol{a}^\top\boldsymbol{\Omega}\boldsymbol{a}$; therefore it follows from theorem A.5.8.

□

PROBLEM 579. *Assume $\boldsymbol{\Omega}$ nonnegative definite symmetric, $\boldsymbol{x}$ satisfies $\boldsymbol{x} = \boldsymbol{\Omega}\boldsymbol{a}$ for some $\boldsymbol{a}$, and $\boldsymbol{R}$ is such that $\boldsymbol{R}\boldsymbol{x}$ is defined. Show that*

$$\boldsymbol{x}^\top\boldsymbol{R}^\top(\boldsymbol{R}\boldsymbol{\Omega}\boldsymbol{R}^\top)^-\boldsymbol{R}\boldsymbol{x} \leq \boldsymbol{x}^\top\boldsymbol{\Omega}^-\boldsymbol{x} \tag{A.5.7}$$

ANSWER. Follows from

$$\max_{\boldsymbol{h}} \frac{(\boldsymbol{h}^\top\boldsymbol{R}\boldsymbol{x})^2}{\boldsymbol{h}^\top\boldsymbol{R}\boldsymbol{\Omega}\boldsymbol{R}^\top\boldsymbol{h}} \leq \max_{\boldsymbol{g}} \frac{(\boldsymbol{g}^\top\boldsymbol{x})^2}{\boldsymbol{g}^\top\boldsymbol{\Omega}\boldsymbol{g}} \tag{A.5.8}$$

because on the term on the lhs maximization is done over the smaller set of $\boldsymbol{g}$ which have the form $\boldsymbol{R}\boldsymbol{h}$. An alternative proof would be to show that $\boldsymbol{\Omega} - \boldsymbol{\Omega}r^\top(\boldsymbol{R}\boldsymbol{\Omega}\boldsymbol{R}^\top)^-\boldsymbol{R}\boldsymbol{\Omega}$ is nnd (it has $\boldsymbol{\Omega}^-$ as g-inverse). □

PROBLEM 580. *Assume $\boldsymbol{\Omega}$ nonnegative definite symmetric. Show that*

$$\max_{\substack{\boldsymbol{g}:\\ \boldsymbol{g}=\boldsymbol{\Omega}\boldsymbol{a}\\ \text{for some } \boldsymbol{a}}} \frac{(\boldsymbol{g}^\top\boldsymbol{x})^2}{\boldsymbol{g}^\top\boldsymbol{\Omega}^-\boldsymbol{g}} = \boldsymbol{x}^\top\boldsymbol{\Omega}\boldsymbol{x}. \tag{A.5.9}$$

ANSWER. Since $\boldsymbol{g} = \boldsymbol{\Omega}\boldsymbol{a}$ for some $\boldsymbol{a}$, maximize over $\boldsymbol{a}$ instead of $\boldsymbol{g}$. This reduces it to theorem A.5.8:

$$\max_{\boldsymbol{g}:\ \boldsymbol{g}=\boldsymbol{\Omega}\boldsymbol{a} \text{ for some } \boldsymbol{a}} \frac{(\boldsymbol{g}^\top\boldsymbol{x})^2}{\boldsymbol{g}^\top\boldsymbol{\Omega}^-\boldsymbol{g}} = \max_{\boldsymbol{a}} \frac{(\boldsymbol{a}^\top\boldsymbol{\Omega}\boldsymbol{x})^2}{\boldsymbol{a}^\top\boldsymbol{\Omega}\boldsymbol{a}} = \boldsymbol{x}^\top\boldsymbol{\Omega}\boldsymbol{x} \tag{A.5.10}$$

□

THEOREM A.5.9. *Let $\boldsymbol{\Omega}$ be symmetric and nonnegative definite, and $\boldsymbol{x}$ an arbitrary vector. Then $\boldsymbol{\Omega} - \boldsymbol{x}\boldsymbol{x}^\top$ is nonnegative definite if and only if the following two conditions hold: $\boldsymbol{x}$ can be written in the form $\boldsymbol{x} = \boldsymbol{\Omega}\boldsymbol{a}$ for some $\boldsymbol{a}$, and $\boldsymbol{x}^\top\boldsymbol{\Omega}^-\boldsymbol{x} \leq 1$ for one (and therefore for all) g-inverses $\boldsymbol{\Omega}^-$ of $\boldsymbol{\Omega}$.*

PROBLEM 581. *Prove theorem A.5.9.*

ANSWER. Assume $\boldsymbol{x} = \boldsymbol{\Omega}\boldsymbol{a}$ and $\boldsymbol{x}^\top\boldsymbol{\Omega}^-\boldsymbol{x} = \boldsymbol{a}^\top\boldsymbol{\Omega}\boldsymbol{a} \leq 1$; then for any $\boldsymbol{g}$, $\boldsymbol{g}^\top(\boldsymbol{\Omega} - \boldsymbol{x}\boldsymbol{x}^\top)\boldsymbol{g}^\top = \boldsymbol{g}^\top\boldsymbol{\Omega}\boldsymbol{g} - \boldsymbol{g}^\top\boldsymbol{\Omega}\boldsymbol{a}\boldsymbol{a}^\top\boldsymbol{\Omega}\boldsymbol{g} \geq \boldsymbol{a}^\top\boldsymbol{\Omega}\boldsymbol{a}\boldsymbol{g}^\top\boldsymbol{\Omega}\boldsymbol{g} - \boldsymbol{g}^\top\boldsymbol{\Omega}\boldsymbol{a}\boldsymbol{a}^\top\boldsymbol{\Omega}\boldsymbol{g} \geq 0$ by theorem A.5.7.

Conversely, assume $\boldsymbol{x}$ cannot be written in the form $\boldsymbol{x} = \boldsymbol{\Omega}\boldsymbol{a}$ for some $\boldsymbol{a}$; then a $\boldsymbol{g}$ exists with $\boldsymbol{g}^\top\boldsymbol{\Omega} = \boldsymbol{o}^\top$ but $\boldsymbol{g}^\top\boldsymbol{x} \neq \boldsymbol{o}$. Then $\boldsymbol{g}^\top(\boldsymbol{\Omega} - \boldsymbol{x}\boldsymbol{x}^\top)\boldsymbol{g}^\top < 0$, therefore not nnd.

Finally assume $\boldsymbol{x}^\top\boldsymbol{\Omega}^-\boldsymbol{x} = \boldsymbol{a}^\top\boldsymbol{\Omega}\boldsymbol{a} > 1$; then $\boldsymbol{a}^\top(\boldsymbol{\Omega} - \boldsymbol{x}\boldsymbol{x}^\top)\boldsymbol{a} = \boldsymbol{a}^\top\boldsymbol{\Omega}\boldsymbol{a} - (\boldsymbol{a}^\top\boldsymbol{\Omega}\boldsymbol{a})^2 < 0$, therefore again not nnd. □

THEOREM A.5.10. *If $\boldsymbol{\Omega}$ and $\boldsymbol{\Sigma}$ are nonnegative definite symmetric, and $\boldsymbol{K}$ a matrix so that $\boldsymbol{\Sigma}\boldsymbol{K}\boldsymbol{\Omega}$ is defined, then*

$$\boldsymbol{K}\boldsymbol{\Omega} = (\boldsymbol{K}\boldsymbol{\Omega}\boldsymbol{K}^\top + \boldsymbol{\Sigma})(\boldsymbol{K}\boldsymbol{\Omega}\boldsymbol{K}^\top + \boldsymbol{\Sigma})^-\boldsymbol{K}\boldsymbol{\Omega}. \tag{A.5.11}$$

Furthermore, $\boldsymbol{\Omega}\boldsymbol{K}^\top(\boldsymbol{K}\boldsymbol{\Omega}\boldsymbol{K}^\top + \boldsymbol{\Sigma})^-\boldsymbol{K}\boldsymbol{\Omega}$ is independent of the choice of g-inverses.

PROBLEM 582. *Prove theorem A.5.10.*

ANSWER. To see that (A.5.11) is a special case of (A.3.3), take any $\boldsymbol{Q}$ with $\boldsymbol{\Omega} = \boldsymbol{Q}\boldsymbol{Q}^\top$ and $\boldsymbol{P}$ with $\boldsymbol{\Sigma} = \boldsymbol{P}\boldsymbol{P}^\top$ and define $\boldsymbol{A} = \begin{bmatrix}\boldsymbol{K}\boldsymbol{Q} & \boldsymbol{P}\end{bmatrix}$. The independence of the choice of g-inverses follows from theorem A.3.1 together with (A.5.11). □

The following was apparently first shown in [**Alb69**] for the special case of the Moore-Penrose pseudoinverse:

THEOREM A.5.11. *The symmetric partitioned matrix $\boldsymbol{\Omega} = \begin{bmatrix}\boldsymbol{\Omega}_{\mathtt{y}\mathtt{y}} & \boldsymbol{\Omega}_{\mathtt{y}\mathtt{z}}\\ \boldsymbol{\Omega}_{\mathtt{y}\mathtt{z}}^\top & \boldsymbol{\Omega}_{\mathtt{z}\mathtt{z}}\end{bmatrix}$ is nonnegative definite if and only if the following conditions hold:*

(A.5.12)

$$\boldsymbol{\Omega}_{\mathtt{y}\mathtt{y}} \quad \textit{and} \quad \boldsymbol{\Omega}_{\mathtt{z}\mathtt{z}.\mathtt{y}} := \boldsymbol{\Omega}_{\mathtt{z}\mathtt{z}} - \boldsymbol{\Omega}_{\mathtt{y}\mathtt{z}}^\top\boldsymbol{\Omega}_{\mathtt{y}\mathtt{y}}^-\boldsymbol{\Omega}_{\mathtt{y}\mathtt{z}} \quad \textit{are both nonnegative definite, and}$$

$$\boldsymbol{\Omega}_{\mathtt{y}\mathtt{z}} = \boldsymbol{\Omega}_{\mathtt{y}\mathtt{y}}\boldsymbol{\Omega}_{\mathtt{y}\mathtt{y}}^-\boldsymbol{\Omega}_{\mathtt{y}\mathtt{z}} \tag{A.5.13}$$

Reminder: It follows from theorem A.3.1 that (A.5.13) holds for some g-inverse if and only if it holds for all, and that, if it holds, $\boldsymbol{\Omega}_{\mathtt{z}\mathtt{z}.\mathtt{y}}$ is independent of the choice of the g-inverse.

Proof of theorem A.5.11: First we prove the necessity of the three conditions in the theorem. If the symmetric partitioned matrix $\boldsymbol{\Omega}$ is nonnegative definite, there exists a $\boldsymbol{R}$ with $\boldsymbol{\Omega} = \boldsymbol{R}^\top\boldsymbol{R}$. Write $\boldsymbol{R} = \begin{bmatrix}\boldsymbol{R}_{\mathtt{y}} & \boldsymbol{R}_{\mathtt{z}}\end{bmatrix}$ to get $\begin{bmatrix}\boldsymbol{\Omega}_{\mathtt{y}\mathtt{y}} & \boldsymbol{\Omega}_{\mathtt{y}\mathtt{z}}\\ \boldsymbol{\Omega}_{\mathtt{y}\mathtt{z}}^\top & \boldsymbol{\Omega}_{\mathtt{z}\mathtt{z}}\end{bmatrix} = \begin{bmatrix}\boldsymbol{R}_{\mathtt{y}}^\top\boldsymbol{R}_{\mathtt{y}} & \boldsymbol{R}_{\mathtt{y}}^\top\boldsymbol{R}_{\mathtt{z}}\\ \boldsymbol{R}_{\mathtt{z}}^\top\boldsymbol{R}_{\mathtt{y}} & \boldsymbol{R}_{\mathtt{z}}^\top\boldsymbol{R}_{\mathtt{z}}\end{bmatrix}$. $\boldsymbol{\Omega}_{\mathtt{y}\mathtt{y}}$ is nonnegative definite because it is equal to $\boldsymbol{R}_{\mathtt{y}}^\top\boldsymbol{R}_{\mathtt{y}}$, and

(A.5.13) follows from (A.5.11): $\boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}}\boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}}^{-}\boldsymbol{\Omega}_{\mathbf{y}\mathbf{z}} = \boldsymbol{R}_{\mathbf{y}}^{\top}\boldsymbol{R}_{\mathbf{y}}(\boldsymbol{R}_{\mathbf{y}}^{\top}\boldsymbol{R}_{\mathbf{y}})^{-}\boldsymbol{R}_{\mathbf{y}}^{\top}\boldsymbol{R}_{\mathbf{z}} = \boldsymbol{R}_{\mathbf{y}}^{\top}\boldsymbol{R}_{\mathbf{z}} = \boldsymbol{\Omega}_{\mathbf{y}\mathbf{z}}$. To show that $\boldsymbol{\Omega}_{\mathbf{z}\mathbf{z}.\mathbf{y}}$ is nonnegative definite, define $\boldsymbol{S} = (\boldsymbol{I}-\boldsymbol{R}_{\mathbf{y}}(\boldsymbol{R}_{\mathbf{y}}^{\top}\boldsymbol{R}_{\mathbf{y}})^{-}\boldsymbol{R}_{\mathbf{y}}^{\top})\boldsymbol{R}_{\mathbf{z}}$. Then $\boldsymbol{S}^{\top}\boldsymbol{S} = \boldsymbol{R}_{\mathbf{z}}^{\top}\big(\boldsymbol{I} - \boldsymbol{R}_{\mathbf{y}}(\boldsymbol{R}_{\mathbf{y}}^{\top}\boldsymbol{R}_{\mathbf{y}})^{-}\boldsymbol{R}_{\mathbf{y}}^{\top}\big)\boldsymbol{R}_{\mathbf{z}} = \boldsymbol{\Omega}_{\mathbf{z}\mathbf{z}.\mathbf{y}}$.

To show sufficiency of the three conditions of theorem A.5.11, assume the symmetric $\begin{bmatrix}\boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}} & \boldsymbol{\Omega}_{\mathbf{y}\mathbf{z}}\\ \boldsymbol{\Omega}_{\mathbf{y}\mathbf{z}}^{\top} & \boldsymbol{\Omega}_{\mathbf{z}\mathbf{z}}\end{bmatrix}$ satisfies them. Pick two matrices $\boldsymbol{Q}$ and $\boldsymbol{S}$ so that $\boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}} = \boldsymbol{Q}^{\top}\boldsymbol{Q}$ and $\boldsymbol{\Omega}_{\mathbf{z}\mathbf{z}.\mathbf{y}} = \boldsymbol{S}^{\top}\boldsymbol{S}$. Then

$$\begin{bmatrix}\boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}} & \boldsymbol{\Omega}_{\mathbf{y}\mathbf{z}}\\ \boldsymbol{\Omega}_{\mathbf{y}\mathbf{z}}^{\top} & \boldsymbol{\Omega}_{\mathbf{z}\mathbf{z}}\end{bmatrix} = \begin{bmatrix}\boldsymbol{Q}^{\top} & \boldsymbol{O}\\ \boldsymbol{\Omega}_{\mathbf{y}\mathbf{z}}^{\top}\boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}}^{-\top}\boldsymbol{Q}^{\top} & \boldsymbol{S}^{\top}\end{bmatrix}\begin{bmatrix}\boldsymbol{Q} & \boldsymbol{Q}\boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}}^{-}\boldsymbol{\Omega}_{\mathbf{y}\mathbf{z}}\\ \boldsymbol{O} & \boldsymbol{S}\end{bmatrix},$$

therefore nonnegative definite.

PROBLEM 583. [**SM86**, A 3.2/11] *Given a positive definite matrix $\boldsymbol{Q}$ and a positive definite $\widetilde{\boldsymbol{Q}}$ with $\boldsymbol{Q}^{*} = \boldsymbol{Q} - \widetilde{\boldsymbol{Q}}$ nonnegative definite.*

• a. *Show that $\widetilde{\boldsymbol{Q}} - \widetilde{\boldsymbol{Q}}\boldsymbol{Q}^{-1}\widetilde{\boldsymbol{Q}}$ is nonnegative definite.*

ANSWER. We know that $\widetilde{\boldsymbol{Q}}^{-1} - \boldsymbol{Q}^{*-1}$ is nnd, therefore $\widetilde{\boldsymbol{Q}}\widetilde{\boldsymbol{Q}}^{-1}\widetilde{\boldsymbol{Q}} - \widetilde{\boldsymbol{Q}}\boldsymbol{Q}^{*-1}\widetilde{\boldsymbol{Q}}$ nnd. □

• b. *This part is more difficult: Show that also $\boldsymbol{Q}^{*} - \boldsymbol{Q}^{*}\boldsymbol{Q}^{-1}\boldsymbol{Q}^{*}$ is nonnegative definite.*

ANSWER. We will write it in a symmetric form from which it is obvious that it is nonnegative definite:

$$\boldsymbol{Q}^{*} - \boldsymbol{Q}^{*}\boldsymbol{Q}^{-1}\boldsymbol{Q}^{*} = \boldsymbol{Q}^{*} - \boldsymbol{Q}^{*}(\widetilde{\boldsymbol{Q}} + \boldsymbol{Q}^{*})^{-1}\boldsymbol{Q}^{*} \tag{A.5.14}$$

$$= \boldsymbol{Q}^{*}(\widetilde{\boldsymbol{Q}} + \boldsymbol{Q}^{*})^{-1}(\widetilde{\boldsymbol{Q}} + \boldsymbol{Q}^{*} - \boldsymbol{Q}^{*}) = \boldsymbol{Q}^{*}(\widetilde{\boldsymbol{Q}} + \boldsymbol{Q}^{*})^{-1}\widetilde{\boldsymbol{Q}} \tag{A.5.15}$$

$$= \widetilde{\boldsymbol{Q}}(\widetilde{\boldsymbol{Q}} + \boldsymbol{Q}^{*})^{-1}(\widetilde{\boldsymbol{Q}} + \boldsymbol{Q}^{*})\widetilde{\boldsymbol{Q}}^{-1}\boldsymbol{Q}^{*}(\widetilde{\boldsymbol{Q}} + \boldsymbol{Q}^{*})^{-1}\widetilde{\boldsymbol{Q}} \tag{A.5.16}$$

$$= \widetilde{\boldsymbol{Q}}\boldsymbol{Q}^{-1}(\boldsymbol{Q}^{*} + \boldsymbol{Q}^{*}\widetilde{\boldsymbol{Q}}^{-1}\boldsymbol{Q}^{*})\boldsymbol{Q}^{-1}\widetilde{\boldsymbol{Q}}. \tag{A.5.17}$$

□

PROBLEM 584. *Given the vector $\boldsymbol{h} \neq \boldsymbol{o}$. For which values of the scalar γ is the matrix $\boldsymbol{I} - \frac{\boldsymbol{h}\boldsymbol{h}^{\top}}{\gamma}$ singular, nonsingular, nonnegative definite, a projection matrix, orthogonal?*

ANSWER. It is nnd iff $\gamma \geq \boldsymbol{h}^{\top}\boldsymbol{h}$, because of theorem A.5.9. One easily verifies that it is orthogonal iff $\gamma = \boldsymbol{h}^{\top}\boldsymbol{h}/2$, and it is a projection matrix iff $\gamma = \boldsymbol{h}^{\top}\boldsymbol{h}$. Now let us prove that it is singular iff $\gamma = \boldsymbol{h}^{\top}\boldsymbol{h}$: if this condition holds, then the matrix annuls $\boldsymbol{h}$; now assume the condition does not hold, i.e., $\gamma \neq \boldsymbol{h}^{\top}\boldsymbol{h}$, and take any $\boldsymbol{x}$ with $(\boldsymbol{I} - \frac{\boldsymbol{h}\boldsymbol{h}^{\top}}{\gamma})\boldsymbol{x} = \boldsymbol{o}$. It follows $\boldsymbol{x} = \boldsymbol{h}\alpha$ where $\alpha = \boldsymbol{h}^{\top}\boldsymbol{x}/\gamma$, therefore $(\boldsymbol{I} - \frac{\boldsymbol{h}\boldsymbol{h}^{\top}}{\gamma})\boldsymbol{x} = \boldsymbol{h}\alpha(1 - \boldsymbol{h}^{\top}\boldsymbol{h}/\gamma)$. Since $\boldsymbol{h} \neq \boldsymbol{o}$ and $1 - \boldsymbol{h}^{\top}\boldsymbol{h}/\gamma \neq 0$ this can only be the null vector if $\alpha = 0$. □

A.6. Projection Matrices

PROBLEM 585. *Show that $\boldsymbol{X}(\boldsymbol{X}^{\top}\boldsymbol{X})^{-}\boldsymbol{X}^{\top}$ is the projection matrix on the range space $\mathrm{R}[\boldsymbol{X}]$ of $\boldsymbol{X}$, i.e., on the space spanned by the columns of $\boldsymbol{X}$. This is true whether or not $\boldsymbol{X}$ has full column rank.*

ANSWER. Idempotence requires theorem A.3.2, and symmetry the invariance under choice of g-inverse. Furthermore one has to show $\boldsymbol{X}(\boldsymbol{X}^{\top}\boldsymbol{X})^{-}\boldsymbol{X}\boldsymbol{a} = \boldsymbol{a}$ holds if and only if $\boldsymbol{a} = \boldsymbol{X}\boldsymbol{b}$ for some $\boldsymbol{b}$. $\Rightarrow$ is clear, and $\Leftarrow$ follows from theorem A.3.2.

□

THEOREM A.6.1. *Let $\boldsymbol{P}$ and $\boldsymbol{Q}$ be projection matrices, i.e., both are symmetric and idempotent. Then the following five conditions are equivalent, each meaning that the space on which $\boldsymbol{P}$ projects is a subspace of the space on which $\boldsymbol{Q}$ projects:*

$$\mathrm{R}[\boldsymbol{P}] \subset \mathrm{R}[\boldsymbol{Q}] \tag{A.6.1}$$

$$\boldsymbol{Q}\boldsymbol{P} = \boldsymbol{P} \tag{A.6.2}$$

$$\boldsymbol{P}\boldsymbol{Q} = \boldsymbol{P} \tag{A.6.3}$$

$$\boldsymbol{Q} - \boldsymbol{P} \quad \textit{projection matrix} \tag{A.6.4}$$

$$\boldsymbol{Q} - \boldsymbol{P} \quad \textit{nonnegative definite.} \tag{A.6.5}$$

(A.6.2) is geometrically trivial. It means: if one first projects on a certain space, and then on a larger space which contains the first space as a subspace, then nothing happens under this second projection because one is already in the larger space. (A.6.3) is geometrically not trivial and worth remembering: if one first projects on a certain space, and then on a smaller space which is a subspace of the first space, then the result is the same as if one had projected directly on the smaller space. (A.6.4) means: the difference $\boldsymbol{Q}-\boldsymbol{P}$ is the projection on the orthogonal complement of $\mathrm{R}[\boldsymbol{P}]$ in $\mathrm{R}[\boldsymbol{Q}]$. And (A.6.5) means: the projection of a vector on the smaller space cannot be longer than that on the larger space.

PROBLEM 586. *Prove theorem A.6.1.*

ANSWER. Instead of going in a circle it is more natural to show (A.6.1) $\iff$ (A.6.2) and (A.6.3) $\iff$ (A.6.2) and then go in a circle for the remaining conditions: (A.6.2), (A.6.3) $\Rightarrow$ (A.6.4) $\Rightarrow$ (A.6.3) $\Rightarrow$ (A.6.5).

