Class Notes Econ 7800 Fall Semester 2003

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

The present version has those chapters relevant for Econ 7800.

Contents

Chapter 1. Syllabus Econ 7800 Fall 2003 ix

Chapter 2. Probability Fields 1
2.1. The Concept of Probability 1
2.2. Events as Sets 6
2.3. The Axioms of Probability 9
2.4. Objective and Subjective Interpretation of Probability 11
2.5. Counting Rules 13
2.6. Relationships Involving Binomial Coefficients 14
2.7. Conditional Probability 15
2.8. Ratio of Probabilities as Strength of Evidence 20
2.9. Bayes Theorem 22
2.10. Independence of Events 23
2.11. How to Plot Frequency Vectors and Probability Vectors 26

Chapter 3. Random Variables 29
3.1. Notation 29
3.2. Digression about Infinitesimals 29
3.3. Definition of a Random Variable 31
3.4. Characterization of Random Variables 32
3.5. Discrete and Absolutely Continuous Probability Measures 35
3.6. Transformation of a Scalar Density Function 36
3.7. Example: Binomial Variable 37
3.8. Pitfalls of Data Reduction: The Ecological Fallacy 39
3.9. Independence of Random Variables 40
3.10. Location Parameters and Dispersion Parameters of a Random Variable 41
3.11. Entropy 46

Chapter 4. Specific Random Variables 55
4.1. Binomial 55
4.2. The Hypergeometric Probability Distribution 58
4.3. The Poisson Distribution 59
4.4. The Exponential Distribution 62
4.5. The Gamma Distribution 63
4.6. The Uniform Distribution 66
4.7. The Beta Distribution 66

4.8. The Normal Distribution 67
4.9. The Chi-Square Distribution 70
4.10. The Lognormal Distribution 70
4.11. The Cauchy Distribution 71

Chapter 5. Chebyshev Inequality, Weak Law of Large Numbers, and Central Limit Theorem 73
5.1. Chebyshev Inequality 73
5.2. The Probability Limit and the Law of Large Numbers 74
5.3. Central Limit Theorem 75

Chapter 6. Vector Random Variables 79
6.1. Expected Value, Variances, Covariances 80
6.2. Marginal Probability Laws 83
6.3. Conditional Probability Distribution and Conditional Mean 84
6.4. The Multinomial Distribution 86
6.5. Independent Random Vectors 87
6.6. Conditional Expectation and Variance 89
6.7. Expected Values as Predictors 91
6.8. Transformation of Vector Random Variables 95

Chapter 7. The Multivariate Normal Probability Distribution 101
7.1. More About the Univariate Case 101
7.2. Definition of Multivariate Normal 102
7.3. Special Case: Bivariate Normal 103
7.4. Multivariate Standard Normal in Higher Dimensions 112

Chapter 8. The Regression Fallacy 117

Chapter 9. A Simple Example of Estimation 125
9.1. Sample Mean as Estimator of the Location Parameter 125
9.2. Intuition of the Maximum Likelihood Estimator 126
9.3. Variance Estimation and Degrees of Freedom 128

Chapter 10. Estimation Principles and Classification of Estimators 137
10.1. Asymptotic or Large-Sample Properties of Estimators 137
10.2. Small Sample Properties 138
10.3. Comparison Unbiasedness Consistency 140
10.4. The Cramer-Rao Lower Bound 143
10.5. Best Linear Unbiased Without Distribution Assumptions 150
10.6. Maximum Likelihood Estimation 152
10.7. Method of Moments Estimators 155
10.8. M-Estimators 155
10.9. Sufficient Statistics and Estimation 155
10.10. The Likelihood Principle 159
10.11. Bayesian Inference 159

Chapter 11. Interval Estimation 163

Chapter 12. Hypothesis Testing 169
12.1. Duality between Significance Tests and Confidence Regions 172
12.2. The Neyman Pearson Lemma and Likelihood Ratio Tests 172
12.3. The Wald, Likelihood Ratio, and Lagrange Multiplier Tests 176

Chapter 13. General Principles of Econometric Modelling 179

Chapter 14. Mean-Variance Analysis in the Linear Model 181
14.1. Three Versions of the Linear Model 181
14.2. Ordinary Least Squares 182
14.3. The Coefficient of Determination 189
14.4. The Adjusted R-Square 194

Chapter 15. Digression about Correlation Coefficients 197
15.1. A Unified Definition of Correlation Coefficients 197

Chapter 16. Specific Datasets 201
16.1. Cobb Douglas Aggregate Production Function 201
16.2. Houthakker's Data 210
16.3. Long Term Data about US Economy 215
16.4. Dougherty Data 216
16.5. Wage Data 216

Chapter 17. The Mean Squared Error as an Initial Criterion of Precision 231
17.1. Comparison of Two Vector Estimators 231

Chapter 18. Sampling Properties of the Least Squares Estimator 235
18.1. The Gauss Markov Theorem 235
18.2. Digression about Minimax Estimators 237
18.3. Miscellaneous Properties of the BLUE 238
18.4. Estimation of the Variance 247
18.5. Mallow's Cp-Statistic as Estimator of the Mean Squared Error 248

Chapter 19. Nonspherical Positive Definite Covariance Matrix 251

Chapter 20. Best Linear Prediction 255
20.1. Minimum Mean Squared Error, Unbiasedness Not Required 255
20.2. The Associated Least Squares Problem 261
20.3. Prediction of Future Observations in the Regression Model 262

Chapter 21. Updating of Estimates When More Observations become Available 267

Chapter 22. Constrained Least Squares 271
22.1. Building the Constraint into the Model 271
22.2. Conversion of an Arbitrary Constraint into a Zero Constraint 272
22.3. Lagrange Approach to Constrained Least Squares 273

22.4. Constrained Least Squares as the Nesting of Two Simpler Models 276
22.5. Solution by Quadratic Decomposition 277
22.6. Sampling Properties of Constrained Least Squares 277
22.7. Estimation of the Variance in Constrained OLS 279
22.8. Inequality Restrictions 282
22.9. Application: Biased Estimators and Pre-Test Estimators 283

Chapter 23. Additional Regressors 285

Chapter 24. Residuals: Standardized, Predictive, "Studentized" 297
24.1. Three Decisions about Plotting Residuals 297
24.2. Relationship between Ordinary and Predictive Residuals 299
24.3. Standardization 302

Chapter 25. Regression Diagnostics 305
25.1. Missing Observations 305
25.2. Grouped Data 305
25.3. Influential Observations and Outliers 306
25.4. Sensitivity of Estimates to Omission of One Observation 308

Chapter 26. Asymptotic Properties of the OLS Estimator 315
26.1. Consistency of the OLS estimator 316
26.2. Asymptotic Normality of the Least Squares Estimator 317

Chapter 27. Least Squares as the Normal Maximum Likelihood Estimate 319

Chapter 28. Random Regressors 325
28.1. Strongest Assumption: Error Term Well Behaved Conditionally on Explanatory Variables 325
28.2. Contemporaneously Uncorrelated Disturbances 326
28.3. Disturbances Correlated with Regressors in Same Observation 327

Chapter 29. The Mahalanobis Distance 329
29.1. Definition of the Mahalanobis Distance 329

Chapter 30. Interval Estimation 333
30.1. A Basic Construction Principle for Confidence Regions 333
30.2. Coverage Probability of the Confidence Regions 337
30.3. Conventional Formulas for the Test Statistics 337
30.4. Interpretation in terms of Studentized Mahalanobis Distance 338

Chapter 31. Three Principles for Testing a Linear Constraint 343
31.1. Mathematical Detail of the Three Approaches 343
31.2. Examples of Tests of Linear Hypotheses 347
31.3. The F-Test Statistic is a Function of the Likelihood Ratio 354
31.4. Tests of Nonlinear Hypotheses 355
31.5. Choosing Between Nonnested Models 355

Chapter 32. Instrumental Variables 357

Appendix A. Matrix Formulas 363
A.1. A Fundamental Matrix Decomposition 363
A.2. The Spectral Norm of a Matrix 364
A.3. Inverses and g-Inverses of Matrices 364
A.4. Deficiency Matrices 365
A.5. Nonnegative Definite Symmetric Matrices 369
A.6. Projection Matrices 373
A.7. Determinants 375
A.8. More About Inverses 376
A.9. Eigenvalues and Singular Value Decomposition 379

Appendix B. Arrays of Higher Rank 383
B.1. Informal Survey of the Notation 383
B.2. Axiomatic Development of Array Operations 385
B.3. An Additional Notational Detail 390
B.4. Equality of Arrays and Extended Substitution 390
B.5. Vectorization and Kronecker Product 391

Appendix C. Matrix Differentiation 401
C.1. First Derivatives 401

Appendix. Bibliography 407

CHAPTER 1

Syllabus Econ 7800 Fall 2003

The class meets Tuesdays and Thursdays 12:25 to 1:45pm in BUC 207. First class Thursday, August 21, 2003; last class Thursday, December 4.

Instructor: Assoc. Prof. Dr. Dr. Hans G. Ehrbar. Hans's office is at 319 BUO, Tel. 581 7797, email `ehrbar@econ.utah.edu` *Office hours:* Monday 10–10:45 am, Thursday 5–5:45 pm or by appointment.

Textbook: There is no obligatory textbook in the Fall Quarter, but detailed class notes are available at `www.econ.utah.edu/ehrbar/ec7800.pdf`, and you can purchase a hardcopy containing the assigned chapters only at the University Copy Center, 158 Union Bldg, tel. 581 8569 (ask for the class materials for Econ 7800).

Furthermore, the following optional texts will be available at the bookstore: Peter Kennedy, *A Guide to Econometrics* (fourth edition), MIT Press, 1998 ISBN 0-262-61140-6.

The bookstore also has available William H. Greene's *Econometric Analysis*, fifth edition, Prentice Hall 2003, ISBN 0-13-066189-9. This is the assigned text for Econ 7801 in the Spring semester 2004, and some of the introductory chapters are already useful for the Fall semester 2003.

The following chapters in the class notes are assigned: 2, 3 (but not section 3.2), 4, 5, 6, 7 (but only until section 7.3), 8, 9, 10, 11, 12, 14, only section 15.1 in chapter 15, in chapter 16, we will perhaps do section 16.1 or 16.4, then in chapter 17 we do section 17.1, then chapter 18 until and including 18.5, and in chapter 22 do sections 22.1, 22.3, 22.6, and 22.7. In chapter 29 only the first section 29.1, finally chapters 30, and section 31.2 in chapter 31.

Summary of the Class: This is the first semester in a two-semester Econometrics field, but it should also be useful for students taking the first semester only as part of their methodology requirement. The course description says: Probability, conditional probability, distributions, transformation of probability densities, sufficient statistics, limit theorems, estimation principles, maximum likelihood estimation, interval estimation and hypothesis testing, least squares estimation, linear constraints.

This class has two focal points: maximum likelihood estimation, and the fundamental concepts of the linear model (regression).

If advanced mathematical concepts are necessary in these theoretical explorations, they will usually be reviewed very briefly before we use them. The class is structured in such a way that, if you allocate enough time, it should be possible to refresh your math skills as you go along.

Here is an overview of the topics to be covered in the Fall Semester. They may not come exactly in the order in which they are listed here

1. Probability fields: Events as sets, set operations, probability axioms, subjective vs. frequentist interpretation, finite sample spaces and counting rules (combinatorics), conditional probability, Bayes theorem, independence, conditional independence.

2. Random Variables: Cumulative distribution function, density function; location parameters (expected value, median) and dispersion parameters (variance).

3. Special Issues and Examples: Discussion of the "ecological fallacy"; entropy; moment generating function; examples (Binomial, Poisson, Gamma, Normal, Chisquare); sufficient statistics.

4. Limit Theorems: Chebyshev inequality; law of large numbers; central limit theorems.

The first Midterm will already be on Thursday, September 18, 2003. It will be closed book, but you are allowed to prepare one sheet with formulas etc. Most of the midterm questions will be similar or identical to the homework questions in the class notes assigned up to that time.

5. Jointly Distributed Random Variables: Joint, marginal, and conditional densities; conditional mean; transformations of random variables; covariance and correlation; sums and linear combinations of random variables; jointly normal variables.

6. Estimation Basics: Descriptive statistics; sample mean and variance; degrees of freedom; classification of estimators.

7. Estimation Methods: Method of moments estimators; least squares estimators. Bayesian inference. Maximum likelihood estimators; large sample properties of MLE; MLE and sufficient statistics; computational aspects of maximum likelihood.

8. Confidence Intervals and Hypothesis Testing: Power functions; Neyman Pearson Lemma; likelihood ratio tests. As example of tests: the run test, goodness of fit test, contingency tables.

The second in-class Midterm will be on Thursday, October 16, 2003.

9. Basics of the "Linear Model." We will discuss the case with nonrandom regressors and a spherical covariance matrix: OLS-BLUE duality, Maximum likelihood estimation, linear constraints, hypothesis testing, interval estimation (t-test, F-test, joint confidence intervals).

The third Midterm will be a takehome exam. You will receive the questions on Tuesday, November 25, 2003, and they are due back at the beginning of class on Tuesday, December 2nd, 12:25 pm. The questions will be similar to questions which you might have to answer in the Econometrics Field exam.

The Final Exam will be given according to the campus-wide examination schedule, which is Wednesday December 10, 10:30–12:30 in the usual classroom. Closed book, but again you are allowed to prepare one sheet of notes with the most important concepts and formulas. The exam will cover material after the second Midterm.

Grading: The three midterms and the final exams will be counted equally. Every week certain homework questions from among the questions in the class notes will

be assigned. It is recommended that you work through these homework questions conscientiously. The answers provided in the class notes should help you if you get stuck. If you have problems with these homeworks despite the answers in the class notes, please write you answer down as far as you get and submit your answer to me; I will look at them and help you out. A majority of the questions in the two in-class midterms and the final exam will be identical to these assigned homework questions, but some questions will be different.

Special circumstances: If there are special circumstances requiring an individualized course of study in your case, please see me about it in the first week of classes.

Hans G. 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. $\begin{array}{ccccc} 11 & 12 & 13 & 14 & \\ 21 & 22 & 23 & & \\ 31 & 32 & & & \\ 41 & & & & \end{array}$, 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_k^n 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_k^n.

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., $P_2^6 = 6 \cdot 5$. Number of arrangements of a given full house over the five dice is $C_2^5 = \frac{5 \cdot 4}{1 \cdot 2}$ (we have to specify the two places taken by the two-of-a-kind outcomes.) Solution is therefore $P_2^6 \cdot C_2^5 / 6^5 = 50/6^4 = 0.03858$. 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)

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

The binomial coefficients also occur in the Binomial Theorem

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

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,
(2.7.5)

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

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

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

Answer.

$$\begin{aligned}\Pr[B|A_1 \cup A_2] &= \frac{\Pr[B \cap (A_1 \cup A_2)]}{\Pr[A_1 \cup A_2]} = \frac{\Pr[(B \cap A_1) \cup (B \cap A_2)]}{\Pr[A_1 \cup A_2]} \\ &= \frac{\Pr[B \cap A_1] + \Pr[B \cap A_2]}{\Pr[A_1] + \Pr[A_2]} = \frac{c\Pr[A_1] + c\Pr[A_2]}{\Pr[A_1] + \Pr[A_2]} = c. \end{aligned} \tag{2.7.10}$$

□

Problem 24. *Show by counterexample that the requirement* $A_1 \cap A_2 = \emptyset$ *is necessary for this result to hold. Hint: use the example in Problem 38 with* $A_1 = \{HH, HT\}$, $A_2 = \{HH, TH\}$, $B = \{HH, TT\}$.

Answer. $\Pr[B|A_1] = 1/2$ and $\Pr[B|A_2] = 1/2$, but $\Pr[B|A_1 \cup A_2] = 1/3$. □

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] =

$$= \frac{3}{3+8}\frac{4}{4+8} + \frac{4}{7+4}\frac{8}{8+4} = \frac{1}{3}. \tag{2.7.11}$$

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 $\frac{4\cdot3\cdot2}{12\cdot11\cdot10} = \frac{24}{1320} = \frac{1}{55}$. 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 $\binom{5}{3} = \frac{5\cdot4\cdot3}{1\cdot2\cdot3} = 10$ 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 $\Pr[B|A]=\Pr[B|A']$. From this follows by Problem 23 that the conditional probability is equal to the unconditional probability $\Pr[B]$, i.e., $\Pr[B]=\Pr[B\cap A]/\Pr[A]$. Therefore we will adopt as definition of independence the so-called *multiplication rule:*

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

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

Problem 35. *3 points Given that* $\Pr[B\cap A]=\Pr[B]\cdot\Pr[A]$ *(i.e.,* B *is independent of* A*), show that* $\Pr[B\cap A']=\Pr[B]\cdot\Pr[A']$ *(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 $\Pr[B|A]=\Pr[B|A']$, 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 $(B\cap A)$ and $(B\cap A')$, it follows $\Pr[B]=\Pr[B\cap A]+\Pr[B\cap A']$ or $\Pr[B\cap A']=\Pr[B]-\Pr[B\cap A]=\Pr[B]-\Pr[B]\Pr[A]=\Pr[B](1-\Pr[A])=\Pr[B]\Pr[A']$. □

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

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

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* $\Pr[B|A]=\Pr[A]$*? e. Are these two events disjoint, i.e., is* $A\cap B=\emptyset$*?*

Answer. Clearly, $\Pr[A]=1/2$. $\Pr[B]=\Pr[B|A]\Pr[A]+\Pr[B|A']\Pr[A']=(4/9)(1/2)+(5/9)(1/2)=1/2$. The events are not independent: $\Pr[B|A]=4/9\neq\Pr[B]$, or $\Pr[A\cap B]=\frac{5}{10}\frac{4}{9}=2/9\neq 1/4$. 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,

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

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 $A=\{1,4\}$, $B=\{2,4\}$, and $C=\{3,4\}$. 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:*

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

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

Answer. $U=\left\{\begin{smallmatrix}HH & HT\\ TH & TT\end{smallmatrix}\right\}$. $A=\{HH,HT\}$, $B=\{HH,TH\}$, $C=\{HH,TT\}$. $\Pr[A]=\frac{1}{2}$, $\Pr[B]=\frac{1}{2}$, $\Pr[C]=\frac{1}{2}$. They are pairwise independent, but $\Pr[A\cap B\cap C]=\Pr[\{HH\}]=\frac{1}{4}\neq\Pr[A]\Pr[B]\Pr[C]$, 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* $A\cap B\subset C$*. Can those events be mutually independent?*

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

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 $A\cap B$ independent of C, and $A\cup B$ independent of C, etc. This is not the case if the events are only pairwise independent. In Problem 39, $A\cap B$ is not independent of C.

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.

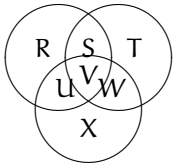


Figure 1. Generic Venn Diagram for 3 Events

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.

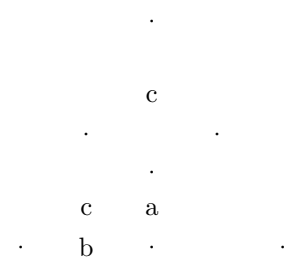


Figure 2. Geometry of an equilateral triangle

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:

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

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 $q \geq 0$, and 0 for $q < 0$.*

ANSWER. If $q \geq 0$ then

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

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

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

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

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

□

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

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

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

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

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 $p \in (0,1)$:

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

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

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

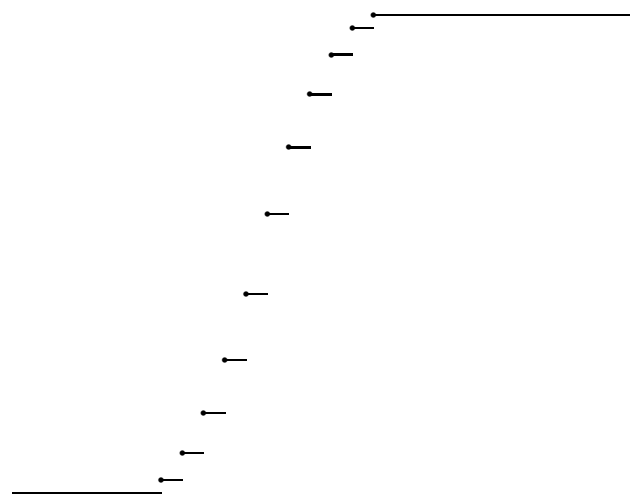


FIGURE 1. Cumulative Distribution Function of Discrete Variable

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 $[12, +\infty)$. □

- 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]. □

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.

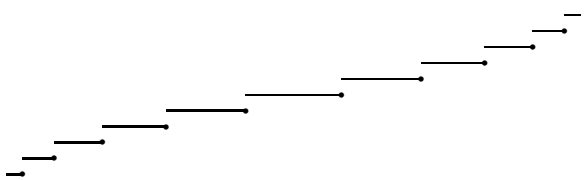


Figure 2. Quantile Function of Discrete Variable

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 140.*

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), $\mathrm{E}[g(\mathsf{x})] = p_1g(x_1) + \cdots + p_ng(x_n)$.

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 169.)*

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

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

but the formula

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

is usually more convenient.

How to calculate with variance?

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

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

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

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* $\mathrm{var}[\mathsf{x}] = \mathrm{E}[\mathsf{x}^2] - (\mathrm{E}[\mathsf{x}])^2$ *from the definition of a variance, which is* $\mathrm{var}[\mathsf{x}] = \mathrm{E}[(\mathsf{x} - \mathrm{E}[\mathsf{x}])^2]$. *Hint: it is convenient to define* $\mu = \mathrm{E}[\mathsf{x}]$. *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 $\mathrm{E}[\mathsf{x}] = \mu$:

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

□

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

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 $\mathrm{E}[\mathsf{y}^2]$ 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* $\mathrm{E}[u(\mathsf{y})]$. *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. $\mathrm{E}[u(\mathsf{y})] = 12\,\mathrm{E}[\mathsf{y}] - \mathrm{E}[\mathsf{y}^2] = 12\,\mathrm{E}[\mathsf{y}] - \mathrm{var}[\mathsf{y}] - (\mathrm{E}[\mathsf{y}])^2$. 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., $\mathrm{H}(p_1, \ldots, p_n) \leq \mathrm{H}(1/n, \ldots, 1/n)$.
- 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 $\mathsf{W}_1 = \mathsf{X}_{11} \cup \cdots \cup \mathsf{X}_{1k_1}$, $\mathsf{W}_2 = \mathsf{X}_{21} \cup \cdots \cup \mathsf{X}_{2k_2}$, $\ldots$, $\mathsf{W}_m = \mathsf{X}_{m1} \cup \cdots \cup \mathsf{X}_{mk_m}$, with corresponding probabilities $w_1 = p_{11} + \cdots + p_{1k_1}$, $w_2 = p_{21} + \cdots + p_{2k_2}$, $\ldots$, $w_m = p_{m1} + \cdots + p_{mk_m}$, then

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

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.,

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

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

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

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.

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

□

- 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* x=*a?" This splits the probability in half. If the answer to the first question is no, then the second question can be "Is* x=*b?" The third question, if it is necessary, can then be: "Is* 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 $0 \leq 0$; if $v > 0$ and $w = 0$ the equation reads $0 \leq v$, and if $w > 0$ and $v = 0$ then the equation reads $w - w\log w \leq +\infty$. □

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:

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

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.

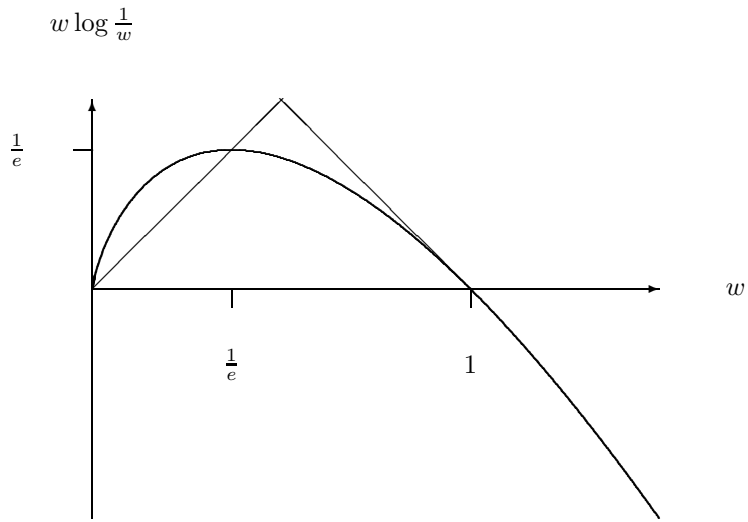


FIGURE 3. $\eta : w \mapsto w \log \frac{1}{w}$ is continuous at 0, and concave everywhere

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* $c_1 \leq c_2 \leq \cdots \leq c_n$*. 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* $p_1, p_2, \ldots p_n$ *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:

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

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

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

□

• 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

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

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,* $\sum \frac{\mathrm{x}_i}{m} c_i$*, as an estimate of the missing* d *in the above procedure, i.e., choose that* λ *which satisfies* $f(\lambda) = \sum \frac{\mathrm{x}_i}{m} c_i$*. 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 $x_1, \ldots, x_n$ 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

Specific Random Variables

4.1. 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 75. *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

$$\Pr[\mathsf{t}{=}n] = pq^{n-1} \qquad (n = 1, 2, \ldots). \tag{4.1.1}$$

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:

$$1 + q + q^2 + q^3 + \cdots = s \quad \text{Now multiply by } q: \tag{4.1.2}$$

$$q + q^2 + q^3 + \cdots = qs \quad \text{Now subtract and write } 1 - q = p: \tag{4.1.3}$$

$$1 = ps \tag{4.1.4}$$

Equation (4.1.4) 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 (4.1.4) for s:

$$s = \frac{1}{p} \quad \text{or} \quad 1 + q + q^2 + q^3 + q^4 \cdots = \sum_{k=0}^{\infty} q^k = \frac{1}{1-q} \tag{4.1.5}$$

and differentiate both sides with respect to q:

$$1 + 2q + 3q^2 + 4q^3 + \cdots = \sum_{k=1}^{\infty} kq^{k-1} = \frac{1}{(1-q)^2} = \frac{1}{p^2}. \tag{4.1.6}$$

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

Problem 76. *Assume* t *is a geometric random variable with parameter* p, *i.e., it has the values* $k = 1, 2, \ldots$ *with probabilities*

$$p_{\mathsf{t}}(k) = pq^{k-1}, \text{ where } q = 1 - p. \tag{4.1.7}$$

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 77. t *is a geometric random variable as in the preceding problem. In order to compute* $\operatorname{var}[\mathsf{t}]$ *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*

$$\sum_{k=2}^{\infty} k(k-1)q^{k-2} = \frac{2}{(1-q)^3} \tag{4.1.8}$$

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 (4.1.6). □

• c. *Use a. and b. to derive*

$$\mathrm{E}[\mathsf{t}(\mathsf{t}-1)] = \frac{2q}{p^2} \tag{4.1.9}$$

Answer.

$$\sum_{k=2}^{\infty} k(k-1)pq^{k-1} = pq\sum_{k=2}^{\infty} k(k-1)q^{k-2} = pq\frac{2}{(1-q)^3} = \frac{2q}{p^2}. \tag{4.1.10}$$

□

• d. *Use c. and the fact that* $\mathrm{E}[\mathsf{t}] = 1/p$ *to derive*

$$\mathrm{var}[\mathsf{t}] = \frac{q}{p^2}. \tag{4.1.11}$$

Answer.

$$\mathrm{var}[\mathsf{t}] = \mathrm{E}[\mathsf{t}^2] - (\mathrm{E}[\mathsf{t}])^2 = \mathrm{E}[\mathsf{t}(\mathsf{t}-1)] + \mathrm{E}[\mathsf{t}] - (\mathrm{E}[\mathsf{t}])^2 = \frac{2q}{p^2} + \frac{1}{p} - \frac{1}{p^2} = \frac{q}{p^2}. \tag{4.1.12}$$

□

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:

$$\Pr[\mathsf{t}{=}n] = \binom{n-1}{r-1}p^r q^{n-r}. \tag{4.1.13}$$

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:

$$\mathrm{E}[\mathsf{t}] = \frac{r}{p} \qquad \mathrm{var}[\mathsf{t}] = \frac{rq}{p^2} \tag{4.1.14}$$

Some authors define the negative binomial as the number of failures before the rth success. Their formulas will look slightly different than ours.

Problem 78. *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 79. *(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 $\binom{n+k-1}{n-1}2^{1-n-k}$. □

4.2. 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

$$\Pr[\text{Sample of } m \text{ elements has exactly } y \text{ white balls}] = \frac{\binom{w}{y}\binom{n-w}{m-y}}{\binom{n}{m}}. \tag{4.2.1}$$

Problem 80. *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 (4.2.1) *if sampling is done with replacement.*

Answer.

$$\left(\frac{w}{n}\right)^y \left(\frac{n-w}{n}\right)^{m-y} \binom{m}{y} \tag{4.2.2}$$

□

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

$$\mathrm{var}[\mathsf{y}] = m\frac{w}{n}\frac{(n-w)}{n}\frac{(n-m)}{(n-1)}. \tag{4.2.3}$$

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].

4.3. 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

$$\Pr[\mathsf{x}{=}k] = \binom{n}{k}\left(\frac{\lambda t}{n}\right)^k \left(1-\frac{\lambda t}{n}\right)^{(n-k)} \tag{4.3.1}$$

$$= \frac{1}{k!}\frac{n(n-1)\cdots(n-k+1)}{n^k}(\lambda t)^k\left(1-\frac{\lambda t}{n}\right)^n\left(1-\frac{\lambda t}{n}\right)^{-k} \tag{4.3.2}$$

$$\to \frac{(\lambda t)^k}{k!}e^{-\lambda t} \quad \text{for } n\to\infty \text{ while } k \text{ remains constant} \tag{4.3.3}$$

(4.3.3) is the limit because the second and the last term in (4.3.2) $\to 1$. 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 $k=1$):

$$\mathrm{E}[\mathsf{x}] = e^{-\lambda t}\sum_{k=1}^{\infty} k\frac{(\lambda t)^k}{k!} = \lambda t e^{-\lambda t}\sum_{k=1}^{\infty}\frac{(\lambda t)^{k-1}}{(k-1)!} = \lambda t. \tag{4.3.4}$$

This is the same as the expected value of the discrete approximations.

Problem 81. x *follows a Poisson distribution, i.e.,*

$$\Pr[\mathsf{x}{=}k] = \frac{(\lambda t)^k}{k!}e^{-\lambda t} \quad \textit{for } k = 0, 1, \ldots. \tag{4.3.5}$$

• a. *2 points Show that* $\mathrm{E}[\mathsf{x}] = \lambda t$.

Answer. See (4.3.4). □

• 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 $k=2$:

$$\mathrm{E}[\mathsf{x}(\mathsf{x}-1)] = e^{-\lambda t}\sum_{k=2}^{\infty} k(k-1)\frac{(\lambda t)^k}{k!} = (\lambda t)^2 e^{-\lambda t}\sum_{k=2}^{\infty}\frac{(\lambda t)^{k-2}}{(k-2)!} = (\lambda t)^2. \tag{4.3.6}$$

From this follows

$$\mathrm{var}[\mathsf{x}] = \mathrm{E}[\mathsf{x}^2] - (\mathrm{E}[\mathsf{x}])^2 = \mathrm{E}[\mathsf{x}(\mathsf{x}-1)] + \mathrm{E}[\mathsf{x}] - (\mathrm{E}[\mathsf{x}])^2 = (\lambda t)^2 + \lambda t - (\lambda t)^2 = \lambda t. \tag{4.3.7}$$

□

The Poisson distribution can be used as an approximation to the Binomial distribution when n large, p small, and np moderate.