(A.6.1) $\Rightarrow$ (A.6.2): $\mathrm{R}[\boldsymbol{P}] \subset \mathrm{R}[\boldsymbol{Q}]$ means that for every $\boldsymbol{c}$ exists a $\boldsymbol{d}$ with $\boldsymbol{P}\boldsymbol{c} = \boldsymbol{Q}\boldsymbol{d}$. Therefore far all $\boldsymbol{c}$ follows $\boldsymbol{Q}\boldsymbol{P}\boldsymbol{c} = \boldsymbol{Q}\boldsymbol{Q}\boldsymbol{d} = \boldsymbol{Q}\boldsymbol{d} = \boldsymbol{P}\boldsymbol{c}$, i.e., $\boldsymbol{Q}\boldsymbol{P} = \boldsymbol{P}$.

(A.6.2) $\Rightarrow$ (A.6.1): if $\boldsymbol{P}\boldsymbol{c} = \boldsymbol{Q}\boldsymbol{P}\boldsymbol{c}$ for all $\boldsymbol{c}$, then clearly $\mathrm{R}[\boldsymbol{P}] \subset \mathrm{R}[\boldsymbol{Q}]$.

(A.6.2) $\Rightarrow$ (A.6.3) by symmetry of $\boldsymbol{P}$ and $\boldsymbol{Q}$: If $\boldsymbol{Q}\boldsymbol{P} = \boldsymbol{P}$ then $\boldsymbol{P}\boldsymbol{Q} = \boldsymbol{P}^{\top}\boldsymbol{Q}^{\top} = (\boldsymbol{Q}\boldsymbol{P})^{\top} = \boldsymbol{P}^{\top} = \boldsymbol{P}$.

(A.6.3) $\Rightarrow$ (A.6.2) follows in exactly the same way: If $\boldsymbol{P}\boldsymbol{Q} = \boldsymbol{P}$ then $\boldsymbol{Q}\boldsymbol{P} = \boldsymbol{Q}^{\top}\boldsymbol{P}^{\top} = (\boldsymbol{P}\boldsymbol{Q})^{\top} = \boldsymbol{P}^{\top} = \boldsymbol{P}$.

(A.6.2), (A.6.3) $\Rightarrow$ (A.6.4): Symmetry of $\boldsymbol{Q}-\boldsymbol{P}$ clear, and $(\boldsymbol{Q}-\boldsymbol{P})(\boldsymbol{Q}-\boldsymbol{P}) = \boldsymbol{Q}-\boldsymbol{P}-\boldsymbol{P}+\boldsymbol{P} = \boldsymbol{Q}-\boldsymbol{P}$.

(A.6.4) $\Rightarrow$ (A.6.5): $\boldsymbol{c}^\top(\boldsymbol{Q}-\boldsymbol{P})\boldsymbol{c} = \boldsymbol{c}^\top(\boldsymbol{Q}-\boldsymbol{P})^\top(\boldsymbol{Q}-\boldsymbol{P})\boldsymbol{c} \geq 0$.

(A.6.5) $\Rightarrow$ (A.6.3): First show that, if $\boldsymbol{Q}-\boldsymbol{P}$ nnd, then $\boldsymbol{Qc} = \boldsymbol{o}$ implies $\boldsymbol{Pc} = \boldsymbol{o}$. Proof: from $\boldsymbol{Q}-\boldsymbol{P}$ nnd and $\boldsymbol{Qc} = \boldsymbol{o}$ follows $0 \leq \boldsymbol{c}^\top(\boldsymbol{Q}-\boldsymbol{P})\boldsymbol{c} = -\boldsymbol{c}^\top\boldsymbol{Pc} \leq 0$, therefore equality throughout, i.e., $0 = \boldsymbol{c}^\top\boldsymbol{Pc} = \boldsymbol{c}^\top\boldsymbol{P}^\top\boldsymbol{Pc} = \|\boldsymbol{Pc}\|^2$ and therefore $\boldsymbol{Pc} = \boldsymbol{o}$. Secondly: this is also true for matrices: $\boldsymbol{QC} = \boldsymbol{O}$ implies $\boldsymbol{PC} = \boldsymbol{O}$, since it is valid for every column of $\boldsymbol{C}$. Thirdly: Since $\boldsymbol{Q}(\boldsymbol{I}-\boldsymbol{Q}) = \boldsymbol{O}$, it follows $\boldsymbol{P}(\boldsymbol{I}-\boldsymbol{Q}) = \boldsymbol{O}$, which is (A.6.3). □

PROBLEM 587. *If $\boldsymbol{Y} = \boldsymbol{XA}$ for some $\boldsymbol{A}$, show that $\boldsymbol{Y}(\boldsymbol{Y}^\top\boldsymbol{Y})^-\boldsymbol{Y}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^-\boldsymbol{X}^\top = \boldsymbol{Y}(\boldsymbol{Y}^\top\boldsymbol{Y})^-\boldsymbol{Y}^\top$.*

ANSWER. $\boldsymbol{Y} = \boldsymbol{XA}$ means that every column of $\boldsymbol{Y}$ is a linear combination of columns of $\boldsymbol{A}$:

$$\begin{bmatrix}\boldsymbol{y}_1 & \cdots & \boldsymbol{y}_m\end{bmatrix} = \boldsymbol{X}\begin{bmatrix}\boldsymbol{a}_1 & \cdots & \boldsymbol{a}_m\end{bmatrix} = \begin{bmatrix}\boldsymbol{Xa}_1 & \cdots & \boldsymbol{Xa}_m\end{bmatrix}. \tag{A.6.6}$$

Therefore geometrically the statement follows from the fact shown in Problem 585 that the above matrices are projection matrices on the columnn spaces. But it can also be shown algebraically: $\boldsymbol{Y}(\boldsymbol{Y}^\top\boldsymbol{Y})^-\boldsymbol{Y}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^-\boldsymbol{X}^\top = \boldsymbol{Y}(\boldsymbol{Y}^\top\boldsymbol{Y})^-\boldsymbol{A}^\top\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^-\boldsymbol{X}^\top = \boldsymbol{Y}(\boldsymbol{Y}^\top\boldsymbol{Y})^-\boldsymbol{Y}^\top$. □

PROBLEM 588. *(Not eligible for in-class exams) Let $\boldsymbol{Q}$ be a projection matrix (i.e., a symmetric and idempotent matrix) with the property that $\boldsymbol{Q} = \boldsymbol{XAX}^\top$ for some $\boldsymbol{A}$. Define $\tilde{\boldsymbol{X}} = (\boldsymbol{I}-\boldsymbol{Q})\boldsymbol{X}$. Then*

$$\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^-\boldsymbol{X}^\top = \tilde{\boldsymbol{X}}(\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}})^-\tilde{\boldsymbol{X}}^\top + \boldsymbol{Q}. \tag{A.6.7}$$

Hint: this can be done through a geometric argument. If you want to do it algebraically, you might want to use the fact that $(\boldsymbol{X}^\top\boldsymbol{X})^-$ is also a g-inverse of $\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}}$.

ANSWER. Geometric argument: $\boldsymbol{Q}$ is a projector on a subspace of the range space of $\boldsymbol{X}$. The columns of $\tilde{\boldsymbol{X}}$ are projections of the columns of $\boldsymbol{X}$ on the orthogonal complement of the space on which $\boldsymbol{Q}$ projects. The equation which we have to prove shows therefore that the projection on the column space of $\boldsymbol{X}$ is the sum of the projections on the space $\boldsymbol{Q}$ projects on plus the projection on the orthogonal complement of that space in $\boldsymbol{X}$.

Now an algebraic proof: First let us show that $(\boldsymbol{X}^\top\boldsymbol{X})^-$ is a g-inverse of $\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}}$, i.e., let us evaluate

$$\boldsymbol{X}^\top(\boldsymbol{I}-\boldsymbol{Q})\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^-\boldsymbol{X}^\top(\boldsymbol{I}-\boldsymbol{Q})\boldsymbol{X} = \boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^-\boldsymbol{X}^\top\boldsymbol{X} - \boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^-\boldsymbol{X}^\top\boldsymbol{QX} - \boldsymbol{X}^\top\boldsymbol{QX}(\boldsymbol{X}^\top\boldsymbol{X})^- \tag{A.6.8}$$

$$= \boldsymbol{X}^\top\boldsymbol{X} - \boldsymbol{X}^\top\boldsymbol{QX} - \boldsymbol{X}^\top\boldsymbol{QX} + \boldsymbol{X}^\top\boldsymbol{QX} = \boldsymbol{X}^\top(\boldsymbol{I}-\boldsymbol{Q})\boldsymbol{X}. \tag{A.6.9}$$

Only for the fourth term did we need the condition $\boldsymbol{Q} = \boldsymbol{XAX}^\top$:

$$\boldsymbol{X}^\top\boldsymbol{XAX}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^-\boldsymbol{X}^\top\boldsymbol{XAX}^\top\boldsymbol{X} = \boldsymbol{X}^\top\boldsymbol{XAX}^\top\boldsymbol{XAX}^\top\boldsymbol{X} = \boldsymbol{X}^\top\boldsymbol{QQX} = \boldsymbol{X}^\top\boldsymbol{X}. \tag{A.6.10}$$

Using this g-inverse we have

$$\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^-\boldsymbol{X}^\top - \tilde{\boldsymbol{X}}(\boldsymbol{X}^\top\boldsymbol{X})^-\tilde{\boldsymbol{X}}^\top = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^-\boldsymbol{X}^\top - (\boldsymbol{I}-\boldsymbol{Q})\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^-\boldsymbol{X}^\top(\boldsymbol{I}-\boldsymbol{Q}) = \tag{A.6.11}$$

$$= \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^-\boldsymbol{X}^\top - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^-\boldsymbol{X}^\top + \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^-\boldsymbol{X}^\top\boldsymbol{Q} + \boldsymbol{QX}(\boldsymbol{X}^\top\boldsymbol{X})^-\boldsymbol{X}^\top - \boldsymbol{QX}(\boldsymbol{X}^\top\boldsymbol{X})^-\boldsymbol{X}^\top\boldsymbol{Q} = \boldsymbol{X}(\boldsymbol{X} \tag{A.6.12}$$

□

PROBLEM 589. *Given any projection matrix $\boldsymbol{P}$. Show that its ith diagonal element can be written*

$$p_{ii} = \sum_j p_{ij}^2. \tag{A.6.13}$$

ANSWER. From idempotence $\boldsymbol{P} = \boldsymbol{PP}$ follows $p_{ii} = \sum_j p_{ij}p_{ji}$, now use symmetry to get (A.6.13).

□

A.7. Determinants

THEOREM A.7.1. *The determinant of a block-triangular matrix is the product of the determinants of the blocks in the diagonal. In other words,*

$$\begin{vmatrix}\boldsymbol{A} & \boldsymbol{B}\\ \boldsymbol{O} & \boldsymbol{D}\end{vmatrix} = |\boldsymbol{A}|\ |\boldsymbol{D}| \tag{A.7.1}$$

For the proof recall the definition of a determinant. A mapping $\pi\colon \{1,\ldots,n\} \to \{1,\ldots,n\}$ is called a permutation if and only if it is one-to-one if and only if it is onto. Permutations can be classified as even or odd according to whether they can be written as the product of an even or odd number of transpositions. Then the determinant is defined as

$$\det(\boldsymbol{A}) = \sum_{\pi\colon\ \pi \text{ even}} a_{1\pi(1)}\cdots a_{n\pi(n)} - \sum_{\pi\colon\ \pi \text{ odd}} a_{1\pi(1)}\cdots a_{n\pi(n)} \tag{A.7.2}$$

Now assume $\boldsymbol{A}$ is $m \times m$, $1 \leq m < n$. If a $j \leq m$ exists with $\pi(j) > m$ then not all $i \leq m$ can be images of other points $j \leq m$, i.e., there must be at least one $j > m$ with $\pi(j) \leq m$. Therefore, in a block triangular matrix in which all $a_{ij} = 0$ for $i \leq m$, $j > m$, only those permutations give a nonzero product which remain in the two submatrices straddling the diagonal.

THEOREM A.7.2. *If $\boldsymbol{B} = \boldsymbol{AA}^-\boldsymbol{B}$, then the following identity is valid between determinants:*

$$\begin{vmatrix}\boldsymbol{A} & \boldsymbol{B}\\ \boldsymbol{C} & \boldsymbol{D}\end{vmatrix} = |\boldsymbol{A}|\ |\boldsymbol{E}| \quad \textit{where} \quad \boldsymbol{E} = \boldsymbol{D} - \boldsymbol{CA}^-\boldsymbol{B}. \tag{A.7.3}$$

Proof: Postmultiply by a matrix whose determinant, by lemma A.7.1, is one, and then apply lemma A.7.1 once more:

$$\begin{vmatrix}\boldsymbol{A} & \boldsymbol{B}\\ \boldsymbol{C} & \boldsymbol{D}\end{vmatrix} = \begin{vmatrix}\boldsymbol{A} & \boldsymbol{B}\\ \boldsymbol{C} & \boldsymbol{D}\end{vmatrix}\begin{vmatrix}\boldsymbol{I} & -\boldsymbol{A}^-\boldsymbol{B}\\ \boldsymbol{O} & \boldsymbol{I}\end{vmatrix} = \begin{vmatrix}\boldsymbol{A} & \boldsymbol{O}\\ \boldsymbol{C} & \boldsymbol{D}-\boldsymbol{CA}^-\boldsymbol{B}\end{vmatrix} = |\boldsymbol{A}|\ \left|\boldsymbol{D}-\boldsymbol{CA}^-\boldsymbol{B}\right|. \tag{A.7.4}$$

PROBLEM 590. *Show the following counterpart of theorem A.7.2: If* $\boldsymbol{C} = \boldsymbol{D}\boldsymbol{D}^{-}\boldsymbol{C}$, *then the following identity is valid between determinants:*

$$\begin{vmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{vmatrix} = \left|\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-}\boldsymbol{C}\right| \, |\boldsymbol{D}| \ . \tag{A.7.5}$$

ANSWER.

$$\begin{vmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{vmatrix} = \begin{vmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{vmatrix} \begin{vmatrix} \boldsymbol{I} & \boldsymbol{O} \\ -\boldsymbol{D}^{-}\boldsymbol{C} & \boldsymbol{I} \end{vmatrix} = \begin{vmatrix} \boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-}\boldsymbol{C} & \boldsymbol{B} \\ \boldsymbol{O} & \boldsymbol{D} \end{vmatrix} = \left|\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-}\boldsymbol{D}\right| \, |\boldsymbol{D}| \ . \tag{A.7.6}$$

□

PROBLEM 591. *Show that whenever* $\boldsymbol{B}\boldsymbol{C}$ *and* $\boldsymbol{C}\boldsymbol{B}$ *are defined, it follows* $|\boldsymbol{I} - \boldsymbol{B}\boldsymbol{C}| = |\boldsymbol{I} - \boldsymbol{C}\boldsymbol{B}|$

ANSWER. Set $\boldsymbol{A} = \boldsymbol{I}$ and $\boldsymbol{D} = \boldsymbol{I}$ in (A.7.3) and (A.7.5). □

THEOREM A.7.3. *Assume that* $\boldsymbol{d} = \boldsymbol{W}\boldsymbol{W}^{-}\boldsymbol{d}$. *Then*

$$\det(\boldsymbol{W} + \alpha \cdot \boldsymbol{d}\boldsymbol{d}^{\top}) = \det(\boldsymbol{W})(1 + \alpha \boldsymbol{d}^{\top}\boldsymbol{W}^{-}\boldsymbol{d}). \tag{A.7.7}$$

Proof: If $\alpha = 0$, then there is nothing to prove. Otherwise look at the determinant of the matrix

$$\boldsymbol{H} = \begin{bmatrix} \boldsymbol{W} & \boldsymbol{d} \\ \boldsymbol{d}^{\top} & -1/\alpha \end{bmatrix} \tag{A.7.8}$$

Equations (A.7.3) and (A.7.5) give two expressions for it:

$$\det(\boldsymbol{H}) = \det(\boldsymbol{W})(-1/\alpha - \boldsymbol{d}^{\top}\boldsymbol{W}^{-}\boldsymbol{d}) = -\frac{1}{\alpha}\det(\boldsymbol{W} + \alpha\boldsymbol{d}\boldsymbol{d}^{\top}). \tag{A.7.9}$$

A.8. More About Inverses

PROBLEM 592. *Given a partitioned matrix* $\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix}$ *which satisfies* $\boldsymbol{B} = \boldsymbol{A}\boldsymbol{A}^{-}\boldsymbol{B}$ *and* $\boldsymbol{C} = \boldsymbol{C}\boldsymbol{A}^{-}\boldsymbol{A}$. *(These conditions hold for instance, due to theorem A.5.11, if* $\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix}$ *is nonnegative definite symmetric, but it also holds in the nonsymmetric case if* $\boldsymbol{A}$ *is nonsingular, which by theorem A.7.2 is the case if the whole partioned matrix is nonsingular.) Define* $\boldsymbol{E} = \boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-}\boldsymbol{B}$, $\boldsymbol{F} = \boldsymbol{A}^{-}\boldsymbol{B}$, *and* $\boldsymbol{G} = \boldsymbol{C}\boldsymbol{A}^{-}$.

• a. *Prove that in terms of* $\boldsymbol{A}$, $\boldsymbol{E}$, $\boldsymbol{F}$, *and* $\boldsymbol{G}$, *the original matrix can be written as*

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{A}\boldsymbol{F} \\ \boldsymbol{G}\boldsymbol{A} & \boldsymbol{E} + \boldsymbol{G}\boldsymbol{A}\boldsymbol{F} \end{bmatrix} \tag{A.8.1}$$

(this is trivial), and that (this is the nontrivial part)

$$\begin{bmatrix} \boldsymbol{A}^{-} + \boldsymbol{F}\boldsymbol{E}^{-}\boldsymbol{G} & -\boldsymbol{F}\boldsymbol{E}^{-} \\ -\boldsymbol{E}^{-}\boldsymbol{G} & \boldsymbol{E}^{-} \end{bmatrix} \quad \textit{is a g-inverse of} \quad \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix}. \tag{A.8.2}$$

ANSWER. This here is not the shortest proof because I was still wondering if it could be formulated in a more general way. Multiply out but do not yet use the conditions $\boldsymbol{B} = \boldsymbol{A}\boldsymbol{A}^{-}\boldsymbol{B}$ and $\boldsymbol{C} = \boldsymbol{C}\boldsymbol{A}^{-}\boldsymbol{A}$:

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix} \begin{bmatrix} \boldsymbol{A}^{-} + \boldsymbol{F}\boldsymbol{E}^{-}\boldsymbol{G} & -\boldsymbol{F}\boldsymbol{E}^{-} \\ -\boldsymbol{E}^{-}\boldsymbol{G} & \boldsymbol{E}^{-} \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}\boldsymbol{A}^{-} - (\boldsymbol{I} - \boldsymbol{A}\boldsymbol{A}^{-})\boldsymbol{B}\boldsymbol{E}^{-}\boldsymbol{G} & (\boldsymbol{I} - \boldsymbol{A}\boldsymbol{A}^{-})\boldsymbol{B}\boldsymbol{E}^{-} \\ (\boldsymbol{I} - \boldsymbol{E}\boldsymbol{E}^{-})\boldsymbol{G} & \boldsymbol{E}\boldsymbol{E}^{-} \end{bmatrix} \tag{A.8.3}$$

and

$$\begin{bmatrix} \boldsymbol{A}\boldsymbol{A}^{-} - (\boldsymbol{I} - \boldsymbol{A}\boldsymbol{A}^{-})\boldsymbol{B}\boldsymbol{E}^{-}\boldsymbol{G} & (\boldsymbol{I} - \boldsymbol{A}\boldsymbol{A}^{-})\boldsymbol{B}\boldsymbol{E}^{-} \\ (\boldsymbol{I} - \boldsymbol{E}\boldsymbol{E}^{-})\boldsymbol{G} & \boldsymbol{E}\boldsymbol{E}^{-} \end{bmatrix} \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix} = \tag{A.8.4}$$

$$= \begin{bmatrix} \boldsymbol{A} + (\boldsymbol{I} - \boldsymbol{A}\boldsymbol{A}^{-})\boldsymbol{B}\boldsymbol{E}^{-}\boldsymbol{C}(\boldsymbol{I} - \boldsymbol{A}^{-}\boldsymbol{A}) & \boldsymbol{B} - (\boldsymbol{I} - \boldsymbol{A}\boldsymbol{A}^{-})\boldsymbol{B}(\boldsymbol{I} - \boldsymbol{E}^{-}\boldsymbol{E}) \\ \boldsymbol{C} - (\boldsymbol{I} - \boldsymbol{E}\boldsymbol{E}^{-})\boldsymbol{C}(\boldsymbol{I} - \boldsymbol{A}^{-}\boldsymbol{A}) & \boldsymbol{D} \end{bmatrix}$$

One sees that not only the conditions $\boldsymbol{B} = \boldsymbol{A}\boldsymbol{A}^{-}\boldsymbol{B}$ and $\boldsymbol{C} = \boldsymbol{C}\boldsymbol{A}^{-}\boldsymbol{A}$, but also the conditions $\boldsymbol{B} = \boldsymbol{A}\boldsymbol{A}^{-}\boldsymbol{B}$ and $\boldsymbol{C} = \boldsymbol{E}\boldsymbol{E}^{-}\boldsymbol{C}$, or alternatively the conditions $\boldsymbol{B} = \boldsymbol{B}\boldsymbol{E}^{-}\boldsymbol{E}$ and $\boldsymbol{C} = \boldsymbol{C}\boldsymbol{A}^{-}\boldsymbol{A}$ imply the statement. I think one can also work with the conditions $\boldsymbol{A}\boldsymbol{A}^{-}\boldsymbol{B} = \boldsymbol{B}\boldsymbol{D}^{-}\boldsymbol{D}$ and $\boldsymbol{D}\boldsymbol{D}^{-}\boldsymbol{C} = \boldsymbol{C}\boldsymbol{A}^{-}\boldsymbol{A}$. Note that the lower right partition is $\boldsymbol{D}$ no matter what. □

• b. *If* $\begin{bmatrix} \boldsymbol{U} & \boldsymbol{V} \\ \boldsymbol{W} & \boldsymbol{X} \end{bmatrix}$ *is a g-inverse of* $\begin{bmatrix} \boldsymbol{A} & \boldsymbol{A}\boldsymbol{F} \\ \boldsymbol{G}\boldsymbol{A} & \boldsymbol{E} + \boldsymbol{G}\boldsymbol{A}\boldsymbol{F} \end{bmatrix}$, *show that* $\boldsymbol{X}$ *is a g-inverse of* $\boldsymbol{E}$.

ANSWER. The g-inverse condition means

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{A}\boldsymbol{F} \\ \boldsymbol{G}\boldsymbol{A} & \boldsymbol{E} + \boldsymbol{G}\boldsymbol{A}\boldsymbol{F} \end{bmatrix} \begin{bmatrix} \boldsymbol{U} & \boldsymbol{V} \\ \boldsymbol{W} & \boldsymbol{X} \end{bmatrix} \begin{bmatrix} \boldsymbol{A} & \boldsymbol{A}\boldsymbol{F} \\ \boldsymbol{G}\boldsymbol{A} & \boldsymbol{E} + \boldsymbol{G}\boldsymbol{A}\boldsymbol{F} \end{bmatrix} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{A}\boldsymbol{F} \\ \boldsymbol{G}\boldsymbol{A} & \boldsymbol{E} + \boldsymbol{G}\boldsymbol{A}\boldsymbol{F} \end{bmatrix} \tag{A.8.5}$$

The first matrix product evaluated is

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{A}\boldsymbol{F} \\ \boldsymbol{G}\boldsymbol{A} & \boldsymbol{E} + \boldsymbol{G}\boldsymbol{A}\boldsymbol{F} \end{bmatrix} \begin{bmatrix} \boldsymbol{U} & \boldsymbol{V} \\ \boldsymbol{W} & \boldsymbol{X} \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}\boldsymbol{U} + \boldsymbol{A}\boldsymbol{F}\boldsymbol{W} & \boldsymbol{A}\boldsymbol{V} + \boldsymbol{A}\boldsymbol{F}\boldsymbol{X} \\ \boldsymbol{G}\boldsymbol{A}\boldsymbol{U} + \boldsymbol{E}\boldsymbol{W} + \boldsymbol{G}\boldsymbol{A}\boldsymbol{F}\boldsymbol{W} & \boldsymbol{G}\boldsymbol{A}\boldsymbol{V} + \boldsymbol{E}\boldsymbol{X} + \boldsymbol{G}\boldsymbol{A}\boldsymbol{F}\boldsymbol{X} \end{bmatrix}. \tag{A.8.6}$$

The g-inverse condition means therefore

(A.8.7)

$$\begin{bmatrix} \boldsymbol{A}\boldsymbol{U} + \boldsymbol{A}\boldsymbol{F}\boldsymbol{W} & \boldsymbol{A}\boldsymbol{V} + \boldsymbol{A}\boldsymbol{F}\boldsymbol{X} \\ \boldsymbol{G}\boldsymbol{A}\boldsymbol{U} + \boldsymbol{E}\boldsymbol{W} + \boldsymbol{G}\boldsymbol{A}\boldsymbol{F}\boldsymbol{W} & \boldsymbol{G}\boldsymbol{A}\boldsymbol{V} + \boldsymbol{E}\boldsymbol{X} + \boldsymbol{G}\boldsymbol{A}\boldsymbol{F}\boldsymbol{X} \end{bmatrix} \begin{bmatrix} \boldsymbol{A} & \boldsymbol{A}\boldsymbol{F} \\ \boldsymbol{G}\boldsymbol{A} & \boldsymbol{E} + \boldsymbol{G}\boldsymbol{A}\boldsymbol{F} \end{bmatrix} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{A}\boldsymbol{F} \\ \boldsymbol{G}\boldsymbol{A} & \boldsymbol{E} + \boldsymbol{G}\boldsymbol{A}\boldsymbol{F} \end{bmatrix}$$

For the upper left partition this means $\boldsymbol{A}\boldsymbol{U}\boldsymbol{A} + \boldsymbol{A}\boldsymbol{F}\boldsymbol{W}\boldsymbol{A} + \boldsymbol{A}\boldsymbol{V}\boldsymbol{G}\boldsymbol{A} + \boldsymbol{A}\boldsymbol{F}\boldsymbol{X}\boldsymbol{G}\boldsymbol{A} = \boldsymbol{A}$, and for the upper right partition it means $\boldsymbol{A}\boldsymbol{U}\boldsymbol{A}\boldsymbol{F} + \boldsymbol{A}\boldsymbol{F}\boldsymbol{W}\boldsymbol{A}\boldsymbol{F} + \boldsymbol{A}\boldsymbol{V}\boldsymbol{E} + \boldsymbol{A}\boldsymbol{V}\boldsymbol{G}\boldsymbol{A}\boldsymbol{F} + \boldsymbol{A}\boldsymbol{F}\boldsymbol{X}\boldsymbol{E} + \boldsymbol{A}\boldsymbol{F}\boldsymbol{X}\boldsymbol{G}\boldsymbol{A}\boldsymbol{F} = \boldsymbol{A}\boldsymbol{F}$. Postmultiply the upper left equation by $\boldsymbol{F}$ and subtract from the upper right to get $\boldsymbol{A}\boldsymbol{V}\boldsymbol{E} + \boldsymbol{A}\boldsymbol{F}\boldsymbol{X}\boldsymbol{E} = \boldsymbol{O}$. For the lower left we get $\boldsymbol{G}\boldsymbol{A}\boldsymbol{U}\boldsymbol{A} + \boldsymbol{E}\boldsymbol{W}\boldsymbol{A} + \boldsymbol{G}\boldsymbol{A}\boldsymbol{F}\boldsymbol{W}\boldsymbol{A} + \boldsymbol{G}\boldsymbol{A}\boldsymbol{V}\boldsymbol{G}\boldsymbol{A} + \boldsymbol{E}\boldsymbol{X}\boldsymbol{G}\boldsymbol{A} + \boldsymbol{G}\boldsymbol{A}\boldsymbol{F}\boldsymbol{X}\boldsymbol{G}\boldsymbol{A} = \boldsymbol{G}\boldsymbol{A}$. Premultiplication of the upper left equation by $\boldsymbol{G}$ and subtraction gives $\boldsymbol{E}\boldsymbol{W}\boldsymbol{A} + \boldsymbol{E}\boldsymbol{X}\boldsymbol{G}\boldsymbol{A} = \boldsymbol{O}$. For the lower right corner we get $\boldsymbol{G}\boldsymbol{A}\boldsymbol{U}\boldsymbol{A}\boldsymbol{F} + \boldsymbol{E}\boldsymbol{W}\boldsymbol{A}\boldsymbol{F} + \boldsymbol{G}\boldsymbol{A}\boldsymbol{F}\boldsymbol{W}\boldsymbol{A}\boldsymbol{F} + \boldsymbol{G}\boldsymbol{A}\boldsymbol{V}\boldsymbol{E} + \boldsymbol{E}\boldsymbol{X}\boldsymbol{E} + \boldsymbol{G}\boldsymbol{A}\boldsymbol{V}\boldsymbol{G}\boldsymbol{A}\boldsymbol{F} + \boldsymbol{E}\boldsymbol{X}\boldsymbol{G}\boldsymbol{A}\boldsymbol{F} + \boldsymbol{G}\boldsymbol{A}\boldsymbol{F}\boldsymbol{X}\boldsymbol{G}\boldsymbol{A}\boldsymbol{F} = \boldsymbol{E} + \boldsymbol{G}\boldsymbol{A}\boldsymbol{F}$. Since $\boldsymbol{A}\boldsymbol{V}\boldsymbol{E} + \boldsymbol{A}\boldsymbol{F}\boldsymbol{X}\boldsymbol{E} = \boldsymbol{O}$ and $\boldsymbol{E}\boldsymbol{W}\boldsymbol{A} + \boldsymbol{E}\boldsymbol{X}\boldsymbol{G}\boldsymbol{A} = \boldsymbol{O}$, this simplifies to $\boldsymbol{G}\boldsymbol{A}\boldsymbol{U}\boldsymbol{A}\boldsymbol{F} + \boldsymbol{G}\boldsymbol{A}\boldsymbol{F}\boldsymbol{W}\boldsymbol{A}\boldsymbol{F} + \boldsymbol{E}\boldsymbol{X}\boldsymbol{E} + \boldsymbol{G}\boldsymbol{A}\boldsymbol{V}\boldsymbol{G}\boldsymbol{A}\boldsymbol{F} +$

$\boldsymbol{GAFXGAF} = \boldsymbol{E}+\boldsymbol{GAF}$. And if one premultiplies the upper right corner by $\boldsymbol{G}$ and postmultiplies it by $\boldsymbol{F}$ and subtracts it from this one gets $\boldsymbol{EXE} = \boldsymbol{E}$. □

PROBLEM 593. *Show that a g-inverse of the matrix*

$$\begin{bmatrix} \boldsymbol{X}_1^\top \boldsymbol{X}_1 & \boldsymbol{X}_1^\top \boldsymbol{X}_2 \\ \boldsymbol{X}_2^\top \boldsymbol{X}_1 & \boldsymbol{X}_2^\top \boldsymbol{X}_2 \end{bmatrix} \tag{A.8.8}$$

has the form

$$\begin{bmatrix} (\boldsymbol{X}_1^\top \boldsymbol{X}_1)^- + \boldsymbol{D}_1^\top \boldsymbol{X}_2 (\boldsymbol{X}_2^\top \boldsymbol{M}_1 \boldsymbol{X}_2)^- \boldsymbol{X}_2^\top \boldsymbol{D}_1 & -\boldsymbol{D}_1^\top \boldsymbol{X}_2 (\boldsymbol{X}_2^\top \boldsymbol{M}_1 \boldsymbol{X}_2)^- \\ -(\boldsymbol{X}_2^\top \boldsymbol{M}_1 \boldsymbol{X}_2)^- \boldsymbol{X}_2^\top \boldsymbol{D}_1 & (\boldsymbol{X}_2^\top \boldsymbol{M}_1 \boldsymbol{X}_2)^- \end{bmatrix} \tag{A.8.9}$$

where $\boldsymbol{M}_1 = \boldsymbol{I} - \boldsymbol{X}_1(\boldsymbol{X}_1^\top \boldsymbol{X}_1)^- \boldsymbol{X}_1^\top$ *and* $\boldsymbol{D}_1 = \boldsymbol{X}_1(\boldsymbol{X}_1^\top \boldsymbol{X}_1)^-$.

ANSWER. Either show it by multiplying it out, or apply Problem 592. □

PROBLEM 594. *Show that the following are g-inverses:*

$$\begin{bmatrix} \boldsymbol{I} & \boldsymbol{X} \\ \boldsymbol{X}^\top & \boldsymbol{X}^\top \boldsymbol{X} \end{bmatrix}^- = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} \end{bmatrix} \qquad \begin{bmatrix} \boldsymbol{X}^\top \boldsymbol{X} & \boldsymbol{X}^\top \\ \boldsymbol{X} & \boldsymbol{I} \end{bmatrix}^- = \begin{bmatrix} (\boldsymbol{X}^\top \boldsymbol{X})^- & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^- \boldsymbol{X} \end{bmatrix} \tag{A.8.10}$$

ANSWER. Either do it by multiplying it out, or apply problem 592. □

PROBLEM 595. *Assume again* $\boldsymbol{B} = \boldsymbol{AA}^-\boldsymbol{B}$ *and* $\boldsymbol{C} = \boldsymbol{CAA}^-$, *but assume this time that* $\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix}$ *nonsingular. Then* $\boldsymbol{A}$ *is nonsingular,*

$$\text{and if} \quad \begin{bmatrix} \boldsymbol{P} & \boldsymbol{Q} \\ \boldsymbol{R} & \boldsymbol{S} \end{bmatrix} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix}^{-1}, \quad \text{then the determinant} \quad \begin{vmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{vmatrix} = \frac{|\boldsymbol{A}|}{|\boldsymbol{S}|}. \tag{A.8.11}$$

ANSWER. The determinant is, by (A.7.3), $|\boldsymbol{A}|\ |\boldsymbol{E}|$ where $\boldsymbol{E} = \boldsymbol{D} - \boldsymbol{CA}^-\boldsymbol{B}$. By assumption, this determinant is nonzero, therefore also $|\boldsymbol{A}|$ and $|\boldsymbol{E}|$ are nonzero, i.e., $\boldsymbol{A}$ and $\boldsymbol{E}$ are nonsingular. Therefore (A.8.2) reads

$$\begin{bmatrix} \boldsymbol{P} & \boldsymbol{Q} \\ \boldsymbol{R} & \boldsymbol{S} \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}^{-1} + \boldsymbol{FE}^{-1}\boldsymbol{G} & -\boldsymbol{FE}^{-1} \\ -\boldsymbol{E}^{-1}\boldsymbol{G} & \boldsymbol{E}^{-1} \end{bmatrix}, \tag{A.8.12}$$

i.e., $\boldsymbol{S} = \boldsymbol{E}^{-1} = (\boldsymbol{D} - \boldsymbol{CA}^-\boldsymbol{B})^{-1}$. hence $|\boldsymbol{A}|\ |\boldsymbol{E}| = |\boldsymbol{A}|\ /\ |\boldsymbol{S}|$. □

THEOREM A.8.1. *Given a* $m \times n$ *matrix* $\boldsymbol{A}$, *a* $m \times h$ *matrix* $\boldsymbol{B}$, *a* $k \times n$ *matrix* $\boldsymbol{C}$, *and a* $k \times h$ *matrix* $\boldsymbol{D}$ *satisfying* $\boldsymbol{AA}^-\boldsymbol{B} = \boldsymbol{BD}^-\boldsymbol{D}$ *and* $\boldsymbol{DD}^-\boldsymbol{C} = \boldsymbol{CA}^-\boldsymbol{A}$. *Then the following are g-inverses:*

$$\Big(\boldsymbol{A} + \boldsymbol{BD}^-\boldsymbol{C}\Big)^- = \boldsymbol{A}^- - \boldsymbol{A}^-\boldsymbol{B}(\boldsymbol{D} + \boldsymbol{CA}^-\boldsymbol{B})^-\boldsymbol{CA}^- \tag{A.8.13}$$

$$\Big(\boldsymbol{D} + \boldsymbol{CA}^-\boldsymbol{B}\Big)^- = \boldsymbol{D}^- - \boldsymbol{D}^-\boldsymbol{C}(\boldsymbol{A} + \boldsymbol{BD}^-\boldsymbol{C})^-\boldsymbol{BD}^-. \tag{A.8.14}$$

PROBLEM 596. *Prove theorem A.8.1.*

ANSWER. Proof: Define $\boldsymbol{E} = \boldsymbol{D} + \boldsymbol{CA}^-\boldsymbol{B}$. Then it follows from the assumptions that

$$(\boldsymbol{A} + \boldsymbol{BD}^-\boldsymbol{C})(\boldsymbol{A}^- - \boldsymbol{A}^-\boldsymbol{BE}^-\boldsymbol{CA}^-) = \boldsymbol{AA}^- - \boldsymbol{BD}^-\boldsymbol{DE}^-\boldsymbol{CA}^- + \boldsymbol{BD}^-\boldsymbol{CA}^- - \boldsymbol{BD}^-\boldsymbol{CA}^-\boldsymbol{BE}^-\boldsymbol{CA}^- = \tag{A.8.15}$$

$$= \boldsymbol{AA}^- + \boldsymbol{BD}^-(\boldsymbol{I} - \boldsymbol{EE}^-)\boldsymbol{CA}^- \tag{A.8.16}$$

Since $\boldsymbol{AA}^-(\boldsymbol{A} + \boldsymbol{BD}^-\boldsymbol{C}) = \boldsymbol{A} + \boldsymbol{BD}^-\boldsymbol{C}$, we have to show that the second term on the rhs. annulls $(\boldsymbol{A} + \boldsymbol{BD}^-\boldsymbol{C})$. Indeed,

$$\boldsymbol{BD}^-(\boldsymbol{I} - \boldsymbol{EE}^-)\boldsymbol{CA}^-(\boldsymbol{A} + \boldsymbol{BD}^-\boldsymbol{C}) = \tag{A.8.17}$$

$$= \boldsymbol{BD}^-\boldsymbol{CA}^-\boldsymbol{A} + \boldsymbol{BD}^-\boldsymbol{CA}^-\boldsymbol{BD}^-\boldsymbol{C} - \boldsymbol{BD}^-\boldsymbol{EE}^-\boldsymbol{CA}^-\boldsymbol{A} - \boldsymbol{BD}^-\boldsymbol{EE}^-\boldsymbol{CA}^-\boldsymbol{BD}^-\boldsymbol{C} = \tag{A.8.18}$$

$$= \boldsymbol{BD}^-(\boldsymbol{D} + \boldsymbol{CA}^-\boldsymbol{B} - \boldsymbol{EE}^-\boldsymbol{D} - \boldsymbol{EE}^-\boldsymbol{CA}^-\boldsymbol{B})\boldsymbol{D}^-\boldsymbol{C} = \boldsymbol{BD}^-(\boldsymbol{E} - \boldsymbol{EE}^-\boldsymbol{E})\boldsymbol{D}^-\boldsymbol{C} = \boldsymbol{O}. \tag{A.8.19}$$

□

THEOREM A.8.2. *(Sherman-Morrison-Woodbury theorem) Given a* $m \times n$ *matrix* $\boldsymbol{A}$, *a* $m \times 1$ *vector* $\boldsymbol{b}$ *satisfying* $\boldsymbol{AA}^-\boldsymbol{b} = \boldsymbol{b}$, *a* $n \times 1$ *vector* $\boldsymbol{c}$ *satisfying* $\boldsymbol{c}^\top\boldsymbol{AA}^- = \boldsymbol{c}^\top$, *and a scalar* δ. *If* $\boldsymbol{A}^-$ *is a g-inverse of* $\boldsymbol{A}$, *then*

$$\boldsymbol{A}^- - \frac{\boldsymbol{A}^-\boldsymbol{bc}^\top\boldsymbol{A}^-}{\boldsymbol{c}^\top\boldsymbol{A}^-\boldsymbol{b} + \delta} \quad \text{is a g-inverse of} \quad \boldsymbol{A} + \frac{\boldsymbol{bc}^\top}{\delta} \tag{A.8.20}$$

PROBLEM 597. *Prove theorem A.8.2.*

ANSWER. It is a special case of theorem A.8.1. □

THEOREM A.8.3. *For any symmetric nonnegative definite* $r \times r$ *matrix* $\boldsymbol{A}$,

$$(\det \boldsymbol{A})\, e^{-(\operatorname{tr} \boldsymbol{A})} \le e^{-r}, \tag{A.8.21}$$

with equality holding if and only if $\boldsymbol{A} = \boldsymbol{I}$.

PROBLEM 598. *Prove Theorem A.8.3. Hint: Let* $\lambda_1, \ldots, \lambda_r$ *be the eigenvalues of* $\boldsymbol{A}$. *Then* $\det \boldsymbol{A} = \prod_i \lambda_i$, *and* $\operatorname{tr} \boldsymbol{A} = \sum_i \lambda_i$.

ANSWER. Therefore the inequality reads

$$\prod_{i=1}^{r} \lambda_i e^{-\lambda_i} \le e^{-r} \tag{A.8.22}$$

For this it is sufficient to show for each value of λ

$$\lambda e^{-\lambda} \le e^{-1}, \tag{A.8.23}$$

which follows immediately by taking the derivatives: $e^{-\lambda} - \lambda e^{-\lambda} = 0$ gives $\lambda = 1$. The matrix with all eigenvalues being equal to 1 is the identity matrix. □

A.9. Eigenvalues and Singular Value Decomposition

Every symmetric matrix $\boldsymbol{B}$ has real eigenvalues and a system of orthogonal eigenvectors which span the whole space. If one normalizes these eigenvectors and combines them as row vectors into a matrix $\boldsymbol{T}$, then orthonormality means $\boldsymbol{T}\boldsymbol{T}^\top = \boldsymbol{I}$, and since $\boldsymbol{T}$ is square, $\boldsymbol{T}\boldsymbol{T}^\top = \boldsymbol{I}$ also implies $\boldsymbol{T}^\top\boldsymbol{T} = \boldsymbol{I}$, i.e., $\boldsymbol{T}$ is an orthogonal matrix. The existence of a complete set of real eigenvectors is therefore equivalent to the following matrix algebraic result: For every symmetric matrix $\boldsymbol{B}$ there is an orthogonal transformation $\boldsymbol{T}$ so that $\boldsymbol{B}\boldsymbol{T}^\top = \boldsymbol{T}^\top\boldsymbol{\Lambda}$ where $\boldsymbol{\Lambda}$ is a diagonal matrix. Equivalently one could write $\boldsymbol{B} = \boldsymbol{T}^\top\boldsymbol{\Lambda}\boldsymbol{T}$. And if $\boldsymbol{B}$ has rank r, then r of the diagonal elements are nonzero and the others zero. If one removes those eigenvectors from $\boldsymbol{T}$ which belong to the eigenvalue zero, and calls the remaining matrix $\boldsymbol{P}$, one gets the following:

THEOREM A.9.1. *If $\boldsymbol{B}$ is a symmetric $n \times n$ matrix of rank r, then a $r \times n$ matrix $\boldsymbol{P}$ exists with $\boldsymbol{P}\boldsymbol{P}^\top = \boldsymbol{I}$ (any $\boldsymbol{P}$ satisfying this condition which is not a square matrix is called incomplete orthogonal), and $\boldsymbol{B} = \boldsymbol{P}^\top\boldsymbol{\Lambda}\boldsymbol{P}$, where $\boldsymbol{\Lambda}$ is a $r \times r$ diagonal matrix with all diagonal elements nonzero.*

PROOF. Let $\boldsymbol{T}$ be an orthogonal matrix whose rows are eigenvectors of $\boldsymbol{B}$, and partition it $\boldsymbol{T} = \begin{bmatrix}\boldsymbol{P}\\ \boldsymbol{Q}\end{bmatrix}$ where $\boldsymbol{P}$ consists of all eigenvectors with nonzero eigenvalue (there are r of them). The eigenvalue property reads $\boldsymbol{B}\begin{bmatrix}\boldsymbol{P}^\top & \boldsymbol{Q}^\top\end{bmatrix} = \begin{bmatrix}\boldsymbol{P}^\top & \boldsymbol{Q}^\top\end{bmatrix}\begin{bmatrix}\boldsymbol{\Lambda} & \boldsymbol{O}\\ \boldsymbol{O} & \boldsymbol{O}\end{bmatrix}$ therefore by orthogonality $\boldsymbol{T}^\top\boldsymbol{T} = \boldsymbol{I}$ follows $\boldsymbol{B} = \begin{bmatrix}\boldsymbol{P}^\top & \boldsymbol{Q}^\top\end{bmatrix}\begin{bmatrix}\boldsymbol{\Lambda} & \boldsymbol{O}\\ \boldsymbol{O} & \boldsymbol{O}\end{bmatrix}\begin{bmatrix}\boldsymbol{P}\\ \boldsymbol{Q}\end{bmatrix} = \boldsymbol{P}^\top\boldsymbol{\Lambda}\boldsymbol{P}$. Orthogonality also means $\boldsymbol{T}\boldsymbol{T}^\top = \boldsymbol{I}$, i.e., $\begin{bmatrix}\boldsymbol{P}\\ \boldsymbol{Q}\end{bmatrix}\begin{bmatrix}\boldsymbol{P}^\top & \boldsymbol{Q}^\top\end{bmatrix} = \begin{bmatrix}\boldsymbol{I} & \boldsymbol{O}\\ \boldsymbol{O} & \boldsymbol{I}\end{bmatrix}$, therefore $\boldsymbol{P}\boldsymbol{P}^\top = \boldsymbol{I}$. □

PROBLEM 599. *If $\boldsymbol{B}$ is a $n \times n$ symmetric matrix of rank r and $\boldsymbol{B}^2 = \boldsymbol{B}$, i.e., $\boldsymbol{B}$ is a projection, then a $r \times n$ matrix $\boldsymbol{P}$ exists with $\boldsymbol{B} = \boldsymbol{P}^\top\boldsymbol{P}$ and $\boldsymbol{P}\boldsymbol{P}^\top = \boldsymbol{I}$.*

ANSWER. Let $\boldsymbol{t}$ be an eigenvector of the projection matrix $\boldsymbol{B}$ with eigenvalue λ. Then $\boldsymbol{B}^2\boldsymbol{t} = \boldsymbol{B}\boldsymbol{t}$, i.e., $\lambda^2\boldsymbol{t} = \lambda\boldsymbol{t}$, and since $\boldsymbol{t} \neq \boldsymbol{o}$, $\lambda^2 = \lambda$. This is a quadratic equation with solutions $\lambda = 0$ or $\lambda = 1$. The matrix $\boldsymbol{\Lambda}$ from theorem A.9.1, whose diagonal elements are the nonzero eigenvalues, is therefore an identity matrix. □

A theorem similar to A.9.1 holds for arbitrary matrices. It is called the "singular value decomposition":

THEOREM A.9.2. *Let $\boldsymbol{B}$ be a $m \times n$ matrix of rank r. Then $\boldsymbol{B}$ can be expressed as*

$$\boldsymbol{B} = \boldsymbol{P}^\top\boldsymbol{\Lambda}\boldsymbol{Q} \tag{A.9.1}$$

where $\boldsymbol{\Lambda}$ is a $r \times r$ diagonal matrix with positive diagonal elements, and $\boldsymbol{P}\boldsymbol{P}^\top = \boldsymbol{I}$ as well as $\boldsymbol{Q}\boldsymbol{Q}^\top = \boldsymbol{I}$. The diagonal elements of $\boldsymbol{\Lambda}$ are called the singular values of $\boldsymbol{B}$.