Problem 82. *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 83. *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*

$$\frac{(\lambda t)^k}{k!}e^{-\lambda t}. \tag{4.3.8}$$

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:

$$\Pr[\mathsf{y}{=}m|\mathsf{x}{=}n] = \frac{\Pr[\mathsf{y}{=}m \text{ and } \mathsf{x}{=}n]}{\Pr[\mathsf{x}{=}n]}. \tag{4.3.9}$$

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

(4.3.10)

$$\Pr[\mathsf{y}{=}m|\mathsf{x}{=}n] = \frac{\Pr[\mathsf{y}{=}m]\Pr[\mathsf{z}{=}n-m]}{\Pr[\mathsf{x}{=}n]} = \frac{\frac{\lambda^m t^m}{m!}e^{-\lambda t}\frac{\lambda^{n-m}(1-t)^{n-m}}{(n-m)!}e^{-\lambda(1-t)}}{\frac{\lambda^n}{n!}e^{-\lambda}} = \binom{n}{m}t^m(1-t)^{n-m},$$

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 84. *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.,*

$$\Pr[\mathsf{x}{=}j] = \frac{2^j}{j!}e^{-2}. \tag{4.3.11}$$

• 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:*

$$e = 1 + 1 + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \cdots. \tag{4.3.12}$$

• c. *For any integer* $k \geq 0$*, compute the probability that Herbert has to serve exactly* k *customers during the hour.*

Problem 85. *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. $\mathrm{E}[e^{t\mathsf{x}}] = \sum_{k=0}^{\infty} e^{tk}\frac{\lambda^k}{k!}e^{-\lambda} = \sum_{k=0}^{\infty}\frac{(\lambda e^t)^k}{k!}e^{-\lambda} = e^{\lambda e^t}e^{-\lambda} = e^{\lambda(e^t-1)}$. □

4.4. 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. $\mathsf{t}{>}t$ if and only if there are no occurrences in the interval $[0, t]$. Therefore $\Pr[\mathsf{t}{>}t] = \frac{(\lambda t)^0}{0!}e^{-\lambda t} = e^{-\lambda t}$, and hence the cumulative distribution function $F_{\mathsf{t}}(t) = \Pr[\mathsf{t}{\leq}t] = 1 - e^{-\lambda t}$ when $t \geq 0$, and $F_{\mathsf{t}}(t) = 0$ for $t < 0$. The density function is therefore $f_{\mathsf{t}}(t) = \lambda e^{-\lambda t}$ for $t \geq 0$, 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 86. *2 points An exponential random variable* t *with parameter* $\lambda > 0$ *has the density* $f_{\mathsf{t}}(t) = \lambda e^{-\lambda t}$ *for* $t \geq 0$*, and 0 for* $t < 0$*. Use this density to compute the expected value of* t*.*

Answer. $\mathrm{E}[\mathsf{t}] = \int_0^\infty \lambda t e^{-\lambda t}dt = \int_0^\infty uv'dt = uv\Big|_0^\infty - \int_0^\infty u'vdt$, 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 $\mathrm{E}[\mathsf{t}] = -te^{-\lambda t}\Big|_0^\infty + \int_0^\infty e^{-\lambda t}dt = 0 - \frac{1}{\lambda}e^{-\lambda t}|_0^\infty = \frac{1}{\lambda}$. □

Problem 87. *4 points An exponential random variable* t *with parameter* $\lambda > 0$ *has the density* $f_{\mathsf{t}}(t) = \lambda e^{-\lambda t}$ *for* $t \geq 0$*, and 0 for* $t < 0$*. Use this density to compute the expected value of* 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: $\mathrm{E}[\mathsf{t}^2] = (1/\lambda^2)\Gamma(3) = 2/\lambda^2$. Or all from scratch: $\mathrm{E}[\mathsf{t}^2] = \int_0^\infty \lambda t^2 e^{-\lambda t}dt = \int_0^\infty uv'dt = uv\Big|_0^\infty - \int_0^\infty u'vdt$, where $\begin{matrix} u = t^2 & v' = \lambda e^{-\lambda t} \\ u' = 2t & v = -e^{-\lambda t}\end{matrix}$. Therefore $\mathrm{E}[\mathsf{t}^2] = -t^2e^{-\lambda t}\Big|_0^\infty + \int_0^\infty 2te^{-\lambda t}\,dt$. 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 88. *2 points Does the exponential random variable with parameter* $\lambda > 0$*, whose cumulative distribution function is* $F_{\mathsf{t}}(t) = 1 - e^{-\lambda t}$ *for* $t \geq 0$*, and 0 otherwise, have a memory-less property? Compare Problem 76. Formulate this memory-less property and then verify whether it holds or not.*

Answer. Here is the formulation: for $\mathsf{s}{<}\mathsf{t}$ follows $\Pr[\mathsf{t}{>}t|\mathsf{t}{>}s] = \Pr[\mathsf{t}{>}t-s]$. This does indeed hold. Proof: lhs $= \frac{\Pr[\mathsf{t}>t \text{ and } \mathsf{t}>s]}{\Pr[\mathsf{t}>s]} = \frac{\Pr[\mathsf{t}>t]}{\Pr[\mathsf{t}>s]} = \frac{e^{-\lambda t}}{e^{-\lambda s}} = e^{-\lambda(t-s)}$. □

Problem 89. *The random variable* t *denotes the duration of an unemployment spell. It has the exponential distribution, which can be defined by:* $\Pr[\mathsf{t}{>}t] = e^{-\lambda t}$ *for* $t \geq 0$ *(*t *cannot assume negative values).*

• a. *1 point Use this formula to compute the cumulative distribution function* $F_{\mathsf{t}}(t)$ *and the density function* $f_{\mathsf{t}}(t)$

Answer. $F_{\mathsf{t}}(t) = \Pr[\mathsf{t}{\le}t] = 1 - \Pr[\mathsf{t}{>}t] = 1 - e^{-\lambda t}$ for $t \ge 0$, zero otherwise. Taking the derivative gives $f_{\mathsf{t}}(t) = \lambda e^{-\lambda t}$ 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.

$$\Pr[\mathsf{t}{>}t+h|\mathsf{t}{>}t] = \frac{\Pr[\mathsf{t}{>}t+h]}{\Pr[\mathsf{t}{>}t]} = \frac{e^{-\lambda(t+h)}}{e^{-\lambda t}} = e^{-\lambda h} \tag{4.4.1}$$

□

• 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*

$$\Pr[t < \mathsf{t}{\le}t+h] = h f_{\mathsf{t}}(t). \tag{4.4.2}$$

Answer.

$$\Pr[\mathsf{t}{\le}t+h|\mathsf{t}{>}t] = \frac{\Pr[\mathsf{t}{\le}t+h \text{ and } \mathsf{t}{>}t]}{\Pr[\mathsf{t}{>}t]} = \frac{h\, f_{\mathsf{t}}(t)}{1-F_{\mathsf{t}}(t)} = \frac{h\,\lambda e^{-\lambda t}}{e^{-\lambda t}} = h\,\lambda. \tag{4.4.3}$$

□

4.5. The Gamma Distribution

The time until the *second* occurrence of a Poisson event is a random variable which we will call $\mathsf{t}^{(2)}$. Its cumulative distribution function is $F_{\mathsf{t}^{(2)}}(t) = \Pr[\mathsf{t}^{(2)}{\le}t] = 1 - \Pr[\mathsf{t}^{(2)}{>}t]$. But $\mathsf{t}^{(2)}{>}t$ means: there are either zero or one occurrences in the time between 0 and t; therefore $\Pr[\mathsf{t}^{(2)}{>}t] = \Pr[\mathsf{x}{=}0]+\Pr[\mathsf{x}{=}1] = e^{-\lambda t}+\lambda t e^{-\lambda t}$. Putting it all together gives $F_{\mathsf{t}^{(2)}}(t) = 1-e^{-\lambda t}-\lambda t e^{-\lambda t}$. In order to differentiate the cumulative distribution function we need the product rule of differentiation: $(uv)' = u'v + uv'$. This gives

$$f_{\mathsf{t}^{(2)}}(t) = \lambda e^{-\lambda t} - \lambda e^{-\lambda t} + \lambda^2 t e^{-\lambda t} = \lambda^2 t e^{-\lambda t}. \tag{4.5.1}$$

Problem 90. *3 points Compute the density function of $\mathsf{t}^{(3)}$, the time of the third occurrence of a Poisson variable.*

Answer.

$$\Pr[\mathsf{t}^{(3)}{>}t] = \Pr[\mathsf{x}{=}0] + \Pr[\mathsf{x}{=}1] + \Pr[\mathsf{x}{=}2] \tag{4.5.2}$$

$$F_{\mathsf{t}^{(3)}}(t) = \Pr[\mathsf{t}^{(3)}{\le}t] = 1 - (1+\lambda t + \frac{\lambda^2}{2}t^2)e^{-\lambda t} \tag{4.5.3}$$

$$f_{\mathsf{t}^{(3)}}(t) = \frac{\partial}{\partial t}F_{\mathsf{t}^{(3)}}(t) = -\Big(-\lambda(1+\lambda t+\frac{\lambda^2}{2}t^2) + (\lambda+\lambda^2 t)\Big)e^{-\lambda t} = \frac{\lambda^3}{2}t^2 e^{-\lambda t}. \tag{4.5.4}$$

□

If one asks for the rth occurrence, again all but the last term cancel in the differentiation, and one gets

$$f_{\mathsf{t}^{(r)}}(t) = \frac{\lambda^r}{(r-1)!}t^{r-1}e^{-\lambda t}. \tag{4.5.5}$$

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:

$$\Gamma(r) = \int_0^\infty \lambda^r t^{r-1} e^{-\lambda t} dt. \tag{4.5.6}$$

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 91. *Show that $\Gamma(r)$ as defined in (4.5.6) is independent of λ, i.e., instead of (4.5.6) one can also use the simpler equation*

$$\Gamma(r) = \int_0^\infty t^{r-1}e^{-t}dt. \tag{4.5.7}$$

Problem 92. *3 points Show by partial integration that the Gamma function satisfies $\Gamma(r+1) = r\Gamma(r)$.*

Answer. Start with

$$\Gamma(r+1) = \int_0^\infty \lambda^{r+1}t^r e^{-\lambda t}dt \tag{4.5.8}$$

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 $v' = r\lambda^r t^{r-1}$:

$$\Gamma(r+1) = -\lambda^r t^r e^{-\lambda t}\Big|_0^\infty + \int_0^\infty r\lambda^r t^{r-1}e^{-\lambda t}dt = 0 + r\Gamma(r). \tag{4.5.9}$$

□

Problem 93. *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 $\Gamma(1) = 1$:

$$\Gamma(1) = \int_0^\infty \lambda e^{-\lambda t}dt = -e^{-\lambda t}\Big|_0^\infty = 1 \tag{4.5.10}$$

and then, assuming that $\Gamma(r) = (r-1)!$ Problem 92 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 141.

Therefore the following defines a density function, called the Gamma density with parameter r and λ, for all $r > 0$ and $\lambda > 0$:

$$f(x) = \frac{\lambda^r}{\Gamma(r)}x^{r-1}e^{-\lambda x} \quad \text{for} \quad x \ge 0, \quad 0 \text{ otherwise.} \tag{4.5.11}$$

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 94. *4 points Compute the moment generating function of the Gamma distribution.*

ANSWER.

$$m_{\mathsf{x}}(t) = \mathrm{E}[e^{t\mathsf{x}}] = \int_0^\infty e^{tx}\frac{\lambda^r}{\Gamma(r)}x^{r-1}e^{-\lambda x}dx \tag{4.5.12}$$

$$= \frac{\lambda^r}{(\lambda-t)^r}\int_0^\infty \frac{(\lambda-t)^r x^{r-1}}{\Gamma(r)}e^{-(\lambda-t)x}\,dx \tag{4.5.13}$$

$$= \left(\frac{\lambda}{\lambda-t}\right)^r \tag{4.5.14}$$

since the integrand in (4.5.12) is the density function of a Gamma distribution with parameters r and $\lambda - t$. □

PROBLEM 95. *2 points The density and moment generating functions of a Gamma variable* x *with parameters* $r > 0$ *and* $\lambda > 0$ *are*

$$f_{\mathsf{x}}(x) = \frac{\lambda^r}{\Gamma(r)}x^{r-1}e^{-\lambda x} \quad \textit{for} \quad x \geq 0, \quad 0 \textit{ otherwise.} \tag{4.5.15}$$

$$m_{\mathsf{x}}(t) = \left(\frac{\lambda}{\lambda-t}\right)^r. \tag{4.5.16}$$

Show the following: If x *has a Gamma distribution with parameters* r *and* 1*, then* $\mathsf{v} = \mathsf{x}/\lambda$ *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 x; its density is $\frac{1}{\Gamma(r)}x^{r-1}e^{-x}$. If $\mathsf{x} = \lambda\mathsf{v}$, then $\frac{dx}{dv} = \lambda$, and the absolute value is also λ. Therefore the density of v is $\frac{\lambda^r}{\Gamma(r)}v^{r-1}e^{-\lambda v}$. Solution using the mgf:

$$m_{\mathsf{x}}(t) = \mathrm{E}[e^{t\mathsf{x}}] = \left(\frac{1}{1-t}\right)^r \tag{4.5.17}$$

$$m_{\mathsf{v}}(t)\,\mathrm{E}[e^{t\mathsf{v}}] = \mathrm{E}[e^{(t/\lambda)\mathsf{x}}] = \left(\frac{1}{1-(t/\lambda)}\right)^r = \left(\frac{\lambda}{\lambda-t}\right)^r \tag{4.5.18}$$

but this last expression can be recognized to be the mgf of a Gamma with r and λ. □

PROBLEM 96. *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* (4.5.14) *without proof*

ANSWER.

$$\left(\frac{\lambda}{\lambda-t}\right)^r\left(\frac{\lambda}{\lambda-t}\right)^p = \left(\frac{\lambda}{\lambda-t}\right)^{r+p}. \tag{4.5.19}$$

□

PROBLEM 97. *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:

$$\frac{d}{dt}\left(\frac{\lambda}{\lambda-t}\right)^r = \frac{r}{\lambda}\left(\frac{\lambda}{\lambda-t}\right)^{r+1}, \tag{4.5.20}$$

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 $\mathrm{var}[\mathsf{t}] = \mathrm{E}[\mathsf{t}^2] - (\mathrm{E}[\mathsf{t}])^2 = r/\lambda^2$. □

4.6. The Uniform Distribution

PROBLEM 98. *Let* x *be uniformly distributed in the interval* $[a, b]$*, i.e., the density function of* x *is a constant for* $a \leq x \leq b$*, 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. $\mathrm{var}[\mathsf{x}] = \mathrm{E}[\mathsf{x}^2] - (\mathrm{E}[\mathsf{x}])^2 = \frac{a^2+ab+b^2}{3} - \frac{(a+b)^2}{4} = \frac{4a^2+4ab+4b^2}{12} - \frac{3a^2+6ab+3b^2}{12} = \frac{(b-a)^2}{12}$. □

4.7. 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 99. 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* $\mathsf{u}\leq q$ *and* $\mathsf{v}\leq r$ *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: $\Pr[\mathsf{u} \leq q] = 1 - \Pr[\mathsf{u} > q] = 1 - (1-q)^2 = 2q - q^2$. $f_{\mathsf{v}}(v) = 2v$ Therefore $f_{\mathsf{u}}(u) = 2 - 2u$

$$\mathrm{E}[\mathsf{u}] = \int_0^1 (2-2u)u\,du = \left(u^2 - \frac{2u^3}{3}\right)\Big|_0^1 = \frac{1}{3}. \tag{4.7.1}$$

For v it is: $\Pr[\mathsf{v} \leq r] = r^2$; this is at the same time the cumulative distribution function. Therefore the density function is $f_{\mathsf{v}}(v) = 2v$ for $0 \leq v \leq 1$ and 0 elsewhere.

$$\mathrm{E}[\mathsf{v}] = \int_0^1 v2v\,dv = \frac{2v^3}{3}\Big|_0^1 = \frac{2}{3}. \tag{4.7.2}$$

□

4.8. 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

$$f_{\mathsf{y}}(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}. \tag{4.8.1}$$

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 100. *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:

$$m_{\mathsf{z}}(t) = \mathrm{E}[e^{t\mathsf{z}}] = \int_{-\infty}^{+\infty} e^{tz} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz. \tag{4.8.2}$$

To solve this integral, complete the square in the exponent:

$$tz - \frac{z^2}{2} = \frac{t^2}{2} - \frac{1}{2}(z-t)^2; \tag{4.8.3}$$

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:

$$m_{\mathsf{z}}(t) = e^{\frac{t^2}{2}} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z-t)^2} dz = e^{\frac{t^2}{2}}, \tag{4.8.4}$$

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

$$m_{\mathsf{x}}(t) = \mathrm{E}[e^{(\mu+\sigma z)t}] = e^{\mu t}\,\mathrm{E}[e^{\sigma z t}] = e^{(\mu t + \sigma^2 t^2/2)}. \tag{4.8.5}$$

Problem 101. *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*

$$\alpha\mathsf{x} + \beta\mathsf{y} \sim \mathsf{N}(\alpha\mu_{\mathsf{x}} + \beta\mu_{\mathsf{y}}, \alpha^2\sigma_{\mathsf{x}}^2 + \beta^2\sigma_{\mathsf{y}}^2). \tag{4.8.6}$$

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 $\beta\mathsf{y}$:

$$m_{\alpha\mathsf{x}+\beta\mathsf{y}}(t) = e^{\mu_{\mathsf{x}}\alpha t + \sigma_{\mathsf{x}}^2\alpha^2 t^2/2} e^{\mu_{\mathsf{y}}\beta t + \sigma_{\mathsf{y}}^2\beta^2 t^2/2} = e^{(\mu_{\mathsf{x}}\alpha + \mu_{\mathsf{y}}\beta)t + (\sigma_{\mathsf{x}}^2\alpha^2 + \sigma_{\mathsf{y}}^2\beta^2)t^2/2}, \tag{4.8.7}$$

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

$$\mathrm{E}[\mathsf{z}|\mathsf{z}{>}z] = \frac{f_{\mathsf{z}}(z)}{1 - F_{\mathsf{z}}(z)}, \qquad \mathrm{var}[\mathsf{z}|\mathsf{z}{>}z] = 1 - \mu(\mu - z), \text{ where } \mu = \mathrm{E}[\mathsf{z}|\mathsf{z}{>}z]. \tag{4.8.8}$$

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 102. *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:

$$p = \Pr[\mathsf{y} > \$6800 | \mathsf{y} > \$6300] = \frac{\Pr[\mathsf{y} > \$6800]}{\Pr[\mathsf{y} > \$6300]} = \frac{1 - \Pr[\mathsf{y} \leq \$6800]}{1 - \Pr[\mathsf{y} \leq \$6300]}. \tag{4.8.9}$$

Again use the standard normal $\mathsf{z} = (\mathsf{y} - \$6000)/\$1000$. As before, $\mathsf{y} \leq \$6800 \iff \mathsf{z} \leq .8$, and $\mathsf{y} \leq \$6300 \iff \mathsf{z} \leq .3$. Therefore

$$p = \frac{1 - \Pr[\mathsf{z} \leq .8]}{1 - \Pr[\mathsf{z} \leq .3]} = \frac{.2119}{.3821} = .5546. \tag{4.8.10}$$

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*

$$\Pr[\mathsf{y} > \$6300 + m | \mathsf{y} > \$6300] = \Pr[\mathsf{y} \leq \$6300 + m | \mathsf{y} > \$6300] = 1/2. \tag{4.8.11}$$

Answer. Obviously, $m \geq \$0$. Therefore

$$\Pr[\mathsf{y} > \$6300 + m | \mathsf{y} > \$6300] = \frac{\Pr[\mathsf{y} > \$6300 + m]}{\Pr[\mathsf{y} > \$6300]} = \frac{1}{2}, \tag{4.8.12}$$

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.

$$\begin{aligned}
\mathrm{E}[\mathsf{y} - 6300 | \mathsf{y} \geq 6300] &= \mathrm{E}[6000 + 1000\mathsf{z} - 6300 | 6000 + 1000\mathsf{z} \geq 6300] && (4.8.13)\\
&= \mathrm{E}[1000\mathsf{z} - 300 | 1000\mathsf{z} \geq 300] && (4.8.14)\\
&= \mathrm{E}[1000\mathsf{z} | \mathsf{z} \geq 0.3] - 300 && (4.8.15)\\
&= 1000\,\mathrm{E}[\mathsf{z} | \mathsf{z} \geq 0.3] - 300 && (4.8.16)\\
&= 1000\frac{f(0.3)}{1 - \Psi(0.3)} - 300 = 698 > 575. && (4.8.17)
\end{aligned}$$

□

4.9. 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 $\frac{dz}{dq}$:

$$\begin{aligned}
f_{\mathsf{q}}(q) &= \frac{d}{dq}\big(2F_{\mathsf{z}}(\sqrt{q}) - 1\big) = \frac{d}{dq}\big(2F_{\mathsf{z}}(z) - 1\big) && (4.9.1)\\
&= 2\frac{dF_{\mathsf{z}}}{dz}(z)\frac{dz}{dq} = \frac{2}{\sqrt{2\pi}}e^{-z^2/2}\frac{1}{2\sqrt{q}} && (4.9.2)\\
&= \frac{1}{\sqrt{2\pi q}}e^{-q/2}. && (4.9.3)
\end{aligned}$$

Now remember the Gamma function. Since $\Gamma(1/2) = \sqrt{\pi}$ (Proof in Problem 141), one can rewrite (4.9.3) as

$$f_{\mathsf{q}}(q) = \frac{(1/2)^{1/2}q^{-1/2}e^{-q/2}}{\Gamma(1/2)}, \tag{4.9.4}$$

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 96) this gives a Gamma variable with parameters $r = p/2$ and $\lambda = 1/2$.

$$\text{If} \quad \mathsf{q} \sim \chi^2_p \qquad \text{then} \quad \mathrm{E}[\mathsf{q}] = p \qquad \text{and} \quad \mathrm{var}[\mathsf{q}] = 2p \tag{4.9.5}$$

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.

4.10. 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

$$\frac{1}{x\sqrt{2\pi\sigma^2}}e^{-(\ln x - \mu/\sigma^2)/2} \tag{4.10.1}$$

[**Cow77**, pp. 82–87] has an excellent discussion of the properties of the lognormal for income distributions.

4.11. The Cauchy Distribution

Problem 103. *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)*

$$f_{\mathtt{x}}(x) = \frac{1}{\pi(1+x^2)} \qquad \mathrm{E}[e^{it\mathtt{x}}] = \exp(-|t|). \tag{4.11.1}$$

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.

$$\mathrm{E}\left[\exp\left(it\frac{\mathtt{x}+\mathtt{y}}{2}\right)\right] = \mathrm{E}\left[\exp\left(i\frac{t}{2}\mathtt{x}\right)\exp\left(i\frac{t}{2}\mathtt{y}\right)\right] = \exp(-\left|\frac{t}{2}\right|)\exp(-\left|\frac{t}{2}\right|) = \exp(-|t|). \tag{4.11.2}$$

□

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 5

Chebyshev Inequality, Weak Law of Large Numbers, and Central Limit Theorem

5.1. 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

$$\Pr\big[|\mathsf{y}-\mu|\geq k\sigma\big] \leq \frac{1}{k^2}. \tag{5.1.1}$$

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

$$\Pr[|\mathsf{z}|\geq k] = \sum_{i:\ |z_i|\geq k} p(z_i). \tag{5.1.2}$$

Now multiply by k^2:

$$k^2\Pr[|\mathsf{z}|\geq k] = \sum_{i:\ |z_i|\geq k} k^2 p(z_i) \tag{5.1.3}$$

$$\leq \sum_{i:\ |z_i|\geq k} z_i^2 p(z_i) \tag{5.1.4}$$

$$\leq \sum_{\text{all i}} z_i^2 p(z_i) = \operatorname{var}[\mathsf{z}] = 1. \tag{5.1.5}$$

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 (5.1.1) becomes an equal sign.

Problem 104. [**HT83**, p. 316] *Let y be the number of successes in n trials of a Bernoulli experiment with success probability p. Show that*

$$\Pr\Big(\Big|\frac{\mathsf{y}}{n}-p\Big|<\varepsilon\Big) \geq 1-\frac{1}{4n\varepsilon^2}. \tag{5.1.6}$$

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

$$\Pr\Big(\Big|\frac{\mathsf{y}}{n}-p\Big|\geq k\sqrt{\frac{pq}{n}}\Big) \leq \frac{1}{k^2}. \tag{5.1.7}$$

Setting $\varepsilon=k\sqrt{pq/n}$, therefore $1/k^2=pq/n\varepsilon^2$ one can rewerite (5.1.7) as

$$\Pr\Big(\Big|\frac{\mathsf{y}}{n}-p\Big|\geq\varepsilon\Big) \leq \frac{pq}{n\varepsilon^2}. \tag{5.1.8}$$

Now note that $pq\leq 1/4$ whatever their values are. □

Problem 105. *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. □

5.2. 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 $\delta>0$

$$\lim_{n\to\infty}\Pr\big[|\mathsf{y}_n-\mathsf{y}|\geq\delta\big]=0. \tag{5.2.1}$$

One can also say that the probability limit of y_n is y, in formulas

$$\operatorname*{plim}_{n\to\infty}\mathsf{y}_n=\mathsf{y}. \tag{5.2.2}$$

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 106. *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*

$$\lim_{n\to\infty} \Pr\big[|\bar{\mathsf{y}}_n - \mathrm{E}[\mathsf{y}]| \geq \delta\big] = 0 \text{ for all } \delta > 0, \tag{5.2.3}$$

from the Chebyshev inequality

$$\Pr[|\mathsf{x} - \mu| \geq k\sigma] \leq \frac{1}{k^2} \qquad \text{where } \mu = \mathrm{E}[\mathsf{x}] \text{ and } \sigma^2 = \mathrm{var}[\mathsf{x}] \tag{5.2.4}$$

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 107. *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,*

$$\#\{\mathsf{y}_i\colon\ |\mathsf{y}_i - \bar{\mathsf{y}}| \geq k\mathsf{s}\} \leq \frac{n}{k^2}. \tag{5.2.5}$$

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.

$$\#\{\mathsf{y}_i\colon\ \mathsf{y}_i \in S\} = n\Pr[S] \tag{5.2.6}$$

□

• 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. □

5.3. 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

$$\frac{\bar{\mathsf{y}}_n - \mu}{\sigma/\sqrt{n}} \tag{5.3.1}$$

converges to a $\mathsf{N}(0,1)$. This is a different concept of convergence than the probability limit, it is convergence in distribution.

PROBLEM 108. *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 109,

$$\frac{\bar{\mathsf{y}}_n - \mu}{\sigma/\sqrt{n}} = \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{\mathsf{y}_i - \mu}{\sigma} = \frac{1}{\sqrt{n}}\sum_{i=1}^n \mathsf{z}_i \quad \text{where } \mathsf{z}_i = \frac{\mathsf{y}_i - \mu}{\sigma}. \tag{5.3.2}$$

Let m_3, m_4, etc., be the third, fourth, etc., moments of z_i; then the m.g.f. of z_i is

$$m_{\mathsf{z}_i}(t) = 1 + \frac{t^2}{2!} + \frac{m_3 t^3}{3!} + \frac{m_4 t^4}{4!} + \cdots \tag{5.3.3}$$

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):

$$\left(1 + \frac{t^2}{2!n} + \frac{m_3 t^3}{3!\sqrt{n^3}} + \frac{m_4 t^4}{4!n^2} + \cdots\right)^n = \left(1 + \frac{w_n}{n}\right)^n \tag{5.3.4}$$

where

$$w_n = \frac{t^2}{2!} + \frac{m_3 t^3}{3!\sqrt{n}} + \frac{m_4 t^4}{4!n} + \cdots. \tag{5.3.5}$$

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 109. *Show in detail that* $\frac{\bar{y}_n-\mu}{\sigma/\sqrt{n}} = \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{y_i-\mu}{\sigma}$.

ANSWER. Lhs $= \frac{\sqrt{n}}{\sigma}\left(\left(\frac{1}{n}\sum_{i=1}^n y_i\right)-\mu\right) = \frac{\sqrt{n}}{\sigma}\left(\left(\frac{1}{n}\sum_{i=1}^n y_i\right)-\left(\frac{1}{n}\sum_{i=1}^n \mu\right)\right) = \frac{\sqrt{n}}{\sigma}\frac{1}{n}\left(\sum_{i=}^n \mu\right) =$ rhs. □

PROBLEM 110. *3 points Explain verbally clearly what the law of large numbers means, what the Central Limit Theorem means, and what their difference is.*

PROBLEM 111. *(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 $x_i^2 = x_i$, $\text{var}[x_i] = \text{E}[x_i] - (\text{E}[x_i])^2 = \frac{1}{6} - \frac{1}{36} = \frac{5}{36}$. Define $x = \sum_{i=1}^{180} x_i$. We need $\Pr[x \geq 25]$. 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. $\text{E}[x] = 180 \cdot (1/6) = 30$ and $\text{var}[x] = 180 \cdot (5/36) = 25$. Therefore define $y \sim \text{N}(30, 25)$. The CLT says that $\Pr[x \geq 25] \approx \Pr[y \geq 25]$. Now $y \geq 25 \iff y - 30 \geq -5 \iff y - 30 \leq +5 \iff (y-30)/5 \leq 1$. But $z = (y-30)/5$ is a standard Normal, therefore $\Pr[(y-30)/5 \leq 1] = F_z(1)$, 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 $\Pr[x \geq 25] = \Pr[x > 24]$. Therefore $\Pr[y \geq 25]$ and $\Pr[y > 24]$ are two alternative good approximations; but the best is $\Pr[y \geq 24.5] = .8643$. This is the continuity correction. □

CHAPTER 6

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.,

$$\mathsf{x}\colon U \ni \omega \mapsto x(\omega) \in \mathbb{R} \qquad \text{and} \qquad \mathsf{y}\colon U \ni \omega \mapsto y(\omega) \in \mathbb{R}. \tag{6.0.6}$$

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

$$F_{\mathsf{x},\mathsf{y}}(x,y) = \Pr[\begin{bmatrix}\mathsf{x}\\ \mathsf{y}\end{bmatrix} \le \begin{bmatrix}x\\ y\end{bmatrix}] = \Pr[\mathsf{x} \le x \text{ and } \mathsf{y} \le y]. \tag{6.0.7}$$

For discrete random variables, for which the cumulative distribution function is a step function, the joint probability mass function provides the same information:

$$p_{\mathsf{x},\mathsf{y}}(x,y) = \Pr[\begin{bmatrix}\mathsf{x}\\ \mathsf{y}\end{bmatrix} = \begin{bmatrix}x\\ y\end{bmatrix}] = \Pr[\mathsf{x}{=}x \text{ and } \mathsf{y}{=}y]. \tag{6.0.8}$$

Problem 112. *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:

(6.0.9)

		Second Flip					Second Flip		
		H	T	sum			H	T	sum
First	H	.25	.25	.50	First	H	.50	.00	.50
Flip	T	.25	.25	.50	Flip	T	.00	.50	.50
	sum	.50	.50	1.00		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:

$$f_{\mathsf{x},\mathsf{y}}(x,y) = \frac{\partial^2}{\partial x\,\partial y} F_{\mathsf{x},\mathsf{y}}(x,y). \tag{6.0.10}$$

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 $B \subset \mathbb{R}^2$:

$$\Pr[\begin{bmatrix}\mathsf{x}\\ \mathsf{y}\end{bmatrix} \in B] = \int\int_B f(x,y)\,dx\,dy, \tag{6.0.11}$$

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) $|dV|$,

$$\Pr[\begin{bmatrix}\mathsf{x}\\ \mathsf{y}\end{bmatrix} \in dV_{x,y}] = f(x,y)\,|dV|. \tag{6.0.12}$$

The vertical bars here do not mean the absolute value but the volume of the argument inside.