PROOF. If $\boldsymbol{P}^\top\boldsymbol{\Lambda}\boldsymbol{Q}$ is the svd of $\boldsymbol{B}$ then $\boldsymbol{P}^\top\boldsymbol{\Lambda}\boldsymbol{Q}\boldsymbol{Q}^\top\boldsymbol{\Lambda}\boldsymbol{P} = \boldsymbol{P}^\top\boldsymbol{\Lambda}^2\boldsymbol{Q}$ is the eigenvalue decomposition of $\boldsymbol{B}\boldsymbol{B}^\top$. We will use this fact to construct $\boldsymbol{P}$ and $\boldsymbol{Q}$, and then verify condition (A.9.1). $\boldsymbol{P}$ and $\boldsymbol{Q}$ have r rows each, write them

$$\boldsymbol{P} = \begin{bmatrix}\boldsymbol{p}_1^\top\\ \vdots\\ \boldsymbol{p}_r^\top\end{bmatrix} \quad\text{and}\quad \boldsymbol{Q} = \begin{bmatrix}\boldsymbol{q}_1^\top\\ \vdots\\ \boldsymbol{q}_r^\top\end{bmatrix}. \tag{A.9.2}$$

Then the $\boldsymbol{p}_i$ are orthonormal eigenvectors of $\boldsymbol{B}\boldsymbol{B}^\top$ corresponding to the nonzero eigenvalues λ_i^2, and $\boldsymbol{q}_i = \boldsymbol{B}^\top\boldsymbol{p}_i\lambda_i^{-1}$. The proof that this definition is symmetric is left as exercise problem 600 below.

Now find $\boldsymbol{p}_{r+1}, \dots, \boldsymbol{p}_m$ such that $\boldsymbol{p}_1, \dots, \boldsymbol{p}_m$ is a complete set of orthonormal vectors, i.e., $\boldsymbol{p}_1\boldsymbol{p}_1^\top + \cdots + \boldsymbol{p}_m\boldsymbol{p}_m^\top = \boldsymbol{I}$. Then

$$\boldsymbol{B} = (\boldsymbol{p}_1\boldsymbol{p}_1^\top + \cdots + \boldsymbol{p}_m\boldsymbol{p}_m^\top)\boldsymbol{B} \tag{A.9.3}$$

$$= (\boldsymbol{p}_1\boldsymbol{p}_1^\top + \cdots + \boldsymbol{p}_r\boldsymbol{p}_r^\top)\boldsymbol{B} \quad \text{because } \boldsymbol{p}_i^\top\boldsymbol{B} = \boldsymbol{o}^\top \text{ for } i > r \tag{A.9.4}$$

$$= (\boldsymbol{p}_1\boldsymbol{q}_1^\top\lambda_1 + \cdots + \boldsymbol{p}_r\boldsymbol{q}_r^\top\lambda_r) = \boldsymbol{P}^\top\boldsymbol{\Lambda}\boldsymbol{Q}. \tag{A.9.5}$$

□

PROBLEM 600. *Show that the $\boldsymbol{q}_i$ are orthonormal eigenvectors of $\boldsymbol{B}^\top\boldsymbol{B}$ corresponding to the same eigenvalues λ_i^2.*

ANSWER.

$$\boldsymbol{q}_i^\top\boldsymbol{q}_j = \lambda_i^{-1}\boldsymbol{p}_i^\top\boldsymbol{B}\boldsymbol{B}^\top\boldsymbol{p}_j\lambda_j^{-1} = \lambda_i^{-1}\boldsymbol{p}_i^\top\boldsymbol{p}_j\lambda_j^2\lambda_j^{-1} = \delta_{ij} \quad \text{Kronecker symbol} \tag{A.9.6}$$

$$\boldsymbol{B}^\top\boldsymbol{B}\boldsymbol{q}_i = \boldsymbol{B}^\top\boldsymbol{B}\boldsymbol{B}^\top\boldsymbol{p}_i\lambda_i^{-1} = \boldsymbol{B}^\top\boldsymbol{p}_i\lambda_i = \boldsymbol{q}_i\lambda_i^2 \tag{A.9.7}$$

□

PROBLEM 601. *Show that $\boldsymbol{B}\boldsymbol{q}_i = \lambda_i\boldsymbol{p}_i$ and $\boldsymbol{B}^\top\boldsymbol{p}_i = \lambda_i\boldsymbol{q}_i$.*

ANSWER. The second condition comes from the definition $\boldsymbol{q}_i = \boldsymbol{B}^\top\boldsymbol{p}_i\lambda_i^{-1}$, and premultiply this definition by $\boldsymbol{B}$ to get $\boldsymbol{B}\boldsymbol{q}_i = \boldsymbol{B}\boldsymbol{B}^\top\boldsymbol{p}_i\lambda_i^{-1} = \lambda^2\boldsymbol{p}_i\lambda_i^{-1} = \lambda\boldsymbol{p}_i$. □

Let $\boldsymbol{P}_0$ and $\boldsymbol{Q}_0$ be such that $\begin{bmatrix}\boldsymbol{P}\\ \boldsymbol{P}_0\end{bmatrix}$ and $\begin{bmatrix}\boldsymbol{Q}\\ \boldsymbol{Q}_0\end{bmatrix}$ are orthogonal. Then the singular value decomposition can also be written in the full form, in which the matrix in the middle is $m \times n$:

$$\boldsymbol{B} = \begin{bmatrix}\boldsymbol{P}^\top & \boldsymbol{P}_0^\top\end{bmatrix}\begin{bmatrix}\boldsymbol{\Lambda} & \boldsymbol{O}\\ \boldsymbol{O} & \boldsymbol{O}\end{bmatrix}\begin{bmatrix}\boldsymbol{Q}\\ \boldsymbol{Q}_0\end{bmatrix} \tag{A.9.8}$$

PROBLEM 602. *Let λ_1 be the biggest diagonal element of $\boldsymbol{\Lambda}$, and let $\boldsymbol{c}$ and $\boldsymbol{d}$ be two vectors with the properties that $\boldsymbol{c}^\top \boldsymbol{B}\boldsymbol{d}$ is defined and $\boldsymbol{c}^\top \boldsymbol{c} = 1$ as well as $\boldsymbol{d}^\top \boldsymbol{d} = 1$. Show that $\boldsymbol{c}^\top \boldsymbol{B}\boldsymbol{d} \leq \lambda_1$. The other singular values maximize among those who are orthogonal to the prior maximizers.*

ANSWER. $\boldsymbol{c}^\top \boldsymbol{B}\boldsymbol{d} = \boldsymbol{c}^\top \boldsymbol{P}^\top \boldsymbol{\Lambda}\boldsymbol{Q}\boldsymbol{d} = \boldsymbol{h}^\top \boldsymbol{\Lambda}\boldsymbol{k}$ where we call $\boldsymbol{P}\boldsymbol{c} = \boldsymbol{h}$ and $\boldsymbol{Q}\boldsymbol{d} = \boldsymbol{k}$. By Cauchy-Schwartz (A.5.1), $(\boldsymbol{h}^\top \boldsymbol{\Lambda}\boldsymbol{k})^2 \leq (\boldsymbol{h}^\top \boldsymbol{\Lambda}\boldsymbol{h})(\boldsymbol{k}^\top \boldsymbol{\Lambda}\boldsymbol{k})$. Now $(\boldsymbol{h}^\top \boldsymbol{\Lambda}\boldsymbol{k}) = \sum \lambda_{ii} h_i^2 \leq \sum \lambda_{11} h_i^2 = \lambda_{11} \boldsymbol{h}^\top \boldsymbol{h}$. Now we only have to show that $\boldsymbol{h}^\top \boldsymbol{h} \leq 1$: $1 - \boldsymbol{h}^\top \boldsymbol{h} = \boldsymbol{c}^\top \boldsymbol{c} - \boldsymbol{c}^\top \boldsymbol{P}^\top \boldsymbol{P}\boldsymbol{c} = \boldsymbol{c}^\top (\boldsymbol{I} - \boldsymbol{P}^\top \boldsymbol{P})\boldsymbol{c} = \boldsymbol{c}^\top (\boldsymbol{I} - \boldsymbol{P}^\top \boldsymbol{P})(\boldsymbol{I} - \boldsymbol{P}^\top \boldsymbol{P})\boldsymbol{c} \geq 0$, here we used that $\boldsymbol{P}\boldsymbol{P}^\top = \boldsymbol{I}$, therefore $\boldsymbol{P}^\top \boldsymbol{P}$ idempotent, therefore also $\boldsymbol{I} - \boldsymbol{P}^\top \boldsymbol{P}$ idempotent. $\square$

APPENDIX B

Arrays of Higher Rank

This chapter was presented at the Array Programming Languages Conference in Berlin, on July 24, 2000.

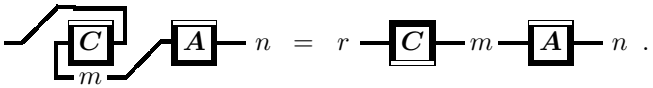
Besides scalars, vectors, and matrices, also higher arrays are necessary in statistics; for instance, the "covariance matrix" of a random matrix is really an array of rank 4, etc. Usually, such higher arrays are avoided in the applied sciences because of the difficulties to write them on a two-dimensional sheet of paper. The following symbolic notation makes the structure of arrays explicit without writing them down element by element. It is hoped that this makes arrays easier to understand, and that this notation leads to simple high-level user interfaces for programming languages manipulating arrays.

B.1. Informal Survey of the Notation

Each array is symbolized by a rectangular tile with arms sticking out, similar to a molecule. Tiles with one arm are vectors, those with two arms matrices, those with more arms are arrays of higher rank (or "valence" as in [**SS35**], [**Mor73**], and [**MS86**, p. 12]), and those without arms are scalars. The arrays considered here are rectangular, not "ragged," therefore in addition to their rank we only need to know the dimension of each arm; it can be thought of as the number of fingers associated with this arm. Arrays can only hold hands (i.e., "contract" along two arms) if the hands have the same number of fingers.

Sometimes it is convenient to write the dimension of each arm at the end of the arm, i.e., a $m \times n$ matrix $\boldsymbol{A}$ can be represented as m —[$\boldsymbol{A}$]— n . Matrix products are represented by joining the obvious arms: if $\boldsymbol{B}$ is $n \times q$, then the matrix product $\boldsymbol{AB}$ is m —[$\boldsymbol{A}$]— n —[$\boldsymbol{B}$]— q or, in short, —[$\boldsymbol{A}$]—[$\boldsymbol{B}$]—. The notation allows the reader to always tell which arm is which, even if the arms are not marked. If

m —[$\boldsymbol{C}$]— r is $m \times r$, then the product $\boldsymbol{C}^\top \boldsymbol{A}$ is

(B.1.1) $\boldsymbol{C}^\top \boldsymbol{A} =$ 

In the second representation, the tile representing $\boldsymbol{C}$ is turned by 180 degrees. Since the white part of the frame of $\boldsymbol{C}$ is at the bottom, not on the top, one knows that the West arm of $\boldsymbol{C}$, not its East arm, is concatenated with the West arm of $\boldsymbol{A}$. The transpose of m —[$\boldsymbol{C}$]— r is r —[$\boldsymbol{C}$]— m , i.e., it is not a different entity but the same entity in a different position. The order in which the elements are arranged on the page (or in computer memory) is not a part of the definition of the array itself. Likewise, there is no distinction between row vectors and column vectors.

Vectors are usually, but not necessarily, written in such a way that their arm points West (column vector convention). If —[$\boldsymbol{a}$] and —[$\boldsymbol{b}$] are vectors, their scalar product $\boldsymbol{a}^\top \boldsymbol{b}$ is the concatenation [$\boldsymbol{a}$]—[$\boldsymbol{b}$] which has no free arms, i.e., it is a scalar, and their outer product $\boldsymbol{a}\boldsymbol{b}^\top$ is —[$\boldsymbol{a}$] [$\boldsymbol{b}$]—, which is a matrix. Juxtaposition of tiles represents the outer product, i.e., the array consisting of all the products of elements of the arrays represented by the tiles placed side by side.

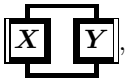
The trace of a square matrix —[$\boldsymbol{Q}$]— is the concatenation [$\boldsymbol{Q}$] , which is a scalar since no arms are sticking out. In general, concatenation of two arms of the same tile represents *contraction*, i.e., summation over equal values of the indices associated with these two arms. This notation makes it obvious that $\operatorname{tr} \boldsymbol{XY} = \operatorname{tr} \boldsymbol{YX}$, because by definition there is no difference between [$\boldsymbol{X}$]—[$\boldsymbol{Y}$] and [$\boldsymbol{Y}$]—[$\boldsymbol{X}$] . Also [$\boldsymbol{X}$] [$\boldsymbol{Y}$] or [$\boldsymbol{X}$]—[$\boldsymbol{Y}$] etc. represent the same array (here array of rank zero, i.e., scalar). Each of these tiles can be evaluated in essentially two different ways. One way is

(1) Juxtapose the tiles for $\boldsymbol{X}$ and $\boldsymbol{Y}$, i.e., form their outer product, which is an array of rank 4 with typical element $x_{mp} y_{qn}$.
(2) Connect the East arm of $\boldsymbol{X}$ with the West arm of $\boldsymbol{Y}$. This is a contraction, resulting in an array of rank 2, the matrix product $\boldsymbol{XY}$, with typical element $\sum_p x_{mp} y_{pn}$.
(3) Now connect the West arm of $\boldsymbol{X}$ with the East arm of $\boldsymbol{Y}$. The result of this second contraction is a scalar, the trace $\operatorname{tr} \boldsymbol{XY} = \sum_{p,m} x_{mp} y_{pm}$.

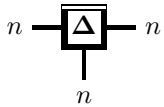
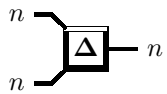
An alternative sequence of operations evaluating this same graph would be

(1) Juxtapose the tiles for $\boldsymbol{X}$ and $\boldsymbol{Y}$.
(2) Connect the West arm of $\boldsymbol{X}$ with the East arm of $\boldsymbol{Y}$ to get the matrix product $\boldsymbol{Y}\boldsymbol{X}$.
(3) Now connect the East arm of $\boldsymbol{X}$ with the West arm of $\boldsymbol{Y}$ to get $\operatorname{tr}\boldsymbol{Y}\boldsymbol{X}$.

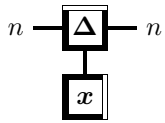
The result is the same, the notation does not specify which of these alternative evaluation paths is meant, and a computer receiving commands based on this notation can choose the most efficient evaluation path. Probably the most efficient evaluation path is given by (B.2.8) below: take the element-by-element product of $\boldsymbol{X}$ with the transpose of $\boldsymbol{Y}$, and add all the elements of the resulting matrix.

If the user specifies $\operatorname{tr}(\boldsymbol{X}\boldsymbol{Y})$, the computer is locked into one evaluation path: it first has to compute the matrix product $\boldsymbol{X}\boldsymbol{Y}$, even if $\boldsymbol{X}$ is a column vector and $\boldsymbol{Y}$ a row vector and it would be much more efficient to compute it as $\operatorname{tr}(\boldsymbol{Y}\boldsymbol{X})$, and then form the trace, i.e., throw away all off-diagonal elements. If the trace is specified as  , the computer can choose the most efficient of a number of different evaluation paths transparently to the user. This advantage of the graphical notation is of course even more important if the graphs are more complex.

There is also the "diagonal" array, which in the case of rank 3 can be written

(B.1.2)  or 

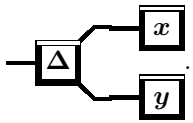
or similar configurations. It has 1's down the main diagonal and 0's elsewhere. It can be used to construct the diagonal matrix $\operatorname{diag}(\boldsymbol{x})$ of a vector (the square matrix with the vector in the diagonal and zeros elsewhere) as

(B.1.3) $\operatorname{diag}(\boldsymbol{x}) =$  ,

the diagonal vector of a square matrix (i.e., the vector containing its diagonal elements) as

(B.1.4) ,

and the "Hadamard product" (element-by-element product) of two vectors $\boldsymbol{x} * \boldsymbol{y}$ as

(B.1.5) $\boldsymbol{x} * \boldsymbol{y} =$  .

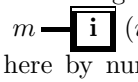
All these are natural operations involving vectors and matrices, but the usual matrix notation cannot represent them and therefore ad-hoc notation must be invented for it. In our graphical representation, however, they all can be built up from a small number of atomic operations, which will be enumerated in Section B.2.

Each such graph can be evaluated in a number of different ways, and all these evaluations give the same result. In principle, each graph can be evaluated as follows: form the outer product of all arrays involved, and then contract along all those pairs of arms which are connected. For practical implementations it is more efficient to develop functions which connect two arrays along one or several of their arms without first forming outer products, and to perform the array concatenations recursively in such a way that contractions are done as early as possible. A computer might be programmed to decide on the most efficient construction path for any given array.

B.2. Axiomatic Development of Array Operations

The following sketch shows how this axiom system might be built up. Since I am an economist I do not plan to develop the material presented here any further. Others are invited to take over. If you are interested in working on this, I would be happy to hear from you; email me at `ehrbar@econ.utah.edu`

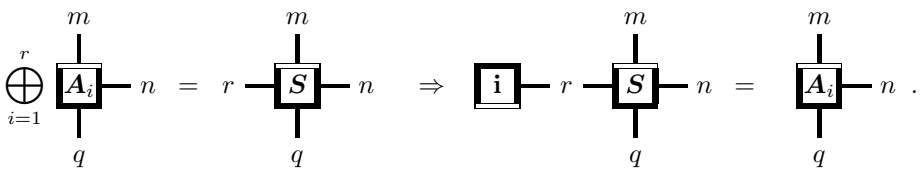
There are two kinds of special arrays: unit vectors and diagonal arrays.

For every natural number $m \geq 1$, *m unit vectors* m  $(i = 1, \ldots, m)$ exist. Despite the fact that the unit vectors are denoted here by numbers, there is no intrinsic ordering among them; they might as well have the names "red, green, blue, ..." (From (B.2.4) and other axioms below it will follow that each unit vector can be represented as a m-vector with 1 as one of the components and 0 elsewhere.)

For every rank ≥ 1 and dimension $n \geq 1$ there is a unique *diagonal array* denoted by $\boldsymbol{\Delta}$. Their main properties are (B.2.1) and (B.2.2). (This and the other axioms must be formulated in such a way that it will be possible to show that the diagonal arrays of rank 1 are the "vectors of ones" $\boldsymbol{\iota}$ which have 1 in every component; diagonal arrays of rank 2 are the identity matrices; and for higher ranks, all arms of a diagonal array have the same dimension, and their $ijk\cdots$ element is 1 if $i = j = k = \cdots$ and 0 otherwise.) Perhaps it makes sense to define the diagonal array of rank 0 and dimension n to be the scalar n, and to declare all arrays which are everywhere 0-dimensional to be diagonal.

There are only three operations of arrays: their outer product, represented by writing them side by side, contraction, represented by the joining of arms, and the direct sum, which will be defined now:

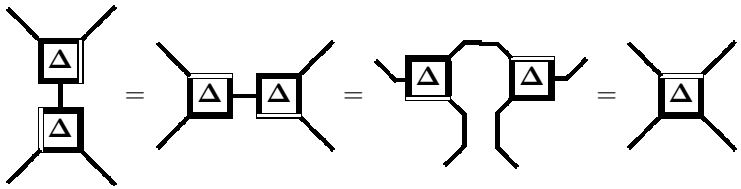
The direct sum is the operation by which a vector can be built up from scalars, a matrix from its row or column vectors, an array of rank 3 from its layers, etc. The direct sum of a set of r similar arrays (i.e., arrays which have the same number of arms, and corresponding arms have the same dimensions) is an array which has one additional arm, called the reference arm of the direct sum. If one "saturates" the reference arm with the ith unit vector, one gets the ith original array back, and this property defines the direct sum uniquely:



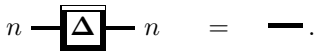
It is impossible to tell which is the first summand and which the second, direct sum is an operation defined on finite sets of arrays (where different elements of a set may be equal to each other in every respect but still have different identities).

There is a broad rule of associativity: the order in which outer products and contractions are performed does not matter, as long as the at the end, the right arms are connected with each other. And there are distributive rules involving (contracted) outer products and direct sums.

Additional rules apply for the special arrays. If two different diagonal arrays join arms, the result is again a diagonal array. For instance, the following three concatenations of diagonal three-way arrays are identical, and they all evaluate to the (for a given dimension) unique diagonal array or rank 4:

(B.2.1) 

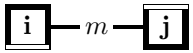
The diagonal array of rank 2 is neutral under concatenation, i.e., it can be written as

(B.2.2) 

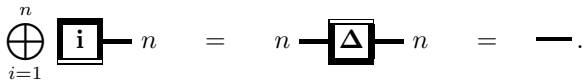
because attaching it to any array will not change this array. (B.2.1) and (B.2.2) make it possible to represent diagonal arrays simply as the branching points of several arms.

This will make the array notation even simpler. However in the present introductory article, all diagonal arrays will be shown explicitly, and the vector of ones will be denoted m — $\boldsymbol{\iota}$ instead of m — $\boldsymbol{\Delta}$ or perhaps m — $\boldsymbol{\delta}$.

Unit vectors concatenate as follows:

(B.2.3)  $= \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases}$

and the direct sum of all unit vectors is the diagonal array of rank 2:

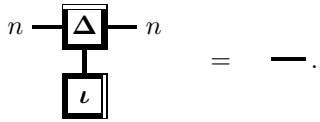
(B.2.4) 

I am sure there will be modifications if one works it all out in detail, but if done right, the number of axioms should be fairly small. Element-by-element addition of arrays is not an axiom because it can be derived: if one saturates the reference arm of a direct sum with the vector of ones, one gets the element-by-element sum of the arrays in this direct sum. Multiplication of an array by a scalar is also contained in the above system of axioms: it is simply the outer product with an array of rank zero.

Problem 603. *Show that the saturation of an arm of a diagonal array with the vector of ones is the same as dropping this arm.*

Answer. Since the vector of ones is the diagonal array of rank 1, this is a special case of the general concantenation rule for diagonal arrays. □

Problem 604. *Show that the diagonal matrix of the vector of ones is the identity matrix, i.e.,*

(B.2.5) 

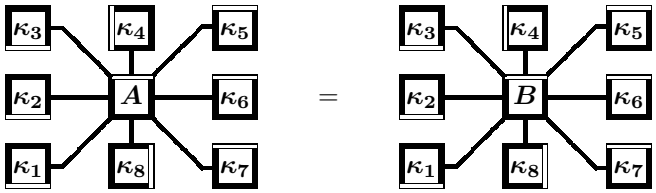
Answer. In view of (B.2.2), this is a special case of Problem 603. □

Problem 605. *A trivial array operation is the addition of an arm of dimension 1; for instance, this is how a n-vector can be turned into a $n \times 1$ matrix. Is this operation contained in the above system of axioms?*

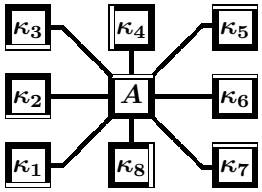
Answer. It is a special case of the direct sum: the direct sum of one array only, the only effect of which is the addition of the reference arm. □

From (B.2.4) and (B.2.2) follows that every array of rank k can be represented as a direct sum of arrays of rank $k-1$, and recursively, as iterated direct sums of those scalars which one gets by saturating all arms with unit vectors. Hence the

following "extensionality property": if the arrays $\boldsymbol{A}$ and $\boldsymbol{B}$ are such that for all possible conformable choices of unit vectors $\kappa_1 \cdots \kappa_8$ follows

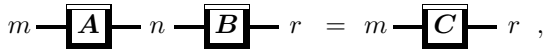
(B.2.6) 

then $\boldsymbol{A} = \boldsymbol{B}$. This is why the saturation of an array with unit vectors can be considered one of its "elements," i.e.,

(B.2.7) 

$= \quad a_{\kappa_1\kappa_2\kappa_3\kappa_4\kappa_5\kappa_6\kappa_7\kappa_8}.$

From (B.2.3) and (B.2.4) follows that the concatenation of two arrays by joining one or more pairs of arms consists in forming all possible products and summing over those subscripts (arms) which are joined to each other. For instance, if



then $c_{\mu\rho} = \sum_{\nu=1}^{n} a_{\mu\nu} b_{\nu\rho}$. This is one of the most basic facts if one thinks of arrays as collections of elements. From this point of view, the proposed notation is simply a graphical elaboration of Einstein's summation convention. But in the holistic approach taken by the proposed system of axioms, which is informed by category theory, it is an implication; it comes at the end, not the beginning.

Instead of considering arrays as bags filled with elements, with the associated false problem of specifying the order in which the elements are packed into the bag, this notation and system of axioms consider each array as an abstract entity, associated with a certain finite graph. These entities can be operated on as specified in the axioms, but the only time they lose their abstract character is when they are fully saturated, i.e., concatenated with each other in such a way that no free arms are left: in this case they become scalars. An array of rank 1 is not the same as a vector, although it can be *represented* as a vector—after an ordering of its elements has been specified. This ordering is not part of the definition of the array itself. (Some vectors, such as time series, have an intrinsic ordering, but I am speaking here of the simplest case where they do not.) Also the ordering of the arms is not specified, and the order in which a set of arrays is packed into its direct sum is not specified

either. These axioms therefore make a strict distinction between the abstract entities themselves (which the user is interested in) and their various representations (which the computer worries about).

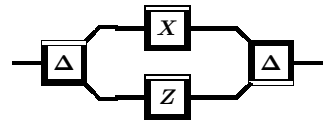
Maybe the following examples may clarify these points. If you specify a set of colors as $\{\text{red}, \text{green}, \text{blue}\}$, then this representation has an ordering built in: red comes first, then green, then blue. However this ordering is not part of the definition of the set; $\{\text{green}, \text{red}, \text{blue}\}$ is the same set. The two notations are two different representations of the same set. Another example: mathematicians usually distinguish between the outer products $\boldsymbol{A} \otimes \boldsymbol{B}$ and $\boldsymbol{B} \otimes \boldsymbol{A}$; there is a "natural isomorphism" between them but they are two different objects. In the system of axioms proposed here these two notations are two different representations of the same object, as in the set example. This object is represented by a graph which has $\boldsymbol{A}$ and $\boldsymbol{B}$ as nodes, but it is not apparent from this graph which node comes first. Interesting conceptual issues are involved here. The proposed axioms are quite different than e.g. [**Mor73**].

Problem 606. *The trace of the product of two matrices can be written as*

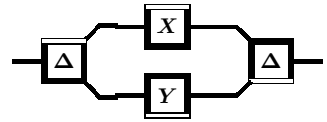
$$\operatorname{tr}(\boldsymbol{X}\boldsymbol{Y}) = \boldsymbol{\iota}^\top (\boldsymbol{X} * \boldsymbol{Y}^\top)\boldsymbol{\iota}. \tag{B.2.8}$$

I.e., one forms the element-by-element product of $\boldsymbol{X}$ *and* $\boldsymbol{Y}^\top$ *and takes the sum of all the elements of the resulting matrix. Use tile notation to show that this gives indeed* $\operatorname{tr}(\boldsymbol{X}\boldsymbol{Y})$.

Answer. In analogy with (B.1.5), the Hadamard product of the two matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$, i.e., their element by element multiplication, is

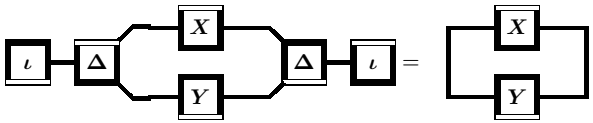
$\boldsymbol{X} * \boldsymbol{Z} =$ 

If $\boldsymbol{Z} = \boldsymbol{Y}^\top$, one gets

$\boldsymbol{X} * \boldsymbol{Y}^\top =$ 

.

Therefore one gets, using (B.2.5):

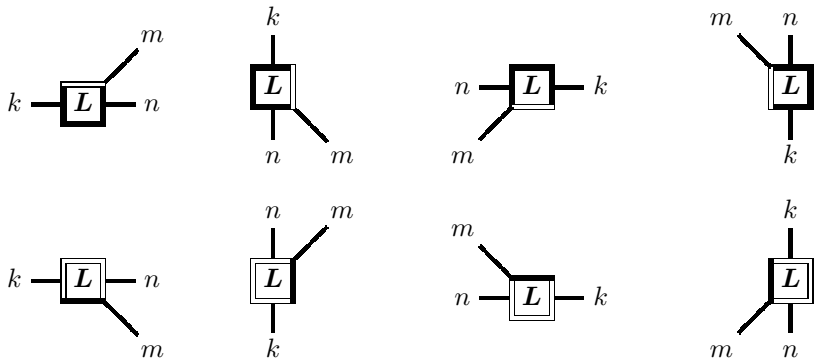
$\boldsymbol{\iota}^\top (\boldsymbol{X} * \boldsymbol{Y}^\top)\boldsymbol{\iota} =$ 

$= \operatorname{tr}(\boldsymbol{X}\boldsymbol{Y})$

□

B.3. An Additional Notational Detail

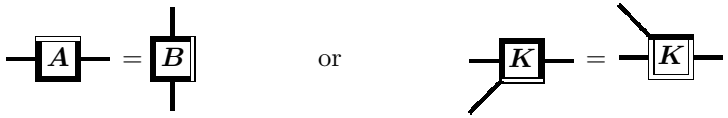
Besides turning a tile by 90, 180, or 270 degrees, the notation proposed here also allows to flip the tile over. The tile (here drawn without its arms) is simply the tile laid on its face; i.e., those parts of the frame, which are black on the side visible to the reader, are white on the opposite side and vice versa. If one flips a tile, the arms appear in a mirror-symmetric manner. For a matrix, flipping over is equivalent to turning by 180 degrees, i.e., there is no difference between the matrix $\boldsymbol{A}$ and the matrix $\boldsymbol{A}$. Since sometimes one and sometimes the other notation seems more natural, both will be used. For higher arrays, flipping over arranges the arms in a different fashion, which is sometimes convenient in order to keep the graphs uncluttered. It will be especially useful for differentiation. If one allows turning in 90 degree increments and flipping, each array can be represented in eight different positions, as shown here with a hypothetical array of rank 3:



The black-and-white pattern at the edge of the tile indicates whether and how much the tile has been turned and/or flipped over, so that one can keep track which arm is which. In the above example, the arm with dimension k will always be called the West arm, whatever position the tile is in.

B.4. Equality of Arrays and Extended Substitution

Given the flexibility of representing the same array in various positions for concatenation, specific conventions are necessary to determine when two such arrays in generalized positions are equal to each other. Expressions like



are not allowed. The arms on both sides of the equal sign must be parallel, in order to make it clear which arm corresponds to which. A permissible way to write the above expressions would therefore be

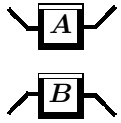
One additional benefit of this tile notation is the ability to substitute arrays with different numbers of arms into an equation. This is also a necessity since the number of possible arms is unbounded. This multiplicity can only be coped with because each arm in an identity written in this notation can be replaced by a bundle of many arms.

Extended substitution also makes it possible to extend definitions familiar from matrices to higher arrays. For instance we want to be able to say that the array $\boldsymbol{\Omega}$ is symmetric if and only if $\boldsymbol{\Omega} = \boldsymbol{\Omega}$. This notion of symmetry is not limited to arrays of rank 2. The arms of this array may symbolize not just a single arm, but whole bundles of arms; for instance an array of the form $\boldsymbol{\Sigma}$ satisfying $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}$ is symmetric according to this definition, and so is every scalar. Also the notion of a nonnegative definite matrix, or of a matrix inverse or generalized inverse, or of a projection matrix, can be extended to arrays in this way.

B.5. Vectorization and Kronecker Product

One conventional generally accepted method to deal with arrays of rank > 2 is the Kronecker product. If $\boldsymbol{A}$ and $\boldsymbol{B}$ are both matrices, then the outer product in tile notation is

(B.5.1)



Since this is an array of rank 4, there is no natural way to write its elements down on a sheet of paper. This is where the Kronecker product steps in. The Kronecker product of two matrices is their outer product written again as a matrix. Its definition includes a protocol how to arrange the elements of an array of rank 4 as a matrix. Alongside the Kronecker product, also the vectorization operator is useful, which is

a protocol how to arrange the elements of a matrix as a vector, and also the so-called "commutation matrices" may become necessary. Here are the relevant definitions:

B.5.1. Vectorization of a Matrix. If $\boldsymbol{A}$ is a matrix, then $\operatorname{vec}(\boldsymbol{A})$ is the vector obtained by stacking the column vectors on top of each other, i.e.,

$$\text{(B.5.2)} \qquad \text{if} \quad \boldsymbol{A} = \begin{bmatrix} \boldsymbol{a}_1 & \cdots & \boldsymbol{a}_n \end{bmatrix} \quad \text{then} \quad \operatorname{vec}(\boldsymbol{A}) = \begin{bmatrix} \boldsymbol{a}_1 \\ \vdots \\ \boldsymbol{a}_n \end{bmatrix}.$$

The vectorization of a matrix is merely a different arrangement of the elements of the matrix on paper, just as the transpose of a matrix.

Problem 607. *Show that* $\operatorname{tr}(\boldsymbol{B}^\top \boldsymbol{C}) = (\operatorname{vec} \boldsymbol{B})^\top \operatorname{vec} \boldsymbol{C}$.

Answer. Both sides are $\sum b_{ji} c_{ji}$. (B.5.28) is a proof in tile notation which does not have to look at the matrices involved element by element. □

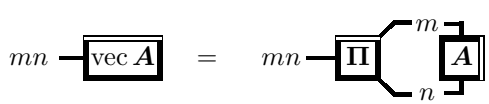
By the way, a better protocol for vectorizing would have been to assemble all rows into one long row vector and then converting it into a column vector. In other words

$$\text{if} \quad \boldsymbol{B} = \begin{bmatrix} \boldsymbol{b}_1^\top \\ \vdots \\ \boldsymbol{b}_m^\top \end{bmatrix} \quad \text{then } \operatorname{vec}(\boldsymbol{B}) \text{ should have been defined as} \quad \begin{bmatrix} \boldsymbol{b}_1 \\ \vdots \\ \boldsymbol{b}_m \end{bmatrix}.$$

The usual protocol of stacking the columns is inconsistent with the lexicograpical ordering used in the Kronecker product. Using the alternative definition, equation (B.5.19) which will be discussed below would be a little more intelligible; it would read

$$\operatorname{vec}(\boldsymbol{ABC}) = (\boldsymbol{A} \otimes \boldsymbol{C}^\top) \operatorname{vec} \boldsymbol{B} \qquad \text{with the alternative definition of vec}$$

and also the definition of vectorization in tile notation would be a little less awkward; instead of (B.5.24) one would have



But this is merely a side remark; we will use the conventional definition (B.5.2) throughout.

B.5.2. Kronecker Product of Matrices. Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be two matrices, say $\boldsymbol{A}$ is $m \times n$ and $\boldsymbol{B}$ is $r \times q$. Their Kronecker product $\boldsymbol{A} \otimes \boldsymbol{B}$ is the $mr \times nq$ matrix which in partitioned form can be written

$$\text{(B.5.3)} \qquad \boldsymbol{A} \otimes \boldsymbol{B} = \begin{bmatrix} a_{11}\boldsymbol{B} & \cdots & a_{1n}\boldsymbol{B} \\ \vdots & \ddots & \vdots \\ a_{m1}\boldsymbol{B} & \cdots & a_{mn}\boldsymbol{B} \end{bmatrix}$$

This convention of how to write the elements of an array of rank 4 as a matrix is not symmetric, so that usually $\boldsymbol{A} \otimes \boldsymbol{C} \neq \boldsymbol{C} \otimes \boldsymbol{A}$. Both Kronecker products represent the same abstract array, but they arrange it differently on the page. However, in many other respects, the Kronecker product maintains the properties of outer products.

Problem 608. [**The71**, pp. 303–306] *Prove the following simple properties of the Kronecker product:*

$$\text{(B.5.4)} \qquad (\boldsymbol{A} \otimes \boldsymbol{B})^\top = \boldsymbol{A}^\top \otimes \boldsymbol{B}^\top$$

$$\text{(B.5.5)} \qquad (\boldsymbol{A} \otimes \boldsymbol{B}) \otimes \boldsymbol{C} = \boldsymbol{A} \otimes (\boldsymbol{B} \otimes \boldsymbol{C})$$

$$\text{(B.5.6)} \qquad \boldsymbol{I} \otimes \boldsymbol{I} = \boldsymbol{I}$$

$$\text{(B.5.7)} \qquad (\boldsymbol{A} \otimes \boldsymbol{B})(\boldsymbol{C} \otimes \boldsymbol{D}) = \boldsymbol{AC} \otimes \boldsymbol{BD}$$

$$\text{(B.5.8)} \qquad (\boldsymbol{A} \otimes \boldsymbol{B})^{-1} = \boldsymbol{A}^{-1} \otimes \boldsymbol{B}^{-1}$$

$$\text{(B.5.9)} \qquad (\boldsymbol{A} \otimes \boldsymbol{B})^{-} = \boldsymbol{A}^{-} \otimes \boldsymbol{B}^{-}$$

$$\text{(B.5.10)} \qquad \boldsymbol{A} \otimes (\boldsymbol{B} + \boldsymbol{C}) = \boldsymbol{A} \otimes \boldsymbol{B} + \boldsymbol{A} \otimes \boldsymbol{C}$$

$$\text{(B.5.11)} \qquad (\boldsymbol{A} + \boldsymbol{B}) \otimes \boldsymbol{C} = \boldsymbol{A} \otimes \boldsymbol{C} + \boldsymbol{B} \otimes \boldsymbol{C}$$

$$\text{(B.5.12)} \qquad (c\boldsymbol{A}) \otimes \boldsymbol{B} = \boldsymbol{A} \otimes (c\boldsymbol{B}) = c(\boldsymbol{A} \otimes \boldsymbol{B})$$

$$\text{(B.5.13)} \qquad \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix} \otimes \boldsymbol{B} = \begin{bmatrix} \boldsymbol{A}_{11} \otimes \boldsymbol{B} & \boldsymbol{A}_{12} \otimes \boldsymbol{B} \\ \boldsymbol{A}_{21} \otimes \boldsymbol{B} & \boldsymbol{A}_{22} \otimes \boldsymbol{B} \end{bmatrix}$$

$$\text{(B.5.14)} \qquad \operatorname{rank}(\boldsymbol{A} \otimes \boldsymbol{B}) = (\operatorname{rank} \boldsymbol{A})(\operatorname{rank} \boldsymbol{B})$$

$$\text{(B.5.15)} \qquad \operatorname{tr}(\boldsymbol{A} \otimes \boldsymbol{B}) = (\operatorname{tr} \boldsymbol{A})(\operatorname{tr} \boldsymbol{B})$$

If a is a 1×1 matrix, then

$$\text{(B.5.16)} \qquad a \otimes \boldsymbol{B} = \boldsymbol{B} \otimes a = a\boldsymbol{B}$$

$$\text{(B.5.17)} \qquad \det(\boldsymbol{A} \otimes \boldsymbol{B}) = (\det(\boldsymbol{A}))^n (\det(\boldsymbol{B}))^k$$

where $\boldsymbol{A}$ is $k \times k$ and $\boldsymbol{B}$ is $n \times n$.

Answer. For the determinant use the following facts: if $\boldsymbol{a}$ is an eigenvector of $\boldsymbol{A}$ with eigenvalue α and $\boldsymbol{b}$ is an eigenvector of $\boldsymbol{B}$ with eigenvalue β, then $\boldsymbol{a} \otimes \boldsymbol{b}$ is an eigenvector of $\boldsymbol{A} \otimes \boldsymbol{B}$ with eigenvalue $\alpha\beta$. The determinant is the product of all eigenvalues (multiple eigenvalues being counted several times). Count how many there are.

An alternative approach would be to write $\boldsymbol{A}\otimes\boldsymbol{B}=(\boldsymbol{A}\otimes\boldsymbol{I})(\boldsymbol{I}\otimes\boldsymbol{B})$ and then to argue that $\det(\boldsymbol{A}\otimes\boldsymbol{I})=(\det(\boldsymbol{A}))^n$ and $\det(\boldsymbol{I}\otimes\boldsymbol{B})=(\det(\boldsymbol{B}))^k$.

The formula for the rank can be shown using $\operatorname{rank}(\boldsymbol{A})=\operatorname{tr}(\boldsymbol{A}\boldsymbol{A}^-)$. compare Problem 566. □

Problem 609. *2 points* [**JHG+88**, pp. 962–4] *Write down the Kronecker product of*

$$\boldsymbol{A}=\begin{bmatrix}1&3\\2&0\end{bmatrix}\qquad\text{and}\qquad\boldsymbol{B}=\begin{bmatrix}2&2&0\\1&0&3\end{bmatrix}. \tag{B.5.18}$$

Show that $\boldsymbol{A}\otimes\boldsymbol{B}\neq\boldsymbol{B}\otimes\boldsymbol{A}$. Which other facts about the outer product do not carry over to the Kronecker product?

Answer.

$$\boldsymbol{A}\otimes\boldsymbol{B}=\begin{bmatrix}2&2&0&6&6&0\\1&0&3&3&0&9\\4&4&0&0&0&0\\2&0&6&0&0&0\end{bmatrix}\qquad\boldsymbol{B}\otimes\boldsymbol{A}=\begin{bmatrix}2&6&2&6&0&0\\4&0&4&0&0&0\\1&3&0&0&3&9\\2&0&0&0&6&0\end{bmatrix}$$

Partitioning of the matrix on the right does not carry over. □

Problem 610. [**JHG+88**, p. 965] *Show that*

$$\operatorname{vec}(\boldsymbol{ABC})=(\boldsymbol{C}^\top\otimes\boldsymbol{A})\operatorname{vec}(\boldsymbol{B}). \tag{B.5.19}$$

Answer. Assume $\boldsymbol{A}$ is $k\times m$, $\boldsymbol{B}$ is $m\times n$, and $\boldsymbol{C}$ is $n\times p$. Write $\boldsymbol{A}=\begin{bmatrix}\boldsymbol{a}_1^\top\\\vdots\\\boldsymbol{a}_k^\top\end{bmatrix}$ and $\boldsymbol{B}=\begin{bmatrix}\boldsymbol{b}_1&\cdots&\boldsymbol{b}_n\end{bmatrix}$. Then $(\boldsymbol{C}^\top\otimes\boldsymbol{A})\operatorname{vec}\boldsymbol{B}=$

$$=\begin{bmatrix}c_{11}\boldsymbol{A}&c_{21}\boldsymbol{A}&\cdots&c_{n1}\boldsymbol{A}\\c_{12}\boldsymbol{A}&c_{22}\boldsymbol{A}&\cdots&c_{n2}\boldsymbol{A}\\\vdots&\vdots&\ddots&\vdots\\c_{1p}\boldsymbol{A}&c_{2p}\boldsymbol{A}&\cdots&c_{np}\boldsymbol{A}\end{bmatrix}\begin{bmatrix}\boldsymbol{b}_1\\\vdots\\\boldsymbol{b}_n\end{bmatrix}=\begin{bmatrix}\begin{bmatrix}c_{11}\boldsymbol{a}_1^\top\boldsymbol{b}_1+c_{21}\boldsymbol{a}_1^\top\boldsymbol{b}_2+\cdots+c_{n1}\boldsymbol{a}_1^\top\boldsymbol{b}_n\\c_{11}\boldsymbol{a}_2^\top\boldsymbol{b}_1+c_{21}\boldsymbol{a}_2^\top\boldsymbol{b}_2+\cdots+c_{n1}\boldsymbol{a}_2^\top\boldsymbol{b}_n\\\vdots\\c_{11}\boldsymbol{a}_k^\top\boldsymbol{b}_1+c_{21}\boldsymbol{a}_k^\top\boldsymbol{b}_2+\cdots+c_{n1}\boldsymbol{a}_k^\top\boldsymbol{b}_n\end{bmatrix}\\\begin{bmatrix}c_{12}\boldsymbol{a}_1^\top\boldsymbol{b}_1+c_{22}\boldsymbol{a}_1^\top\boldsymbol{b}_2+\cdots+c_{n2}\boldsymbol{a}_1^\top\boldsymbol{b}_n\\c_{12}\boldsymbol{a}_2^\top\boldsymbol{b}_1+c_{22}\boldsymbol{a}_2^\top\boldsymbol{b}_2+\cdots+c_{n2}\boldsymbol{a}_2^\top\boldsymbol{b}_n\\\vdots\\c_{12}\boldsymbol{a}_k^\top\boldsymbol{b}_1+c_{22}\boldsymbol{a}_k^\top\boldsymbol{b}_2+\cdots+c_{n2}\boldsymbol{a}_k^\top\boldsymbol{b}_n\end{bmatrix}\\\vdots\\\begin{bmatrix}c_{1p}\boldsymbol{a}_1^\top\boldsymbol{b}_1+c_{2p}\boldsymbol{a}_1^\top\boldsymbol{b}_2+\cdots+c_{np}\boldsymbol{a}_1^\top\boldsymbol{b}_n\\c_{1p}\boldsymbol{a}_2^\top\boldsymbol{b}_1+c_{2p}\boldsymbol{a}_2^\top\boldsymbol{b}_2+\cdots+c_{np}\boldsymbol{a}_2^\top\boldsymbol{b}_n\\\vdots\\c_{1p}\boldsymbol{a}_k^\top\boldsymbol{b}_1+c_{2p}\boldsymbol{a}_k^\top\boldsymbol{b}_2+\cdots+c_{np}\boldsymbol{a}_k^\top\boldsymbol{b}_n\end{bmatrix}\end{bmatrix}.$$

One obtains the same result by vectorizing the matrix

$$\boldsymbol{ABC}=\begin{bmatrix}\boldsymbol{a}_1^\top\boldsymbol{b}_1&\boldsymbol{a}_1^\top\boldsymbol{b}_2&\cdots&\boldsymbol{a}_1^\top\boldsymbol{b}_n\\\boldsymbol{a}_2^\top\boldsymbol{b}_1&\boldsymbol{a}_2^\top\boldsymbol{b}_2&\cdots&\boldsymbol{a}_2^\top\boldsymbol{b}_n\\\vdots&\vdots&\ddots&\vdots\\\boldsymbol{a}_k^\top\boldsymbol{b}_1&\boldsymbol{a}_k^\top\boldsymbol{b}_2&\cdots&\boldsymbol{a}_k^\top\boldsymbol{b}_n\end{bmatrix}\begin{bmatrix}c_{11}&c_{12}&\cdots&c_{1p}\\c_{21}&c_{22}&\cdots&c_{2p}\\\vdots&\vdots&\ddots&\vdots\\c_{n1}&c_{n2}&\cdots&c_{np}\end{bmatrix}=$$

$$=\begin{bmatrix}\boldsymbol{a}_1^\top\boldsymbol{b}_1c_{11}+\boldsymbol{a}_1^\top\boldsymbol{b}_2c_{21}+\cdots+\boldsymbol{a}_1^\top\boldsymbol{b}_nc_{n1}&\boldsymbol{a}_1^\top\boldsymbol{b}_1c_{12}+\boldsymbol{a}_1^\top\boldsymbol{b}_2c_{22}+\cdots+\boldsymbol{a}_1^\top\boldsymbol{b}_nc_{n2}&\cdots&\boldsymbol{a}_1^\top\boldsymbol{b}_1c_{1p}+\boldsymbol{a}_1^\top\boldsymbol{b}_2c_{2p}+\cdots+\boldsymbol{a}_1^\top\boldsymbol{b}_nc_{np}\\\boldsymbol{a}_2^\top\boldsymbol{b}_1c_{11}+\boldsymbol{a}_2^\top\boldsymbol{b}_2c_{21}+\cdots+\boldsymbol{a}_2^\top\boldsymbol{b}_nc_{n1}&\boldsymbol{a}_2^\top\boldsymbol{b}_1c_{12}+\boldsymbol{a}_2^\top\boldsymbol{b}_2c_{22}+\cdots+\boldsymbol{a}_2^\top\boldsymbol{b}_nc_{n2}&\cdots&\boldsymbol{a}_2^\top\boldsymbol{b}_1c_{1p}+\boldsymbol{a}_2^\top\boldsymbol{b}_2c_{2p}+\cdots+\boldsymbol{a}_2^\top\boldsymbol{b}_nc_{np}\\\vdots&\vdots&\ddots&\vdots\\\boldsymbol{a}_k^\top\boldsymbol{b}_1c_{11}+\boldsymbol{a}_k^\top\boldsymbol{b}_2c_{21}+\cdots+\boldsymbol{a}_k^\top\boldsymbol{b}_nc_{n1}&\boldsymbol{a}_k^\top\boldsymbol{b}_1c_{12}+\boldsymbol{a}_k^\top\boldsymbol{b}_2c_{22}+\cdots+\boldsymbol{a}_k^\top\boldsymbol{b}_nc_{n2}&\cdots&\boldsymbol{a}_k^\top\boldsymbol{b}_1c_{1p}+\boldsymbol{a}_k^\top\boldsymbol{b}_2c_{2p}+\cdots+\boldsymbol{a}_k^\top\boldsymbol{b}_nc_{np}\end{bmatrix}.$$

The main challenge in this automatic proof is to fit the many matrix rows, columns, and single elements involved on the same sheet of paper. Among the shuffling of matrix entries, it is easy to lose track of how the result comes about. Later, in equation (B.5.29), a compact and intelligible proof will be given in tile notation.