6.1. Expected Value, Variances, Covariances

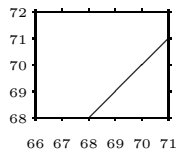
To get the expected value of a function of x and y, one simply has to put this function together with the density function into the integral, i.e., the formula is

$$\mathrm{E}[g(\mathsf{x},\mathsf{y})] = \int\int_{\mathbb{R}^2} g(x,y) f_{\mathsf{x},\mathsf{y}}(x,y)\,dx\,dy. \tag{6.1.1}$$

Problem 113. *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* $\begin{bmatrix} u \\ v \end{bmatrix}$ *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 $\begin{bmatrix} y \\ z \end{bmatrix}$ we will also write $\mathbf{x}$ (bold face). Vectors are always considered to be column vectors. The expected value of a random vector is a vector of constants, notation

$$\mathcal{E}[\mathbf{x}] = \begin{bmatrix} \mathrm{E}[x_1] \\ \vdots \\ \mathrm{E}[x_n] \end{bmatrix} \tag{6.1.2}$$

For two random variables x and y, their *covariance* is defined as

$$\mathrm{cov}[x, y] = \mathrm{E}\big[(x - \mathrm{E}[x])(y - \mathrm{E}[y])\big] \tag{6.1.3}$$

Computation rules with covariances are

$$\mathrm{cov}[x, z] = \mathrm{cov}[z, x] \quad \mathrm{cov}[x, x] = \mathrm{var}[x] \quad \mathrm{cov}[x, \alpha] = 0 \tag{6.1.4}$$

$$\mathrm{cov}[x + y, z] = \mathrm{cov}[x, z] + \mathrm{cov}[y, z] \quad \mathrm{cov}[\alpha x, y] = \alpha\,\mathrm{cov}[x, y] \tag{6.1.5}$$

Problem 114. *3 points Using* (6.1.3) *prove the following formula:*

$$\mathrm{cov}[x, y] = \mathrm{E}[xy] - \mathrm{E}[x]\,\mathrm{E}[y]. \tag{6.1.6}$$

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 $\mathrm{E}[x] = \mu$ and $\mathrm{E}[y] = \nu$:

$$\begin{aligned} \mathrm{cov}[x, y] &= \mathrm{E}\big[(x - \mathrm{E}[x])(y - \mathrm{E}[y])\big] \\ &= \mathrm{E}\big[xy - x\,\mathrm{E}[y] - \mathrm{E}[x]y + \mathrm{E}[x]\,\mathrm{E}[y]\big] \\ &= \mathrm{E}[xy] - \mathrm{E}[x]\,\mathrm{E}[y] - \mathrm{E}[x]\,\mathrm{E}[y] + \mathrm{E}[x]\,\mathrm{E}[y] \\ &= \mathrm{E}[xy] - \mathrm{E}[x]\,\mathrm{E}[y]. \end{aligned} \qquad \begin{aligned} \mathrm{cov}[x, y] &= \mathrm{E}[(x - \mu)(y - \nu)] \\ &= \mathrm{E}[xy - x\nu - \mu y + \mu\nu] \\ &= \mathrm{E}[xy] - \mu\nu - \mu\nu + \mu\nu \\ &= \mathrm{E}[xy] - \mu\nu. \end{aligned} \tag{6.1.7}$$

□

Problem 115. *1 point Using* (6.1.6) *prove the five computation rules with covariances* (6.1.4) *and* (6.1.5).

Problem 116. *Using the computation rules with covariances, show that*

$$\mathrm{var}[x + y] = \mathrm{var}[x] + 2\,\mathrm{cov}[x, y] + \mathrm{var}[y]. \tag{6.1.8}$$

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 $\mathcal{V}[\mathbf{x}]$. Its formal definition is

$$\mathcal{V}[\mathbf{x}] = \mathcal{E}\big[(\mathbf{x} - \mathcal{E}[\mathbf{x}])(\mathbf{x} - \mathcal{E}[\mathbf{x}])^\top\big], \tag{6.1.9}$$

but we can look at it simply as the matrix of all variances and covariances, for example

$$\mathcal{V}[\begin{bmatrix} x \\ y \end{bmatrix}] = \begin{bmatrix} \mathrm{var}[x] & \mathrm{cov}[x, y] \\ \mathrm{cov}[y, x] & \mathrm{var}[y] \end{bmatrix}. \tag{6.1.10}$$

An important computation rule for the covariance matrix is

$$\mathcal{V}[\mathbf{x}] = \boldsymbol{\Psi} \Rightarrow \mathcal{V}[\boldsymbol{A}\mathbf{x}] = \boldsymbol{A}\boldsymbol{\Psi}\boldsymbol{A}^\top. \tag{6.1.11}$$

Problem 117. *4 points Let* $\mathbf{x} = \begin{bmatrix} y \\ z \end{bmatrix}$ *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*

$$\mathcal{V}[\boldsymbol{A}\mathbf{x}] = \boldsymbol{A}\boldsymbol{\Psi}\boldsymbol{A}^\top. \tag{6.1.12}$$

Hint: You need to multiply matrices, and to use the following computation rules for covariances:

(6.1.13)

$$\mathrm{cov}[x + y, z] = \mathrm{cov}[x, z] + \mathrm{cov}[y, z] \quad \mathrm{cov}[\alpha x, y] = \alpha\,\mathrm{cov}[x, y] \quad \mathrm{cov}[x, x] = \mathrm{var}[x].$$

Answer. $\mathcal{V}[\boldsymbol{A}\mathbf{x}] =$

$$\mathcal{V}[\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} y \\ z \end{bmatrix}\right)] = \mathcal{V}[\begin{bmatrix} ay + bz \\ cy + dz \end{bmatrix}] = \begin{bmatrix} \mathrm{var}[ay + bz] & \mathrm{cov}[ay + bz, cy + dz] \\ \mathrm{cov}[cy + dz, ay + bz] & \mathrm{var}[cy + dz] \end{bmatrix}$$

On the other hand, $\boldsymbol{A}\boldsymbol{\Psi}\boldsymbol{A}^\top =$

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} \mathrm{var}[y] & \mathrm{cov}[y, z] \\ \mathrm{cov}[y, z] & \mathrm{var}[z] \end{bmatrix}\begin{bmatrix} a & c \\ b & d \end{bmatrix} = \begin{bmatrix} a\,\mathrm{var}[y] + b\,\mathrm{cov}[y, z] & a\,\mathrm{cov}[y, z] + b\,\mathrm{var}[z] \\ c\,\mathrm{var}[y] + d\,\mathrm{cov}[y, z] & c\,\mathrm{cov}[y, z] + d\,\mathrm{var}[z] \end{bmatrix}\begin{bmatrix} a & c \\ b & d \end{bmatrix}$$

Multiply out and show that it is the same thing. □

Since the variances are nonnegative, one can see from equation (6.1.11) 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 118. *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

$$\operatorname{corr}[\mathsf{x}, \mathsf{y}] = \frac{\operatorname{cov}[\mathsf{x}, \mathsf{y}]}{\sqrt{\operatorname{var}[\mathsf{x}]\operatorname{var}[\mathsf{y}]}}. \tag{6.1.14}$$

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

$$(\operatorname{cov}[\mathsf{x}, \mathsf{y}])^2 \leq \operatorname{var}[\mathsf{x}]\operatorname{var}[\mathsf{y}]. \tag{6.1.15}$$

PROBLEM 119. *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.

$$\operatorname{var}[a\mathsf{y} - \mathsf{z}] = a^2\operatorname{var}[\mathsf{y}] - 2a\operatorname{cov}[\mathsf{y}, \mathsf{z}] + \operatorname{var}[\mathsf{z}] \tag{6.1.16}$$

$$\text{First order condition:} \quad 0 = 2a\operatorname{var}[\mathsf{y}] - 2\operatorname{cov}[\mathsf{y}, \mathsf{z}] \tag{6.1.17}$$

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:

$$0 \leq \operatorname{var}[a^*\mathsf{y} - \mathsf{z}] = \frac{(\operatorname{cov}[\mathsf{y}, \mathsf{z}])^2}{\operatorname{var}[\mathsf{y}]} - \frac{2(\operatorname{cov}[\mathsf{y}, \mathsf{z}])^2}{\operatorname{var}[\mathsf{y}]} + \operatorname{var}[\mathsf{z}] \tag{6.1.18}$$

$$0 \leq -(\operatorname{cov}[\mathsf{y}, \mathsf{z}])^2 + \operatorname{var}[\mathsf{y}]\operatorname{var}[\mathsf{z}]. \tag{6.1.19}$$

This proves (6.1.15) 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 (6.1.15) holds trivially.

□

6.2. 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:

$$p_{\mathsf{x}}(x) = \sum_{y:\, p(x,y)\neq 0} p_{\mathsf{x},\mathsf{y}}(x, y). \tag{6.2.1}$$

For density functions, the following argument can be given:

$$\Pr[\mathsf{x} \in dV_x] = \Pr\Big[\begin{bmatrix}\mathsf{x}\\ \mathsf{y}\end{bmatrix} \in dV_x \times \mathbb{R}\Big]. \tag{6.2.2}$$

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

$$\Pr[\mathsf{x} \in dV_x] = \sum_i \Pr\Big[\begin{bmatrix}\mathsf{x}\\ \mathsf{y}\end{bmatrix} \in dV_x \times dV_{y_i}\Big] \tag{6.2.3}$$

$$= \sum_i f_{\mathsf{x},\mathsf{y}}(x, y_i)|dV_x||dV_{y_i}| \tag{6.2.4}$$

$$= |dV_x| \sum_i f_{\mathsf{x},\mathsf{y}}(x, y_i)|dV_{y_i}|. \tag{6.2.5}$$

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

$$f_{\mathsf{x}}(x) = \int_{y=-\infty}^{y=+\infty} f_{\mathsf{x},\mathsf{y}}(x, y)\, dy. \tag{6.2.6}$$

In other words, one has to "integrate out" the variable which one is not interested in.

6.3. 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:

$$p_{\mathsf{y}|\mathsf{x}}(y, x) = \Pr[\mathsf{y}{=}y|\mathsf{x}{=}x] = \frac{\Pr[\mathsf{y}{=}y \text{ and } \mathsf{x}{=}x]}{\Pr[\mathsf{x}{=}x]} = \frac{p_{\mathsf{x},\mathsf{y}}(x, y)}{p_{\mathsf{x}}(x)}. \tag{6.3.1}$$

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 $\mathsf{x} \in dV_x$:

$$\Pr[\mathsf{y} \in dV_y | \mathsf{x} \in dV_x] = \frac{\Pr[\mathsf{y} \in dV_y \text{ and } \mathsf{x} \in dV_x]}{\Pr[\mathsf{x} \in dV_x]} \tag{6.3.2}$$

$$= \frac{f_{\mathsf{x},\mathsf{y}}(x,y)|dV_x||dV_y|}{f_{\mathsf{x}}(x)|dV_x|} \tag{6.3.3}$$

$$= \frac{f_{\mathsf{x},\mathsf{y}}(x,y)}{f_{\mathsf{x}}(x)}|dV_y|. \tag{6.3.4}$$

This no longer depends on dV_x, only on its location x. The conditional density is therefore

$$f_{\mathsf{y}|\mathsf{x}}(y,x) = \frac{f_{\mathsf{x},\mathsf{y}}(x,y)}{f_{\mathsf{x}}(x)}. \tag{6.3.5}$$

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 (6.3.5) follows also that the joint density function is the product of the conditional times the marginal density functions.

Problem 120. *2 points The conditional density is the joint divided by the marginal:*

$$f_{\mathsf{y}|\mathsf{x}}(y,x) = \frac{f_{\mathsf{x},\mathsf{y}}(x,y)}{f_{\mathsf{x}}(x)}. \tag{6.3.6}$$

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,

$$\int_{y=-\infty}^{+\infty} f_{\mathsf{y}|\mathsf{x}=x}(y,x)\,dy = \frac{\int_{y=-\infty}^{+\infty} f_{\mathsf{x},\mathsf{y}}(x,y)\,dy}{f_{\mathsf{x}}(x)} = \frac{f_{\mathsf{x}}(x)}{f_{\mathsf{x}}(x)} = 1 \tag{6.3.7}$$

because of the formula for the marginal:

$$f_{\mathsf{x}}(x) = \int_{y=-\infty}^{+\infty} f_{\mathsf{x},\mathsf{y}}(x,y)\,dy \tag{6.3.8}$$

You see that formula (6.3.6) divides the joint density exactly by the right number which makes the integral equal to 1. □

Problem 121. [**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* $\{\max(\mathsf{x},\mathsf{y}) \le 0.5$ *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* $\{\max(\mathsf{x},\mathsf{y}) \le 0.5$ *and* $\min(\mathsf{x},\mathsf{y}) \le 0.4\}$.

Answer. It is $\Pr[\max(\mathsf{x},\mathsf{y}) \le 0.5] - \Pr[\max(\mathsf{x},\mathsf{y}) \le 0.5 \text{ and } \min(\mathsf{x},\mathsf{y}) > 0.4]$, i.e., the area of the square from 0 to 0.5 minus the square we just had, i.e., 0.24. □

• c. *Compute* $\Pr[\max(\mathsf{x},\mathsf{y}) \le 0.5 | \min(\mathsf{x},\mathsf{y}) \le 0.4]$.

Answer.

$$\frac{\Pr[\max(\mathsf{x},\mathsf{y}) \le 0.5 \text{ and } \min(\mathsf{x},\mathsf{y}) \le 0.4]}{\Pr[\min(\mathsf{x},\mathsf{y}) \le 0.4]} = \frac{0.24}{1-0.36} = \frac{0.24}{0.64} = \frac{3}{8}. \tag{6.3.9}$$

□

• d. *Compute the joint cumulative distribution function of* u *and* v.

Answer. One good way is to do it geometrically: for arbitrary $0 \le u, v \le 1$ draw the area $\{\mathsf{u} \le u \text{ and } \mathsf{v} \le v\}$ and then derive its size. If $u \le v$ then $\Pr[\mathsf{u} \le u \text{ and } \mathsf{v} \le v] = \Pr[\mathsf{v} \le v] - \Pr[\mathsf{u} \le u \text{ and } \mathsf{v} > v] = v^2 - (v-u)^2 = 2uv - u^2$. If $u \ge v$ then $\Pr[\mathsf{u} \le u \text{ and } \mathsf{v} \le v] = \Pr[\mathsf{v} \le v] = v^2$. □

• 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

$$f_{\mathsf{u},\mathsf{v}}(u,v) = \begin{cases} 2 & \text{if } 0 \le u \le v \le 1 \\ 0 & \text{otherwise} \end{cases} \tag{6.3.10}$$

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

$$f_{\mathsf{u}}(u) = \int_{v=u}^{1} 2\,dv = 2v\Big|_u^1 = 2 - 2u \quad \text{if } 0 \le u \le 1, \qquad \text{and 0 otherwise.} \tag{6.3.11}$$

□

• 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:

$$f_{\mathsf{v}|\mathsf{u}=u}(v) = \begin{cases} \frac{1}{1-u} & \text{for } 0 \le u \le v \le 1 \\ 0 & \text{otherwise.} \end{cases} \tag{6.3.12}$$

□

6.4. 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

$$\Pr[\mathsf{x}_1 = x_1, \ldots, \mathsf{x}_r = x_r] = \frac{m!}{x_1! \cdots x_r!} p_1^{x_1} p_2^{x_2} \cdots p_r^{x_r} \tag{6.4.1}$$

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 122. *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*

(6.4.2)
$$\boldsymbol{\mu} = \mathcal{E}[\begin{bmatrix}\mathsf{x}_1\\ \vdots\\ \mathsf{x}_r\end{bmatrix}] = n\begin{bmatrix}p_1\\ p_2\\ \vdots\\ p_r\end{bmatrix} \quad \text{and} \quad \boldsymbol{\Psi} = \mathcal{V}[\begin{bmatrix}\mathsf{x}_1\\ \vdots\\ \mathsf{x}_r\end{bmatrix}] = n\begin{bmatrix}p_1-p_1^2 & -p_1p_2 & \cdots & -p_1p_r\\ -p_2p_1 & p_2-p_2^2 & \cdots & -p_2p_r\\ \vdots & \vdots & \ddots & \vdots\\ -p_rp_1 & -p_rp_2 & \cdots & p_r-p_r^2\end{bmatrix}.$$

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

(6.4.3)
$$n\begin{bmatrix}p_1-p_1^2 & -p_1p_2 & \cdots & -p_1p_r\\ -p_2p_1 & p_2-p_2^2 & \cdots & -p_2p_r\\ \vdots & \vdots & \ddots & \vdots\\ -p_rp_1 & -p_rp_2 & \cdots & p_r-p_r^2\end{bmatrix}\begin{bmatrix}1\\1\\ \vdots\\1\end{bmatrix} = \begin{bmatrix}0\\0\\ \vdots\\0\end{bmatrix}$$

□

6.5. 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 123. *4 points Prove that, if x and y are independent, then $\operatorname{E}[\mathsf{x}\mathsf{y}] = \operatorname{E}[\mathsf{x}]\operatorname{E}[\mathsf{y}]$ and therefore $\operatorname{cov}[\mathsf{x},\mathsf{y}] = 0$. (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: $\operatorname{E}[\mathsf{x}\mathsf{y}] = \operatorname{E}\Big[\operatorname{E}[\mathsf{x}\mathsf{y}|\mathsf{x}]\Big] = \operatorname{E}\Big[\mathsf{x}\operatorname{E}[\mathsf{y}|\mathsf{x}]\Big] =$ now independence is needed $= \operatorname{E}\Big[\mathsf{x}\operatorname{E}[\mathsf{y}]\Big] = \operatorname{E}[\mathsf{x}]\operatorname{E}[\mathsf{y}]$. A counterexample is given in Problem 139. □

PROBLEM 124. *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 $\operatorname{cov}[\mathsf{x},\mathsf{y}] = 0$, 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 $\mathsf{x}\mathsf{y} = \mathsf{i}_{\mathsf{A}\cap\mathsf{B}}$. If $\operatorname{cov}[\mathsf{x},\mathsf{y}] = \operatorname{E}[\mathsf{x}\mathsf{y}] - \operatorname{E}[\mathsf{x}]\operatorname{E}[\mathsf{y}] = \Pr[\mathsf{A}\cap\mathsf{B}] - \Pr[\mathsf{A}]\Pr[\mathsf{B}] = 0$, then A and B are independent. □

PROBLEM 125. *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 $\operatorname{E}[\mathsf{x}_1\mathsf{x}_2\mathsf{y}] \neq \operatorname{E}[\mathsf{x}_1\mathsf{x}_2]\operatorname{E}[\mathsf{y}]$ therefore they are not independent. □

PROBLEM 126. *4 points Prove that, if x and y are independent, then $\operatorname{var}[\mathsf{x}\mathsf{y}] = (\operatorname{E}[\mathsf{x}])^2\operatorname{var}[\mathsf{y}] + (\operatorname{E}[\mathsf{y}])^2\operatorname{var}[\mathsf{x}] + \operatorname{var}[\mathsf{x}]\operatorname{var}[\mathsf{y}]$.*

Answer. Start with result and replace all occurrences of $\operatorname{var}[z]$ with $\operatorname{E}[z^2]-\operatorname{E}[z]^2$, then multiply out: $\operatorname{E}[x]^2(\operatorname{E}[y^2]-\operatorname{E}[y]^2)+\operatorname{E}[y]^2(\operatorname{E}[x^2]-\operatorname{E}[x]^2)+(\operatorname{E}[x^2]-\operatorname{E}[x]^2)(\operatorname{E}[y^2]-\operatorname{E}[y]^2)=\operatorname{E}[x^2]\operatorname{E}[y^2]-\operatorname{E}[x]^2\operatorname{E}[y]^2=\operatorname{E}[(xy)^2]-\operatorname{E}[xy]^2$. □

6.6. 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

$$\operatorname{E}[y|x{=}x]=\frac{\int y\, f_{x,y}(x,y)\,dy}{f_x(x)}=:g(x). \tag{6.6.1}$$

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 $\operatorname{E}[y|x]$ without specifying x. This is, by definition, the random variable $g(x)$ which one gets by plugging x into g; it assigns to every outcome $\omega\in U$ the conditional expectation of y given $x{=}x(\omega)$.

Since $\operatorname{E}[y|x]$ 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:

$$\operatorname{E}\big[\operatorname{E}[y|x]\big]=\operatorname{E}[y]. \tag{6.6.2}$$

Proof (for the case that the densities exist):

$$\begin{aligned}\operatorname{E}\big[\operatorname{E}[y|x]\big]=\operatorname{E}[g(x)]&=\int\frac{\int y\, f_{x,y}(x,y)\,dy}{f_x(x)}f_x(x)\,dx\\&=\int\int y\, f_{x,y}(x,y)\,dy\,dx=\operatorname{E}[y].\end{aligned} \tag{6.6.3}$$

Problem 127. *Let x and y be two jointly distributed variables. For every fixed value x, $\operatorname{var}[y|x=x]$ is the variance of y under the conditional distribution, and $\operatorname{var}[y|x]$ is this variance as a random variable, namely, as a function of x.*

• a. *1 point Prove that*

$$\operatorname{var}[y|x]=\operatorname{E}[y^2|x]-(\operatorname{E}[y|x])^2. \tag{6.6.4}$$

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

$$\operatorname{var}[y]=\operatorname{E}[y^2]-(\operatorname{E}[y])^2 \tag{6.6.5}$$

applied to the conditional density given $x=x$. 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*

$$\operatorname{var}[y]=\operatorname{var}\big[\operatorname{E}[y|x]\big]+\operatorname{E}\big[\operatorname{var}[y|x]\big], \tag{6.6.6}$$

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 $\operatorname{E}[(\operatorname{E}[y|x])^2]-(\operatorname{E}[\operatorname{E}[y|x]])^2$, and the second term, due to (6.6.4), becomes $\operatorname{E}[\operatorname{E}[y^2|x]]-\operatorname{E}[(\operatorname{E}[y|x])^2]$. If one adds, the two $\operatorname{E}[(\operatorname{E}[y|x])^2]$ cancel out, and the other two terms can be simplified by the law of iterated expectations to give $\operatorname{E}[y^2]-(\operatorname{E}[y])^2$. □

• 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*

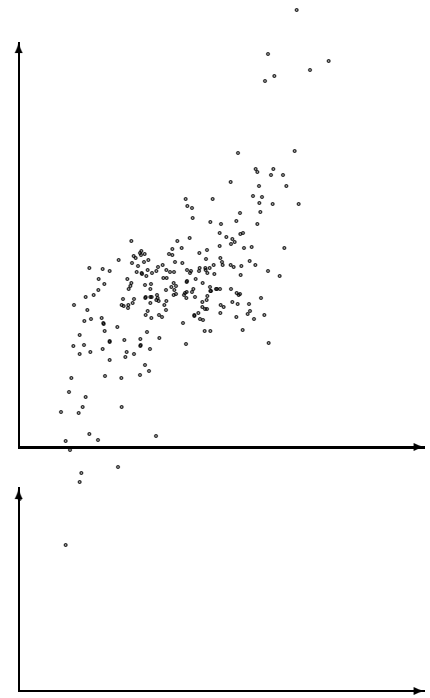
$$\frac{\operatorname{E}\big[\operatorname{var}[y|x]\big]}{\operatorname{var}\big[\operatorname{E}[y|x]\big]} \tag{6.6.7}$$

can be used to determine whether the regression function $\operatorname{E}[y|x]$ 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 128. *The figure on page 90 shows 250 independent observations of the random vector $\left[\begin{smallmatrix}x\\y\end{smallmatrix}\right]$.*



• a. *2 points Draw in by hand the approximate location of $\mathcal{E}[\left[\begin{smallmatrix}x\\y\end{smallmatrix}\right]]$ and the graph of $\operatorname{E}[y|x]$. 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* $\mathrm{E}[\mathsf{y}|\mathsf{x}]$ *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 y are independent; then the graph of $\mathrm{E}[\mathsf{y}|\mathsf{x}]$ is a horizontal line at the height of the unconditional expectation $\mathrm{E}[\mathsf{y}]$. 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 90 was generated?*

Problem 129. *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. □

6.7. Expected Values as Predictors

Expected values and conditional expected values have optimal properties as predictors.

Problem 130. *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

$$\begin{aligned}(\mathsf{y}-a)^2 &= \big((\mathsf{y}-\mu)-(a-\mu)\big)^2\\ &= (\mathsf{y}-\mu)^2 - 2(\mathsf{y}-\mu)(a-\mu) + (a-\mu)^2;\end{aligned} \tag{6.7.1}$$

$$\text{therefore}\quad \mathrm{E}[(\mathsf{y}-a)^2] = \mathrm{E}[(\mathsf{y}-\mu)^2] - 0 + (a-\mu)^2$$

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 131. *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:*

$$\mathrm{E}[\big(\mathsf{y}-h(\mathbf{x})\big)^2] \geq \mathrm{E}[\big(\mathsf{y}-\mathrm{E}[\mathsf{y}|\mathbf{x}]\big)^2]. \tag{6.7.2}$$

For this proof and the proofs required in Problems 132 and 133, 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.

$$\begin{aligned}\mathrm{E}[\big(\mathsf{y}-h(\mathbf{x})\big)^2] &= \mathrm{E}\Big[\big(\mathsf{y}-\mathrm{E}[\mathsf{y}|\mathbf{x}]-(h(\mathbf{x})-\mathrm{E}[\mathsf{y}|\mathbf{x}])\big)^2\Big]\\ &= \mathrm{E}[(\mathsf{y}-\mathrm{E}[\mathsf{y}|\mathbf{x}])^2] - 2\,\mathrm{E}[(\mathsf{y}-\mathrm{E}[\mathsf{y}|\mathbf{x}])(h(\mathbf{x})-\mathrm{E}[\mathsf{y}|\mathbf{x}])] + \mathrm{E}[(h(\mathbf{x})-\mathrm{E}[\mathsf{y}|\mathbf{x}])^2].\end{aligned} \tag{6.7.3}$$

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

$$\mathrm{E}[(\mathsf{y}-\mathrm{E}[\mathsf{y}|\mathbf{x}])(h(\mathbf{x})-\mathrm{E}[\mathsf{y}|\mathbf{x}])] = \mathrm{E}\Big[\mathrm{E}[(\mathsf{y}-\mathrm{E}[\mathsf{y}|\mathbf{x}])(h(\mathbf{x})-\mathrm{E}[\mathsf{y}|\mathbf{x}])|\mathbf{x}]\Big] \tag{6.7.4}$$

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 (6.7.4) 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 132. *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:*

$$\mathrm{E}[\varepsilon|\mathbf{x}] = 0 \tag{6.7.5}$$

$$\mathrm{E}[\varepsilon] = 0 \tag{6.7.6}$$

$$\mathrm{E}[\mathsf{x}_i\varepsilon|\mathbf{x}] = 0 \quad \textit{for all } i,\ 1 \leq i \leq j \tag{6.7.7}$$

$$\mathrm{E}[\mathsf{x}_i\varepsilon] = 0 \quad \textit{for all } i,\ 1 \leq i \leq j \tag{6.7.8}$$

$$\mathrm{cov}[\mathsf{x}_i,\varepsilon] = 0 \quad \textit{for all } i,\ 1 \leq i \leq j. \tag{6.7.9}$$

Interpretation of (6.7.9)*:* 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. (6.7.5): $\mathrm{E}[\varepsilon|\mathbf{x}] = \mathrm{E}[\mathrm{y}|\mathbf{x}] - \mathrm{E}\big[\mathrm{E}[\mathrm{y}|\mathbf{x}]|\mathbf{x}\big] = 0$ since $\mathrm{E}[\mathrm{y}|\mathbf{x}]$ 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.)

(6.7.6) follows from (6.7.5) (i.e., (6.7.5) is stronger than (6.7.6)): 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 (6.7.5): $\mathrm{E}[\varepsilon] = \mathrm{E}[\mathrm{y} - \mathrm{E}[\mathrm{y}|\mathbf{x}]] = 0$. Either way you need the law of iterated expectations for this.

(6.7.7): $\mathrm{E}[\mathrm{x}_i\varepsilon|\mathbf{x}] = \mathrm{x}_i\mathrm{E}[\varepsilon|\mathbf{x}] = 0$.

(6.7.8): $\mathrm{E}[\mathrm{x}_i\varepsilon] = \mathrm{E}\big[\mathrm{E}[\mathrm{x}_i\varepsilon|\mathbf{x}]\big] = \mathrm{E}[0] = 0$; or in one swoop: $\mathrm{E}[\mathrm{x}_i\varepsilon] = \mathrm{E}\big[\mathrm{x}_i\mathrm{y} - \mathrm{x}_i\mathrm{E}[\mathrm{y}|\mathbf{x}]\big] = \mathrm{E}\big[\mathrm{x}_i\mathrm{y} - \mathrm{E}[\mathrm{x}_i\mathrm{y}|\mathbf{x}]\big] = \mathrm{E}[\mathrm{x}_i\mathrm{y}] - \mathrm{E}[\mathrm{x}_i\mathrm{y}] = 0$. The following "proof" is not correct: $\mathrm{E}[\mathrm{x}_i\varepsilon] = \mathrm{E}[\mathrm{x}_i]\,\mathrm{E}[\varepsilon] = \mathrm{E}[\mathrm{x}_i]\cdot 0 = 0$. x_i and ε are generally not independent, therefore the multiplication rule $\mathrm{E}[\mathrm{x}_i\varepsilon] = \mathrm{E}[\mathrm{x}_i]\,\mathrm{E}[\varepsilon]$ cannot be used. Of course, the following "proof" does not work either: $\mathrm{E}[\mathrm{x}_i\varepsilon] = \mathrm{x}_i\,\mathrm{E}[\varepsilon] = \mathrm{x}_i \cdot 0 = 0$. x_i is a random variable and $\mathrm{E}[\mathrm{x}_i\varepsilon]$ is a constant; therefore $\mathrm{E}[\mathrm{x}_i\varepsilon] = \mathrm{x}_i\,\mathrm{E}[\varepsilon]$ cannot hold.