□

The dispersion of a random matrix $\mathbf{Y}$ is often given as the matrix $\mathcal{V}[\operatorname{vec}\mathbf{Y}]$, where the vectorization is usually not made explicit, i.e., this matrix is denoted $\mathcal{V}[\mathbf{Y}]$.

Problem 611. *If $\mathcal{V}[\operatorname{vec}\mathbf{Y}]=\boldsymbol{\Sigma}\otimes\boldsymbol{\Omega}$ and $\boldsymbol{P}$ and $\boldsymbol{Q}$ are matrices of constants, show that $\mathcal{V}[\operatorname{vec}\boldsymbol{PYQ}]=(\boldsymbol{Q}^\top\boldsymbol{\Sigma}\boldsymbol{Q})\otimes(\boldsymbol{P}\boldsymbol{\Omega}\boldsymbol{P}^\top)$.*

Answer. Apply (B.5.19): $\mathcal{V}[\operatorname{vec}\boldsymbol{PYQ}]=\mathcal{V}[(\boldsymbol{Q}^\top\otimes\boldsymbol{P})\operatorname{vec}\mathbf{Y}]=(\boldsymbol{Q}^\top\otimes\boldsymbol{P})(\boldsymbol{\Sigma}\otimes\boldsymbol{\Omega})(\boldsymbol{Q}\otimes\boldsymbol{P}^\top)$. Now apply (B.5.7). □

Problem 612. *2 points If $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ are vectors, then show that $\operatorname{vec}(\boldsymbol{\alpha}\boldsymbol{\gamma}^\top)=\boldsymbol{\gamma}\otimes\boldsymbol{\alpha}$.*

Answer. One sees this by writing down the matrices, or one can use (B.5.19) with $\boldsymbol{A}=\boldsymbol{\alpha}$, $\boldsymbol{B}=1$, the 1×1 matrix, and $\boldsymbol{C}=\gamma^\top$. □

Problem 613. *2 points If $\boldsymbol{\alpha}$ is a nonrandom vector and $\boldsymbol{\delta}$ a random vector, show that $\mathcal{V}[\boldsymbol{\delta}\otimes\boldsymbol{\alpha}]=\mathcal{V}[\boldsymbol{\delta}]\otimes(\boldsymbol{\alpha}\boldsymbol{\alpha}^\top)$.*

Answer.

$$\boldsymbol{\delta}\otimes\boldsymbol{\alpha} = \begin{bmatrix}\boldsymbol{\alpha}\delta_1\\ \vdots\\ \boldsymbol{\alpha}\delta_n\end{bmatrix} \qquad \mathcal{V}[\boldsymbol{\delta}\otimes\boldsymbol{\alpha}] = \begin{bmatrix}\boldsymbol{\alpha}\operatorname{var}[\delta_1]\boldsymbol{\alpha}^\top & \boldsymbol{\alpha}\operatorname{cov}[\delta_1,\boldsymbol{\delta}_2]\boldsymbol{\alpha}^\top & \cdots & \boldsymbol{\alpha}\operatorname{cov}[\delta_1,\boldsymbol{\delta}_n]\boldsymbol{\alpha}^\top\\ \boldsymbol{\alpha}\operatorname{cov}[\delta_2,\boldsymbol{\delta}_1]\boldsymbol{\alpha}^\top & \boldsymbol{\alpha}\operatorname{var}[\delta_2]\boldsymbol{\alpha}^\top & \cdots & \boldsymbol{\alpha}\operatorname{cov}[\delta_2,\boldsymbol{\delta}_n]\boldsymbol{\alpha}^\top\\ \vdots & \vdots & \ddots & \vdots\\ \boldsymbol{\alpha}\operatorname{cov}[\delta_n,\boldsymbol{\delta}_1]\boldsymbol{\alpha}^\top & \boldsymbol{\alpha}\operatorname{cov}[\delta_n,\boldsymbol{\delta}_2]\boldsymbol{\alpha}^\top & \cdots & \boldsymbol{\alpha}\operatorname{cov}[\delta_n,\boldsymbol{\delta}_n]\boldsymbol{\alpha}^\top\end{bmatrix} =$$

$$= \begin{bmatrix}\operatorname{var}[\delta_1]\boldsymbol{\alpha}\boldsymbol{\alpha}^\top & \operatorname{cov}[\delta_1,\boldsymbol{\delta}_2]\boldsymbol{\alpha}\boldsymbol{\alpha}^\top & \cdots & \operatorname{cov}[\delta_1,\boldsymbol{\delta}_n]\boldsymbol{\alpha}\boldsymbol{\alpha}^\top\\ \operatorname{cov}[\delta_2,\boldsymbol{\delta}_1]\boldsymbol{\alpha}\boldsymbol{\alpha}^\top & \operatorname{var}[\delta_2]\boldsymbol{\alpha}\boldsymbol{\alpha}^\top & \cdots & \operatorname{cov}[\delta_2,\boldsymbol{\delta}_n]\boldsymbol{\alpha}\boldsymbol{\alpha}^\top\\ \vdots & \vdots & \ddots & \vdots\\ \operatorname{cov}[\delta_n,\boldsymbol{\delta}_1]\boldsymbol{\alpha}\boldsymbol{\alpha}^\top & \operatorname{cov}[\delta_n,\boldsymbol{\delta}_2]\boldsymbol{\alpha}\boldsymbol{\alpha}^\top & \cdots & \operatorname{cov}[\delta_n,\boldsymbol{\delta}_n]\boldsymbol{\alpha}\boldsymbol{\alpha}^\top\end{bmatrix} = \mathcal{V}[\boldsymbol{\delta}]\otimes\boldsymbol{\alpha}\boldsymbol{\alpha}^\top$$

□

B.5.3. The Commutation Matrix. Besides the Kronecker product and the vectorization operator, also the "commutation matrix" [**MN88**, pp. 46/7], [**Mag88**, p. 35] is needed for certain operations involving arrays of higher rank. Assume $\boldsymbol{A}$ is $m\times n$. Then the commutation matrix $\boldsymbol{K}^{(m,n)}$ is the $mn\times mn$ matrix which transforms $\operatorname{vec}\boldsymbol{A}$ into $\operatorname{vec}(\boldsymbol{A}^\top)$:

$$\boldsymbol{K}^{(m,n)}\operatorname{vec}\boldsymbol{A} = \operatorname{vec}(\boldsymbol{A}^\top) \tag{B.5.20}$$

The main property of the commutation matrix is that it allows to commute the Kronecker product. For any $m\times n$ matrix $\boldsymbol{A}$ and $r\times q$ matrix $\boldsymbol{B}$ follows

$$\boldsymbol{K}^{(r,m)}(\boldsymbol{A}\otimes\boldsymbol{B})\boldsymbol{K}^{(n,q)} = \boldsymbol{B}\otimes\boldsymbol{A} \tag{B.5.21}$$

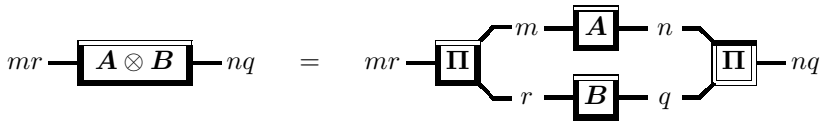
Problem 614. *Use* (B.5.20) *to compute* $\boldsymbol{K}^{(2,3)}$.

Answer.

$$\boldsymbol{K}^{(2,3)} = \begin{bmatrix}1&0&0&0&0&0\\0&0&1&0&0&0\\0&0&0&0&1&0\\0&1&0&0&0&0\\0&0&0&1&0&0\\0&0&0&0&0&1\end{bmatrix} \tag{B.5.22}$$

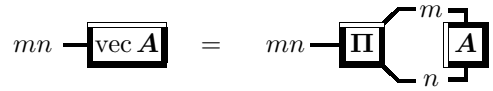
□

B.5.4. Kronecker Product and Vectorization in Tile Notation. The Kronecker product of m — $\boxed{\boldsymbol{A}}$ — n and r — $\boxed{\boldsymbol{B}}$ — q is the following concatenation of $\boldsymbol{A}$ and $\boldsymbol{B}$ with members of a certain family of three-way arrays $\boldsymbol{\Pi}^{(i,j)}$:

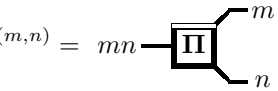
(B.5.23) 

Strictly speaking we should have written $\boldsymbol{\Pi}^{(m,r)}$ and $\boldsymbol{\Pi}^{(n,q)}$ for the two $\boldsymbol{\Pi}$-arrays in (B.5.23), but the superscripts can be inferred from the context: the first superscript is the dimension of the Northeast arm, and the second that of the Southeast arm.

Vectorization uses a member of the same family $\boldsymbol{\Pi}^{(m,n)}$ to convert the matrix n — $\boxed{\boldsymbol{A}}$ — m into the vector

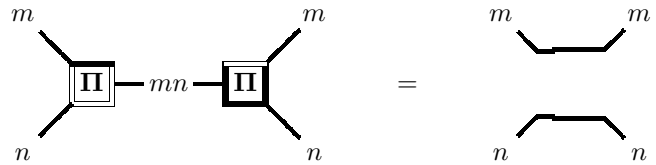
(B.5.24) 

This equation is a little awkward because the $\boldsymbol{A}$ is here a $n\times m$ matrix, while elsewhere it is a $m\times n$ matrix. It would have been more consistent with the lexicographical ordering used in the Kronecker product to define vectorization as the stacking of the row vectors; then some of the formulas would have looked more natural.

The array $\boldsymbol{\Pi}^{(m,n)} =$ 

exists for every $m\geq 1$ and $n\geq 1$. The dimension of the West arm is always the product of the dimensions of the two East arms. The elements of $\boldsymbol{\Pi}^{(m,n)}$ will be given in (B.5.30) below; but first I will list three important properties of these arrays and give examples of their application.

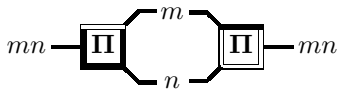
First of all, each $\boldsymbol{\Pi}^{(m,n)}$ satisfies

(B.5.25) 

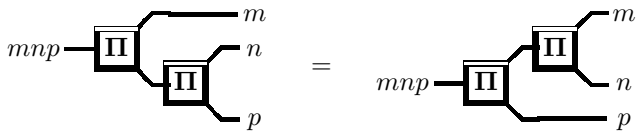
.

Let us discuss the meaning of (B.5.25) in detail. The lefthand side of (B.5.25) shows the concatenation of two copies of the three-way array $\boldsymbol{\Pi}^{(m,n)}$ in a certain way that yields a 4-way array. Now look at the righthand side. The arm m — m by itself (which was bent only in order to remove any doubt about which arm to the left of the equal sign corresponds to which arm to the right) represents the neutral element under concatenation (i.e., the $m\times m$ identity matrix). Writing two arrays next to each other without joining any arms represents their outer product, i.e., the array whose rank is the sum of the ranks of the arrays involved, and whose elements are all possible products of elements of the first array with elements of the second array.

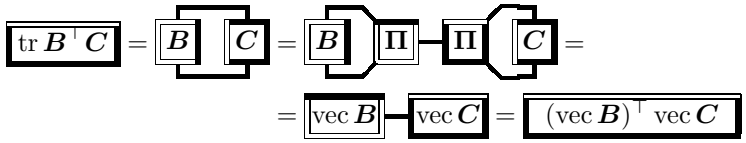
The second identity satisfied by $\mathbf{\Pi}^{(m,n)}$ is

(B.5.26) 

Finally, there is also associativity:

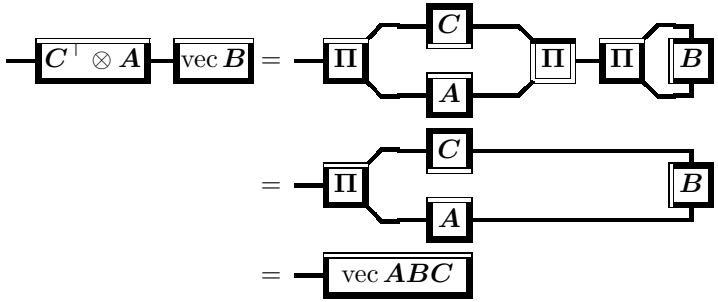
(B.5.27) 

Here is the answer to Problem 607 in tile notation:



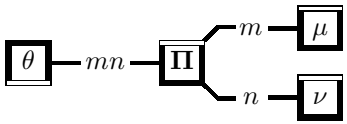
(B.5.28)

Equation (B.5.25) was central for obtaining the result. The answer to Problem 610 also relies on equation (B.5.25):



(B.5.29)

B.5.5. Looking Inside the Kronecker Arrays. It is necessary to open up the arrays from the $\boxed{\mathbf{\Pi}}$-family and look at them "element by element," in order to verify (B.5.23), (B.5.24), (B.5.25), (B.5.26), and (B.5.27). The elements of $\mathbf{\Pi}^{(m,n)}$, which can be written in tile notation by saturating the array with unit vectors, are

(B.5.30) $\pi^{(m,n)}_{\theta\mu\nu} =$ 

$= \begin{cases} 1 & \text{if } \theta = (\mu-1)n + \nu \\ 0 & \text{otherwise.} \end{cases}$

Note that for every θ there is exactly one μ and one ν such that $\pi^{(m,n)}_{\theta\mu\nu} = 1$; for all other values of μ and ν, $\pi^{(m,n)}_{\theta\mu\nu} = 0$.

Writing $\boxed{\boldsymbol{\nu}}\!-\!\boxed{\boldsymbol{A}}\!-\!\boxed{\boldsymbol{\mu}} = a_{\nu\mu}$ and $\boxed{\boldsymbol{\theta}}\!-\!\boxed{\operatorname{vec}\boldsymbol{A}} = c_\theta$, (B.5.24) reads

$$c_\theta = \sum_{\mu,\nu} \pi^{(m,n)}_{\theta\mu\nu} a_{\nu\mu}, \tag{B.5.31}$$

which coincides with definition (B.5.2) of $\operatorname{vec}\boldsymbol{A}$.

One also checks that (B.5.23) is (B.5.3). Calling $\boldsymbol{A}\otimes\boldsymbol{B} = \boldsymbol{C}$, it follows from (B.5.23) that

$$c_{\phi\theta} = \sum_{\mu,\nu,\rho,\kappa} \pi^{(m,r)}_{\phi\mu\rho} a_{\mu\nu} b_{\rho\kappa} \pi^{(n,q)}_{\theta\nu\kappa}. \tag{B.5.32}$$

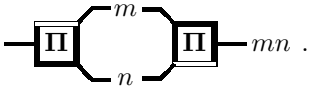
For $1 \le \phi \le r$ one gets a nonzero $\pi^{(m,r)}_{\phi\mu\rho}$ only for $\mu = 1$ and $\rho = \phi$, and for $1 \le \theta \le q$ one gets a nonzero $\pi^{(n,q)}_{\theta\nu\kappa}$ only for $\nu = 1$ and $\kappa = \theta$. Therefore $c_{\phi\theta} = a_{11} b_{\phi\theta}$ for all elements of matrix $\boldsymbol{C}$ with $\phi \le r$ and $\theta \le q$. Etc.

The proof of (B.5.25) uses the fact that for every θ there is exactly one μ and one ν such that $\pi^{(m,n)}_{\theta\mu\nu} \neq 0$:

$$\sum_{\theta=1}^{\theta=mn} \pi^{(m,n)}_{\theta\mu\nu} \pi^{(m,n)}_{\theta\omega\sigma} = \begin{cases} 1 & \text{if } \mu = \omega \text{ and } \nu = \sigma \\ 0 & \text{otherwise} \end{cases} \tag{B.5.33}$$

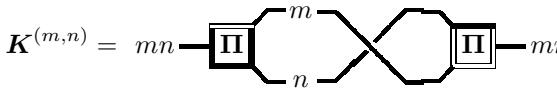
Similarly, (B.5.26) and (B.5.27) can be shown by elementary but tedious proofs. The best verification of these rules is their implementation in a computer language, see Section **??** below.

B.5.6. The Commutation Matrix in Tile Notation. The simplest way to represent the commutation matrix $\boldsymbol{K}^{(m,n)}$ in a tile is

$\boldsymbol{K}^{(m,n)} =$ 

(B.5.34)

This should not be confused with the lefthand side of (B.5.26): $\boldsymbol{K}^{(m,n)}$ is composed of $\mathbf{\Pi}^{(m,n)}$ on its West and $\mathbf{\Pi}^{(n,m)}$ on its East side, while (B.5.26) contains $\mathbf{\Pi}^{(m,n)}$ twice. We will therefore use the following representation, mathematically equivalent to (B.5.34), which makes it easier to see the effects of $\boldsymbol{K}^{(m,n)}$:

$\boldsymbol{K}^{(m,n)} =$ 

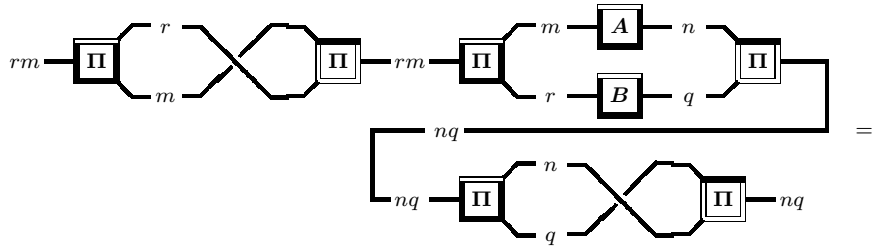
(B.5.35)

PROBLEM 615. *Using the definition* (B.5.35) *show that* $\boldsymbol{K}^{(m,n)}\boldsymbol{K}^{(n,m)} = \boldsymbol{I}_{mn}$, *the* $mn \times mn$ *identity matrix.*

ANSWER. You will need (B.5.25) and (B.5.26). □

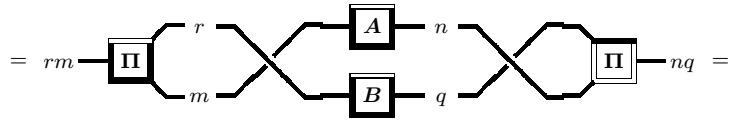
PROBLEM 616. *Prove* (B.5.21) *in tile notation.*

ANSWER. Start with a tile representation of $\boldsymbol{K}^{(r,m)}(\boldsymbol{A} \otimes \boldsymbol{B})\boldsymbol{K}^{(n,q)}$:

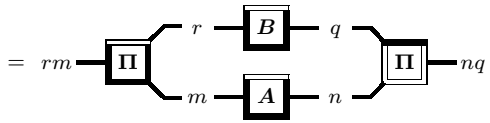


=

Now use (B.5.25) twice to get



=



.

□

APPENDIX C

Matrix Differentiation

C.1. First Derivatives

Let us first consider the scalar case and then generalize from there. The derivative of a function f is often written

$$\frac{dy}{dx} = f'(x) \tag{C.1.1}$$

Multiply through by dx to get $dy = f'(x)\,dx$. In order to see the meaning of this equation, we must know the definition $dy = f(x+dx) - f(x)$. Therefore one obtains $f(x+dx) = f(x) + f'(x)\,dx$. If one holds x constant and only varies dx this formula shows that in an infinitesimal neighborhood of x, the function f is an *affine* function of dx, i.e., a linear function of dx with a constant term: $f(x)$ is the intercept, i.e., the value for $dx = 0$, and $f'(x)$ is the slope parameter.

Now let us transfer this argument to vector functions $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x})$. Here $\boldsymbol{y}$ is a n-vector and $\boldsymbol{x}$ a m-vector, i.e., $\boldsymbol{f}$ is a n-tuple of functions of m variables each

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} f_1(x_1, \ldots, x_m) \\ \vdots \\ f_n(x_1, \ldots, x_m) \end{bmatrix} \tag{C.1.2}$$

One may also say, $\boldsymbol{f}$ is a n-vector, each element of which depends on $\boldsymbol{x}$. Again, under certain differentiability conditions, it is possible to write this function infinitesimally as an affine function, i.e., one can write

$$\boldsymbol{f}(\boldsymbol{x} + \boldsymbol{dx}) = \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{A}\boldsymbol{dx}. \tag{C.1.3}$$

Here the coefficient of $\boldsymbol{dx}$ is no longer a scalar but necessarily a *matrix* $\boldsymbol{A}$ (whose elements again depend on $\boldsymbol{x}$). $\boldsymbol{A}$ is called the *Jacobian matrix* of $\boldsymbol{f}$. The Jacobian matrix generalizes the concept of a derivative to vectors. Instead of a prime denoting the derivative, as in $f'(x)$, one writes $\boldsymbol{A} = \boldsymbol{D}\boldsymbol{f}$.

Problem 617. *2 points If f is a scalar function of a vector argument $\boldsymbol{x}$, is its Jacobian matrix $\boldsymbol{A}$ a row vector or a column vector? Explain why this must be so.*

The Jacobian $\boldsymbol{A}$ defined in this way turns out to have a very simple functional form: its elements are the partial derivatives of all components of $\boldsymbol{f}$ with respect to all components of $\boldsymbol{x}$:

$$a_{ij} = \frac{\partial f_i}{\partial x_j}. \tag{C.1.4}$$

Since in this matrix $\boldsymbol{f}$ acts as column and $\boldsymbol{x}$ as a row vector, this matrix can be written, using matrix differentiation notation, as $\boldsymbol{A}(\boldsymbol{x}) = \partial\boldsymbol{f}(\boldsymbol{x})/\partial\boldsymbol{x}^\top$.

Strictly speaking, *matrix* notation can be used for matrix differentiation only if we differentiate a column vector (or scalar) with respect to a row vector (or scalar), or if we differentiate a scalar with respect to a matrix or a matrix with respect to a scalar. If we want to differentiate matrices with respect to vectors or vectors with respect to matrices or matrices with respect to each other, we need the tile notation for arrays. A different, much less enlightening approach is to first "vectorize" the matrices involved. Both of those methods will be discussed later.

If the dependence of $\boldsymbol{y}$ on $\boldsymbol{x}$ can be expressed in terms of matrix operations or more general array concatenations, then some useful matrix differentiation rules exist.

The simplest matrix differentiation rule, for $f(\boldsymbol{x}) = \boldsymbol{w}^\top\boldsymbol{x}$ with

$$\boldsymbol{w} = \begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix} \quad \text{and} \quad \boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \tag{C.1.5}$$

is

$$\partial\boldsymbol{w}^\top\boldsymbol{x}/\partial\boldsymbol{x}^\top = \boldsymbol{w}^\top \tag{C.1.6}$$

Here is the proof of (C.1.6):

$$\begin{aligned} \frac{\partial\boldsymbol{w}^\top\boldsymbol{x}}{\partial\boldsymbol{x}^\top} &= \begin{bmatrix} \frac{\partial}{\partial x_1}(w_1x_1 + \cdots + w_nx_n) & \cdots & \frac{\partial}{\partial x_n}(w_1x_1 + \cdots + w_nx_n) \end{bmatrix} \\ &= \begin{bmatrix} w_1 & \cdots & w_n \end{bmatrix} = \boldsymbol{w}^\top \end{aligned}$$

The second rule, for $f(\boldsymbol{x}) = \boldsymbol{x}^\top\boldsymbol{M}\boldsymbol{x}$ and $\boldsymbol{M}$ symmetric, is:

$$\partial\boldsymbol{x}^\top\boldsymbol{M}\boldsymbol{x}/\partial\boldsymbol{x}^\top = 2\boldsymbol{x}^\top\boldsymbol{M}. \tag{C.1.7}$$

To show (C.1.7), write

$$\begin{array}{rcccccccccc}
\boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x} & = & x_1 m_{11} x_1 & + & x_1 m_{12} x_2 & + & \cdots & + & x_1 m_{1n} x_n & + \\
 & + & x_2 m_{21} x_1 & + & x_2 m_{22} x_2 & + & \cdots & + & x_2 m_{2n} x_n & + \\
 & & \vdots & & \vdots & & & & \vdots & \\
 & + & x_n m_{n1} x_1 & + & x_n m_{n2} x_2 & + & \cdots & + & x_n m_{nn} x_n &
\end{array}$$

and take the partial derivative of this sum with respect to each of the x_i. For instance, differentiation with respect to x_1 gives

$$\begin{array}{rcccccccccc}
\partial \boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x}/\partial x_1 & = & 2m_{11}x_1 & + & m_{12}x_2 & + & \cdots & + & m_{1n}x_n & + \\
 & + & x_2 m_{21} & + & & & & & & \\
 & & \vdots & & & & & & & \\
 & + & x_n m_{n1} & & & & & & &
\end{array}$$

Now split the upper diagonal element, writing it as $m_{11}x_1 + x_1 m_{11}$, to get

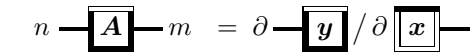
$$\begin{array}{rcccccccccc}
 & = & m_{11}x_1 & + & m_{12}x_2 & + & \cdots & + & m_{1n}x_n & + \\
 & + & x_1 m_{11} & + & & & & & & \\
 & + & x_2 m_{21} & + & & & & & & \\
 & & \vdots & & & & & & & \\
 & + & x_n m_{n1} & & & & & & &
\end{array}$$

The sum of the elements in the first row is the first element of the column vector $\boldsymbol{M}\boldsymbol{x}$, and the sum of the elements in the column underneath is the first element of the row vector $\boldsymbol{x}^\top \boldsymbol{M}$. Overall this has to be arranged as a row vector, since we differentiate with respect to $\partial \boldsymbol{x}^\top$, therefore we get

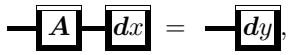
(C.1.8) $$\partial \boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x}/\partial \boldsymbol{x}^\top = \boldsymbol{x}^\top(\boldsymbol{M} + \boldsymbol{M}^\top).$$

This is true for arbitrary $\boldsymbol{M}$, and for symmetric $\boldsymbol{M}$, it simplifies to (C.1.7). The formula for symmetric $\boldsymbol{M}$ is all we need, since a quadratic form with an unsymmetric $\boldsymbol{M}$ is identical to that with the symmetric $(\boldsymbol{M} + \boldsymbol{M}^\top)/2$.

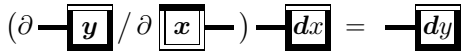
Here is the tile notation for matrix differentiation: If n — $\boxed{\boldsymbol{y}}$ depends on m — $\boxed{\boldsymbol{x}}$, then

(C.1.9) 

is that array which satisfies

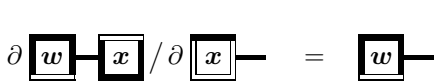
(C.1.10) 

i.e.,

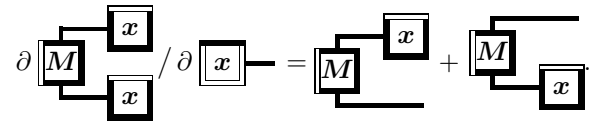
(C.1.11) 

Extended substitutability applies here: n — $\boxed{\boldsymbol{y}}$ and m — $\boxed{\boldsymbol{x}}$ are not necessarily vectors; the arms with dimension m and n can represent different bundles of several arms.

In tiles, (C.1.6) is

(C.1.12) 

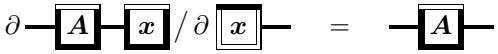
and (C.1.8) is

(C.1.13) 

In (C.1.6) and (C.1.7), we took the derivatives of scalars with respect to vectors. The simplest example of a derivative of a vector with respect to a vector is a linear function. This gives the most basic matrix differentiation rule: If $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ is a linear vector function, then its derivative is that same linear vector function:

(C.1.14) $$\partial \boldsymbol{A}\boldsymbol{x}/\partial \boldsymbol{x}^\top = \boldsymbol{A},$$

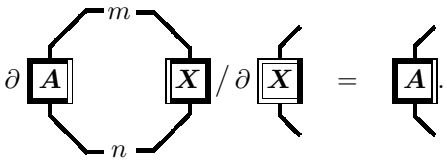
or in tiles

(C.1.15) 

Problem 618. *Show that*

(C.1.16) $$\frac{\partial \operatorname{tr} \boldsymbol{A}\boldsymbol{X}}{\partial \boldsymbol{X}^\top} = \boldsymbol{A}.$$

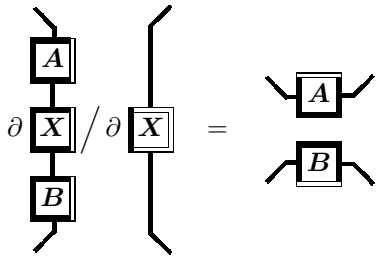
In tiles it reads

(C.1.17) 

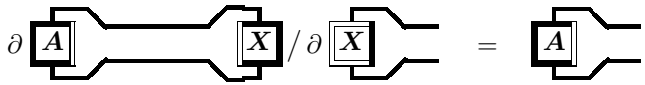
Answer. $\operatorname{tr}(\boldsymbol{A}\boldsymbol{X}) = \sum_{i,j} a_{ij} x_{ji}$ i.e., the coefficient of x_{ji} is a_{ij}. □

Here is a differentiation rule for a matrix with respect to a matrix, first written element by element, and then in tiles: If $\boldsymbol{Y} = \boldsymbol{AXB}$, i.e., $y_{im} = \sum_{j,k} a_{ij} x_{jk} b_{km}$, then $\frac{\partial y_{im}}{\partial x_{jk}} = a_{ij} a_{km}$, because for every fixed i and m this sum contains only one term which has x_{jk} in it, namely, $a_{ij} x_{jk} b_{km}$. In tiles:

(C.1.18)



Equations (C.1.17) and (C.1.18) can be obtained from (C.1.12) and (C.1.15) by extended substitution, since a bundle of several arms can always be considered as one arm. For instance, (C.1.17) can be written

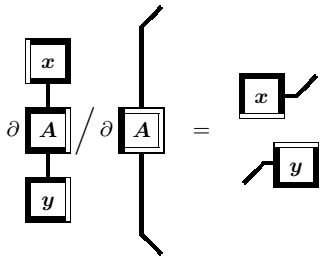


and this is a special case of (C.1.12), since the two parallel arms can be treated as one arm. With a better development of the logic underlying this notation, it will not be necessary to formulate them as separate theorems; all matrix differentiation rules given so far are trivial applications of (C.1.15).

Problem 619. *As a special case of* (C.1.18) *show that* $\frac{\partial x^\top \boldsymbol{A} y}{\partial \boldsymbol{A}^\top} = y x^\top$.

Answer.

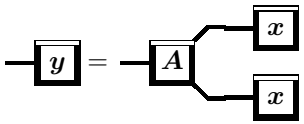
(C.1.19)



□

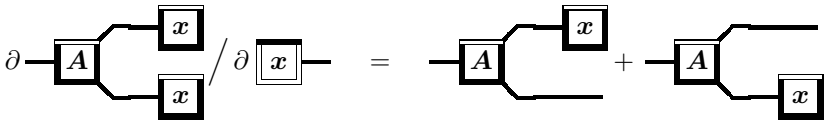
Here is a basic differentiation rule for *bilinear* array concatenations: if

(C.1.20)



then one gets the following simple generalization of (C.1.13):

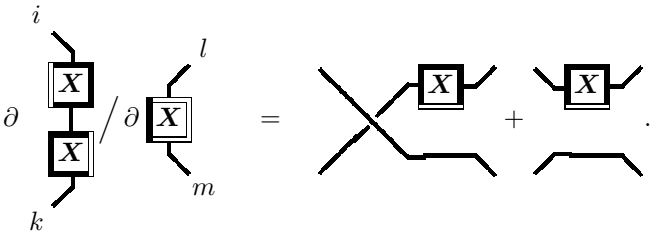
(C.1.21)



Proof. $y_i = \sum_{j,k} a_{ijk} x_j x_k$. For a given i, this has x_p^2 in the term $a_{ipp} x_p^2$, and it has x_p in the terms $a_{ipk} x_p x_k$ where $p \neq k$, and in $a_{ijp} x_j x_p$ where $j \neq p$. The derivatives of these terms are $2a_{ipp} x_p + \sum_{k \neq p} a_{ipk} x_k + \sum_{j \neq p} a_{ijp} x_j$, which simplifies to $\sum_k a_{ipk} x_k + \sum_j a_{ijp} x_j$. This is the i, p-element of the matrix on the rhs of (C.1.21). □

But there are also other ways to have the array $\boldsymbol{X}$ occur twice in a concatenation $\boldsymbol{Y}$. If $\boldsymbol{Y} = \boldsymbol{X}^\top \boldsymbol{X}$ then $y_{ik} = \sum_j x_{ji} x_{jk}$ and therefore $\partial y_{ik} / \partial x_{lm} = 0$ if $m \neq i$ and $m \neq k$. Now assume $m = i \neq k$: $\partial y_{ik} / \partial x_{li} = \partial x_{li} x_{lk} / \partial x_{li} = x_{lk}$. Now assume $m = k \neq i$: $\partial y_{ik} / \partial x_{lk} = \partial x_{li} x_{lk} / \partial x_{lk} = x_{li}$. And if $m = k = i$ then one gets the sum of the two above: $\partial y_{ii} / \partial x_{li} = \partial x_{li}^2 / \partial x_{li} = 2x_{li}$. In tiles this is

(C.1.22)



This rule is helpful for differentiating the multivariate Normal likelihood function.

A computer implementation of this tile notation should contain algorithms to automatically take the derivatives of these array concatenations.

Here are some more matrix differentiation rules:

Chain rule: If $\boldsymbol{g} = \boldsymbol{g}(\boldsymbol{\eta})$ and $\boldsymbol{\eta} = \boldsymbol{\eta}(\boldsymbol{\beta})$ are two vector functions, then

$$\partial \boldsymbol{g} / \partial \boldsymbol{\beta}^\top = \partial \boldsymbol{g} / \partial \boldsymbol{\eta}^\top \cdot \partial \boldsymbol{\eta} / \partial \boldsymbol{\beta}^\top \tag{C.1.23}$$

For instance, the linear least squares objective function is $\mathsf{SSE} = (\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}) = \hat{\boldsymbol{\varepsilon}}^\top \hat{\boldsymbol{\varepsilon}}$ where $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}$. Application of the chain rule gives $\partial \mathsf{SSE} / \partial \boldsymbol{\beta}^\top = \partial \mathsf{SSE} / \partial \hat{\boldsymbol{\varepsilon}}^\top \cdot \partial \hat{\boldsymbol{\varepsilon}} / \partial \boldsymbol{\beta}^\top = 2\hat{\boldsymbol{\varepsilon}}^\top (-\boldsymbol{X})$ which is the same result as in (18.2.2).

If $\boldsymbol{A}$ is nonsingular then

$$\frac{\partial \log \det \boldsymbol{A}}{\partial \boldsymbol{A}^\top} = \boldsymbol{A}^{-1} \tag{C.1.24}$$

Proof in [**Gre97**, pp. 52/3].

Bibliography

[AD75] J. Aczél and Z. Daróczy. *On Measures of Information and their Characterizations*. Academic Press, 1975.

[AG97] Gianni Amisano and Carlo Giannini. *Topics in Structural VAR Econometrics*. Springer-Verlag, Berlin, New York, 2nd, rev. and enl. edition, 1997.

[Alb69] Arthur E. Albert. Conditions for positive and negative semidefiniteness in terms of pseudoinverses. *SIAM (Society for Industrial and Applied Mathematics) Journal of Applied Mathematics*, 17:434–440, 1969.

[Alb72] Arthur E. Albert. *Regression and the Moore-Penrose Pseudoinverse*. Academic Press, New York and London, 1972.

[Ame85] Takeshi Amemiya. *Advanced Econometrics*. Harvard University Press, 1985.

[Ame94] Takeshi Amemiya. *Introduction to Statistics and Econometrics*. Harvard University Press, Cambridge, MA, 1994.

[And66] T. W. Anderson. The choice of the degree of a polynomial regression as a multiple decision problem. *Annals of Mathematical Statistics*, 33(1):255–265, 1966.

[And71] T. W. Anderson. *The Statistical Analysis of Time Series*. Wiley, New York, 1971.

[AO85] T. W. Anderson and I. Olkin. Maximum-likelihood estimation of the parameters of a multivariate normal distribution. *Linear Algebra and its Applications*, 70:147–171, 1985.

[Ati62] M. Atiqullah. The estimation of residual variance in quadratically balanced least squares problems and the robustness of the F-test. *Biometrika*, 49:83–91, 1962.

[BA97] Adrian W. Bowman and Adelchi Azzalini. *Applied Smoothing Techniques for Data Analysis*. Clarendon, Oxford, 1997.

[Bar63] G. A. Barnard. The logic of least squares. *Journal of the Royal Statistical Society, Series B*, 25:124–127, 1963.

[Bar64] A. J. Baranchik. Multiple regression and estimation of the mean of a multivariate normal distribution. Technical Report 51, Department of Statistics, Stanford University, Stanford, 1964.

[Bar74] Yonathan Bard. *Nonlinear Parameter Estimation*. Academic Press, 1974.

[Bar82] Vic Barnett. *Comparative Statistical Inference*. Wiley, New York, 1982.

[BCS96] Andreas Buja, Dianne Cook, and Deborah F. Swayne. Interactive high-dimensional data visualization. *Journal of Computational and Graphical Statistics*, 5:78–99, 1996. A companion video tape can be borrowed from the ASA Statistical Graphics Section, lending library.

[BCW96] Richard A. Becker, John M. Chambers, and Allan R. Wilks. *The New S Language: A Programming Environment for Data Analysis and Graphics*. Chapman and Hall, 1996. Reprint of the 1988 Wadsworth edition.

[BD77] Peter J. Bickel and Kjell A. Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics*. Holden-Day, San Francisco, 1977.

[Ber91] Ernst R. Berndt. *The Practice of Econometrics: Classic and Contemporary*. Addison-Wesley, Reading, Massachusetts, 1991.

[BF85] Leo Breiman and Jerome H. Friedman. Estimating optimal transformations for multiple regression and correlation. *JASA*, 80(391):580–619, 1985.

[BF91] F. A. G. den Butter and M. M. G. Fase. *Seasonal Adjustment as a Practical Problem*. North-Holland, 1991.

[Bha78] Roy Bhaskar. *A Realist Theory of Science*. Harvester Wheatsheaf, London and New York, second edition, 1978.

[Bha93] Roy Bhaskar. *Dialectic: The Pulse of Freedom*. Verso, London, New York, 1993.

[BJ76] George E. P. Box and Gwilym M. Jenkins. *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco, revised edition, 1976.

[BKW80] David A. Belsley, Edwin Kuh, and Roy E. Welsch. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley, New York, 1980.

[Bla73] R. C. Blattberg. Evaluation of the power of the Durbin-Watson statistic for non-first order serial correlation alternatives. *Review of Economics and Statistics*, 55:508–515, August 1973.

[BLR99] Luc Bauwens, Michel Lubrano, and Jean-Fran¸ois Richard. *Bayesian Inference in Dynamic Econometric Models*. Oxford University Press, 1999. Data sets at http://www.core.ucl.ac.be/econometrics/index.htm.

[BM78] Charles M. Beach and James G. MacKinnon. A maximum likelihood procedure for regression with autocorrelated errors. *Econometrica*, 46(1):51–58, 1978.

[BQ89] Olivier Jean Blanchard and Danny Quah. The dynamic effect of demand and supply disturbances. *American Economic Review*, 79:655–673, 1989.

[Bra68] James Vandiver Bradley. *Distribution-Free Statistical Tests*. Prentice-Hall, Englewood Cliffs, N.J., 1968.

[BT99] Kaye E. Basford and J. W. Tukey. *Graphical Analysis of Multiresponse Data: Illustrated with a Plant Breeding Trial*. Interdisciplinary Statistics. Chapman & Hall/CRC, Boca Raton, Fla., 1999.

[Buj90] Andreas Buja. Remarks on functional canonical variates, alternating least squares methods and ACE. *The Annals of Statistics*, 18(3):1032–1069, 1990.

[Bur98] Patrick J. Burns. S poetry. `www.seanet.com/~pburns/Spoetry`, 1998.

[Cam89] Mike Camden. *The Data Bundle*. New Zealand Statistical Association, Wellington, New Zealand, 1989.

[CB97] Dianne Cook and Andreas Buja. Manual controls for high-dimensional data projections. *Journal of Computational and Graphical Statistics*, 1997.

[CBCH97] Dianne Cook, Andreas Buja, J. Cabrera, and H. Hurley. Grand tour and projection pursuit. *Journal of Computational and Graphical Statistics*, 2(3):225–250, 1997.

[CCCM81] M. Cameron, K. D. Collerson, W. Compston, and R. Morton. The statistical analysis and interpretation of imperfectly-fitted Rb-Sr isochrons from polymetamorphic terrains. *Geochimica et Geophysica Acta*, 45:1087–1097, 1981.

[CD28] Charles W. Cobb and Paul H. Douglas. A theory of production. *American Economic Review*, 18(1, Suppl.):139–165, 1928. J.

[CD97] Wojciech W. Charemza and Derek F. Deadman. *New Directions in Econometric Practice: General to Specific Modelling, Cointegration, and Vector Autoregression*. Edward Elgar, Cheltenham, UK; Lynne, NH, 2nd ed. edition, 1997.

[CH93] John M. Chambers and Trevor Hastie, editors. *Statistical Models in S*. Chapman and Hall, 1993.

[Cha96] B. G. Charlton. Should epidemiologists be pragmatists, biostatisticians, or clinical scientists? *Epidemiology*, 7(5):552–4, 1996.

[Cho60] G. C. Chow. Tests of equality between sets of coefficients in two linear regressions. *Econometrica*, 28:591–605, July 1960.

[Chr87] Ronald Christensen. *Plane Answers to Complex Questions; The Theory of Linear Models*. Springer-Verlag, New York, 1987.

[Coh50] A. C. Cohen. Estimating the mean and variance of normal populations from singly and doubly truncated samples. *Annals of Mathematical Statistics*, pages 557–569, 1950.

[Col89] Andrew Collier. *Scientific Realism and Socialist Thought*. Harvester Wheatsheaf and Lynne Rienner, Hertfordshire, U.K. and Boulder, Colorado, 1989.

[Coo77] R. Dennis Cook. Detection of influential observations in linear regression. *Technometrics*, 19(1):15–18, February 1977.

[Coo98] R. Dennis Cook. *Regression Graphics: Ideas for Studying Regressions through Graphics*. Series in Probability and Statistics. Wiley, New York, 1998.

[Cor69] J. Cornfield. The Bayesian outlook and its applications. *Biometrics*, 25:617–657, 1969.

[Cow77] Frank Alan Cowell. *Measuring Inequality: Techniques for the Social Sciences*. Wiley, New York, 1977.

[CP77] Samprit Chatterjee and Bertram Price. *Regression Analysis by Example*. Wiley, New York, 1977.

[CR88] Raymond J. Carroll and David Ruppert. *Transformation and Weighting in Regression*. Chapman and Hall, London and New York, 1988.

[Cra43] A. T. Craig. Note on the independence of certain quadratic forms. *Annals of Mathematical Statistics*, 14:195, 1943.

[Cra83] J. G. Cragg. More efficient estimation in the presence of heteroskedasticity of unknown form. *Econometrica*, 51:751–63, 1983.

[Cra91] Jan Salomon Cramer. *An Introduction of the Logit Model for Economists*. Edward Arnold, London, 1991.

[CT91] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Series in Telecommunications. Wiley, New York, 1991.

[CW99] R. Dennis Cook and Sanford Weisberg. *Applied Regression Including Computing and Graphics*. Wiley, 1999.

[Dav75] P. J. Davis. *Interpolation and Approximation*. Dover Publications, New York, 1975.

[Daw79a] A. P. Dawid. Conditional independence in statistical theory. *JRSS(B)*, 41(1):1–31, 1979.

[Daw79b] A. P. Dawid. Some misleading arguments involving conditional independence. *JRSS(B)*, 41(2):249–252, 1979.

[Daw80] A. P. Dawid. Conditional independence for statistical operations. *Annals of Statistics*, 8:598–617, 1980.

[Dea92] Angus Deaton. *Understanding Consumption*. Clarendon Press, Oxford, 1992.

[DH94] B. R. Davis and R. J. Hardy. Data monitoring in clinical trials: The case for stochastic curtailment. *J. Clinical Epidemiology*, 47(9):1033–42, 1994.