(6.7.9): $\mathrm{cov}[\mathrm{x}_i, \varepsilon] = \mathrm{E}[\mathrm{x}_i\varepsilon] - \mathrm{E}[\mathrm{x}_i]\,\mathrm{E}[\varepsilon] = 0 - \mathrm{E}[\mathrm{x}_i] \cdot 0 = 0$. □

• 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., $\mathrm{E}[\varepsilon^2] = \mathrm{E}\big[\mathrm{E}[\varepsilon^2|\mathbf{x}]\big] =$ constant (does not depend on $\mathbf{x}$). □

Problem 133. *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* $\mathrm{c} = \mathrm{f}(\mathrm{y}_1, \ldots, \mathrm{y}_n)$*. 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* $\mathrm{y}_1, \ldots, \mathrm{y}_n$ *are known to him. Therefore in period* t*, when he only knows the values of* $\mathrm{y}_1, \ldots, \mathrm{y}_t$*, but not yet the future values, the consumer decides to consume the amount* $\mathrm{c}_t = \mathrm{E}[\mathrm{c}|\mathrm{y}_1, \ldots, \mathrm{y}_t]$*, which is the best possible prediction of* c *given the information available to him. Show that in this situation,* $\mathrm{c}_{t+1} - \mathrm{c}_t$ *is uncorrelated with all* $\mathrm{y}_1, \ldots, \mathrm{y}_t$*. 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:*

$$\mathrm{E}\big[\mathrm{E}[\mathrm{x}|\mathbf{y}, \mathbf{z}]\big|\mathbf{y}\big] = \mathrm{E}[\mathrm{x}|\mathbf{y}]. \tag{6.7.10}$$

Here is an explanation of (6.7.10)*:* $\mathrm{E}[\mathrm{x}|\mathbf{y}]$ *is the best predictor of* x *based on information set* $\mathbf{y}$*.* $\mathrm{E}[\mathrm{x}|\mathbf{y}, \mathbf{z}]$ *is the best predictor of* x *based on the extended information set consisting of* $\mathbf{y}$ *and* $\mathbf{z}$*.* $\mathrm{E}\big[\mathrm{E}[\mathrm{x}|\mathbf{y}, \mathbf{z}]\big|\mathbf{y}\big]$ *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* $\mathrm{E}[\mathrm{x}|\mathbf{y}]$*, i.e.,* (6.7.10) *says therefore that I cannot predict how I will change my mind after better information becomes available.*

Answer. In (6.7.10) set $\mathrm{x} = \mathrm{c} = f(\mathrm{y}_1, \ldots, \mathrm{y}_t, \mathrm{y}_{t+1}, \ldots, \mathrm{y}_n)$, $\mathbf{y} = [\mathrm{y}_1, \ldots, \mathrm{y}_t]^\top$, and $\mathbf{z} = \mathrm{y}_{t+1}$ to get

$$\mathrm{E}\big[\mathrm{E}[\mathrm{c}|\mathrm{y}_1, \ldots, \mathrm{y}_{t+1}]\big|\mathrm{y}_1, \ldots, \mathrm{y}_t\big] = \mathrm{E}[\mathrm{c}|\mathrm{y}_1, \ldots, \mathrm{y}_t]. \tag{6.7.11}$$

Writing c_t for $\mathrm{E}[\mathrm{c}|\mathrm{y}_1, \ldots, \mathrm{y}_t]$, this becomes $\mathrm{E}[\mathrm{c}_{t+1}|\mathrm{y}_1, \ldots, \mathrm{y}_t] = \mathrm{c}_t$, i.e., c_t is not only the best predictor of c, but also that of c_{t+1}. The change in consumption $\mathrm{c}_{t+1} - \mathrm{c}_t$ is therefore the prediction error, which is uncorrelated with the conditioning variables, as shown in Problem 132. □

Problem 134. *3 points Show that for any two random variables* x *and* y *whose covariance exists, the following equation holds:*

$$\mathrm{cov}[\mathrm{x}, \mathrm{y}] = \mathrm{cov}\big[\mathrm{x}, \mathrm{E}[\mathrm{y}|\mathrm{x}]\big] \tag{6.7.12}$$

Note: Since $\mathrm{E}[\mathrm{y}|\mathrm{x}]$ *is the best predictor of* y *based on the observation of* x*,* (6.7.12) *can also be written as*

$$\mathrm{cov}\big[\mathrm{x}, (\mathrm{y} - \mathrm{E}[\mathrm{y}|\mathrm{x}])\big] = 0, \tag{6.7.13}$$

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 (6.1.6) to the righthand side of (6.7.12):

$$\mathrm{cov}\big[\mathrm{x}, \mathrm{E}[\mathrm{y}|\mathrm{x}]\big] = \mathrm{E}\big[\mathrm{x}\mathrm{E}[\mathrm{y}|\mathrm{x}]\big] - \mathrm{E}[\mathrm{x}]\,\mathrm{E}\big[\mathrm{E}[\mathrm{y}|\mathrm{x}]\big] = \mathrm{E}\big[\mathrm{E}[\mathrm{x}\mathrm{y}|\mathrm{x}]\big] - \mathrm{E}[\mathrm{x}]\,\mathrm{E}[\mathrm{y}] = \mathrm{E}[\mathrm{x}\mathrm{y}] - \mathrm{E}[\mathrm{x}]\,\mathrm{E}[\mathrm{y}] = \mathrm{cov}[\mathrm{x}, \mathrm{y}]. \tag{6.7.14}$$

The tricky part here is to see that $\mathrm{x}\mathrm{E}[\mathrm{y}|\mathrm{x}] = \mathrm{E}[\mathrm{x}\mathrm{y}|\mathrm{x}]$.

□

Problem 135. *Assume* x *and* y *have a joint density function* $f_{\mathrm{x},\mathrm{y}}(x, y)$ *which is symmetric about the* x*-axis, i.e.,*

$$f_{\mathrm{x},\mathrm{y}}(x, y) = f_{\mathrm{x},\mathrm{y}}(x, -y).$$

Also assume that variances and covariances exist. Show that $\mathrm{cov}[\mathrm{x}, \mathrm{y}] = 0$*. Hint: one way to do it is to look at* $\mathrm{E}[\mathrm{y}|\mathrm{x}]$*.*

Answer. We know that $\mathrm{cov}[\mathrm{x}, \mathrm{y}] = \mathrm{cov}\big[\mathrm{x}, \mathrm{E}[\mathrm{y}|\mathrm{x}]\big]$. Furthermore, from symmetry follows $\mathrm{E}[\mathrm{y}|\mathrm{x}] = 0$. Therefore $\mathrm{cov}[\mathrm{x}, \mathrm{y}] = \mathrm{cov}[\mathrm{x}, 0] = 0$. Here is a detailed proof of $\mathrm{E}[\mathrm{y}|\mathrm{x}] = 0$: $\mathrm{E}[\mathrm{y}|\mathrm{x}{=}x] = \int_{-\infty}^{\infty} y\frac{f_{\mathrm{x},\mathrm{y}}(x,y)}{f_{\mathrm{x}}(x)}dy$. Now substitute $z = -y$, then also $dz = -dy$, and the boundaries of integration are reversed:

$$\mathrm{E}[\mathrm{y}|\mathrm{x}{=}x] = \int_{\infty}^{-\infty} z\frac{f_{\mathrm{x},\mathrm{y}}(x, -z)}{f_{\mathrm{x}}(x)}dz = \int_{\infty}^{-\infty} z\frac{f_{\mathrm{x},\mathrm{y}}(x, z)}{f_{\mathrm{x}}(x)}dz = -\,\mathrm{E}[\mathrm{y}|\mathrm{x}{=}x]. \tag{6.7.15}$$

One can also prove directly under this presupposition $\mathrm{cov}[\mathrm{x}, \mathrm{y}] = \mathrm{cov}[\mathrm{x}, -\mathrm{y}]$ and therefore it must be zero.

□

Problem 136. [**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* $\mathrm{p} = \mathrm{m} + \gamma\mathrm{x}$*, 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 $\mathfrak{m}$ *and* $\mathfrak{x}$ *and the value of* γ*. But they are unable to observe* $\mathfrak{m}$ *and* $\mathfrak{x}$*, they can only observe* $\mathfrak{p}$*. Then the best predictor of* $\mathfrak{x}$ *using* $\mathfrak{p}$ *is the conditional expectation* $\mathrm{E}[\mathfrak{x}|\mathfrak{p}]$.

• a. *Assume you are one of these agents and you observe* $\mathfrak{p} = p$*. How great would you predict* $\mathfrak{x}$ *to be, i.e., what is the value of* $\mathrm{E}[\mathfrak{x}|\mathfrak{p} = p]$*?*

Answer. It is, according to formula (7.3.18), $\mathrm{E}[\mathfrak{x}|\mathfrak{p} = p] = \mu_{\mathfrak{x}} + \frac{\mathrm{cov}(\mathfrak{x},\mathfrak{p})}{\mathrm{var}(\mathfrak{p})}(p - \mathrm{E}[\mathfrak{p}])$. Now $\mathrm{E}[\mathfrak{p}] = \mu_{\mathfrak{m}} + \gamma\mu_{\mathfrak{x}}$, $\mathrm{cov}[\mathfrak{x},\mathfrak{p}] = \mathrm{cov}[\mathfrak{x},\mathfrak{m}] + \gamma\,\mathrm{cov}[\mathfrak{x},\mathfrak{x}] = \gamma\sigma^2_{\mathfrak{x}}$, and $\mathrm{var}(\mathfrak{p}) = \sigma^2_{\mathfrak{m}} + \gamma^2\sigma^2_{\mathfrak{x}}$. Therefore

$$\mathrm{E}[\mathfrak{x}|\mathfrak{p} = p] = \mu_{\mathfrak{x}} + \frac{\gamma\sigma^2_{\mathfrak{x}}}{\sigma^2_{\mathfrak{m}} + \gamma^2\sigma^2_{\mathfrak{x}}}(p - \mu_{\mathfrak{m}} - \gamma\mu_{\mathfrak{x}}). \tag{6.7.16}$$

□

• b. *Define the prediction error* $\varepsilon = \mathfrak{x} - \mathrm{E}[\mathfrak{x}|\mathfrak{p}]$*. Compute expected value and variance of* ε*.*

Answer.

$$\varepsilon = \mathfrak{x} - \mu_{\mathfrak{x}} - \frac{\gamma\sigma^2_{\mathfrak{x}}}{\sigma^2_{\mathfrak{m}} + \gamma^2\sigma^2_{\mathfrak{x}}}(\mathfrak{p} - \mu_{\mathfrak{m}} - \gamma\mu_{\mathfrak{x}}). \tag{6.7.17}$$

This has zero expected value, and its variance is

$$\mathrm{var}[\varepsilon] = \mathrm{var}[\mathfrak{x}] + \Big(\frac{\gamma\sigma^2_{\mathfrak{x}}}{\sigma^2_{\mathfrak{m}} + \gamma^2\sigma^2_{\mathfrak{x}}}\Big)^2 \mathrm{var}[\mathfrak{p}] - 2\Big(\frac{\gamma\sigma^2_{\mathfrak{x}}}{\sigma^2_{\mathfrak{m}} + \gamma^2\sigma^2_{\mathfrak{x}}}\Big)\,\mathrm{cov}[\mathfrak{x},\mathfrak{p}] = \tag{6.7.18}$$

$$= \sigma^2_{\mathfrak{x}} + \frac{\gamma^2(\sigma^2_{\mathfrak{x}})^2}{\sigma^2_{\mathfrak{m}} + \gamma^2\sigma^2_{\mathfrak{x}}} - 2\frac{\gamma^2(\sigma^2_{\mathfrak{x}})^2}{\sigma^2_{\mathfrak{m}} + \gamma^2\sigma^2_{\mathfrak{x}}} \tag{6.7.19}$$

$$= \frac{\sigma^2_{\mathfrak{x}}\sigma^2_{\mathfrak{m}}}{\sigma^2_{\mathfrak{m}} + \gamma^2\sigma^2_{\mathfrak{x}}} = \frac{\sigma^2_{\mathfrak{x}}}{1 + \gamma^2\sigma^2_{\mathfrak{x}}/\sigma^2_{\mathfrak{m}}}. \tag{6.7.20}$$

□

• c. *In an attempt to fine tune the economy, the central bank increases* $\sigma^2_{\mathfrak{m}}$*. Does that increase or decrease* $\mathrm{var}(\varepsilon)$*?*

Answer. From (6.7.20) follows that it increases the variance. □

6.8. 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 $\begin{bmatrix}\mathfrak{x}\\ \mathfrak{y}\end{bmatrix}$ is known, set $\begin{bmatrix}\mathfrak{x}\\ \mathfrak{y}\end{bmatrix} = \boldsymbol{t}(\mathfrak{u},\mathfrak{v})$. Here $\boldsymbol{t}$ is a vector-valued function, (i.e., it could be written $\boldsymbol{t}(\mathfrak{u},\mathfrak{v}) = \begin{bmatrix}\mathfrak{q}(\mathfrak{u},\mathfrak{v})\\ \mathfrak{r}(\mathfrak{u},\mathfrak{v})\end{bmatrix}$, 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 $\mathfrak{x}$ and $\mathfrak{y}$ is described by a probability mass function, then the joint probability mass function of $\mathfrak{u}$ and $\mathfrak{v}$ can simply be obtained by substituting $\boldsymbol{t}$ into the joint probability mass function of $\mathfrak{x}$ and $\mathfrak{y}$ (and it is zero for any values which are not in A):

$$p_{\mathfrak{u},\mathfrak{v}}(u,v) = \Pr\Big[\begin{bmatrix}\mathfrak{u}\\ \mathfrak{v}\end{bmatrix} = \begin{bmatrix}u\\ v\end{bmatrix}\Big] = \Pr\big[\boldsymbol{t}(\mathfrak{u},\mathfrak{v}) = \boldsymbol{t}(u,v)\big] = \Pr\Big[\begin{bmatrix}\mathfrak{x}\\ \mathfrak{y}\end{bmatrix} = \boldsymbol{t}(u,v)\Big] = p_{\mathfrak{x},\mathfrak{y}}\big(\boldsymbol{t}(u,v)\big). \tag{6.8.1}$$

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

$$\Big\{\begin{bmatrix}\mathfrak{u}\\ \mathfrak{v}\end{bmatrix} \in dV_{u,v}\Big\} = \{\boldsymbol{t}(\mathfrak{u},\mathfrak{v}) \in \boldsymbol{t}(dV)_{x,y}\}. \tag{6.8.2}$$

Therefore

$$f_{\mathfrak{u},\mathfrak{v}}(u,v)|dV_{u,v}| = \Pr\Big[\begin{bmatrix}\mathfrak{u}\\ \mathfrak{v}\end{bmatrix} \in dV_{u,v}\Big] = \Pr[\boldsymbol{t}(\mathfrak{u},\mathfrak{v}) \in \boldsymbol{t}(dV)_{x,y}] = f_{\mathfrak{x},\mathfrak{y}}(\boldsymbol{t}(u,v))|\boldsymbol{t}(dV)_{x,y}| = \tag{6.8.3}$$

$$= f_{\mathfrak{x},\mathfrak{y}}(\boldsymbol{t}(u,v))\frac{|\boldsymbol{t}(dV)_{x,y}|}{|dV_{u,v}|}|dV_{u,v}|. \tag{6.8.4}$$

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

$$J = \begin{vmatrix}\frac{\partial\mathfrak{x}}{\partial\mathfrak{u}} & \frac{\partial\mathfrak{x}}{\partial\mathfrak{v}}\\ \frac{\partial\mathfrak{y}}{\partial\mathfrak{u}} & \frac{\partial\mathfrak{y}}{\partial\mathfrak{v}}\end{vmatrix} = \begin{vmatrix}\frac{\partial\mathfrak{q}}{\partial u}(u,v) & \frac{\partial\mathfrak{q}}{\partial v}(u,v)\\ \frac{\partial\mathfrak{r}}{\partial u}(u,v) & \frac{\partial\mathfrak{r}}{\partial v}(u,v)\end{vmatrix}. \tag{6.8.5}$$

Remember, $\mathfrak{u},\mathfrak{v}$ are the new and $\mathfrak{x},\mathfrak{y}$ 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:

$$(\mathfrak{x},\mathfrak{y}) = \boldsymbol{t}(\mathfrak{u},\mathfrak{v}) \quad \text{one-to-one} \quad \Rightarrow \quad f_{\mathfrak{u},\mathfrak{v}}(u,v) = f_{\mathfrak{x},\mathfrak{y}}\big(\boldsymbol{t}(u,v)\big)\,|J| \quad \text{where} \quad J = \begin{vmatrix}\frac{\partial\mathfrak{x}}{\partial\mathfrak{u}} & \frac{\partial\mathfrak{x}}{\partial\mathfrak{v}}\\ \frac{\partial\mathfrak{y}}{\partial\mathfrak{u}} & \frac{\partial\mathfrak{y}}{\partial\mathfrak{v}}\end{vmatrix}.$$

Problem 137. *Let* $\mathfrak{x}$ *and* $\mathfrak{y}$ *be two random variables with joint density function* $f_{\mathfrak{x},\mathfrak{y}}(x,y)$*.*

• a. *3 points Define* $\mathfrak{u} = \mathfrak{x} + \mathfrak{y}$*. Derive the joint density function of* $\mathfrak{u}$ *and* $\mathfrak{y}$*.*

Answer. You have to express the "old" $\mathfrak{x}$ and $\mathfrak{y}$ as functions of the "new" $\mathfrak{u}$ and $\mathfrak{y}$:

$$\begin{aligned}\mathfrak{x} &= \mathfrak{u} - \mathfrak{y}\\ \mathfrak{y} &= \mathfrak{y}\end{aligned} \quad \text{or} \quad \begin{bmatrix}\mathfrak{x}\\ \mathfrak{y}\end{bmatrix} = \begin{bmatrix}1 & -1\\ 0 & 1\end{bmatrix}\begin{bmatrix}\mathfrak{u}\\ \mathfrak{y}\end{bmatrix} \quad \text{therefore} \quad J = \begin{vmatrix}\frac{\partial\mathfrak{x}}{\partial\mathfrak{u}} & \frac{\partial\mathfrak{x}}{\partial\mathfrak{y}}\\ \frac{\partial\mathfrak{y}}{\partial\mathfrak{u}} & \frac{\partial\mathfrak{y}}{\partial\mathfrak{y}}\end{vmatrix} = \begin{vmatrix}1 & -1\\ 0 & 1\end{vmatrix} = 1.$$

Therefore

$$f_{\mathfrak{u},\mathfrak{y}}(u,y) = f_{\mathfrak{x},\mathfrak{y}}(u - y, y). \tag{6.8.6}$$

□

• b. *1 point Derive from this the following formula computing the density function $f_{\mathsf{u}}(u)$ of the sum $\mathsf{u} = \mathsf{x} + \mathsf{y}$ from the joint density function $f_{\mathsf{x},\mathsf{y}}(x, y)$ of x and y.*

$$f_{\mathsf{u}}(u) = \int_{y=-\infty}^{y=\infty} f_{\mathsf{x},\mathsf{y}}(u - y, y)dy. \tag{6.8.7}$$

Answer. Write down the joint density of u and y and then integrate y out, i.e., take its integral over y from $-\infty$ to $+\infty$:

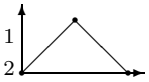
$$f_{\mathsf{u}}(u) = \int_{y=-\infty}^{y=\infty} f_{\mathsf{u},\mathsf{y}}(u, y)dy = \int_{y=-\infty}^{y=\infty} f_{\mathsf{x},\mathsf{y}}(u - y, y)dy. \tag{6.8.8}$$

i.e., one integrates over all $\begin{bmatrix} x \\ y \end{bmatrix}$ with $x + y = u$. □

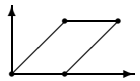
Problem 138. *6 points Let x and y be independent and uniformly distributed over the interval $[0, 1]$. Compute the density function of $\mathsf{u} = \mathsf{x} + \mathsf{y}$ and draw its graph. Hint: you may use formula (6.8.7) for the density of the sum of two jointly distributed random variables. An alternative approach would be to first compute the cumulative distribution function $\Pr[\mathsf{x} + \mathsf{y} \leq u]$ for all u.*

Answer. Using equation (6.8.7):

$$f_{\mathsf{x}+\mathsf{y}}(u) = \int_{-\infty}^{\infty} f_{\mathsf{x},\mathsf{y}}(u - y, y)\, dy = \begin{cases} u & \text{for } 0 \leq u \leq 1 \\ 2 - u & \text{for } 1 \leq u \leq 2 \\ 0 & \text{otherwise.} \end{cases} \tag{6.8.9}$$



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 $f_{\mathsf{x},\mathsf{y}}(u - v, v)$ 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: $f_{\mathsf{x},\mathsf{y}}(x, y) = 1$ 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 $f_{\mathsf{x},\mathsf{y}}(u - y, y) = 1$ 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 $f_{\mathsf{x},\mathsf{y}}(u - y, y) = 1$ iff $u - 1 \leq y \leq 1$. □

Problem 139. *Assume $\left[\begin{smallmatrix} \mathsf{x} \\ \mathsf{y} \end{smallmatrix}\right]$ is uniformly distributed on a round disk around the origin with radius 10.*

• a. *4 points Derive the joint density, the marginal density of x, and the conditional density of y given x=x.*

• b. *3 points Now let us go over to polar coordinates r and ϕ, which satisfy*

$$\begin{aligned} \mathsf{x} &= \mathsf{r}\cos\phi \\ \mathsf{y} &= \mathsf{r}\sin\phi \end{aligned}, \quad \textit{i.e., the vector transformation } \boldsymbol{t} \textit{ is } \quad \boldsymbol{t}\left(\begin{bmatrix} \mathsf{r} \\ \phi \end{bmatrix}\right) = \begin{bmatrix} \mathsf{r}\cos\phi \\ \mathsf{r}\sin\phi \end{bmatrix}. \tag{6.8.10}$$

Which region in $\left(\begin{smallmatrix} \mathsf{r} \\ \phi \end{smallmatrix}\right)$-space is necessary to cover $\left(\begin{smallmatrix} \mathsf{x} \\ \mathsf{y} \end{smallmatrix}\right)$-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* $\operatorname{cov}[\mathsf{x}, \mathsf{y}]$.

• d. *2 points Compute the conditional variance* $\operatorname{var}[\mathsf{y}|\mathsf{x}{=}x]$.

• e. *2 points Are x and y independent?*

Problem 140. [**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 $\mathsf{u}_1(\omega), \mathsf{u}_2(\omega), \mathsf{u}_3(\omega)$. Here u_1, u_2, and u_3 are assumed to be* independent *random variables with the following cumulative distribution functions:*

$$\Pr[\mathsf{u}_i \leq u] = F_i(u) = \exp\bigl(-\exp(\mu_i - u)\bigr), \qquad i = 1, 2, 3. \tag{6.8.11}$$

*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 $\mathsf{u}_3(\omega) \geq \mathsf{u}_1(\omega)$ and $\mathsf{u}_3(\omega) \geq \mathsf{u}_2(\omega)$. Let I be a function of three arguments such that $I(\mathsf{u}_1, \mathsf{u}_2, \mathsf{u}_3)$ is the indicator function of the event that one randomly chooses an individual ω who likes cars best, i.e.,

$$I(u_1, u_2, u_3) = \begin{cases} 1 & \textit{if } u_1 \leq u_3 \textit{ and } u_2 \leq u_3 \\ 0 & \textit{otherwise.} \end{cases} \tag{6.8.12}$$

Then $\Pr[car] = \mathrm{E}[I(\mathsf{u}_1, \mathsf{u}_2, \mathsf{u}_3)]$. The following steps have the purpose to compute this probability:

• b. *2 points For any fixed number u, define $g(u) = \mathrm{E}[I(\mathsf{u}_1, \mathsf{u}_2, \mathsf{u}_3)|\mathsf{u}_3 = u]$. Show that*

$$g(u) = \exp\bigl(-\exp(\mu_1 - u) - \exp(\mu_2 - u)\bigr). \tag{6.8.13}$$

• 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*

$$\Pr[car] = \frac{\exp \mu_3}{\exp \mu_1 + \exp \mu_2 + \exp \mu_3}. \tag{6.8.14}$$

CHAPTER 7

The Multivariate Normal Probability Distribution

7.1. 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

$$f_{\mathsf{z}}(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}. \tag{7.1.1}$$

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

$$\int_{y=-\infty}^{y=\infty} \int_{x=-\infty}^{x=\infty} \frac{1}{2\pi} e^{-\frac{x^2+y^2}{2}}\,dx\,dy = \int_{\phi=0}^{2\pi} \int_{r=0}^{\infty} \frac{1}{2\pi} e^{-\frac{r^2}{2}} r\,dr\,d\phi. \tag{7.1.2}$$

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 141. *6 points The Gamma function can be defined as* $\Gamma(r) = \int_0^\infty x^{r-1} e^{-x}$ *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:

$$\int_0^\infty \frac{1}{\sqrt{x}} e^{-x}\,dx = \sqrt{2}\int_0^\infty e^{-z^2/2}\,dz = \frac{1}{\sqrt{2}}\int_{-\infty}^{\infty} e^{-z^2/2}\,dz = \frac{\sqrt{2\pi}}{\sqrt{2}} = \sqrt{\pi}. \tag{7.1.3}$$

□

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:)

$$f_{\mathsf{x}}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = (2\pi\sigma^2)^{-1/2} \exp\bigl(-(x-\mu)^2/2\sigma^2\bigr). \tag{7.1.4}$$

Problem 142. *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*

$$f_{\mathsf{y}}(y_1, \ldots, y_n; \theta) = g\bigl(T(y_1, \ldots, y_n); \theta\bigr) \cdot h(y_1, \ldots, y_n). \tag{7.1.5}$$

Answer. The joint density function can be written (factorization indicated by $\cdot$):

(7.1.6)

$$(2\pi)^{-n/2} \exp\Bigl(-\frac{1}{2}\sum_{i=1}^n (y_i-\mu)^2\Bigr) = (2\pi)^{-n/2} \exp\Bigl(-\frac{1}{2}\sum_{i=1}^n (y_i-\bar{y})^2\Bigr) \cdot \exp\Bigl(-\frac{n}{2}(\bar{y}-\mu)^2\Bigr) = h(y_1, \ldots, y_n) \cdot g(\bar{y};$$

□

7.2. 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 $\mathbf{y}$ 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 $\boldsymbol{A}\mathbf{y} + \boldsymbol{b}$ 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.

7.3. 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:

$$f_{\mathsf{x},\mathsf{y}}(x,y) = \frac{1}{2\pi\sigma^2}\exp\Big(-\frac{1}{2\sigma^2}(x^2+y^2)\Big). \tag{7.3.1}$$

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

$$f_{\mathsf{u},\mathsf{v}}(u,v) = \frac{1}{2\pi\sigma^2}\left|\det(\boldsymbol{P}^{-1})\right|\exp\Big(-\frac{1}{2\sigma^2}\begin{bmatrix}u-\mu\\ v-\nu\end{bmatrix}^\top \boldsymbol{P}^{-1\top}\boldsymbol{P}^{-1}\begin{bmatrix}u-\mu\\ v-\nu\end{bmatrix}\Big). \tag{7.3.2}$$

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

$$f_{\mathsf{u},\mathsf{v}}(u,v) = \frac{1}{2\pi\sigma^2}\frac{1}{\sqrt{\det(\boldsymbol{\Psi})}}\exp\Big(-\frac{1}{2\sigma^2}\begin{bmatrix}u-\mu\\ v-\nu\end{bmatrix}^\top \boldsymbol{\Psi}^{-1}\begin{bmatrix}u-\mu\\ v-\nu\end{bmatrix}\Big). \tag{7.3.3}$$

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 $\sigma^2\boldsymbol{\Psi}$:

$$f_{\boldsymbol{\mathsf{x}}}(\boldsymbol{x}) = (2\pi\sigma^2)^{-n/2}(\det\boldsymbol{\Psi})^{-1/2}\exp\Big(-\frac{1}{2\sigma^2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Psi}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\Big). \tag{7.3.4}$$

Problem 143. *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 144. *3 points All vectors in this question are $n\times 1$ column vectors. Let $\boldsymbol{\mathsf{y}} = \boldsymbol{\alpha}+\boldsymbol{\varepsilon}$, 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.,*

$$\mathcal{V}[\boldsymbol{T}\boldsymbol{\varepsilon}] = \sigma^2\boldsymbol{I}_n. \tag{7.3.5}$$

Show that in this case the density function of $\boldsymbol{\mathsf{y}}$ is

$$f_{\boldsymbol{\mathsf{y}}}(\boldsymbol{y}) = (2\pi\sigma^2)^{-n/2}\left|\det(\boldsymbol{T})\right|\exp\Big(-\frac{1}{2\sigma^2}\big(\boldsymbol{T}(\boldsymbol{y}-\boldsymbol{\alpha})\big)^\top\boldsymbol{T}(\boldsymbol{y}-\boldsymbol{\alpha})\Big). \tag{7.3.6}$$

Hint: define $\boldsymbol{z} = \boldsymbol{T}\boldsymbol{\varepsilon}$, write down the density function of $\boldsymbol{z}$, and make a transformation between $\boldsymbol{z}$ and $\boldsymbol{\mathsf{y}}$.

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 $\boldsymbol{\mathsf{y}}$, whose density function we want to know. $\boldsymbol{z} = \boldsymbol{T}(\boldsymbol{\mathsf{y}}-\boldsymbol{\alpha})$ or

$$z_1 = t_{11}(y_1-\alpha_1)+t_{12}(y_2-\alpha_2)+\cdots+t_{1n}(y_n-\alpha_n) \tag{7.3.7}$$

$$\vdots \tag{7.3.8}$$

$$z_n = t_{n1}(y_1-\alpha_1)+t_{n2}(y_1-\alpha_2)+\cdots+t_{nn}(y_n-\alpha_n) \tag{7.3.9}$$

therefore the Jacobian determinant is $\det(\boldsymbol{T})$. This gives the result. □

7.3.1. Most Natural Form of Bivariate Normal Density.

Problem 145. *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 $\boldsymbol{\Psi} = \mathcal{V}[\begin{bmatrix}\mathsf{u}\\ \mathsf{v}\end{bmatrix}]$ in terms of the standard deviations σ_{u} and σ_{v} and the correlation coefficient ρ.*

• b. *1 point Show that the inverse of a 2×2 matrix has the following form:*

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}. \tag{7.3.10}$$

• c. *2 points Show that*

$$\mathsf{q}^2 = \begin{bmatrix} \mathsf{u} - \mu & \mathsf{v} - \nu \end{bmatrix} \boldsymbol{\Psi}^{-1} \begin{bmatrix} \mathsf{u} - \mu \\ \mathsf{v} - \nu \end{bmatrix} \tag{7.3.11}$$

$$= \frac{1}{1 - \rho^2} \Big(\frac{(\mathsf{u} - \mu)^2}{\sigma_{\mathsf{u}}^2} - 2\rho \frac{\mathsf{u} - \mu}{\sigma_{\mathsf{u}}} \frac{\mathsf{v} - \nu}{\sigma_{\mathsf{v}}} + \frac{(\mathsf{v} - \nu)^2}{\sigma_{\mathsf{v}}^2} \Big). \tag{7.3.12}$$

• d. *2 points Show the following quadratic decomposition:*

$$\mathsf{q}^2 = \frac{(\mathsf{u} - \mu)^2}{\sigma_{\mathsf{u}}^2} + \frac{1}{(1 - \rho^2)\sigma_{\mathsf{v}}^2} \Big(\mathsf{v} - \nu - \rho \frac{\sigma_{\mathsf{v}}}{\sigma_{\mathsf{u}}} (\mathsf{u} - \mu) \Big)^2. \tag{7.3.13}$$

• e. *1 point Show that* (7.3.13) *can also be written in the form*

$$\mathsf{q}^2 = \frac{(\mathsf{u} - \mu)^2}{\sigma_{\mathsf{u}}^2} + \frac{\sigma_{\mathsf{u}}^2}{\sigma_{\mathsf{u}}^2 \sigma_{\mathsf{v}}^2 - (\sigma_{\mathsf{uv}})^2} \Big(\mathsf{v} - \nu - \frac{\sigma_{\mathsf{uv}}}{\sigma_{\mathsf{u}}^2} (\mathsf{u} - \mu) \Big)^2. \tag{7.3.14}$$

• f. *1 point Show that $d = \sqrt{\det \boldsymbol{\Psi}}$ can be split up, not additively but multiplicatively, as follows: $d = \sigma_{\mathsf{u}} \cdot \sigma_{\mathsf{v}} \sqrt{1 - \rho^2}$.*

• g. *1 point Using these decompositions of d and q^2, show that the density function $f_{\mathsf{u},\mathsf{v}}(u, v)$ reads*

(7.3.15)

$$\frac{1}{\sqrt{2\pi\sigma_{\mathsf{u}}^2}} \exp\Big(-\frac{(u - \mu)^2}{2\sigma_{\mathsf{u}}^2} \Big) \cdot \frac{1}{\sqrt{2\pi\sigma_{\mathsf{v}}^2}\sqrt{1 - \rho^2}} \exp\Big(-\frac{\big((v - \nu) - \rho\frac{\sigma_{\mathsf{v}}}{\sigma_{\mathsf{u}}}(u - \mu)\big)^2}{2(1 - \rho^2)\sigma_{\mathsf{v}}^2} \Big).$$

The second factor in (7.3.15) is the density of a $\mathsf{N}(\rho\frac{\sigma_{\mathsf{v}}}{\sigma_{\mathsf{u}}}u, (1 - \rho^2)\sigma_{\mathsf{v}}^2)$ evaluated at v, and the first factor does not depend on v. Therefore if I integrate v out to get the marginal density of u, this simply gives me the first factor. The conditional density of v given $\mathsf{u} = u$ 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: $f_{\mathsf{u},\mathsf{v}}(u, v) = f_{\mathsf{u}}(u) \cdot f_{\mathsf{v}|\mathsf{u}}(v; u)$.

From this decomposition one can draw the following conclusions:

- $\mathsf{u} \sim \mathsf{N}(0, \sigma_{\mathsf{u}}^2)$ is normal and, by symmetry, v is normal as well. Note that u (or v) can be chosen to be any nonzero linear combination of x and 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 v conditionally on $\mathsf{u} = u$ is the second term on the rhs of (7.3.15), i.e., it is normal too.

- The conditional mean is

$$\mathrm{E}[\mathsf{v}|\mathsf{u} = u] = \rho\frac{\sigma_{\mathsf{v}}}{\sigma_{\mathsf{u}}}u, \tag{7.3.16}$$

i.e., it is a linear function of u. If the (unconditional) means are not zero, then the conditional mean is

$$\mathrm{E}[\mathsf{v}|\mathsf{u} = u] = \mu_{\mathsf{v}} + \rho\frac{\sigma_{\mathsf{v}}}{\sigma_{\mathsf{u}}}(u - \mu_{\mathsf{u}}). \tag{7.3.17}$$

Since $\rho = \frac{\mathrm{cov}[\mathsf{u},\mathsf{v}]}{\sigma_{\mathsf{u}}\sigma_{\mathsf{v}}}$, (7.3.17) can als be written as follows:

$$\mathrm{E}[\mathsf{v}|\mathsf{u} = u] = \mathrm{E}[\mathsf{v}] + \frac{\mathrm{cov}[\mathsf{u}, \mathsf{v}]}{\mathrm{var}[\mathsf{u}]}(u - \mathrm{E}[\mathsf{u}]) \tag{7.3.18}$$

- The conditional variance is the same whatever value of u was chosen: its value is

$$\mathrm{var}[\mathsf{v}|\mathsf{u} = u] = \sigma_{\mathsf{v}}^2(1 - \rho^2), \tag{7.3.19}$$

which can also be written as

$$\mathrm{var}[\mathsf{v}|\mathsf{u} = u] = \mathrm{var}[\mathsf{v}] - \frac{(\mathrm{cov}[\mathsf{u}, \mathsf{v}])^2}{\mathrm{var}[\mathsf{u}]}. \tag{7.3.20}$$

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.

7.3.2. Level Lines of the Normal Density.

Problem 146. *8 points Define the angle $\delta = \arccos(\rho)$, i.e, $\rho = \cos\delta$. In terms of δ, the covariance matrix* (**??**) *has the form*

$$\boldsymbol{\Psi} = \begin{bmatrix} \sigma_{\mathsf{u}}^2 & \sigma_{\mathsf{u}}\sigma_{\mathsf{v}}\cos\delta \\ \sigma_{\mathsf{u}}\sigma_{\mathsf{v}}\cos\delta & \sigma_{\mathsf{v}}^2 \end{bmatrix} \tag{7.3.21}$$

Show that for all ϕ, the vector

$$\boldsymbol{x} = \begin{bmatrix} r\,\sigma_{\mathsf{u}}\cos\phi \\ r\,\sigma_{\mathsf{v}}\cos(\phi + \delta) \end{bmatrix} \tag{7.3.22}$$

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 (7.3.22) *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 147. *The ellipse in Figure 1 contains all the points x, y for which*

$$\begin{bmatrix} x-1 & y-1 \end{bmatrix} \begin{bmatrix} 0.5 & -0.25 \\ -0.25 & 1 \end{bmatrix}^{-1} \begin{bmatrix} x-1 \\ y-1 \end{bmatrix} \leq 6 \tag{7.3.23}$$

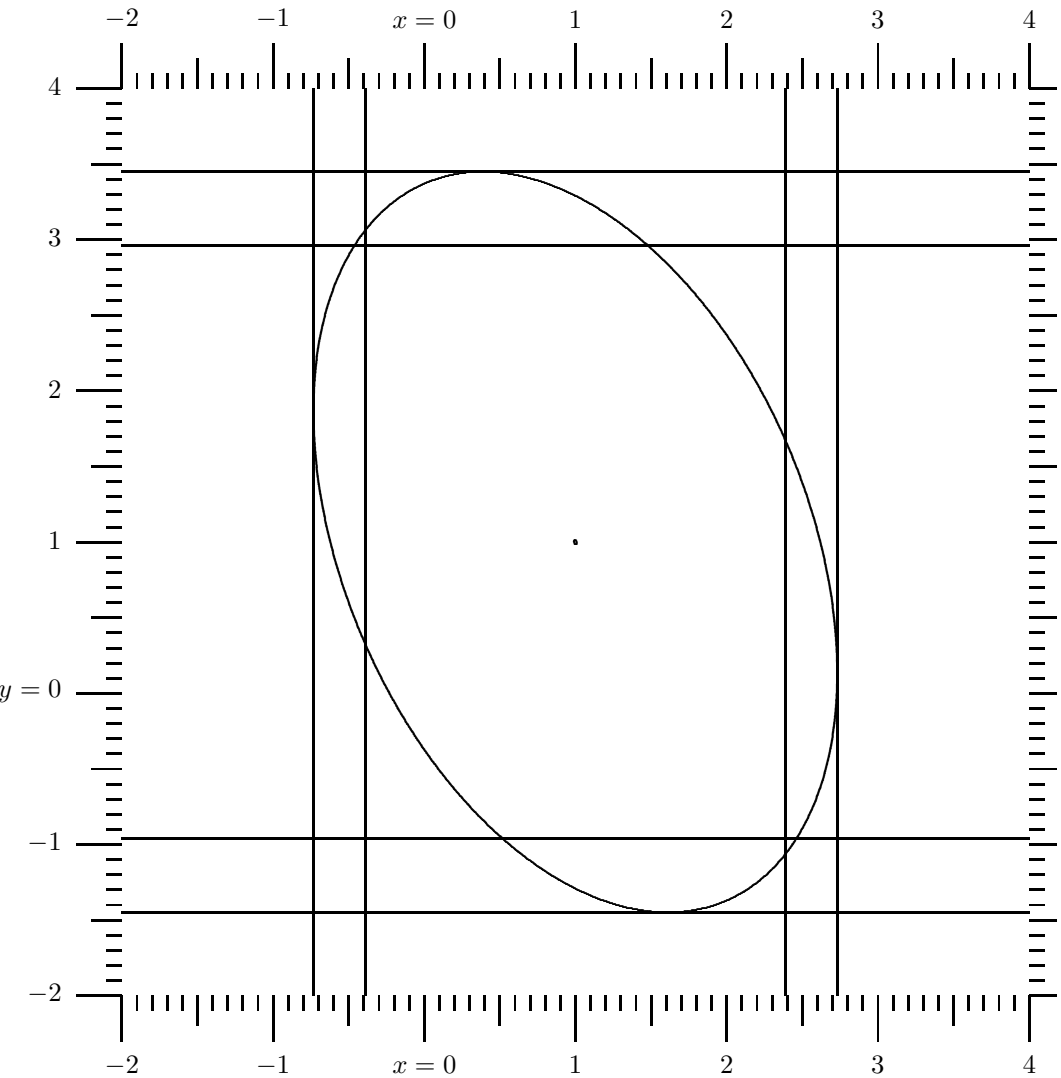


FIGURE 1. Level Line for Normal Density

• a. *3 points Compute the probability that a random variable*

$$\begin{bmatrix} \mathsf{x} \\ \mathsf{y} \end{bmatrix} \sim N\left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 0.5 & -0.25 \\ -0.25 & 1 \end{bmatrix}\right) \tag{7.3.24}$$

falls into this ellipse. Hint: you should apply equation (7.4.9)*. 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.*

• 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*

$$|a(x-1) + b(y-1)| \leq \sqrt{6}\,\sqrt{a^2 \operatorname{var}[x] + 2ab \operatorname{cov}[x,y] + b^2 \operatorname{var}[y]}. \tag{7.3.25}$$

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*

$$\boldsymbol{x} = \begin{bmatrix} \mathsf{x} \\ \mathsf{y} \end{bmatrix} \quad \boldsymbol{\mu} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \boldsymbol{\Psi} = \begin{bmatrix} 0.5 & -0.25 \\ -0.25 & 1 \end{bmatrix} \quad \boldsymbol{a} = \begin{bmatrix} a \\ b \end{bmatrix} \tag{7.3.26}$$

Equation (7.3.23) *in matrix notation says: the ellipse contains all the points for which*

$$(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Psi}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \leq 6. \tag{7.3.27}$$

Show that the band defined by inequality (7.3.25) *contains all the points for which*

$$\frac{\left(\boldsymbol{a}^\top (\boldsymbol{x} - \boldsymbol{\mu})\right)^2}{\boldsymbol{a}^\top \boldsymbol{\Psi} \boldsymbol{a}} \leq 6. \tag{7.3.28}$$

• h. *2 points Inequality* (7.3.28) *can also be written as:*

$$(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{a} (\boldsymbol{a}^\top \boldsymbol{\Psi} \boldsymbol{a})^{-1} \boldsymbol{a}^\top (\boldsymbol{x} - \boldsymbol{\mu}) \leq 6 \tag{7.3.29}$$

or alternatively

$$\begin{bmatrix} x-1 & y-1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} \left(\begin{bmatrix} a & b \end{bmatrix} \boldsymbol{\Psi}^{-1} \begin{bmatrix} a \\ b \end{bmatrix} \right)^{-1} \begin{bmatrix} x-1 \\ y-1 \end{bmatrix} \begin{bmatrix} a & b \end{bmatrix} \leq 6. \tag{7.3.30}$$

Show that the matrix

$$\boldsymbol{\Omega} = \boldsymbol{\Psi}^{-1} - \boldsymbol{a} (\boldsymbol{a}^\top \boldsymbol{\Psi} \boldsymbol{a})^{-1} \boldsymbol{a}^\top \tag{7.3.31}$$

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* $\boldsymbol{x} - \boldsymbol{\mu}$,

$$(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Psi}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \geq (\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{a}(\boldsymbol{a}^\top \boldsymbol{\Psi}\boldsymbol{a})^{-1}\boldsymbol{a}^\top(\boldsymbol{x} - \boldsymbol{\mu}). \tag{7.3.32}$$

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* (7.3.32) *is an equality, and if* $\alpha = \pm\sqrt{6/\boldsymbol{a}^\top\boldsymbol{\Psi}\boldsymbol{a}}$, *then both sides in* (7.3.32) *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.*

7.3.3. Miscellaneous Exercises.

Problem 148. *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 y; $\boldsymbol{\Psi}_7$ too large, x has larger variance than y; $\boldsymbol{\Psi}_8$ not positive definite; $\boldsymbol{\Psi}_9$ not symmetric; $\boldsymbol{\Psi}_{10}$ 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 149. *Let* x *and* y *be two independent standard normal random variables, and let* u *and* v *be bivariate normal with mean zero, variances* $\sigma_{\mathsf{u}}^2 = \sigma_{\mathsf{v}}^2 = 1$, *and correlation coefficient* $\rho \neq 0$. *Let* $f_{\mathsf{x},\mathsf{y}}$ *and* $f_{\mathsf{u},\mathsf{v}}$ *be the corresponding density functions, i.e.,*

$$f_{\mathsf{x},\mathsf{y}}(a,b) = \frac{1}{2\pi}\exp(-\frac{a^2+b^2}{2}) \quad f_{\mathsf{u},\mathsf{v}}(a,b) = \frac{1}{2\pi\sqrt{1-\rho^2}}\exp(-a^2+b^2-2\rho a\frac{b}{2(1-\rho^2)}).$$

Assume the random variables a *and* b *are defined by the following experiment: You flip a fair coin; if it shows head, then you observe* x *and* y *and give* a *the value*

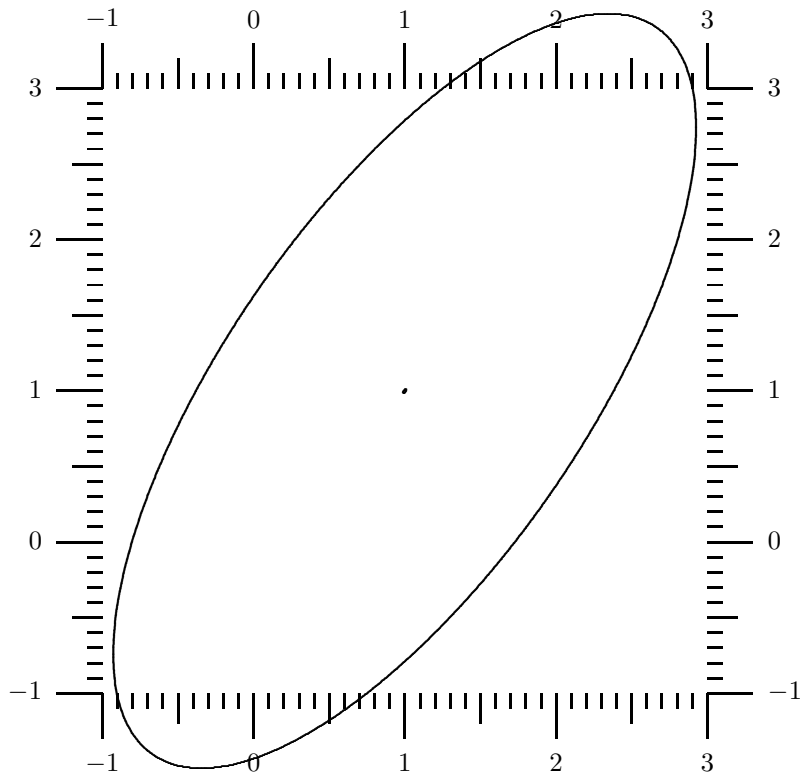


Figure 2. Level Line of Bivariate Normal Density, see Problem 148

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*

$$f_{\mathtt{a},\mathtt{b}}(a,b) = \frac{1}{2} f_{\mathtt{x},\mathtt{y}}(a,b) + \frac{1}{2} f_{\mathtt{u},\mathtt{v}}(a,b). \tag{7.3.33}$$

Hint: first show the corresponding equation for the cumulative distribution functions.

Answer. Following this hint:

$$F_{\mathtt{a},\mathtt{b}}(a,b) = \Pr[\mathtt{a} \le a \text{ and } \mathtt{b} \le b] = \tag{7.3.34}$$

$$= \Pr[\mathtt{a} \le a \text{ and } \mathtt{b} \le b|\text{head}]\Pr[\text{head}] + \Pr[\mathtt{a} \le a \text{ and } \mathtt{b} \le b|\text{tail}]\Pr[\text{tail}] \tag{7.3.35}$$

$$= F_{\mathtt{x},\mathtt{y}}(a,b)\frac{1}{2} + F_{\mathtt{u},\mathtt{v}}(a,b)\frac{1}{2}. \tag{7.3.36}$$

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 (7.3.15) for $f_{\mathtt{u},\mathtt{v}}$, i.e., write

$$f_{\mathtt{u},\mathtt{v}}(a,b) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{a^2}{2}\right) \cdot \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} \exp\left(-\frac{(b-\rho a)^2}{2(1-\rho^2)}\right). \tag{7.3.37}$$

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

$$f_{\mathtt{b}|\mathtt{a}=0}(b) = \sqrt{2\pi} f_{\mathtt{a},\mathtt{b}}(0,b) = \frac{1}{2}\frac{1}{\sqrt{2\pi}} \exp\frac{-b^2}{2} + \frac{1}{2}\frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} \exp\frac{-b^2}{2(1-\rho^2)}. \tag{7.3.38}$$

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 150. *This is* [**HT83**, 4.8-6 on p. 263] *with variance* σ^2 *instead of* 1*: Let* $\mathtt{x}$ *and* $\mathtt{y}$ *be independent normal with mean 0 and variance* σ^2*. Go over to polar coordinates* $\mathtt{r}$ *and* ϕ*, which satisfy*

$$\begin{aligned} \mathtt{x} &= \mathtt{r}\cos\phi \\ \mathtt{y} &= \mathtt{r}\sin\phi. \end{aligned} \tag{7.3.39}$$

• 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

$$J = \begin{vmatrix} \frac{\partial \mathtt{x}}{\partial \mathtt{r}} & \frac{\partial \mathtt{x}}{\partial \phi} \\ \frac{\partial \mathtt{y}}{\partial \mathtt{r}} & \frac{\partial \mathtt{y}}{\partial \phi} \end{vmatrix} = \begin{vmatrix} \cos\phi & -\mathtt{r}\sin\phi \\ \sin\phi & \mathtt{r}\cos\phi \end{vmatrix} = \big((\cos\phi)^2 + (\sin\phi)^2\big)\mathtt{r} = \mathtt{r}. \tag{7.3.40}$$

□

• b. *2 points Find the joint probability density function of* $\mathtt{r}$ *and* ϕ*. Also indicate the area in* (r,ϕ) *space in which it is nonzero.*

Answer. $f_{\mathtt{x},\mathtt{y}}(x,y) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2}$; therefore $f_{\mathtt{r},\phi}(r,\phi) = \frac{1}{2\pi\sigma^2} r e^{-r^2/2\sigma^2}$ for $0 \le r < \infty$ and $0 \le \phi < 2\pi$. □

• c. *3 points Find the marginal distributions of* $\mathtt{r}$ *and* ϕ*. Hint: for one of the integrals it is convenient to make the substitution* $q = r^2/2\sigma^2$*.*

Answer. $f_{\mathtt{r}}(r) = \frac{1}{\sigma^2} r e^{-r^2/2\sigma^2}$ 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* $\mathtt{r}$ *and* ϕ *independent?*

Answer. Yes, because joint density function is the product of the marginals. □

7.4. 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 $\begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix}$ is jointly normal, and every component of $\boldsymbol{x}$ is independent of every component of $\boldsymbol{y}$, then the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ are already independent. Not surprised? You should be, see Problem 125.

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 7.4.1. 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 $f_{\boldsymbol{x}}(\boldsymbol{z}) = (2\pi\sigma^2)^{-n/2}\exp(-\boldsymbol{z}^\top\boldsymbol{z}/2\sigma^2)$.

The following property of the multivariate standard normal distributions is basic:

THEOREM 7.4.2. *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 $\boldsymbol{\mathsf{x}} = \boldsymbol{P}\boldsymbol{z}$ is a multivariate standard normal m-vector with the same variance σ^2, and $\boldsymbol{z}^\top\boldsymbol{z} - \boldsymbol{\mathsf{x}}^\top\boldsymbol{\mathsf{x}} \sim \sigma^2\chi^2_{p-m}$ independent of $\boldsymbol{\mathsf{x}}$.*

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 $\boldsymbol{\mathsf{y}} = \boldsymbol{T}\boldsymbol{z}$, i.e., $\boldsymbol{z} = \boldsymbol{T}^\top\boldsymbol{\mathsf{y}}$. Then $\boldsymbol{z}^\top\boldsymbol{z} = \boldsymbol{\mathsf{y}}^\top\boldsymbol{T}\boldsymbol{T}^\top\boldsymbol{\mathsf{y}} = \boldsymbol{\mathsf{y}}^\top\boldsymbol{\mathsf{y}}$, and the Jacobian of the transformation from $\boldsymbol{\mathsf{y}}$ to $\boldsymbol{z}$ has absolute value one. Therefore the density function of $\boldsymbol{\mathsf{y}}$ is $(2\pi\sigma^2)^{-n/2}\exp(-\boldsymbol{y}^\top\boldsymbol{y}/2\sigma^2)$, which means $\boldsymbol{\mathsf{y}}$ 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 $i \neq j$. Therefore also any subvector of $\boldsymbol{\mathsf{y}}$, such as $\boldsymbol{\mathsf{x}}$, is standard normal. Since $\boldsymbol{z}^\top\boldsymbol{z} - \boldsymbol{\mathsf{x}}^\top\boldsymbol{\mathsf{x}} = \boldsymbol{\mathsf{y}}^\top\boldsymbol{\mathsf{y}} - \boldsymbol{\mathsf{x}}^\top\boldsymbol{\mathsf{x}}$ is the sum of the squares of those elements of $\boldsymbol{\mathsf{y}}$ which are not in $\boldsymbol{\mathsf{x}}$, it follows that it is an independent $\sigma^2\chi^2_{p-m}$. □

PROBLEM 151. *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

$$m_{\boldsymbol{z}}(\boldsymbol{t}) = \mathrm{E}[\exp(\boldsymbol{t}^\top\boldsymbol{z})] \tag{7.4.1}$$

$$= (2\pi\sigma^2)^{n/2}\int\cdots\int \exp(-\frac{1}{2\sigma^2}\boldsymbol{z}^\top\boldsymbol{z})\,\exp(\boldsymbol{t}^\top\boldsymbol{z})\,dz_1\cdots dz_n \tag{7.4.2}$$

$$= (2\pi\sigma^2)^{n/2}\int\cdots\int \exp(-\frac{1}{2\sigma^2}(\boldsymbol{z}-\sigma^2\boldsymbol{t})^\top(\boldsymbol{z}-\sigma^2\boldsymbol{t}) + \frac{\sigma^2}{2}\boldsymbol{t}^\top\boldsymbol{t})\,dz_1\cdots dz_n \tag{7.4.3}$$

$$= \exp(\frac{\sigma^2}{2}\boldsymbol{t}^\top\boldsymbol{t}) \quad \text{since first part of integrand is density function.} \tag{7.4.4}$$

□

THEOREM 7.4.3. *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 $\boldsymbol{\mathsf{x}} = \boldsymbol{T}\boldsymbol{z}$; it is standard normal by theorem 7.4.2. 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

$$\mathrm{E}[\exp(\mathsf{q}t)] = \mathrm{E}[\exp(t\sum_{i=1}^r \lambda_i\mathsf{x}_i^2)] \tag{7.4.5}$$

$$= \mathrm{E}[\exp(t\lambda_1\mathsf{x}_1^2)]\cdots\mathrm{E}[\exp(t\lambda_r\mathsf{x}_r^2)] \tag{7.4.6}$$

$$= (1-2\lambda_1\sigma^2t)^{-1/2}\cdots(1-2\lambda_r\sigma^2t)^{-1/2}. \tag{7.4.7}$$

By assumption this is equal to $(1-2\sigma^2t)^{-k/2}$ with some integer $k \geq 1$. Taking squares and inverses one obtains

$$(1-2\lambda_1\sigma^2t)\cdots(1-2\lambda_r\sigma^2t) = (1-2\sigma^2t)^k. \tag{7.4.8}$$

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 7.4.3 one can derive a characterization of all the quadratic forms of multivariate normal variables with arbitrary covariance matrices that are χ^2's. Assume $\boldsymbol{\mathsf{y}}$ 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 $(\boldsymbol{\mathsf{y}}-\boldsymbol{\mu})^\top\boldsymbol{\Omega}(\boldsymbol{\mathsf{y}}-\boldsymbol{\mu}) \sim \sigma^2\chi^2_k$ iff

$$\boldsymbol{\Psi}\boldsymbol{\Omega}\boldsymbol{\Psi}\boldsymbol{\Omega}\boldsymbol{\Psi} = \boldsymbol{\Psi}\boldsymbol{\Omega}\boldsymbol{\Psi}, \tag{7.4.9}$$

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 7.4.3. This is the reason why the minimum value of the SSE has a $\sigma^2\chi^2$ distribution, see (27.0.10).
- $\boldsymbol{\Psi}$ 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 147).
- $\boldsymbol{\Psi}$ singular and $\boldsymbol{\Omega} = \boldsymbol{\Psi}^-$, its g-inverse. The multinomial distribution has a singular covariance matrix, and equation (**??**) 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 7.4.4. *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 7.4.5. *Let* $\mathbf{x}$ *be* p*-variate standard normal with variance* σ^2*, and* $\mathbf{x}^\top\mathbf{x} = \sum_{i=1}^k \mathbf{x}^\top \boldsymbol{P}_i \mathbf{x}$*. Then for the quadratic forms* $\mathbf{x}^\top \boldsymbol{P}_i \mathbf{x}$ *to be independently distributed as* $\sigma^2\chi^2$*, any one of the following three equivalent conditions is necessary and sufficient:*

$$\boldsymbol{P}_i^2 = \boldsymbol{P}_i \qquad \textit{for all } i \tag{7.4.10}$$

$$\boldsymbol{P}_i\boldsymbol{P}_j = \boldsymbol{O} \qquad i \neq j \tag{7.4.11}$$

$$\sum_{i=1}^k \operatorname{rank}(\boldsymbol{P}_i) = p \tag{7.4.12}$$

CHAPTER 8

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.,

$$\mathsf{x} = \mathsf{z} + \varepsilon \tag{8.0.13}$$

$$\mathsf{y} = \mathsf{z} + \delta \tag{8.0.14}$$

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 $\operatorname{var}[\mathsf{x}] = \operatorname{var}[\mathsf{y}]$, (7.3.17) 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 152 illustrates this:

Problem 152. *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 $\operatorname{E}[\mathsf{x}] = 100$; $\operatorname{E}[\mathsf{y}] = 100$; $\operatorname{var}[\mathsf{x}] = 20 + 5 = 25$; $\operatorname{var}[\mathsf{y}] = 5 + 4 = 9$; $\operatorname{cov}[\mathsf{x}, \mathsf{y}] = 10$; $\operatorname{corr}[\mathsf{x}, \mathsf{y}] = 10/15 = 2/3$. In matrix notation

$$\begin{bmatrix} \mathsf{x} \\ \mathsf{y} \end{bmatrix} \sim \mathsf{N}\Big[\begin{bmatrix} 100 \\ 100 \end{bmatrix}, \begin{bmatrix} 25 & 10 \\ 10 & 9 \end{bmatrix}\Big] \tag{8.0.15}$$

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* (7.3.18) *for this.*

Answer.

$$\mathrm{E}[\mathsf{y}|\mathsf{x}{=}x] = 100 + \frac{10}{25}(x-100) = 60 + \frac{2}{5}x \tag{8.0.16}$$

$$\mathrm{E}[x|\mathsf{y}{=}y] = 100 + \frac{10}{9}(y-100) = -\frac{100}{9} + \frac{10}{9}y. \tag{8.0.17}$$

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* $\mathsf{w} = \frac{6}{10}\mathsf{z} + 40$, $\mathsf{z} \sim \mathsf{N}(100, \frac{100}{6})$, $\varepsilon \sim \mathsf{N}(0, \frac{50}{6})$, $\delta \sim \mathsf{N}(0, 3)$. *Is this compatible with the data?*

Answer. Yes, it is compatible: $\mathrm{E}[\mathsf{x}] = \mathrm{E}[\mathsf{z}]+\mathrm{E}[\varepsilon] = 100$; $\mathrm{E}[\mathsf{y}] = \mathrm{E}[\mathsf{w}]+\mathrm{E}[\delta] = \frac{6}{10}100+40 = 100$; $\mathrm{var}[\mathsf{x}] = \frac{100}{6} + \frac{50}{6} = 25$; $\mathrm{var}[\mathsf{y}] = \left(\frac{6}{10}\right)^2 \mathrm{var}[\mathsf{z}] + \mathrm{var}[\delta] = \frac{63}{100}\frac{100}{6} + 3 = 9$; $\mathrm{cov}[\mathsf{x}, \mathsf{y}] = \frac{6}{10}\mathrm{var}[\mathsf{z}] = 10$. □

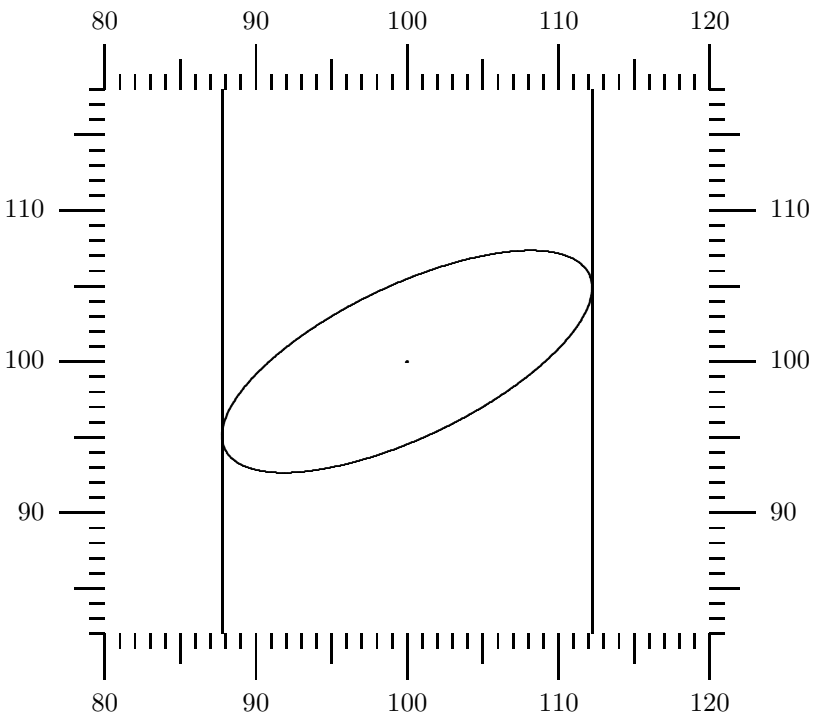


Figure 1. Ellipse containing 95% of the probability mass of test results x and y

• 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 **??** 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 153 shows that in this latter case, the OLS estimate is not consistent, but other estimates exist that are consistent.