[Dhr86] Phoebus J. Dhrymes. Limited dependent variables. In Zvi Griliches and Michael D. Intriligator, editors, *Handbook of Econometrics*, volume 3, chapter 27, pages 1567–1631. North-Holland, Amsterdam, 1986.

[DL91] Gerard Dumenil and Dominique Levy. The U.S. economy since the Civil War: Sources and construction of the series. Technical report, CEPREMAP, LAREA-CEDRA, December 1991.

[DM93] Russell Davidson and James G. MacKinnon. *Estimation and Inference in Econometrics*. Oxford University Press, New York, 1993.

[dMG89] Neil de Marchi and Christopher Gilbert, editors. *History and Methodology of Econometrics*. Clarendon Press, 1989.

[Dou92] Christopher Dougherty. *Introduction to Econometrics*. Oxford University Press, Oxford, 1992.

[Dow99] Paul Downward. *Pricing Theory in Post-Keynesian Economics: A Realist Approach*. New Directions in Modern Economics. Elgar, Cheltenham, UK; Northamption, MA, USA, 1999.

[DP20] R. E. Day and W. M. Persons. An index of the physical volume of production. *Review of Economic Statistsics*, II:309–37, 361–67, 1920.

[DW50] J. Durbin and G. Watson. Testing for serial correlation in least squares regression—I. *Biometrika*, 37:409–428, 1950.

[DW51] J. Durbin and G. Watson. Testing for serial correlation in least squares regression—II. *Biometrika*, 38:159–178, 1951.

[DW71] J. Durbin and G. Watson. Testing for serial correlation in least squares regression—III. *Biometrika*, 58:1–42, 1971.

[Efr82] Bradley Efron. *The Jackknife, the Bootstrap, and Other Resampling Plans*. SIAM (Society for Industrial and Applies Mathematics), Philadelphia, PA, 1982.

[Ell95] Rebecca J. Elliott. *Learning SAS in the Computer Lab*. Cuxbury Press, Belmont, California, 1995.

[End95] Walter Enders. *Applied Econometric Time Series*. Wiley, New York, 1995.

[ESS97] Klaus Edel, Karl August Schäffer, and Winfried Stier, editors. *Analyse saisonaler Zeitreihen*. Number 134 in Wirtschaftwissenschaftliche Beiträge. Physica-Verlag, Heidelberg, 1997.

[ET93] Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, 1993.

[Eub88] Randall L. Eubank. *Spline Smoothing and Nonparametric Regression*. Marcel Dekker, New York, 1988.

[Eve94] Brian Everitt. *A Handbook of Statistical Analyses Using S-Plus*. Chapman & Hall, 1994.

[Far80] R. W. Farebrother. The Durbin-Watson test for serial correlation when there is no intercept in the regression. *Econometrica*, 48:1553–1563, September 1980.

[Fis] R. A. Fisher. Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society*, 22.

[Fri57] Milton Friedman. *A Theory of the Consumption Function*. Princeton University Press, 1957.

[FS81] J. H. Friedman and W. Stuetzle. Projection pursuit regression. *JASA*, 76:817–23, 1981.

[FS91] Milton Friedman and Anna J. Schwarz. Alternative approaches to analyzing economic data. *American Economic Review*, 81(1):39–49, March 1991.

[FT74] J. H. Friedman and J. W. Tukey. A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, C-23:881–90, 1974.

[FW74] George M. Furnival and Jr. Wilson, Robert W. Regression by leaps and bounds. *Technometrics*, 16:499–511, 1974.

[Gas88] Joseph L. Gastwirth. *Statistical Reasoning in Law and Public Policy*. Statistical modeling and decision science. Academic Press, Boston, 1988.

[GC92] Jean Dickinson Gibbons and S. Chakraborti. *Nonparametric Statistical Inference*. Marcel Dekker, 3rd edition, 1992.

[GG95] Joseph L. Gastwirth and S. W. Greenhouse. Biostatistical concepts and methods in the legal setting. *Statistics in Medicine*, 14:1641–53, 1995.

[GJM96] Amos Golan, George Judge, and Douglas Miller. *Maximum Entropy Econometrics: Robust Estimation with Limited Data*. Wiley, Chichester, England, 1996.

[Gra76] Franklin A. Graybill. *Theory and Application of the Linear Model*. Duxbury Press, North Sciutate, Mass., 1976.

[Gra83] Franklin A. Graybill. *Matrices with Applications in Statistics*. Wadsworth and Brooks/Cole, Pacific Grove, CA, second edition, 1983.

[Gra89] Clive William John Granger. *Forecasting in Business and Economics*. Academic Press, second edition, 1989.

[Gre93] William H. Greene. *Econometric Analysis*. Macmillan, New York, second edition, 1993.

[Gre97] William H. Greene. *Econometric Analysis*. Prentice Hall, Upper Saddle River, NJ, third edition, 1997.

[Gri67] Zvi Griliches. Distributed lags: A survey. *Econometrica*, 35:16–49, 1967.

[Gri79] R. C. Grimson. The clustering of disease. *Mathematical Biosciences*, 46:257–78, 1979.

[Gum58] E. J. Gumbel. *Statistics of Extremes*. Columbia University Press, New York, 1958.

[Gut94] Robert Guttmann. *How Credit Money Shapes the Economy*. Sharpe, Armonk, 1994.

[Hal78] Robert E. Hall. Stochastic implications of the life cycle-permanent income hypothesis: Theory and evidence. *Journal of Political Economy*, pages 971–987, December 1978.

[Ham94] James D. Hamilton. *Time Series Analysis*. Princeton University Press, 1994.

[Har76] A. C. Harvey. Estimating regression models with multiplicative heteroscedasticity. *Econometrica*, 44:461–465, 1976.

[Hay00] Fumio Hayashi. *Econometrics*. Princeton University Press, 2000.

[HC70] Robert V. Hogg and Allen T. Craig. *Introduction to Mathematical Statistics*. Macmillan, third edition, 1970.

[Hen95] David F. Hendry. *Dynamic Econometrics*. Oxford University Press, Oxford, New York, 1995.

[HH68] C. Hildreth and J. Houck. Some estimators for a linear model with random coefficients. *Journal of the American Statistical Association*, 63:584–95, 1968.

[HH71] Gerald J. Hahn and Richard W. Hendrickson. A table of percentage points of the distribution of the largest absolute value of k Student t variates and its applications. *Biometrica*, 58(2):323–332, 1971.

[HK70a] Arthur E. Hoerl and R. W. Kennard. Ridge regression: Application to non-orthogonal problems. *Technometrics*, 12:69–82, 1970.

[HK70b] Arthur E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for non-orthogonal problems. *Technometrics*, 12:55–67, 1970.

[HK79] James M. Henle and Eugene M. Kleinberg. *Infinitesimal Calculus*. MIT Press, 1979.

[HM89] David F. Hendry and Mary Morgan. A re-analysis of confluence analysis. *Oxford Economic Papers*, pages 35–52, 1989. Reprinted in [**dMG89**, pp. 35–52].

[Hol86] Paul W. Holland. Statistics and causal inference. *JASA*, 81(396):945–960, 1986.

[Hou51] H. S. Houthakker. Some calculations on electricity consumption in Great Britain. *Journal of the Royal Statistical Society (A)*, (114 part III):351–371, 1951. J.

[HR97] Omar F. Hamouda and J. C. R. Rowley, editors. *The Reappraisal of Econometrics*, volume 9 of *Foundations of Probability, Econometrics and Economic Games*. Elgar, Cheltenham, UK; Lyme, US, 1997.

[Hsu38] P. L. Hsu. On the best unbiased quadratic estimate of variance. *Statistical Research Memoirs*, 2:91–104, 1938. Issued by the Department of Statistics, University of London, University College.

[HT83] Robert V. Hogg and Elliot A. Tanis. *Probability and Statistical Inference*. Macmillan, second edition, 1983.

[HT90] Trevor J. Hastie and Robert J. Tibshirani. *Generalized Additive Models*. Chapman and Hall, 1990.

[HVdP02] Ben J. Hejdra and Frederick Van der Ploeg. *Foundations of Modern Macroeconomics*. Oxford University Press, 2002.

[Hyl92] Svend Hylleberg, editor. *Modelling Seasonality*. Oxford University Press, 1992.

[JGH+85] George G. Judge, William E. Griffiths, R. Carter Hill, Helmut Lütkepohl, and Tsoung-Chao Lee. *The Theory and Practice of Econometrics*. Wiley, New York, second edition, 1985.

[JHG+88] George G. Judge, R. Carter Hill, William E. Griffiths, Helmut Lütkepohl, and Tsoung-Chao Lee. *Introduction to the Theory and Practice of Econometrics*. Wiley, New York, second edition, 1988.

[JK70] Norman Johnson and Samuel Kotz. *Continuous Univariate Distributions*, volume 1. Houghton Mifflin, Boston, 1970.

[JL97] B. D. Javanovic and P. S. Levy. A look at the rule of three. *American Statistician*, 51(2):137–9, 1997.

[JS61] W. James and C. Stein. Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 361–379. University of California Press, Berkeley, 1961.

[JW88] Richard A. Johnson and Dean W. Wichern. *Applied Multivariate Statistical Analysis*. Prentice Hall, 1988.

[KA69] J. Koerts and A. P. J. Abramanse. *On the Theory and Application of the General Linear Model*. Rotterdam University Press, Rotterdam, 1969.

[Kal82] R. E. Kalman. System identification from noisy data. In A. R. Bednarek and L. Cesari, editors, *Dynamical Systems*, volume II, pages 135–164. Academic Press, New York, 1982.

[Kal83] R. E. Kalman. Identifiability and modeling in econometrics. In P. R. Krisnaiah, editor, *Developments in Statistics*, volume 4. Academic Press, New York, 1983.

[Kal84] R. E. Kalman. We can do something about multicollinearity! *Communications in Statistics, Theory and Methods*, 13(2):115–125, 1984.

[Kap89] Jagat Narain Kapur. *Maximum Entropy Models in Science and Engineering*. Wiley, 1989.

[KG80] William G. Kennedy and James E. Gentle. *Statistical Computing*. Dekker, New York, 1980.

[Khi57] R. T. Khinchin. *Mathematical Foundations of Information Theory*. Dover Publications, New York, 1957.

[Kim] Kim. *Introduction of Factor Analysis*.

[Kin81] M. King. The Durbin-Watson test for serial correlation: Bounds for regressions with trend and/or seasonal dummy variables. *Econometrica*, 49:1571–1581, 1981.

[KM78] Jae-On Kim and Charles W. Mueller. *Factor Analysis: Statistical Methods and Practical Issues*. Sage, 1978.

[Kme86] Jan Kmenta. *Elements of Econometrics*. Macmillan, New York, second edition, 1986.

[Knu81] Donald E. Knuth. *Seminumerical Algorithms*, volume 2 of *The Art of Computer Programming*. Addison-Wesley, second edition, 1981.

[Knu98] Donald E. Knuth. *Seminumerical Algorithms*, volume 2 of *The Art of Computer Programming*. Addison-Wesley, third edition, 1998.

[Krz88] W. J. Krzanowski. *Principles of Multivariate Analysis: A User's Persective*. Clarendon Press, Oxford, 1988.

[KS79] Sir Maurice Kendall and Alan Stuart. *The Advanced Theory of Statistics*, volume 2. Griffin, London, fourth edition, 1979.

[Ksh19] Anant M. Kshirsagar. *Multivariate Analysis*. Marcel Dekker, New York and Basel, 19??

[Lan69] H. O. Lancaster. *The Chi-Squared Distribution*. Wiley, 1969.

[Lar82] Harold Larson. *Introduction to Probability and Statistical Inference*. Wiley, 1982.

[Law89] Tony Lawson. Realism and instrumentalism in the development of econometrics. *Oxford Economic Papers*, 41:236–258, 1989. Reprinted in [**dMG89**] and [**HR97**].

[Lea75] Edward E. Leamer. A result on the sign of the restricted least squares estimator. *Journal of Econometrics*, 3:387–390, 1975.

[LN81] D. V. Lindley and M. R. Novick. The role of exchangeability in inference. *Annals of Statistics*, 9(1):45–58, 1981.

[Loa99] Clive Loader. *Local Regression and Likelihood*. Statistics and Computing. Springer, New York, 1999.

[Ḿ99] László Mátyás, editor. *Generalized Method of Moments Estimation*. Cambridge University Press, 1999.

[Mad88] G. S. Maddala. *Introduction to Econometrics*. Macmillan, New York, 1988.

[Mag88] Jan R. Magnus. *Linear Structures*. Oxford University Press, New York, 1988.

[Mal70] E. Malinvaud. *Statistical Methods of Econometrics*. North-Holland, Amsterdam, second edition, 1970.

[Mal78] E. Malinvaud. *Méthodes Mathématiques de l'Économetrie*. Dunod, troisième edition, 1978.

[Mal80] E. Malinvaud. *Statistical Methods of Econometrics*. North-Holland, Amsterdam, third edition, 1980.

[Mil90] Alan J. Miller. *Subset Selection in Regression*. Chapman and Hall, New York, 1990.

[Mir96] Jeff A. Miron. *The Economics of Seasonal Cycles*. MIT Press, Cambridge, Massachusetts, 1996.

[MN88] Jan R. Magnus and Heinz Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Wiley, Chichester, 1988.

[MN89] Peter McCullagh and John Asworth Nelder. *Generalized Linear Models*. Chapman and Hall, second edition, 1989.

[Mor65] A. Q. Morton. The authorship of Greek prose (with discussion). *Journal of the Royal Statistical Society, Series A*, 128:169–233, 1965.

[Mor73] Trenchard More, Jr. Axioms and theorems for a theory of arrays. *IBM Journal of Research and Development*, 17(2):135–175, March 1973.

[Mor02] Jamie Morgan. The global power of orthodox economics. *Journal of Critical Realism*, 1(2):7–34, May 2002.

[MR91] Ieke Moerdijk and Gonzalo E. Reyes. *Models for Smooth Infinitesimal Analysis*. Springer-Verlag, New York, 1991.

[MS86] Parry Hiram Moon and Domina Eberle Spencer. *Theory of Holors; A Generalization of Tensors*. Cambridge University Press, 1986.

[MT98] Allan D. R. McQuarrie and Chih-Ling Tsai. *Regression and Time Series Model Selection*. World Scientific, Singapore, 1998.

[Mul72] S. A. Mulaik. *The Foundations of Factor Analysis*. McGraw-Hill, New York, 1972.

[NT92] Sharon-Lise Normand and David Tritchler. Parameter updating in a Bayes network. *JASA Journal of the American Statistical Association*, 87(420):1109–1115, December 1992.

[Qua90] D. Quah. Permanent and transitory movements in labor income: An explanation for 'excess smoothness' in consumption. *Journal of Political Economy*, 98:449–475, 1990.

[Rao52] C. Radhakrishna Rao. Some theorems on minimum variance estimation. *Sankhyā*, 12:27–42, 1952.

[Rao62] C. Radhakrishna Rao. A note on a generalized inverse of a matrix with applications to problems in mathematical statistics. *Journal of the Royal Statistical Society, Series B*, 24:152–158, 1962.

[Rao73] C. Radhakrishna Rao. *Linear Statistical Inference and Its Applications*. Wiley, New York, second edition, 1973.

[Rao97] C. Radhakrishna Rao. *Statistics and Truth: Putting Chance to Work*. World Scientific, Singapore, second edition, 1997.

[Rei89] Rolf-Dieter Reiss. *Approximate Distributions of Order Statistics*. Springer-Verlag, New York, 1989.

[Rei93] Gregory C. Reinsel. *Elements of Multivariate Time Series Analysis*. Springer-Verlag, New York, 1993.

[Rén70] Alfred Rényi. *Foundations of Probability*. Holden-Day, San Francisco, 1970.

[Ric] J Rice. *Mathematical Statistics and Data Analysis*. Wadsworth.

[Rie77] E. Rietsch. The maximum entropy approach to inverse problems. *Journal of Geophysics*, 42, 1977.

[Rie85] E. Rietsch. On an alleged breakdown of the maximum-entropy principle. In C. Ray Smith and Jr W. T. Grandy, editors, *Maximum-Entropy and Bayesian Methods in Inverse Problems*, pages 67–82. D. Reidel, Dordrecht, Boston, Lancaster, 1985.

[Rip96] Brian D. Ripley. A short guide to XGobi. Statistics Department Mimeo, January 1996.

[Rob70] Herbert Robbins. Statistical methods related to the law of the iterated logarithm. *Annals of Mathematical Statistics*, 41:1397–1409, 1970.

[Rob74] Abraham Robinson. *Non-Standard Analysis*. North Holland, Amsterdam, 1974.

[Ron02] Amit Ron. Regression analysis and the philosophy of social science: A critical realist view. *Journal of Critical Realism*, 1(1):119–142, November 2002.

[Roy97] Richard M. Royall. *Statistical evidence: A Likelihood Paradigm*. Number 71 in Monographs on Statistics and Applied Probability. Chapman & Hall, London; New York, 1997.

[Ruu00] Paul A. Ruud. *An Introduction to Classical Econometric Theory*. Oxford University Press, Oxford and New York, 2000.

[RZ78] L. S. Robertson and P. L. Zador. Driver education and fatal crash invovlement of teenage drivers. *American Journal of Public Health*, 68:959–65, 1978.

[SAS85] SAS Institute Inc., Cary, NC. *SAS User's Guide: Statistics*, version 5 edition edition, 1985.

[SCB91] Deborah F. Swayne, Dianne Cook, and Andreas Buja. Xgobi: Interactive dynamic graphics in the X windows system with a link to S. *ASA Proceedings of the Section on Statistical Graphics*, pages 1–8, 1991.

[Sch59] Henry Scheffé. *The Analysis of Variance*. Wiley, New York, 1959.

[Sch97] Manfred R. Schroeder. *Number Theory in Science and Communication*. Number 7 in Information Sciences. Springer-Verlag, Berlin Heidelberg New York, 3rd edition, 1997.

[Scl68] Stanley L. Sclove. Inproved estimators for coefficients in linear regression. *Journal of the American Statistical Association*, 63:595–606, 1968.

[Seb77] G. A. F. Seber. *Linear Regression Analysis*. Wiley, New York, 1977.

[Sel58] H. C. Selvin. Durkheim's suicide and problems of empirical research. *American Journal of Sociology*, 63:607–619, 1958.

[SG85] John Skilling and S. F. Gull. Algorithms and applications. In C. Ray Smith and Jr W. T. Grandy, editors, *Maximum-Entropy and Bayesian Methods in Inverse Problems*, pages 83–132. D. Reidel, Dordrecht, Boston, Lancaster, 1985.

[Shi73] R. Shiller. A distributed lag estimator derived from smoothness priors. *Econometrica*, 41:775–778, 1973.

[Sim96] Jeffrey S. Simonoff. *Smoothing Methods in Statistics*. Springer Series in Statistics. Springer, New York, 1996.

[SM86] Hans Schneeweiß and Hans-Joachim Mittag. *Lineare Modelle mit fehlerbehafteten Daten*. Physica Verlag, Heidelberg, Wien, 1986.

[Spe94] Phil Spector. *An Introduction to S and S-Plus*. Duxbury Press, Belmont, California, 1994.

[Spr98] Peter Sprent. *Data Driven Statistical Methods*. Texts in statistical science. Chapman & Hall, London; New York, 1st ed. edition, 1998.

[SS35] J. A. Schouten and Dirk J. Struik. *Einführung in die neuren Methoden der Differentialgeometrie*, volume I. 1935.

[Sta95] William. Stallings. *Protect your Privacy: The PGP User's Guide*. Prentice Hall, Englewood Cliffs, N.J., 1995.

[Sta99] William Stallings. *Cryptography and Network Security: Principles and Practice*. Prentice Hall, Upper Saddle River, N.J., 2nd edition, 1999.

[Ste56] Charles Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 197–206. University of California Press, Berkeley, 1956.

[SW76] Thomes J. Sargent and Neil Wallace. Rational expectations and the theory of economic policy. *Journal of Monetary Economics*, 2:169–183, 1976.

[Sze59] G. Szegö. *Orthogonal Polynomials*. Number 23 in AMS Colloquium Publications. American Mathematical Society, 1959.

[The71] Henri Theil. *Principles of Econometrics*. Wiley, New York, 1971.

[Thi88] Ronald A. Thisted. *Elements of Statistical Computing*. Chapman and Hall, 1988.

[Tib88] Robert Tibshirani. Estimating transformations for regression via additivity and variance stabilization. *JASA*, 83(402):394–405, 1988.

[Tin51] J. Tinbergen. *Econometrics*. George Allen & Unwin Ltd., London, 1951.

[TS61] Henri Theil and A. Schweitzer. The best quadratic estimator of the residual variance in regression analysis. *Statistica Neerlandica*, 15:19–23, 1961.

[Vas76] D. Vasicek. A test for normality based on sample entropy. *JRSS (B)*, 38:54–9, 1976.

[VdBDL83] E. Van der Burg and J. De Leeuw. Nonlinear canonical correlation. *British J. Math. Statist. Psychol.*, 36:54–80, 1983.

[VR99] William N. Venables and Brian D. Ripley. *Modern Applied Statistics with S-Plus*. Springer-Verlag, New York, third edition, 1999.

[VU81] Hrishikesh D. Vinod and Aman Ullah. *Recent Advances in Regression Methods*. Dekker, New York, 1981.

[Wah90] Grace Wahba. *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1990.

[Wal72] Kenneth Wallis. Testing for fourth order autocorrelation in quarterly regression equations. *Econometrica*, 40:617–36, 1972.

[WG68] M. B. Wilk and R. Gnanadesikan. Probability plotting methods for the analysis of data. *Biometrica*, 55:1–17, 1968.

[WH82] B. A. Wichmann and I. D. Hill. Algorithm AS 183: An efficient and portable pseudo-random number generator. *Applied Statistics*, 31:188–190, 1982. Correction in Wichmann/Hill:BS183. See also [**?**] and [**Zei86**].

[WH97] Mike West and Jeff Harrison. *Bayesian Forecasting and Dynamic Models*. Springer-Verlag, second edition, 1997.

[Whi93] K. White. *SHAZAM, Version 7*. Department of Economics, University of British Columbia, Vancouver, 1993.

[Wit85] Uli Wittman. *Das Konzept rationaler Preiserwartungen*, volume 241 of *Lecture Notes in Economics and Mathematical Systems*. Springer, 1985.

[WJ95] M. P. Wand and M. C. Jones. *Kernel Smoothing*, volume 60 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London; New York, 1st ed. edition, 1995.

[WM83] Howard Wainer and Samuel Messick, editors. *Principals of Modern Psychologial Measurement: A Festschrift for Frederic M. Lord*. Associates,, Hillsdale, N.J.: L. Erlbaum, 1983.

[Woo72] L. A. Wood. Modulus of natural rubber cross-linked by dicumyl peroxide. i. experimental observations. *J Res. Nat. Bur. Stand*, 76A:51–59, 1972.

[WW79] Ronald J. Wonnacott and Thomas H. Wonnacott. *Econometrics*. Wiley, New York, second edition, 1979.

[Yul07] G. V. Yule. On the theory of correlation for any number of variables treated by a new system of notation. *Proc. Roy. Soc. London A*, 79:182, 1907.

[Zei86] H. Zeisel. A remark on algorithm AS 183. an efficient and portable random number generator. *Applied Statistics*, 35(1):89, 1986.

[Zel62] Arnold Zellner. An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association*, 57(298):348–368, 1962.

[ZG70] Arnold Zellner and M. Geisel. Analysis of distributed lag models with application to the consumption function. *Econometrica*, 38:865–888, 1970.

[Zim95] Philip R. Zimmermann. *The Official PGP User's Guide*. MIT Press, Cambridge, Mass., 1995.