Problem 153. [**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*

$$\mathsf{y} = \mathsf{y}^{\mathsf{p}} + \mathsf{y}^{\mathsf{t}}, \tag{8.0.18}$$

$$\mathsf{c} = \mathsf{c}^{\mathsf{p}} + \mathsf{c}^{\mathsf{t}}, \tag{8.0.19}$$

$$\mathsf{c}^{\mathsf{p}} = \beta\mathsf{y}^{\mathsf{p}}. \tag{8.0.20}$$

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* $\mathrm{var}[\mathsf{y}^{\mathsf{p}}] = \tau^2_{\mathsf{y}}$. *The transitory components* y^{t} *and* c^{t} *are assumed to be independent of each other and of* y^{p}, *and* $\mathrm{E}[\mathsf{y}^{\mathsf{t}}] = 0$, $\mathrm{var}[\mathsf{y}^{\mathsf{t}}] = \sigma^2_{\mathsf{y}}$, $\mathrm{E}[\mathsf{c}^{\mathsf{t}}] = 0$, *and* $\mathrm{var}[\mathsf{c}^{\mathsf{t}}] = \sigma^2_{\mathsf{c}}$. *Finally, it is assumed that all variables are normally distributed.*

• a. *2 points Given the above information, write down the vector of expected values* $\mathcal{E}\left[\left[\begin{smallmatrix}\mathsf{y}\\ \mathsf{c}\end{smallmatrix}\right]\right]$ *and the covariance matrix* $\mathcal{V}\left[\left[\begin{smallmatrix}\mathsf{y}\\ \mathsf{c}\end{smallmatrix}\right]\right]$ *in terms of the five unknown parameters of the model* μ, β, τ^2_{y}, σ^2_{y}, *and* σ^2_{c}.

Answer.

$$\mathcal{E}\left[\begin{bmatrix}\mathsf{y}\\ \mathsf{c}\end{bmatrix}\right] = \begin{bmatrix}\mu\\ \beta\mu\end{bmatrix} \quad \text{and} \quad \mathcal{V}\left[\begin{bmatrix}\mathsf{y}\\ \mathsf{c}\end{bmatrix}\right] = \begin{bmatrix}\tau^2_{\mathsf{y}} + \sigma^2_{\mathsf{y}} & \beta\tau^2_{\mathsf{y}}\\ \beta\tau^2_{\mathsf{y}} & \beta^2\tau^2_{\mathsf{y}} + \sigma^2_{\mathsf{c}}\end{bmatrix}. \tag{8.0.21}$$

□

• 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*

$$\mathsf{y}^{\mathsf{p}*} = \frac{\sigma_{\mathsf{y}}^2}{\tau_{\mathsf{y}}^2 + \sigma_{\mathsf{y}}^2}\mu + \frac{\tau_{\mathsf{y}}^2}{\tau_{\mathsf{y}}^2 + \sigma_{\mathsf{y}}^2}\mathsf{y}. \tag{8.0.22}$$

Note: here we are guessing income, not yet consumption! Use (7.3.17) *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*

$$\beta = 0.7 \tag{8.0.23}$$

$$\sigma_{\mathsf{y}} = 2{,}000 \tag{8.0.24}$$

$$\sigma_{\mathsf{c}} = 1{,}000 \tag{8.0.25}$$

$$\mu = 12{,}000 \tag{8.0.26}$$

$$\tau_{\mathsf{y}} = 4{,}000. \tag{8.0.27}$$

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*

$$\mathsf{c}^* = \beta\Big(\frac{\sigma_{\mathsf{y}}^2}{\tau_{\mathsf{y}}^2 + \sigma_{\mathsf{y}}^2}\mu + \frac{\tau_{\mathsf{y}}^2}{\tau_{\mathsf{y}}^2 + \sigma_{\mathsf{y}}^2}\mathsf{y}\Big). \tag{8.0.28}$$

Instead of an exact mathematical proof you may also reason out how it can be obtained from (8.0.22). *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:

$$\mathrm{E}[\mathsf{c}|y = 22{,}000] = \mathrm{E}[\mathsf{c}] + \frac{\mathrm{cov}[\mathsf{c}, \mathsf{y}]}{\mathrm{var}[\mathsf{y}]}(22{,}000 - \mathrm{E}[\mathsf{y}]) \tag{8.0.29}$$

$$= \beta\mu + \frac{\beta\tau_{\mathsf{y}}^2}{\tau_{\mathsf{y}}^2 + \sigma_{\mathsf{y}}^2}(22{,}000 - \mu) = 8{,}400 + 0.7\frac{16{,}000{,}000}{20{,}000{,}000}(22{,}000 - 12{,}000) = 14{,}000 \tag{8.0.30}$$

$$\text{or} \quad = \beta\Big(\frac{\sigma_{\mathsf{y}}^2}{\tau_{\mathsf{y}}^2 + \sigma_{\mathsf{y}}^2}\mu + \frac{\tau_{\mathsf{y}}^2}{\tau_{\mathsf{y}}^2 + \sigma_{\mathsf{y}}^2}22{,}000\Big) \tag{8.0.31}$$

$$= 0.7\Big((0.2)(12{,}000) + (0.8)(22{,}000)\Big) = (0.7)(20{,}000) = 14{,}000. \tag{8.0.32}$$

□

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* $\mathbf{y}$ *with the intercept suppressed. This would give the estimator*

$$\hat{\beta} = \frac{\sum \mathsf{c}_i\mathsf{y}_i}{\sum \mathsf{y}_i^2} \tag{8.0.33}$$

Show that the plim *of this estimator is*

$$\mathrm{plim}[\hat{\beta}] = \frac{\mathrm{E}[\mathsf{c}\mathsf{y}]}{\mathrm{E}[\mathsf{y}^2]} \tag{8.0.34}$$

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.

$$\hat{\beta} = \frac{\frac{1}{n}\sum \mathsf{c}_i\mathsf{y}_i}{\frac{1}{n}\sum \mathsf{y}_i^2}; \quad \mathrm{plim}[\hat{\beta}] = \frac{\mathrm{E}[\mathsf{c}\mathsf{y}]}{\mathrm{E}[\mathsf{y}^2]} = \frac{\mathrm{E}[\mathsf{c}]\,\mathrm{E}[\mathsf{y}] + \mathrm{cov}[\mathsf{c}, \mathsf{y}]}{(\mathrm{E}[\mathsf{y}])^2 + \mathrm{var}[\mathsf{y}]} = \frac{\mu\beta\mu + \beta\tau_{\mathsf{y}}^2}{\mu^2 + \tau_{\mathsf{y}}^2 + \sigma_{\mathsf{y}}^2} = \beta\frac{\mu^2 + \tau_{\mathsf{y}}^2}{\mu^2 + \tau_{\mathsf{y}}^2 + \sigma_{\mathsf{y}}^2}.$$

□

• 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* $\mathrm{E}[\mathsf{y}]$, $\mathrm{E}[\mathsf{c}]$, $\mathrm{var}[\mathsf{y}]$, $\mathrm{var}[\mathsf{c}]$, *and* $\mathrm{cov}[\mathsf{y}, \mathsf{c}]$, *and then simply replace the population moments by the sample moments.) Are these consistent estimators?*

Answer. From (8.0.21) follows $\mathrm{E}[\mathsf{c}] = \beta\,\mathrm{E}[\mathsf{y}]$, therefore $\beta = \frac{\mathrm{E}[\mathsf{c}]}{\mathrm{E}[\mathsf{y}]}$. This together with $\mathrm{cov}[\mathsf{y}, \mathsf{c}] = \beta\tau_{\mathsf{y}}^2$ gives $\tau_{\mathsf{y}}^2 = \frac{\mathrm{cov}[\mathsf{y},\mathsf{c}]}{\beta} = \frac{\mathrm{cov}[\mathsf{y},\mathsf{c}]\,\mathrm{E}[\mathsf{y}]}{\mathrm{E}[\mathsf{c}]}$. This together with $\mathrm{var}[\mathsf{y}] = \tau_{\mathsf{y}}^2 + \sigma_{\mathsf{y}}^2$ gives $\sigma_{\mathsf{y}}^2 = \mathrm{var}[\mathsf{y}] - \tau_{\mathsf{y}}^2 = \mathrm{var}[\mathsf{y}] - \frac{\mathrm{cov}[\mathsf{y},\mathsf{c}]\,\mathrm{E}[\mathsf{y}]}{\mathrm{E}[\mathsf{c}]}$. And from the last equation $\mathrm{var}[\mathsf{c}] = \beta^2\tau_{\mathsf{y}}^2 + \sigma_{\mathsf{c}}^2$ one get $\sigma_{\mathsf{c}}^2 = \mathrm{var}[\mathsf{c}] - \frac{\mathrm{cov}[\mathsf{y},\mathsf{c}]\,\mathrm{E}[\mathsf{c}]}{\mathrm{E}[\mathsf{y}]}$. 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* (8.0.28) *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.,*

$$\hat{\beta} = \frac{\sum(\mathsf{y}_i - \bar{\mathsf{y}})(\mathsf{c}_i - \bar{\mathsf{c}})}{\sum(\mathsf{y}_i - \bar{\mathsf{y}})^2} \tag{8.0.35}$$

$$\hat{\alpha} = \bar{\mathsf{c}} - \hat{\beta}\bar{\mathsf{y}} \tag{8.0.36}$$

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 (8.0.28): $\tau_{\mathsf{y}}^2+\sigma_{\mathsf{y}}^2 = \operatorname{var}[\mathsf{y}]$. Therefore the first summand becomes

$$\beta\sigma_{\mathsf{y}}^2\mu\frac{1}{\operatorname{var}[\mathsf{y}]} = \frac{\operatorname{E}[\mathsf{c}]}{\operatorname{E}[\mathsf{y}]}\left(\operatorname{var}[\mathsf{y}]-\frac{\operatorname{cov}[\mathsf{y},\mathsf{c}]\operatorname{E}[\mathsf{y}]}{\operatorname{E}[\mathsf{c}]}\right)\operatorname{E}[\mathsf{y}]\frac{1}{\operatorname{var}[\mathsf{y}]} = \operatorname{E}[\mathsf{c}]\left(1-\frac{\operatorname{cov}[\mathsf{y},\mathsf{c}]\operatorname{E}[\mathsf{y}]}{\operatorname{var}[\mathsf{y}]\operatorname{E}[\mathsf{c}]}\right) = \operatorname{E}[\mathsf{c}]-\frac{\operatorname{cov}[\mathsf{y},\mathsf{c}]\operatorname{E}[\mathsf{y}]}{\operatorname{var}[\mathsf{y}]}$$

But since $\frac{\operatorname{cov}[\mathsf{y},\mathsf{c}]}{\operatorname{var}[\mathsf{y}]} = \hat{\beta}$ and $\hat{\alpha}+\hat{\beta}\operatorname{E}[\mathsf{y}] = \operatorname{E}[\mathsf{c}]$ 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 154. *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:

$$\mathsf{y}_t = \alpha + \lambda\mathsf{y}_{t-1} + \beta\mathsf{m}_t + \mathsf{u}_t \tag{8.0.37}$$

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 $\operatorname{E}[\mathsf{y}_t]$ 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 m_t according to a rule with feedback:*

$$\mathsf{m}_t = g_0 + g_1\mathsf{y}_{t-1} + \varepsilon_t \tag{8.0.38}$$

Show that the following values of g_0 and g_1

$$g_0 = (y^* - \alpha)/\beta \qquad g_1 = -\lambda/\beta \tag{8.0.39}$$

represent an optimal monetary policy, since they bring the expected value of the steady state $\operatorname{E}[\mathsf{y}_t]$ to y^* and minimize the steady state variance $\operatorname{var}[\mathsf{y}_t]$.

• 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* (8.0.37) *can be derived are equations* (8.0.40)–(8.0.42) *below. Equation* (8.0.40) *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).*

$$\mathsf{y}_t = \xi_0 + \xi_1(\mathsf{m}_t - \operatorname{E}_{t-1}\mathsf{m}_t) + \xi_2\mathsf{y}_{t-1} + \mathsf{u}_t \tag{8.0.40}$$

The Fed uses the policy rule

$$\mathsf{m}_t = g_0 + g_1\mathsf{y}_{t-1} + \varepsilon_t \tag{8.0.41}$$

and the agents know this policy rule, therefore

$$\operatorname{E}_{t-1}\mathsf{m}_t = g_0 + g_1\mathsf{y}_{t-1}. \tag{8.0.42}$$

Show that in this system, the parameters g_0 and g_1 have no influence on the time path of 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 y_t on y_{t-1} and m_t in order to estimate the coefficients in* (8.0.37). *Which values of α, λ, and β will such a regression yield?*

CHAPTER 9

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.

9.1. 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 155. *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*

$$\sum_{i=1}^{n}(y_i - \alpha)^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2 + n(\bar{y} - \alpha)^2 \tag{9.1.1}$$

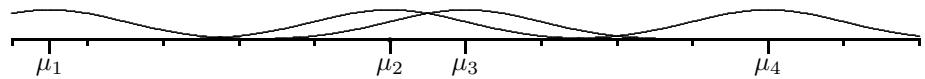


Figure 1. Possible Density Functions for y

Answer.

$$\sum_{i=1}^{n}(y_i - \alpha)^2 = \sum_{i=1}^{n}\big((y_i - \bar{y}) + (\bar{y} - \alpha)\big)^2 \tag{9.1.2}$$

$$= \sum_{i=1}^{n}(y_i - \bar{y})^2 + 2\sum_{i=1}^{n}\big((y_i - \bar{y})(\bar{y} - \alpha)\big) + \sum_{i=1}^{n}(\bar{y} - \alpha)^2 \tag{9.1.3}$$

$$= \sum_{i=1}^{n}(y_i - \bar{y})^2 + 2(\bar{y} - \alpha)\sum_{i=1}^{n}(y_i - \bar{y}) + n(\bar{y} - \alpha)^2 \tag{9.1.4}$$

Since the middle term is zero, (9.1.1) follows. □

Problem 156. *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*

$$(y_1 - \alpha)^2 + (y_2 - \alpha)^2 + \cdots + (y_n - \alpha)^2 = \sum (y_i - \alpha)^2 \tag{9.1.5}$$

is the arithmetic mean $\alpha = \bar{y}$.

Answer. Use (9.1.1). □

Problem 157. *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?*

9.2. 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

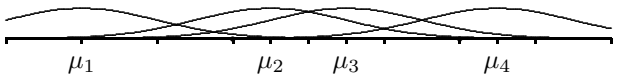


FIGURE 2. Two observations, $\sigma^2 = 1$

FIGURE 3. Two observations, σ^2 unknown

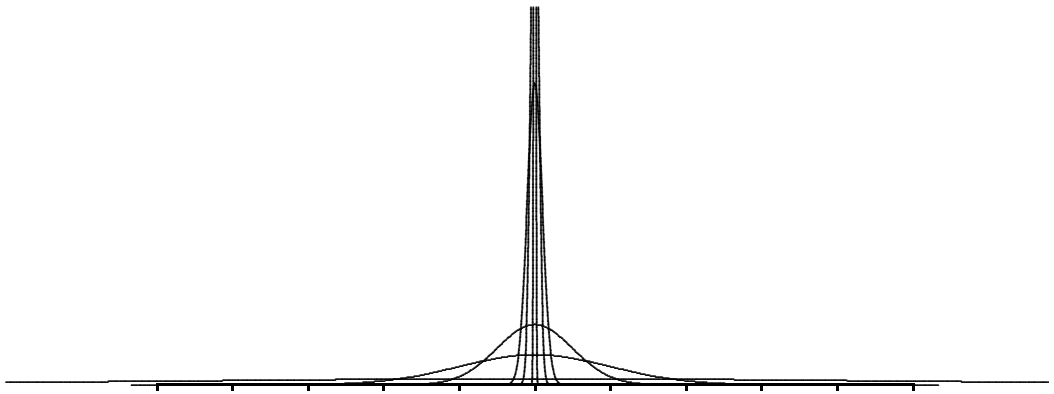


FIGURE 4. Only those centered over the two observations need to be considered

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.

FIGURE 5. Many Observations

9.2.1. 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 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.

9.3. 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:

$$\mathrm{s}_m^2 = \frac{1}{n}\sum(\mathfrak{y}_i - \bar{\mathfrak{y}})^2 \tag{9.3.1}$$

or

$$\mathrm{s}_u^2 = \frac{1}{n-1}\sum(\mathfrak{y}_i - \bar{\mathfrak{y}})^2. \tag{9.3.2}$$

Let us compute the expected value of our two estimators. Equation (9.1.1) with $\alpha = \mathrm{E}[\mathfrak{y}]$ allows us to simplify the sum of squared errors so that it becomes easy to take expected values:

$$\mathrm{E}[\sum_{i=1}^{n}(\mathfrak{y}_i - \bar{\mathfrak{y}})^2] = \sum_{i=1}^{n}\mathrm{E}[(\mathfrak{y}_i - \mu)^2] - n\,\mathrm{E}[(\bar{\mathfrak{y}} - \mu)^2] \tag{9.3.3}$$

$$= \sum_{i=1}^{n}\sigma^2 - n\frac{\sigma^2}{n} = (n-1)\sigma^2. \tag{9.3.4}$$

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

$$\mathsf{s}_u^2=\frac{1}{n-1}\sum_{i=1}^{n}(\mathsf{y}_i-\bar{\mathsf{y}})^2 \tag{9.3.5}$$

then this is an unbiased estimate.

Problem 158. *4 points Show that*

$$\mathsf{s}_u^2=\frac{1}{n-1}\sum_{i=1}^{n}(\mathsf{y}_i-\bar{\mathsf{y}})^2 \tag{9.3.6}$$

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 (9.1.1) with $\alpha=\mathrm{E}[\mathsf{y}]$:

$$\mathrm{E}[\sum_{i=1}^{n}(\mathsf{y}_i-\bar{\mathsf{y}})^2]=\sum_{i=1}^{n}\mathrm{E}[(\mathsf{y}_i-\mu)^2]-n\,\mathrm{E}[(\bar{\mathsf{y}}-\mu)^2] \tag{9.3.7}$$

$$=\sum_{i=1}^{n}\sigma^2-n\frac{\sigma^2}{n}=(n-1)\sigma^2. \tag{9.3.8}$$

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 4.9 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 159 gives a proof of this in the simplest case $n=2$, and Problem 160 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 (4.9.5) to get the variance of s_u^2: it is $\frac{2\sigma^4}{n-1}$.

Problem 159. *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*

$$\mathsf{SSE}=\sum_{i-1}^{2}(\mathsf{y}_i-\bar{\mathsf{y}})^2\sim\sigma^2\chi^2_1 \tag{9.3.9}$$

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.

$$\bar{\mathsf{y}}=\frac{\mathsf{y}_1+\mathsf{y}_2}{2} \tag{9.3.10}$$

$$\mathsf{y}_1-\bar{\mathsf{y}}=\frac{\mathsf{y}_1-\mathsf{y}_2}{2} \tag{9.3.11}$$

$$\mathsf{y}_2-\bar{\mathsf{y}}=-\frac{\mathsf{y}_1-\mathsf{y}_2}{2} \tag{9.3.12}$$

$$(\mathsf{y}_1-\bar{\mathsf{y}})^2+(\mathsf{y}_2-\bar{\mathsf{y}})^2=\frac{(\mathsf{y}_1-\mathsf{y}_2)^2}{4}+\frac{(\mathsf{y}_1-\mathsf{y}_2)^2}{4} \tag{9.3.13}$$

$$=\frac{(\mathsf{y}_1-\mathsf{y}_2)^2}{2}=\sigma^2\left(\frac{\mathsf{y}_1-\mathsf{y}_2}{\sqrt{2\sigma^2}}\right)^2, \tag{9.3.14}$$

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*

$$\begin{bmatrix}\mathsf{y}_1-\bar{\mathsf{y}}\\ \mathsf{y}_2-\bar{\mathsf{y}}\end{bmatrix} \tag{9.3.15}$$

and show that it is singular.

Answer. (9.3.11) and (9.3.12) give

$$\begin{bmatrix}\mathsf{y}_1-\bar{\mathsf{y}}\\ \mathsf{y}_2-\bar{\mathsf{y}}\end{bmatrix}=\begin{bmatrix}\frac{1}{2}&-\frac{1}{2}\\ -\frac{1}{2}&\frac{1}{2}\end{bmatrix}\begin{bmatrix}\mathsf{y}_1\\ \mathsf{y}_2\end{bmatrix}=\boldsymbol{D}\boldsymbol{\mathsf{y}} \tag{9.3.16}$$

and $\mathcal{V}[\boldsymbol{D}\boldsymbol{\mathsf{y}}]=\boldsymbol{D}\,\mathcal{V}[\boldsymbol{\mathsf{y}}]\boldsymbol{D}^\top=\sigma^2\boldsymbol{D}$ because $\mathcal{V}[\boldsymbol{\mathsf{y}}]=\sigma^2\boldsymbol{I}$ 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 160. *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* $z_1^2 + z_2^2 + z_3^2 = y_1^2 + y_2^2 + y_3^2$. *Is this a surprise?*

• f. *1 point Show further that* $s_u^2 = \frac{1}{2}\sum_{i=1}^{3}(y_i - \bar{y})^2 = \frac{1}{2}(z_2^2 + z_3^2)$. *(There is a simple trick!) Conclude from this that* $s_u^2 \sim \frac{\sigma^2}{2}\chi_2^2$, *independent of* $\bar{y}$.

For a matrix-interpretation of what is happening, see equation (7.4.9) together with Problem 161.

PROBLEM 161. *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*

$$\text{sample covariance}(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\boldsymbol{x}^\top \boldsymbol{D}\boldsymbol{y} \tag{9.3.17}$$

In general, one can always find $n-1$ normal variables with variance σ^2, independent of each other and of $\bar{y}$, whose sum of squares is equal to $\sum(y_i - \bar{y})^2$. Simply start with $\bar{y}\sqrt{n}$ 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 $\bar{y}\sqrt{n}$ as its first vector; there are many different ways to do this.

Next let us show that $\bar{y}$ and s_u^2 are statistically independent. This is an advantage. Assume, hypothetically, $\bar{y}$ and s_u^2 were negatively correlated. Then, if the observed value of $\bar{y}$ 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 $\bar{y}$ may be. To prove independence, we will first show that $\bar{y}$ and $y_i - \bar{y}$ are uncorrelated:

$$\operatorname{cov}[\bar{y}, y_i - \bar{y}] = \operatorname{cov}[\bar{y}, y_i] - \operatorname{var}[\bar{y}] \tag{9.3.18}$$

$$= \operatorname{cov}[\frac{1}{n}(y_1 + \cdots + y_i + \cdots + y_n), y_i] - \frac{\sigma^2}{n} = 0 \tag{9.3.19}$$

By normality, $\bar{y}$ is therefore *independent* of $y_i - \bar{y}$ for all i. Since all variables involved are jointly normal, it follows from this that $\bar{y}$ is independent of the vector $\begin{bmatrix} y_1 - \bar{y} & \cdots & y_n - \bar{y} \end{bmatrix}^\top$; 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., $y_i \sim \mathsf{NID}(\mu, \sigma^2)$, then we know that $s_u^2 \sim \frac{\sigma^2}{n-1}\chi_{n-1}^2$. Therefore $\mathrm{E}[s_u^2] = \sigma^2$ and $\operatorname{var}[s_u^2] = 2\sigma^4/(n-1)$. This estimate of σ^2 therefore not only gives us an estimate of the precision of $\bar{y}$, but it has an estimate of its own precision built in.

Interestingly, the MSE of the alternative estimator $s_m^2 = \frac{\sum(y_i - \bar{y})^2}{n}$ 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, $\operatorname{MSE}[t; \theta] = \operatorname{var}[t] + (\mathrm{E}[t - \theta])^2$, 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 $s_m^2 = \frac{n-1}{n}s_u^2$ 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 162. *4 points Assume* $y_i \sim \mathsf{NID}(\mu, \sigma^2)$. *Show that the so-called Theil Schweitzer estimator* [**TS61**]

$$s_t^2 = \frac{1}{n+1}\sum(y_i - \bar{y})^2 \tag{9.3.20}$$

has even smaller MSE *than* s_u^2 *and* s_m^2 *as an estimator of* σ^2.

ANSWER. $s_t^2 = \frac{n-1}{n+1}s_u^2$; 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 163. *3 points Computer assignment: Given 20 independent observations of a random variable* $y \sim \mathsf{N}(\mu, \sigma^2)$. *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*

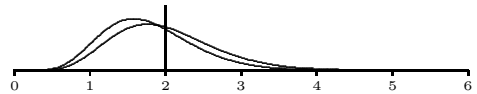


FIGURE 6. Densities of Unbiased and Theil Schweitzer Estimators

of a certain multiple of a Chi-square. (Use the transformation theorem for density functions!)

ANSWER. $s_u^2 \sim \frac{2}{19}\chi^2_{19}$. To express the density of the variable whose density is known by that whose density one wants to know, say $\frac{19}{2}s_u^2 \sim \chi^2_{19}$. Therefore

$$f_{s_u^2}(x) = \frac{19}{2} f_{\chi^2_{19}}(\frac{19}{2}x). \tag{9.3.21}$$

□

• a. *2 points In the same plot, plot the density function of the Theil-Schweitzer estimate* s_t^2 *defined in equation* (9.3.20). *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(theils` 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 164. *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*

$$\frac{n\mu^2 - \sigma^2}{n\mu^2 + \sigma^2} < \alpha < 1 \tag{9.3.22}$$

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:

$$\mathrm{MSE}[\alpha\bar{\mathsf{y}}; \mu] = \mathrm{E}\Big[(\alpha\bar{\mathsf{y}} - \mu)^2\Big] = \mathrm{E}\Big[(\alpha\bar{\mathsf{y}} - \alpha\mu + \alpha\mu - \mu)^2\Big] < \mathrm{MSE}[\bar{\mathsf{y}}; \mu] = \mathrm{var}[\bar{\mathsf{y}}] \tag{9.3.23}$$

$$\alpha^2\sigma^2/n + (1-\alpha)^2\mu^2 < \sigma^2/n \tag{9.3.24}$$

Now simplify it:

$$(1-\alpha)^2\mu^2 < (1-\alpha^2)\sigma^2/n = (1-\alpha)(1+\alpha)\sigma^2/n \tag{9.3.25}$$

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:

$$(1-\alpha)\mu^2 < (1+\alpha)\sigma^2/n \tag{9.3.26}$$

$$\mu^2 - \sigma^2/n < \alpha(\mu^2 + \sigma^2/n) \tag{9.3.27}$$

The answer is therefore

$$\frac{n\mu^2 - \sigma^2}{n\mu^2 + \sigma^2} < \alpha < 1. \tag{9.3.28}$$

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 165. □

PROBLEM 165. [**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*

$$\mathrm{MSE}[\mathsf{t};\theta] = \mathrm{var}[\mathsf{t}] + \big(\mathrm{E}[\mathsf{t}-\theta]\big)^2. \tag{9.3.29}$$

ANSWER. The most elegant proof, which also indicates what to do when θ is random, is:

$$\mathrm{MSE}[\mathsf{t};\theta] = \mathrm{E}\Big[(\mathsf{t}-\theta)^2\Big] = \mathrm{var}[\mathsf{t}-\theta] + (\mathrm{E}[\mathsf{t}-\theta])^2 = \mathrm{var}[\mathsf{t}] + (\mathrm{E}[\mathsf{t}-\theta])^2. \tag{9.3.30}$$

□

• 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*

$$\mathrm{MSE}[a\mathsf{t};\theta] = a^2\,\mathrm{var}[\mathsf{t}] + (a-1)^2\theta^2. \tag{9.3.31}$$

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 (9.3.30). □

• 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*

$$\frac{d}{da}\,\mathrm{MSE}[a\mathsf{t};\theta]\bigg|_{a=1} > 0. \tag{9.3.32}$$

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 (9.3.31) is

$$\frac{d}{da}\,\mathrm{MSE}[a\mathsf{t};\theta] = 2a\,\mathrm{var}[\mathsf{t}] + 2(a-1)\theta^2 \tag{9.3.33}$$

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*

$$a = \frac{\theta^2}{\mathrm{var}[\mathsf{t}] + \theta^2}. \tag{9.3.34}$$

ANSWER. Rewrite (9.3.33) as $2a(\text{var}[\mathsf{t}] + \theta^2) - 2\theta^2$ and set it zero. □

• f. *1 point Assume* t *has an exponential distribution with parameter* $\lambda > 0$*, i.e.,*

$$f_{\mathsf{t}}(t) = \lambda \exp(-\lambda t), \qquad t \geq 0 \quad \text{and} \quad f_{\mathsf{t}}(t) = 0 \quad \text{otherwise.} \tag{9.3.35}$$

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* $\text{var}[\mathsf{t}] = 1/\lambda^2$.

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 $-t^2 \exp(-\lambda t)\Big|_0^\infty + 2\int_0^\infty t \exp(-\lambda t)\, dt = \frac{2}{\lambda}\int_0^\infty \lambda t \exp(-\lambda t)\, d$ $2/\lambda^2$. Therefore $\text{var}[\mathsf{t}] = \text{E}[\mathsf{t}^2] - (\text{E}[\mathsf{t}])^2 = 2/\lambda^2 - 1/\lambda^2 = 1/\lambda^2$. □

• 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 (9.3.34).

$$a = \frac{1/\lambda^2}{\text{var}[\mathsf{t}] + 1/\lambda^2} = \frac{1/\lambda^2}{1/\lambda^2 + 1/\lambda^2} = \frac{1}{2}. \tag{9.3.36}$$

□

• 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

$$a = \frac{1/\lambda^2}{\text{var}[\mathsf{t}] + 1/\lambda^2} = \frac{1/\lambda^2}{1/n\lambda^2 + 1/\lambda^2} = \frac{n}{n+1}, \tag{9.3.37}$$

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* $\text{E}[\chi^2_m] = m$ *and* $\text{var}[\chi^2_m] = 2m$*, 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. $\text{E}[\mathsf{q}] = m\sigma^2$, therefore q/m is the unbiased estimator. Since $\text{var}[\mathsf{q}/m] = 2\sigma^4/m$, it follows from (9.3.34) 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* y *with unknown mean* μ *and standard deviation* σ^2*. The best unbiased estimator of* σ^2 *is* $\frac{1}{n-1}\sum(\mathsf{y}_i - \bar{\mathsf{y}})^2$*, and the maximum likelihood extimator is* $\frac{1}{n}\sum(\mathsf{y}_i - \bar{\mathsf{y}})^2$*. 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* $\sum(\mathsf{y}_i - \bar{\mathsf{y}})^2$*?*

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 $\sum(\mathsf{y}_i - \bar{\mathsf{y}})^2$ is $\frac{1}{n+1}\sum(\mathsf{y}_i - \bar{\mathsf{y}})^2$.

□

• l. *3 points We are still in the model defined in* k*. Which multiple of the sample mean* $\bar{\mathsf{y}}$ *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 $\bar{\mathsf{y}}$. Indeed, $\bar{\mathsf{y}}$ 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 $\bar{\mathsf{y}}$ and s^2_u: 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 $\bar{\mathsf{y}}$ and s^2_u is less efficient than it could be. Since the factorization theorem for sufficient statistics holds even if the parameter θ and its estimate 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

$$f_{\mathbf{y}}(y_1, \ldots, y_n; \mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\Big(-\sum_{i=1}^n (y_i - \mu)^2/2\sigma^2\Big) = \tag{9.3.38}$$

$$= (2\pi\sigma^2)^{-n/2} \exp\Big(-\Big(\sum_{i=1}^n (y_i - \bar{y})^2 - n(\bar{y} - \mu)^2\Big)/2\sigma^2\Big) = \tag{9.3.39}$$

$$= (2\pi\sigma^2)^{-n/2} \exp\Big(-\frac{(n-1)\mathsf{s}^2_u - n(\bar{y} + \mu)^2}{2\sigma^2}\Big). \tag{9.3.40}$$

CHAPTER 10

Estimation Principles and Classification of Estimators

10.1. 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 $\mathtt{t}_n$ (where n is the sample size) of the parameter θ is consistent iff

$$\operatorname*{plim}_{n\to\infty} \mathtt{t}_n = \theta. \tag{10.1.1}$$

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 166. *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 $\mathtt{t}$ is a consistent estimator for θ, and the function g is continuous at the true value of θ, then $g(\mathtt{t})$ 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 $g(\mathtt{t})$ we have to show that for all $\varepsilon > 0$, $\Pr[|g(\mathtt{t}) - g(\theta)| \geq \varepsilon] \to 0$. Choose for the given ε a δ as above, then $|g(\mathtt{t}) - g(\theta)| \geq \varepsilon$ implies $|\mathtt{t} - \theta| \geq \delta$, because all those values of $\mathtt{t}$ for with $|\mathtt{t} - \theta| < \delta$ lead to a $g(\mathtt{t})$ with $|g(\mathtt{t}) - g(\theta)| < \varepsilon$. This logical

implication means that

$$\Pr[|g(\mathtt{t}) - g(\theta)| \geq \varepsilon] \leq \Pr[|\mathtt{t} - \theta| \geq \delta]. \tag{10.1.2}$$

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 167. *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. □

10.2. 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:

$$\Pr[|\mathsf{x}-\theta| \leq |\mathsf{y}-\theta|] = 1 \tag{10.2.1}$$

$$\mathrm{E}[g(\mathsf{x}-\theta)] \leq \mathrm{E}[g(\mathsf{y}-\theta)] \tag{10.2.2}$$

for every continuous function g which is and nonincreasing for $x < 0$ and nondecreasing for $x > 0$

$$\mathrm{E}[g(|\mathsf{x}-\theta|)] \leq \mathrm{E}[g(|\mathsf{y}-\theta|)] \tag{10.2.3}$$

for every continuous and nondecreasing function g

$$\Pr[\{|\mathsf{x}-\theta| > \varepsilon\}] \leq \Pr[\{|\mathsf{y}-\theta| > \varepsilon\}] \quad \text{for every } \varepsilon \tag{10.2.4}$$

$$\mathrm{E}[(\mathsf{x}-\theta)^2] \leq \mathrm{E}[(\mathsf{y}-\theta)^2] \tag{10.2.5}$$

$$\Pr[|\mathsf{x}-\theta| < |\mathsf{y}-\theta|] \geq \Pr[|\mathsf{x}-\theta| > |\mathsf{y}-\theta|] \tag{10.2.6}$$

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.

10.3. 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 10.3.1. An estimator $\hat{\theta}$ of the scalar θ is called *median unbiased* for all $\theta \in \Theta$ iff

$$\Pr[\hat{\theta} < \theta] = \Pr[\hat{\theta} > \theta] = \frac{1}{2} \tag{10.3.1}$$

This concept is always applicable, even for estimators whose expected value does not exist.

PROBLEM 168. *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*

$$\Pr[\mathsf{x} = r] = (1-p)^{r-1}p. \tag{10.3.2}$$

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 169. *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.

$$\mathrm{E}[e^{\lambda\mathsf{t}}] = e^0 \cdot (1-\theta)^2 + e^{\lambda} \cdot 2\theta(1-\theta) + e^{2\lambda} \cdot \theta^2 = \left(1 - \theta + \theta e^{\lambda}\right)^2 \tag{10.3.3}$$

□

PROBLEM 170. *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*

$$\mathsf{t} + \beta(\mathsf{u} - \mathsf{t}). \tag{10.3.4}$$

- b. *2 points By solving the first order condition show that the unbiased linear combination of* t *and* u *which has lowest* MSE *is*

$$\hat{\theta} = \mathsf{t} - \frac{\mathrm{cov}[\mathsf{t}, \mathsf{u}-\mathsf{t}]}{\mathrm{var}[\mathsf{u}-\mathsf{t}]}(\mathsf{u} - \mathsf{t}) \tag{10.3.5}$$

Hint: your arithmetic will be simplest if you start with (10.3.4)*.*

- c. *1 point If* ρ^2 *is the squared correlation coefficient between* t *and* $\mathsf{u}-\mathsf{t}$*, i.e.,*

$$\rho^2 = \frac{(\mathrm{cov}[\mathsf{t}, \mathsf{u}-\mathsf{t}])^2}{\mathrm{var}[\mathsf{t}]\,\mathrm{var}[\mathsf{u}-\mathsf{t}]} \tag{10.3.6}$$

show that $\mathrm{var}[\hat{\theta}] = \mathrm{var}[\mathsf{t}](1-\rho^2)$*.*

- d. *1 point Show that* $\mathrm{cov}[\mathsf{t}, \mathsf{u}-\mathsf{t}] \neq 0$ *implies* $\mathrm{var}[\mathsf{u}-\mathsf{t}] \neq 0$*.*
- e. *2 points Use* (10.3.5) *to show that if* t *is the minimum* MSE *unbiased estimator of* θ*, and* u *another unbiased estimator of* θ*, then*

$$\mathrm{cov}[\mathsf{t}, \mathsf{u}-\mathsf{t}] = 0. \tag{10.3.7}$$

- f. *1 point Use* (10.3.5) *to show also the opposite: if* t *is an unbiased estimator of* θ *with the property that* $\mathrm{cov}[\mathsf{t}, \mathsf{u}-\mathsf{t}] = 0$ *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:

$$\hat{\theta}_n = \begin{cases} \theta & \text{with probability } 1 - \frac{1}{n} \\ n & \text{with probability } \frac{1}{n} \end{cases} \tag{10.3.8}$$

Then $\Pr(\left|\hat{\theta}_n - \theta\right| \geq \varepsilon) \leq \frac{1}{n}$, i.e., the estimator is consistent, but $\mathrm{E}[\hat{\theta}] = \theta\frac{n-1}{n} + 1 \to \theta + 1 \neq 0$.

PROBLEM 171. *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 (10.3.8), say

$$\hat{\theta}_n = \begin{cases} \theta & \text{with probability } 1 - \frac{1}{n} \\ n^2 & \text{with probability } \frac{1}{n} \end{cases} \tag{10.3.9}$$

Here $\Pr(\left|\hat{\theta}_n - \theta\right| \geq \varepsilon) \leq \frac{1}{n}$, i.e., the estimator is consistent, but $\mathrm{E}[\hat{\theta}] = \theta\frac{n-1}{n} + n \to \theta + n$. □

And of course there are estimators which are unbiased but not consistent: simply take the first observation x_1 as an estimator if $\mathrm{E}[\mathsf{x}]$ and ignore all the other observations.

10.4. The Cramer-Rao Lower Bound

Take a scalar random variable y with density function f_{y}. The entropy of y, if it exists, is $\mathrm{H}[\mathsf{y}] = -\mathrm{E}[\log(f_{\mathsf{y}}(\mathsf{y}))]$. 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

$$f_{\mathsf{y}}(y) - f_{\mathsf{y}}(y)\log f_{\mathsf{y}}(y) \leq g(y) - f_{\mathsf{y}}(y)\log g(y). \tag{10.4.1}$$

This holds for every value y, and integrating over y gives $1 - \mathrm{E}[\log f_{\mathsf{y}}(\mathsf{y})] \leq 1 - \mathrm{E}[\log g(\mathsf{y})]$ or

$$\mathrm{E}[\log f_{\mathsf{y}}(\mathsf{y})] \geq \mathrm{E}[\log g(\mathsf{y})]. \tag{10.4.2}$$

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 $\mathrm{E}[\log g(\mathsf{y})]$ 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 (10.4.2) 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 172. *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 $\mathrm{E}_\theta[\mathsf{t}]$ 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 $\mathrm{E}_\circ$ 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 $\mathrm{E}_\circ$ 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

$$\int_{-\infty}^{\infty} \frac{\partial}{\partial\theta} f_{\mathsf{y}}(y;\theta)t(y)\,dy = \frac{\partial}{\partial\theta}\int_{-\infty}^{\infty} f_{\mathsf{y}}(y;\theta)t(y)\,dy = \frac{\partial}{\partial\theta}\,\mathrm{E}_\theta[t(\mathsf{y})] \tag{10.4.3}$$

$$\int_{-\infty}^{\infty} \frac{\partial^2}{(\partial\theta)^2} f_{\mathsf{y}}(y;\theta)t(y)\,dy = \frac{\partial^2}{(\partial\theta)^2}\int_{-\infty}^{\infty} f_{\mathsf{y}}(y;\theta)t(y)\,dy = \frac{\partial^2}{(\partial\theta)^2}\,\mathrm{E}_\theta[t(\mathsf{y})]. \tag{10.4.4}$$

• a. *1 point The* score *is defined as the random variable*

$$q(\mathsf{y};\theta) = \frac{\partial}{\partial\theta}\log f_{\mathsf{y}}(\mathsf{y};\theta). \tag{10.4.5}$$

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

$$q(\mathsf{y};\theta) = \frac{1}{f_{\mathsf{y}}(\mathsf{y};\theta)}\frac{\partial f_{\mathsf{y}}(\mathsf{y};\theta)}{\partial\theta} \tag{10.4.6}$$

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* (**??**)*, show that the score function has the form*

$$q(\mathsf{y};\theta) = \frac{\mathsf{y} - \frac{\partial b(\theta)}{\partial\theta}}{a(\psi)} \tag{10.4.7}$$

ANSWER. This is a simple substitution: if

$$f_{\mathsf{y}}(y;\theta,\psi) = \exp\Bigl(\frac{y\theta - b(\theta)}{a(\psi)} + c(y,\psi)\Bigr), \tag{10.4.8}$$

then

$$\frac{\partial \log f_{\mathsf{y}}(\mathsf{y};\theta,\psi)}{\partial\theta} = \frac{\mathsf{y} - \frac{\partial b(\theta)}{\partial\theta}}{a(\psi)} \tag{10.4.9}$$

□

• c. *3 points If $f_{\mathsf{y}}(y;\theta^\circ)$ is the true density function of y, then we know from* (10.4.2) *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

$$\mathrm{E}_\circ[q(\mathsf{y};\theta)] = \int_{-\infty}^{\infty} q(y;\theta) f_{\mathsf{y}}(y;\theta^\circ)\,dy = \int_{-\infty}^{\infty} \frac{1}{f_{\mathsf{y}}(y;\theta)} \frac{\partial f_{\mathsf{y}}(y;\theta)}{\partial\theta} f_{\mathsf{y}}(y;\theta^\circ)\,dy. \tag{10.4.10}$$

For $\theta = \theta^\circ$ this simplifies:

$$\mathrm{E}_\circ[q(\mathsf{y};\theta^\circ)] = \int_{-\infty}^{\infty} \frac{\partial f_{\mathsf{y}}(y;\theta)}{\partial\theta}\bigg|_{\theta=\theta^\circ} dy = \frac{\partial}{\partial\theta}\int_{-\infty}^{\infty} f_{\mathsf{y}}(y;\theta)\,dy\bigg|_{\theta=\theta^\circ} = \frac{\partial}{\partial\theta} 1 = 0. \tag{10.4.11}$$

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,*

$$\mathrm{E}_\circ[\mathsf{y}] = \frac{\partial b(\theta)}{\partial\theta}\bigg|_{\theta=\theta^\circ} \tag{10.4.12}$$

Answer. Follows from the fact that the score function of the exponential family (10.4.7) has zero expected value. □

• e. *5 points If we differentiate the score, we obtain the* Hessian

$$\mathsf{h}(\theta) = \frac{\partial^2}{(\partial\theta)^2} \log f_{\mathsf{y}}(\mathsf{y};\theta). \tag{10.4.13}$$

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:

$$\mathrm{E}_\circ[\mathsf{h}(\theta^\circ)] = -\mathrm{E}_\circ[\mathsf{q}^2(\theta^\circ)]. \tag{10.4.14}$$

Answer. Start with the definition of the score

$$q(\mathsf{y};\theta) = \frac{\partial}{\partial\theta}\log f_{\mathsf{y}}(\mathsf{y};\theta) = \frac{1}{f_{\mathsf{y}}(\mathsf{y};\theta)}\frac{\partial}{\partial\theta} f_{\mathsf{y}}(\mathsf{y};\theta), \tag{10.4.15}$$

and differentiate the rightmost expression one more time:

$$h(\mathsf{y};\theta) = \frac{\partial}{(\partial\theta)} q(\mathsf{y};\theta) = -\frac{1}{f_{\mathsf{y}}^2(\mathsf{y};\theta)}\left(\frac{\partial}{\partial\theta} f_{\mathsf{y}}(\mathsf{y};\theta)\right)^2 + \frac{1}{f_{\mathsf{y}}(\mathsf{y};\theta)}\frac{\partial^2}{\partial\theta^2} f_{\mathsf{y}}(\mathsf{y};\theta) \tag{10.4.16}$$

$$= -q^2(\mathsf{y};\theta) + \frac{1}{f_{\mathsf{y}}(\mathsf{y};\theta)}\frac{\partial^2}{\partial\theta^2} f_{\mathsf{y}}(\mathsf{y};\theta) \tag{10.4.17}$$

Taking expectations we get

$$\mathrm{E}_\circ[h(\mathsf{y};\theta)] = -\mathrm{E}_\circ[q^2(\mathsf{y};\theta)] + \int_{-\infty}^{+\infty} \frac{1}{f_{\mathsf{y}}(y;\theta)}\left(\frac{\partial^2}{\partial\theta^2} f_{\mathsf{y}}(y;\theta)\right) f_{\mathsf{y}}(y;\theta^\circ)\,dy \tag{10.4.18}$$

Again, for $\theta = \theta^\circ$, we can simplify the integrand and differentiate under the integral sign:

$$\int_{-\infty}^{+\infty} \frac{\partial^2}{\partial\theta^2} f_{\mathsf{y}}(y;\theta)\,dy = \frac{\partial^2}{\partial\theta^2}\int_{-\infty}^{+\infty} f_{\mathsf{y}}(y;\theta)\,dy = \frac{\partial^2}{\partial\theta^2} 1 = 0. \tag{10.4.19}$$

□

• f. *Derive from* (10.4.14) *that, for the exponential dispersion family* (**??**),

$$\mathrm{var}_\circ[\mathsf{y}] = \frac{\partial^2 b(\theta)}{\partial\theta^2} a(\phi)\bigg|_{\theta=\theta^\circ} \tag{10.4.20}$$

Answer. Differentiation of (10.4.7) 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. (10.4.14) says therefore

$$\frac{\partial^2 b(\theta)}{\partial\theta^2}\bigg|_{\theta=\theta^\circ}\frac{1}{a(\phi)} = \mathrm{E}_\circ[\mathsf{q}^2(\theta^\circ)] = \frac{1}{\big(a(\phi)\big)^2}\,\mathrm{var}_\circ[\mathsf{y}] \tag{10.4.21}$$

from which (10.4.20) follows. □

Problem 173.

• a. *Use the results from question 172 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:

$$\mathrm{E}_\circ[\mathsf{q}(\theta)\mathsf{t}] = \int_{-\infty}^{\infty} \frac{1}{f_{\mathsf{y}}(y;\theta)}\frac{\partial f_{\mathsf{y}}(y;\theta)}{\partial\theta} t(y) f_{\mathsf{y}}(y;\theta^\circ)\,dy \tag{10.4.22}$$

If the θ in $\mathsf{q}(\theta)$ is the right parameter value θ° one can simplify:

$$\mathrm{E}_\circ[\mathsf{q}(\theta^\circ)\mathsf{t}] = \int_{-\infty}^{\infty} \frac{\partial f_{\mathsf{y}}(y;\theta)}{\partial\theta}\bigg|_{\theta=\theta^\circ} t(y)\,dy \tag{10.4.23}$$

$$= \frac{\partial}{\partial\theta}\int_{-\infty}^{\infty} f_{\mathsf{y}}(y;\theta) t(y)\,dy\bigg|_{\theta=\theta^\circ} \tag{10.4.24}$$

$$= \frac{\partial}{\partial\theta}\mathrm{E}_\theta[\mathsf{t}]\bigg|_{\theta=\theta^\circ} \tag{10.4.25}$$

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* $\operatorname{cov}_\circ[\mathsf{q}(\theta^\circ), \mathsf{t}] = \frac{\partial}{\partial\theta}\theta = 1$*. From this follows by Cauchy-Schwartz* $\operatorname{var}_\circ[\mathsf{t}]\operatorname{var}_\circ[\mathsf{q}(\theta^\circ)] \geq 1$*, or* $\operatorname{var}_\circ[\mathsf{t}] \geq 1/\operatorname{var}_\circ[\mathsf{q}(\theta^\circ)]$*. Since* $\operatorname{E}_\circ[\mathsf{q}(\theta^\circ)] = 0$*, we know* $\operatorname{var}_\circ[\mathsf{q}(\theta^\circ)] = \operatorname{E}_\circ[\mathsf{q}^2(\theta^\circ)]$*, and since* t *is unbiased, we know* $\operatorname{var}_\circ[\mathsf{t}] = \operatorname{MSE}_\circ[\mathsf{t};\theta^\circ]$*. Therefore the Cauchy-Schwartz inequality reads*

$$\operatorname{MSE}_\circ[\mathsf{t};\theta^\circ] \geq 1/\operatorname{E}_\circ[\mathsf{q}^2(\theta^\circ)]. \tag{10.4.26}$$

This is the Cramer-Rao inequality*. The inverse of the variance of* $\mathsf{q}(\theta^\circ)$*,* $1/\operatorname{var}_\circ[\mathsf{q}(\theta^\circ)] = 1/\operatorname{E}_\circ[\mathsf{q}^2(\theta^\circ)]$*, 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* (10.4.14)*, the Cramer Rao inequality can also be written in the form*

$$\operatorname{MSE}[\mathsf{t};\theta^\circ] \geq -1/\operatorname{E}_\circ[\mathsf{h}(\theta^\circ)]. \tag{10.4.27}$$

(10.4.26) and (10.4.27) 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

$$\operatorname{var}[\mathsf{t}] \geq \frac{1}{\operatorname{E}[\left(\frac{\partial}{\partial\theta}\log f_{\mathsf{y}}(\mathsf{y};\theta)\right)^2]} = \frac{-1}{\operatorname{E}[\frac{\partial^2}{\partial\theta^2}\log f_{\mathsf{y}}(\mathsf{y};\theta)]}. \tag{10.4.28}$$

(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:

$$\operatorname{var}[\mathsf{t}] \geq \frac{1}{\operatorname{E}[\left(\frac{\partial}{\partial\theta}\log f_{\mathbf{y}}(\mathbf{y};\theta)\right)^2]} = \frac{-1}{\operatorname{E}[\frac{\partial^2}{\partial\theta^2}\log f_{\mathbf{y}}(\mathbf{y};\theta)]}. \tag{10.4.29}$$

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 174. *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

$$f_{\mathsf{y}_i}(y) = (2\pi\sigma^2)^{-1/2}\exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \tag{10.4.30}$$

therefore the log likelihood function of the whole vector is

$$\ell(\mathbf{y};\mu) = \sum_{i=1}^n \log f_{\mathsf{y}_i}(\mathsf{y}_i) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^n(\mathsf{y}_i-\mu)^2 \tag{10.4.31}$$

$$\frac{\partial}{\partial\mu}\ell(\mathsf{y};\mu) = \frac{1}{\sigma^2}\sum_{i=1}^n(\mathsf{y}_i-\mu) \tag{10.4.32}$$

In order to apply (10.4.29) you can either square this and take the expected value

$$\operatorname{E}[\left(\frac{\partial}{\partial\mu}\ell(\mathsf{y};\mu)\right)^2] = \frac{1}{\sigma^4}\sum\operatorname{E}[(\mathsf{y}_i-\mu)^2] = n/\sigma^2 \tag{10.4.33}$$

alternatively one may take one more derivative from (10.4.32) to get

$$\frac{\partial^2}{\partial\mu^2}\ell(\mathsf{y};\mu) = -\frac{n}{\sigma^2} \tag{10.4.34}$$

This is constant, therefore equal to its expected value. Therefore the Cramer-Rao Lower Bound says that $\operatorname{var}[\bar{\mathsf{y}}] \geq \sigma^2/n$. This holds with equality. □

Problem 175. *Assume* $\mathsf{y}_i \sim \mathsf{NID}(0,\sigma^2)$ *(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* (4.9.5)*.*

Answer.

$$\operatorname{E}[y_i^2] = \operatorname{var}[y_i] + (\operatorname{E}[y_i])^2 = \sigma^2 + 0 = \sigma^2 \tag{10.4.35}$$

$$z_i = \frac{y_i}{\sigma} \sim \mathsf{NID}(0,1) \tag{10.4.36}$$

$$y_i = \sigma z_i \tag{10.4.37}$$

$$y_i^2 = \sigma^2 z_i^2 \tag{10.4.38}$$

$$\sum_{i=1}^n y_i^2 = \sigma^2\sum_{i=1}^n z_i^2 \sim \sigma^2\chi^2_n \tag{10.4.39}$$

$$\frac{1}{n}\sum_{i=1}^n y_i^2 = \frac{\sigma^2}{n}\sum_{i=1}^n z_i^2 \sim \frac{\sigma^2}{n}\chi^2_n \tag{10.4.40}$$

$$\operatorname{var}\Big[\frac{1}{n}\sum_{i=1}^n y_i^2\Big] = \frac{\sigma^4}{n^2}\operatorname{var}[\chi^2_n] = \frac{\sigma^4}{n^2}2n = \frac{2\sigma^4}{n} \tag{10.4.41}$$

□

• b. *4 points Show that this variance is at the same time the Cramer Rao lower bound.*

Answer.

$$\ell(y,\sigma^2) = \log f_{\mathsf{y}}(y;\sigma^2) = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma^2 - \frac{y^2}{2\sigma^2} \tag{10.4.42}$$

$$\frac{\partial\log f_{\mathsf{y}}}{\partial\sigma^2}(y;\sigma^2) = -\frac{1}{2\sigma^2} + \frac{y^2}{2\sigma^4} = \frac{y^2-\sigma^2}{2\sigma^4} \tag{10.4.43}$$

Since $\dfrac{\mathsf{y}^2-\sigma^2}{2\sigma^4}$ has zero mean, it follows

$$\operatorname{E}[\left(\frac{\partial\log f_{\mathsf{y}}}{\partial\sigma^2}(\mathsf{y};\sigma^2)\right)^2] = \frac{\operatorname{var}[\mathsf{y}^2]}{4\sigma^8} = \frac{1}{2\sigma^4}. \tag{10.4.44}$$

Alternatively, one can differentiate one more time:

$$\frac{\partial^2 \log f_{\mathsf{y}}}{(\partial\sigma^2)^2}(y;\sigma^2) = -\frac{y^2}{\sigma^6} + \frac{1}{2\sigma^4} \tag{10.4.45}$$

$$\mathrm{E}[\frac{\partial^2 \log f_{\mathsf{y}}}{(\partial\sigma^2)^2}(\mathsf{y};\sigma^2)] = -\frac{\sigma^2}{\sigma^6} + \frac{1}{2\sigma^4} = \frac{1}{2\sigma^4} \tag{10.4.46}$$

(10.4.47)

This makes the Cramer Rao lower bound $2\sigma^4/n$. □

PROBLEM 176. *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*

$$p_{\mathsf{x}_i}(x) = \Pr[\mathsf{x}_i = x] = \frac{\lambda^x}{x!}e^{-\lambda} \quad x = 0, 1, 2, \ldots. \tag{10.4.48}$$

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 176. 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

$$\operatorname{var}[\mathsf{t}] \geq \frac{1}{n\,\mathrm{E}[\left(\frac{\partial}{\partial\theta}\ln p_{\mathsf{y}}(\mathsf{y};\theta)\right)^2]} = \frac{-1}{n\,\mathrm{E}[\frac{\partial^2}{\partial\theta^2}\ln p_{\mathsf{y}}(\mathsf{y};\theta)]}. \tag{10.4.49}$$

ANSWER. The Cramer Rao lower bound says no.

$$\log p_{\mathsf{x}}(\mathsf{x};\lambda) = \mathsf{x}\log\lambda - \log\mathsf{x}! - \lambda \tag{10.4.50}$$

$$\frac{\partial \log p_{\mathsf{x}}}{\partial\lambda}(\mathsf{x};\lambda) = \frac{\mathsf{x}}{\lambda} - 1 = \frac{\mathsf{x}-\lambda}{\lambda} \tag{10.4.51}$$

$$\mathrm{E}[\left(\frac{\partial \log p_{\mathsf{x}}}{\partial\lambda}(\mathsf{x};\lambda)\right)^2] = \mathrm{E}[\frac{(\mathsf{x}-\lambda)^2}{\lambda^2}] = \frac{\operatorname{var}[\mathsf{x}]}{\lambda^2} = \frac{1}{\lambda}. \tag{10.4.52}$$

Or alternatively, after (10.4.51) do

$$\frac{\partial^2 \log p_{\mathsf{x}}}{\partial\lambda^2}(\mathsf{x};\lambda) = -\frac{\mathsf{x}}{\lambda^2} \tag{10.4.53}$$

$$-\mathrm{E}[\left(\frac{\partial^2 \log p_{\mathsf{x}}}{\partial\lambda^2}(\mathsf{x};\lambda)\right)] = \frac{\mathrm{E}[\mathsf{x}]}{\lambda^2} = \frac{1}{\lambda}. \tag{10.4.54}$$

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*:

(10.4.55)

$$\boldsymbol{I} = \begin{bmatrix} -n\,\mathrm{E}[\frac{\partial^2\ell}{\partial\theta_1^2}] & \cdots & -n\,\mathrm{E}[\frac{\partial^2\ell}{\partial\theta_1\partial\theta_k}] \\ \vdots & \ddots & \vdots \\ -n\,\mathrm{E}[\frac{\partial^2\ell}{\partial\theta_k\partial\theta_1}] & \cdots & -n\,\mathrm{E}[\frac{\partial^2\ell}{\partial\theta_k^2}] \end{bmatrix} = \begin{bmatrix} n\,\mathrm{E}[(\frac{\partial\ell}{\partial\theta_1})^2] & \cdots & n\,\mathrm{E}[\frac{\partial\ell}{\partial\theta_1}\frac{\partial\ell}{\partial\theta_k}] \\ \vdots & \ddots & \vdots \\ n\,\mathrm{E}[\frac{\partial\ell}{\partial\theta_k}\frac{\partial\ell}{\partial\theta_1}] & \cdots & n\,\mathrm{E}[(\frac{\partial\ell}{\partial\theta_k})^2] \end{bmatrix},$$

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 $\operatorname{var}[\mathsf{t}_i] \geq i^{ii}$.

10.5. 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 177 shows that in some situations one can even dispense with the independence of the observations.

PROBLEM 177. *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.

$$\tilde{\mathsf{y}} = \alpha_1\mathsf{y}_1 + \alpha_2\mathsf{y}_2 \tag{10.5.1}$$

$$\operatorname{var}\tilde{\mathsf{y}} = \alpha_1^2\operatorname{var}[\mathsf{y}_1] + \alpha_2^2\operatorname{var}[\mathsf{y}_2] + 2\alpha_1\alpha_2\operatorname{cov}[\mathsf{y}_1,\mathsf{y}_2] \tag{10.5.2}$$

$$= \sigma^2(\alpha_1^2 + \alpha_2^2 + 2\alpha_1\alpha_2\rho). \tag{10.5.3}$$

Here we used (6.1.14). Unbiasedness means $\alpha_2 = 1 - \alpha_1$, therefore we call $\alpha_1 = \alpha$ and $\alpha_2 = 1 - \alpha$:

$$\operatorname{var}[\tilde{\mathsf{y}}]/\sigma^2 = \alpha^2 + (1-\alpha)^2 + 2\alpha(1-\alpha)\rho \tag{10.5.4}$$

Now sort by the powers of α:

$$= 2\alpha^2(1-\rho) - 2\alpha(1-\rho) + 1 \tag{10.5.5}$$

$$= 2(\alpha^2 - \alpha)(1-\rho) + 1. \tag{10.5.6}$$

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 (10.5.3) to get $\frac{\sigma^2}{2}\big(1+\rho\big)$. □

Problem 178. *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

$$\tilde{\mu} = \alpha_1\mathsf{y}_1 + \alpha_2\mathsf{y}_2 + \gamma \tag{10.5.7}$$

The measurements themselves are unbiased, i.e., $\mathrm{E}[\mathsf{y}_i - \mu] = 0$, therefore

$$\mathrm{E}[\tilde{\mu} - \mu] = (\alpha_1 + \alpha_2 - 1)\,\mathrm{E}[\mu] + \gamma = 0 \tag{10.5.8}$$

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

$$\mathrm{var}[\tilde{\mu} - \mu] = \alpha^2\sigma^2 + (1-\alpha)^2\tau^2 \tag{10.5.9}$$

□

• b. *4 points Define ω^2 by*

$$\frac{1}{\omega^2} = \frac{1}{\sigma^2} + \frac{1}{\tau^2} \qquad \textit{which can be solved to give} \qquad \omega^2 = \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}. \tag{10.5.10}$$

Show that the Best (i.e., minimum MSE*) linear unbiased estimator (BLUE) of μ based on these two measurements is*

$$\hat{\mathsf{y}} = \frac{\omega^2}{\sigma^2}\mathsf{y}_1 + \frac{\omega^2}{\tau^2}\mathsf{y}_2 \tag{10.5.11}$$

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 (10.5.9) takes its minimum value where its derivative with respect of α is zero, i.e., where

$$\frac{\partial}{\partial\alpha}\Big(\alpha^2\sigma^2 + (1-\alpha)^2\tau^2\Big) = 2\alpha\sigma^2 - 2(1-\alpha)\tau^2 = 0 \tag{10.5.12}$$

$$\alpha\sigma^2 = \tau^2 - \alpha\tau^2 \tag{10.5.13}$$

$$\alpha = \frac{\tau^2}{\sigma^2+\tau^2} \tag{10.5.14}$$

In terms of ω one can write

$$\alpha = \frac{\tau^2}{\sigma^2+\tau^2} = \frac{\omega^2}{\sigma^2} \qquad \text{and} \qquad 1-\alpha = \frac{\sigma^2}{\sigma^2+\tau^2} = \frac{\omega^2}{\tau^2}. \tag{10.5.15}$$

□

• c. *2 points Show: the* MSE *of the BLUE ω^2 satisfies the following equation:*

$$\frac{1}{\omega^2} = \frac{1}{\sigma^2} + \frac{1}{\tau^2} \tag{10.5.16}$$

Answer. We already have introduced the notation ω^2 for the quantity defined by (10.5.16); therefore all we have to show is that the MSE or, equivalently, the variance of the estimation error is equal to this ω^2:

$$\mathrm{var}[\tilde{\mu} - \mu] = \Big(\frac{\omega^2}{\sigma^2}\Big)^2\sigma^2 + \Big(\frac{\omega^2}{\tau^2}\Big)^2\tau^2 = \omega^4\Big(\frac{1}{\sigma^2} + \frac{1}{\tau^2}\Big) = \omega^4\frac{1}{\omega^2} = \omega^2 \tag{10.5.17}$$

□

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)

10.6. 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 179. *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

$$f_{\mathsf{y}}(y; \mu_0) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-\mu_0)^2}{2}}. \tag{10.6.1}$$

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

$$L(\mu, \sigma^2; \mathsf{y}_1, \ldots, \mathsf{y}_n) = \prod_{i=1}^{n} f_{\mathsf{y}}(\mathsf{y}_i) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2\sigma^2}\sum(\mathsf{y}_i-\mu)^2}. \tag{10.6.2}$$

Its logarithm is more convenient to maximize:

$$\ell = \ln L(\mu, \sigma^2; \mathsf{y}_1, \ldots, \mathsf{y}_n) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum(\mathsf{y}_i - \mu)^2. \tag{10.6.3}$$

To compute the maximum we need the partial derivatives:

$$\frac{\partial \ell}{\partial \mu} = \frac{1}{\sigma^2}\sum(\mathsf{y}_i - \mu) \tag{10.6.4}$$

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum(\mathsf{y}_i - \mu)^2. \tag{10.6.5}$$

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.,

$$f_{\mathsf{t}}(t; \lambda) = \lambda e^{-\lambda t} \quad (t > 0) \tag{10.6.6}$$

$$L(\mathsf{t}_1, \ldots, \mathsf{t}_n; \lambda) = \lambda^n e^{-\lambda(\mathsf{t}_1 + \cdots + \mathsf{t}_n)} \tag{10.6.7}$$

$$\ell(\mathsf{t}_1, \ldots, \mathsf{t}_n \lambda) = n \ln \lambda - \lambda(\mathsf{t}_1 + \cdots + \mathsf{t}_n) \tag{10.6.8}$$

$$\frac{\partial \ell}{\partial \lambda} = \frac{n}{\lambda} - (\mathsf{t}_1 + \cdots + \mathsf{t}_n). \tag{10.6.9}$$

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 180. *$\mathsf{x}_1, \ldots, \mathsf{x}_m$ 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.

$$\begin{aligned}\ell(\mu_{\mathsf{x}}, \mu_{\mathsf{y}}, \sigma^2) = &-\frac{m}{2}\ln 2\pi - \frac{m}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{m}(\mathsf{x}_i - \mu_{\mathsf{x}})^2 \\ &- \frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{j=1}^{n}(\mathsf{y}_j - \mu_{\mathsf{y}})^2\end{aligned} \tag{10.6.10}$$

$$\frac{\partial \ell}{\partial \mu_{\mathsf{x}}} = -\frac{1}{2\sigma^2}\sum -2(\mathsf{x}_i - \mu_{\mathsf{x}}) \qquad = 0 \quad \text{for} \quad \mu_{\mathsf{x}} = \bar{\mathsf{x}} \tag{10.6.11}$$

□

• b. *2 points Derive the MLE of σ^2, based on the combined samples.*

Answer.

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{m+n}{2\sigma^2} + \frac{1}{2\sigma^4}\Big(\sum_{i=1}^{m}(\mathsf{x}_i - \mu_{\mathsf{x}})^2 + \sum_{j=1}^{n}(\mathsf{y}_j - \mu_{\mathsf{y}})^2\Big) \tag{10.6.12}$$

$$\hat{\sigma}^2 = \frac{1}{m+n}\Big(\sum_{i=1}^{m}(\mathsf{x}_i - \bar{\mathsf{x}})^2 + \sum_{j=1}^{n}(\mathsf{y}_i - \bar{\mathsf{y}})^2\Big). \tag{10.6.13}$$

□

10.7. 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.

10.8. 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.

10.9. Sufficient Statistics and Estimation

Weak Sufficiency Principle: If $\mathbf{x}$ has a p.d.f. $f_{\mathbf{x}}(\boldsymbol{x};\theta)$ and if a sufficient statistic $\mathbf{s}(\mathbf{x})$ 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 $\mathbf{s}(\mathbf{x})$, the rest of the data $\mathbf{x}$ 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 $\mathsf{t}(\mathbf{x})$ be an estimator of θ and $\mathbf{s}(\mathbf{x})$ a sufficient statistic for θ. Then one can get an estimator $\mathsf{t}^*(\mathbf{x})$ of θ which has no worse a MSE than $\mathsf{t}(\mathbf{x})$ by taking expectations conditionally on the sufficient statistic, i.e., $\mathsf{t}^*(\mathbf{x}) = \mathrm{E}[\mathsf{t}(\mathbf{x})|\mathbf{s}(\mathbf{x})]$.

To recapitulate: $\mathsf{t}^*(\mathsf{x})$ is obtained by the following two steps: (1) Compute the conditional expectation $\mathsf{t}^{**}(\boldsymbol{s}) = \mathrm{E}[\mathsf{t}(\mathbf{x})|\mathbf{s}(\mathbf{x}) = \boldsymbol{s}]$, and (2) plug $\mathbf{s}(\mathbf{x})$ into t^{**}, i.e., $\mathsf{t}^*(\mathbf{x}) = \mathsf{t}^{**}(\mathbf{s}(\mathbf{x}))$.

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 181 steps you through the proof.

Problem 181. [**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.

$$\pi = \Pr[\mathsf{y}_i \geq 0] = \Pr[\mathsf{y}_i - \mu \geq -\mu] = \Pr[\mathsf{y}_i - \mu \leq \mu] = \Phi(\mu) \tag{10.9.1}$$

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 jth 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 (10.9.1) 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:

$$\begin{bmatrix} \mathsf{y}_j \\ \bar{\mathsf{y}} \end{bmatrix} \sim N\left(\begin{bmatrix} \mu \\ \mu \end{bmatrix}, \begin{bmatrix} 1 & 1/n \\ 1/n & 1/n \end{bmatrix}\right) \tag{10.9.2}$$

□

• 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 (7.3.18) and (7.3.20).*

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

$$\mathrm{E}[\mathsf{v}|\mathsf{u} = u] = \mathrm{E}[\mathsf{v}] + \frac{\mathrm{cov}[\mathsf{u},\mathsf{v}]}{\mathrm{var}[\mathsf{u}]}(u - \mathrm{E}[\mathsf{u}]) \tag{10.9.3}$$

and variance

$$\mathrm{var}[\mathsf{v}|\mathsf{u} = u] = \mathrm{var}[\mathsf{v}] - \frac{(\mathrm{cov}[\mathsf{u},\mathsf{v}])^2}{\mathrm{var}[\mathsf{u}]}. \tag{10.9.4}$$

Plugging $\mathsf{u} = \bar{\mathsf{y}}$ and $\mathsf{v} = \mathsf{y}_j$ into (7.3.18) and (7.3.20) gives: the conditional distribution of y_j conditionally on $\bar{\mathsf{y}} = \bar{y}$ has mean

$$\mathrm{E}[\mathsf{y}_j|\bar{\mathsf{y}} = \bar{y}] = \mathrm{E}[\mathsf{y}_j] + \frac{\mathrm{cov}[\bar{\mathsf{y}},\mathsf{y}_j]}{\mathrm{var}[\bar{\mathsf{y}}]}(\bar{y} - \mathrm{E}[\bar{\mathsf{y}}]) \tag{10.9.5}$$

$$= \mu + \frac{1/n}{1/n}(\bar{y} - \mu) = \bar{y} \tag{10.9.6}$$

and variance

$$\mathrm{var}[\mathsf{y}_j|\bar{\mathsf{y}} = \bar{y}] = \mathrm{var}[\mathsf{y}_j] - \frac{(\mathrm{cov}[\bar{\mathsf{y}},\mathsf{y}_j])^2}{\mathrm{var}[\bar{\mathsf{y}}]} \tag{10.9.7}$$

$$= 1 - \frac{(1/n)^2}{1/n} = 1 - \frac{1}{n}. \tag{10.9.8}$$

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 (6.6.6) gives a decomposition of $\mathrm{var}[\mathsf{y}_j]$: give it here:*

Answer.

$$\mathrm{var}[\mathsf{y}_j] = \mathrm{var}\Big[\mathrm{E}[\mathsf{y}_j|\bar{\mathsf{y}}]\Big] + \mathrm{E}\Big[\mathrm{var}[\mathsf{y}_j|\bar{\mathsf{y}}]\Big] \tag{10.9.9}$$

$$= \mathrm{var}[\bar{\mathsf{y}}] + \mathrm{E}\left[\frac{n-1}{n}\right] = \frac{1}{n} + \frac{n-1}{n} \tag{10.9.10}$$

□

• 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*

$$\hat{\pi} = \Phi(\bar{\mathsf{y}}\sqrt{n/(n-1)}), \tag{10.9.11}$$

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].

10.10. 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 182. *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 (4.1.13) $\binom{11}{8}\theta^9(1-\theta)^3$. They are proportional, the stopping rule therefore does not matter! □

10.11. 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

$$f(\boldsymbol{y}, \boldsymbol{\theta}) = f(\boldsymbol{\theta})f(\boldsymbol{y}|\boldsymbol{\theta}) \tag{10.11.1}$$

$$= f(\boldsymbol{y})f(\boldsymbol{\theta}|\boldsymbol{y}). \tag{10.11.2}$$

Therefore

$$f(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{f(\boldsymbol{\theta})f(\boldsymbol{y}|\boldsymbol{\theta})}{f(\boldsymbol{y})}. \tag{10.11.3}$$

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 (10.11.3) is usually written as ($\propto$ means "proportional to")

$$f(\boldsymbol{\theta}|\boldsymbol{y}) \propto f(\boldsymbol{\theta})f(\boldsymbol{y}|\boldsymbol{\theta}); \tag{10.11.4}$$

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 11

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 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 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.,

$$\Pr\big[R(\mathsf{y}) \ni \boldsymbol{\phi}_0 | \boldsymbol{\phi} = \boldsymbol{\phi}_0\big] \geq 1-\alpha \quad \text{for all } \boldsymbol{\phi}_0 \in \Omega. \tag{11.0.5}$$

The important thing to remember about this definition is that these regions $R(\mathsf{y})$ 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 $\mathsf{y} \sim \mathsf{N}(\boldsymbol{\iota}\mu, \sigma^2\boldsymbol{I})$, 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 (9.1.1), 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

$$\frac{\sum(y_i - a)^2}{\sum(y_i - \bar{y})^2} \leq c \tag{11.0.6}$$

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 (11.0.6) is equivalent to

$$\frac{\sum(y_i - a)^2 - \sum(y_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} \leq c - 1 \tag{11.0.7}$$

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 (11.0.6):

$$\frac{n(\bar{y} - a)^2}{\sum(y_i - \bar{y})^2} \leq c - 1 \tag{11.0.8}$$

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

$$\Pr\Big[\frac{n(\bar{\mathsf{y}} - \mu)^2}{\sum(\mathsf{y}_i - \bar{\mathsf{y}})^2} \leq c - 1\Big] \tag{11.0.9}$$

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.:

$$\frac{(\bar{\mathsf{y}}-\mu)^2}{\frac{1}{n}\frac{1}{n-1}\sum(\mathsf{y}_i-\bar{\mathsf{y}})^2}\sim \mathsf{F}_{1,n-1} \tag{11.0.10}$$

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:

$$\frac{\bar{\mathsf{y}}-\mu}{\sqrt{\frac{1}{n}}\sqrt{\frac{1}{n-1}\sum(\mathsf{y}_i-\bar{\mathsf{y}})^2}}\sim \mathsf{t}_{n-1} \tag{11.0.11}$$

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:

$$R(\boldsymbol{y})=\{u\in\mathbb{R}:\ |u-\bar{y}|\leq N_{(\alpha/2)}\sigma_{\bar{\mathsf{y}}}\}. \tag{11.0.12}$$

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:

$$\Pr[|\mathsf{z}|\geq N_{(\alpha/2)}]=\alpha. \tag{11.0.13}$$

From this follows the coverage probability:

$$\Pr[\mathsf{R}(\boldsymbol{\mathsf{y}})\ni\mu]=\Pr[|\mu-\bar{\mathsf{y}}|\leq N_{(\alpha/2)}\sigma_{\bar{\mathsf{y}}}] \tag{11.0.14}$$

$$=\Pr[|(\mu-\bar{\mathsf{y}})/\sigma_{\bar{\mathsf{y}}}|\leq N_{(\alpha/2)}]=\Pr[|-\mathsf{z}|\leq N_{(\alpha/2)}]=1-\alpha \tag{11.0.15}$$

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 183. *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*

$$R(\boldsymbol{y})=\{u\in\mathbb{R}:\ |u-\bar{y}|\leq t_{(n-1;\alpha/2)}s_{\bar{\mathsf{y}}}\}. \tag{11.0.16}$$

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 $\operatorname{std.dev}[\mathbf{y}]=\sqrt{\operatorname{var}[\mathbf{y}]}$.

• 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

$$s_{\bar{\mathsf{y}}}=s/\sqrt{n}=\sqrt{\sum\frac{(\mathsf{y}_i-\bar{\mathsf{y}})^2}{n(n-1)}} \tag{11.0.17}$$

□

• b. *2 points Compute the coverage probability of the interval* (11.0.16).

Answer. The coverage probability is

$$\Pr[\mathsf{R}(\mathbf{y})\ni\mu]=\Pr[\left|\mu-\bar{\mathsf{y}}\right|\leq t_{(n-1;\alpha/2)}s_{\bar{\mathsf{y}}}] \tag{11.0.18}$$

$$=\Pr[\left|\frac{\mu-\bar{\mathsf{y}}}{s_{\bar{\mathsf{y}}}}\right|\leq t_{(n-1;\alpha/2)}] \tag{11.0.19}$$

$$=\Pr[\left|\frac{(\mu-\bar{\mathsf{y}})/\sigma_{\bar{\mathsf{y}}}}{s_{\bar{\mathsf{y}}}/\sigma_{\bar{\mathsf{y}}}}\right|\leq t_{(n-1;\alpha/2)}] \tag{11.0.20}$$

$$=\Pr[\left|\frac{(\bar{\mathsf{y}}-\mu)/\sigma_{\bar{\mathsf{y}}}}{s/\sigma}\right|\leq t_{(n-1;\alpha/2)}] \tag{11.0.21}$$

$$=1-\alpha, \tag{11.0.22}$$

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 (11.0.22) follows from (30.4.8).

□

• 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.*

TABLE 1. Percentiles of Student's t Distribution. Table entry x satisfies $\Pr[\mathrm{t}_n \le x] = p$.

	$p =$					
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. In our situation

$$\frac{\bar{\mathrm{x}} - \mu}{\mathrm{s}/\sqrt{n}} \sim \mathrm{t}_3 \tag{11.0.23}$$

According to table 1, for $b = 3.182$ follows

$$\Pr[\mathrm{t}_3 \le b] = 0.975 \tag{11.0.24}$$

therefore

$$\Pr[\mathrm{t}_3 > b] = 0.025 \tag{11.0.25}$$

and by symmetry of the t-distribution

$$\Pr[\mathrm{t}_3 < -b] = 0.025 \tag{11.0.26}$$

Now subtract (11.0.26) from (11.0.24) to get

$$\Pr[-b \le \mathrm{t}_3 \le b] = 0.95 \tag{11.0.27}$$

or

$$\Pr[|\mathrm{t}_3| \le b] = 0.95 \tag{11.0.28}$$

or, plugging in the formula for t_3,

$$\Pr\Big[\Big|\frac{\bar{\mathrm{x}} - \mu}{\mathrm{s}/\sqrt{n}}\Big| \le b\Big] = .95 \tag{11.0.29}$$

$$\Pr[|\bar{\mathrm{x}} - \mu| \le b\mathrm{s}/\sqrt{n}] = .95 \tag{11.0.30}$$

$$\Pr[-b\mathrm{s}/\sqrt{n} \le \mu - \bar{\mathrm{x}} \le b\mathrm{s}/\sqrt{n}] = .95 \tag{11.0.31}$$

$$\Pr[\bar{\mathrm{x}} - b\mathrm{s}/\sqrt{n} \le \mu \le \bar{\mathrm{x}} + b\mathrm{s}/\sqrt{n}] = .95 \tag{11.0.32}$$

the confidence interval is therefore $\big[\bar{\mathrm{x}} - b\mathrm{s}/\sqrt{n}, \bar{\mathrm{x}} + b\mathrm{s}/\sqrt{n}\big]$. 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 184. *Using* R, *construct 20 samples of 12 observation each from a* $\mathrm{N}(0,1)$ *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-halfwid` *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=5`

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(stdr
  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 12

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 185. *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 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 $\bar{\mathsf{y}} \geq c$, where c is chosen in such a way that the significance level is 5%, for $n = 100$.

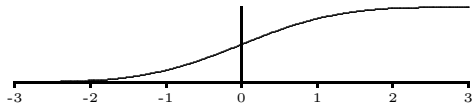


FIGURE 1. Eventually this Figure will show the Power function of a one-sided normal test, i.e., the probability of error of type one as a function of μ; right now this is simply the cdf of a Standard Normal

The 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 186. *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. □

12.1. 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

$$\Pr\big[C(\boldsymbol{\phi}_0)|\boldsymbol{\phi}=\boldsymbol{\phi}_0\big] \leq \alpha \quad \text{for all } \boldsymbol{\phi}_0 \in \Omega. \tag{12.1.1}$$

Mathematically, confidence regions and such families of tests are one and the same thing: if one has a confidence region $R(\mathbf{y})$, 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 187. *Show that with these definitions, equations* (11.0.5) *and* (12.1.1) *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 $\Pr[R(\mathbf{y}) \in \boldsymbol{\phi}_0|\boldsymbol{\phi}=\boldsymbol{\phi}_0] = 1-\Pr[C(\boldsymbol{\phi}_0)|\boldsymbol{\phi}=\boldsymbol{\phi}_0] \geq 1-\alpha$. □

This duality is discussed in [**BD77**, pp. 177–182].

12.2. 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 $\bar{\mathrm{y}} \geq \mu_0$, why not one of the form $\mu_1 \leq \bar{\mathrm{y}} \leq \mu_2$, or maybe not use the mean but decide to build if $\mathrm{y}_1 \geq \mu_3$? 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 $\bar{\mathrm{y}} \geq \mu_0$ 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 $\mathsf{y}_1, \dots, \mathsf{y}_n$ 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:

$$\mathsf{C} = \{y_1, \dots, y_n \colon L(y_1, \dots, y_n; \theta_1) \geq kL(y_1, \dots, y_n; \theta_0)\}. \tag{12.2.1}$$

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

$$\Pr[\mathsf{U}|\theta_0] = \Pr[\mathsf{W}|\theta_0]. \tag{12.2.2}$$

Also

$$\Pr[\mathsf{U}|\theta_1] \geq k \Pr[\mathsf{U}|\theta_0], \tag{12.2.3}$$

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

$$\Pr[\mathsf{W}|\theta_1] \leq k \Pr[\mathsf{W}|\theta_0]. \tag{12.2.4}$$

Linking those two inequalities and the equality gives

$$\Pr[\mathsf{W}|\theta_1] \leq k \Pr[\mathsf{W}|\theta_0] = k \Pr[\mathsf{U}|\theta_0] \leq \Pr[\mathsf{U}|\theta_1], \tag{12.2.5}$$

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

$$C = \{y_1, \dots, y_n \colon \Big(\frac{1}{\sqrt{2\pi}}\Big)^n e^{-\frac{1}{2}((y_1-t)^2+\cdots+(y_n-t)^2)} \geq k\Big(\frac{1}{\sqrt{2\pi}}\Big)^n e^{-\frac{1}{2}(y_1^2+\cdots+y_n^2)}\} \tag{12.2.6}$$

$$= \{y_1, \dots, y_n \colon -\frac{1}{2}((y_1-t)^2 + \cdots + (y_n-t)^2) \geq \ln k - \frac{1}{2}(y_1^2 + \cdots + y_n^2)\} \tag{12.2.7}$$

$$= \{y_1, \dots, y_n \colon t(y_1 + \cdots + y_n) - \frac{t^2 n}{2} \geq \ln k\} \tag{12.2.8}$$

$$= \{y_1, \dots, y_n \colon \bar{y} \geq \frac{\ln k}{nt} + \frac{t}{2}\} \tag{12.2.9}$$

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 188. *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*

$$C = \{y_1, \dots, y_4 \colon L(y_1, \dots, y_4; \mu = 1) \geq 3.633 \cdot L(y_1, \dots, y_4; \mu = 0)\}. \tag{12.2.10}$$

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:

$$(2\pi)^{-2}\exp -\frac{1}{2}\Big((y_1-1)^2+\cdots+(y_4-1)^2\Big) \geq 3.633\cdot(2\pi)^{-2}\exp -\frac{1}{2}\Big(y_1^2+\cdots+y_4^2\Big) \tag{12.2.11}$$

$$-\frac{1}{2}\Big((y_1-1)^2+\cdots+(y_4-1)^2\Big) \geq \ln(3.633)-\frac{1}{2}\Big(y_1^2+\cdots+y_4^2\Big) \tag{12.2.12}$$

$$y_1+\cdots+y_4-2 \geq 1.290 \tag{12.2.13}$$

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

$$C = \{x_1, \dots, x_n \colon \frac{\sup_{\theta\in\Omega} L(x_1,\dots,x_n;\theta)}{\sup_{\theta\in\omega} L(x_1,\dots,x_n;\theta)} \geq k\} \tag{12.2.14}$$

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.

12.3. 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 $\mathsf{N}(\boldsymbol{\theta}, \boldsymbol{\Psi})$. 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:

$$\text{G.W.} = \boldsymbol{h}(\tilde{\theta})^\top \Big\{ \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{\theta}^\top}\Big|_{\tilde{\theta}} \hat{\boldsymbol{\Psi}} \frac{\partial \boldsymbol{h}^\top}{\partial \boldsymbol{\theta}}\Big|_{\tilde{\theta}} \Big\}^{-1} \boldsymbol{h}(\tilde{\theta}) \tag{12.3.1}$$

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

$$\frac{\partial \boldsymbol{h}}{\partial \boldsymbol{\theta}^\top}\Big|_{\tilde{\theta}} \hat{\boldsymbol{\Psi}} \frac{\partial \boldsymbol{h}^\top}{\partial \boldsymbol{\theta}}\Big|_{\tilde{\theta}} \tag{12.3.2}$$

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 $f_{\boldsymbol{\mathsf{x}}}(\boldsymbol{x};\boldsymbol{\theta})$ of $\boldsymbol{\mathsf{x}}$ 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

$$LRT = -2\log\frac{\max_{\boldsymbol{h}(\boldsymbol{\theta})=\boldsymbol{o}} f_{\mathsf{y}}(\mathsf{y};\boldsymbol{\theta})}{\max_{\boldsymbol{\theta}} f_{\mathsf{y}}(\mathsf{y};\boldsymbol{\theta})} = 2(\log f_{\mathsf{y}}(\mathsf{y},\hat{\theta}) - \log f_{\mathsf{y}}(\mathsf{y},\bar{\theta})) \tag{12.3.3}$$

It rejects if imposing the constraint reduces the attained level of the likelihood function too much.

The Wald test has the form

$$\text{Wald} = -\boldsymbol{h}(\hat{\theta})^\top \Big\{ \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{\theta}^\top}\Big|_{\hat{\theta}} \Big(\frac{\partial^2 \log f(\mathsf{y};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\Big|_{\hat{\theta}} \Big)^{-1} \frac{\partial \boldsymbol{h}^\top}{\partial \boldsymbol{\theta}}\Big|_{\hat{\theta}} \Big\}^{-1} \boldsymbol{h}(\hat{\theta}) \tag{12.3.4}$$

To understand this formula, note that $-\Big(\mathcal{E}\Big[\frac{\partial^2 \log f(\mathsf{y};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\Big]\Big)^{-1}$ 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

$$\text{Score} = -\frac{\partial \log f(\mathsf{y};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}\Big|_{\bar{\theta}} \Big(\frac{\partial^2 \log f(\mathsf{y};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\Big|_{\bar{\theta}} \Big)^{-1} \frac{\partial \log f(\mathsf{y};\boldsymbol{\theta})^\top}{\partial \boldsymbol{\theta}}\Big|_{\bar{\theta}} \tag{12.3.5}$$

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 $-\mathcal{E}\left[\frac{\partial^2 \log f(\mathbf{y};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\right]$ is the covariance matrix of the score, i.e., here we take the score twice and divide it by its estimated covariance matrix.

CHAPTER 13

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 189. *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 14

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).

14.1. 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 9, 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

$$\boldsymbol{\mathsf{y}} = \boldsymbol{\iota}\mu + \boldsymbol{\varepsilon} \qquad \boldsymbol{\varepsilon} \sim (\boldsymbol{o}, \sigma^2\boldsymbol{I}) \tag{14.1.1}$$

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

$$\begin{bmatrix}\mathsf{y}_1\\ \vdots \\ \mathsf{y}_n\end{bmatrix} = \begin{bmatrix}1\\ \vdots \\ 1\end{bmatrix}\alpha + \begin{bmatrix}x_1\\ \vdots \\ x_n\end{bmatrix}\beta + \begin{bmatrix}\varepsilon_1\\ \vdots \\ \varepsilon_n\end{bmatrix} = \begin{bmatrix}1 & x_1\\ \vdots & \vdots \\ 1 & x_n\end{bmatrix}\begin{bmatrix}\alpha\\ \beta\end{bmatrix} + \begin{bmatrix}\varepsilon_1\\ \vdots \\ \varepsilon_n\end{bmatrix} \tag{14.1.2}$$

or

$$\boldsymbol{\mathsf{y}} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \boldsymbol{\varepsilon} \sim (\boldsymbol{o}, \sigma^2\boldsymbol{I}) \tag{14.1.3}$$

Problem 190. *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 (14.1.3), 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

14.2. Ordinary Least Squares

In the model $\boldsymbol{\mathsf{y}} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, 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

$$\mathsf{SSE} = (\boldsymbol{\mathsf{y}} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{\mathsf{y}} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{\mathsf{y}}^\top\boldsymbol{\mathsf{y}} - 2\boldsymbol{\mathsf{y}}^\top\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta}. \tag{14.2.1}$$

Problem 156 shows that in model 1, this principle yields the arithmetic mean.

Problem 191. *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 (14.2.1) to get

$$\partial\mathsf{SSE}/\partial\boldsymbol{\beta}^\top = -2\mathbf{y}^\top\boldsymbol{X} + 2\boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{X}. \tag{14.2.2}$$

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"

$$\boldsymbol{X}^\top\mathbf{y} = \boldsymbol{X}^\top\boldsymbol{X}\hat{\boldsymbol{\beta}}. \tag{14.2.3}$$

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 (14.2.3) by $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\mathbf{y}. \tag{14.2.4}$$

If the columns of $\boldsymbol{X}$ are not linearly independent, then (14.2.3) 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 194).

Problem 192. *4 points Using the matrix differentiation rules*

$$\partial\boldsymbol{w}^\top\boldsymbol{x}/\partial\boldsymbol{x}^\top = \boldsymbol{w}^\top \tag{14.2.5}$$

$$\partial\boldsymbol{x}^\top\boldsymbol{M}\boldsymbol{x}/\partial\boldsymbol{x}^\top = 2\boldsymbol{x}^\top\boldsymbol{M} \tag{14.2.6}$$

for symmetric $\boldsymbol{M}$, compute the least-squares estimate $\hat{\boldsymbol{\beta}}$ which minimizes

$$\mathsf{SSE} = (\mathbf{y}-\boldsymbol{X}\boldsymbol{\beta})^\top(\mathbf{y}-\boldsymbol{X}\boldsymbol{\beta}) \tag{14.2.7}$$

You are allowed to assume that $\boldsymbol{X}^\top\boldsymbol{X}$ has an inverse.

Answer. First you have to multiply out

$$(\mathbf{y}-\boldsymbol{X}\boldsymbol{\beta})^\top(\mathbf{y}-\boldsymbol{X}\boldsymbol{\beta}) = \mathbf{y}^\top\mathbf{y} - 2\mathbf{y}^\top\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta}. \tag{14.2.8}$$

The matrix differentiation rules (14.2.5) and (14.2.6) allow us to differentiate (14.2.8) to get

$$\partial\mathsf{SSE}/\partial\boldsymbol{\beta}^\top = -2\mathbf{y}^\top\boldsymbol{X} + 2\boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{X}. \tag{14.2.9}$$

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"

$$\boldsymbol{X}^\top\mathbf{y} = \boldsymbol{X}^\top\boldsymbol{X}\hat{\boldsymbol{\beta}}. \tag{14.2.10}$$

Since $\boldsymbol{X}^\top\boldsymbol{X}$ has an inverse, one can premultiply both sides of (14.2.10) by $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\mathbf{y}. \tag{14.2.11}$$

□

Problem 193. *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 194. *3 points In this Problem we do not assume that $\boldsymbol{X}$ has full column rank, it may be arbitrary.*

• a. *The normal equation* (14.2.3) *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*

$$(\mathbf{y}-\boldsymbol{X}\boldsymbol{\beta})^\top(\mathbf{y}-\boldsymbol{X}\boldsymbol{\beta}) = (\mathbf{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\mathbf{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}) + (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^\top\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}). \tag{14.2.12}$$

Answer. This is true even if $\boldsymbol{X}$ has deficient rank, and it will be shown here in this general case. To prove (14.2.12), write (14.2.1) 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 (14.2.3), 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* (14.2.12).

Answer. (14.2.12) 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 (14.2.12) that for any solution $\boldsymbol{\beta}^*$ of the minimization, $\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{\beta}^*-\hat{\boldsymbol{\beta}}) = \boldsymbol{o}$. Use (14.2.3) 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

$$\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\varepsilon}} \tag{14.2.13}$$

Problem 195. *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, $\hat{\varepsilon} = \mathsf{y} - \boldsymbol{X}\hat{\beta} = \mathsf{y} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}\mathsf{y} = \big(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}\big)\mathsf{y}$. Idempotent, i.e. $\boldsymbol{M}\boldsymbol{M} = \boldsymbol{M}$:

(14.2.14)
$$\boldsymbol{M}\boldsymbol{M} = \big(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\big)\big(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\big) \quad = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top -$$

□

Problem 196. *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 197. *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 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 $\hat{\mathsf{y}}$ is the projection of the vector 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 198. *"Simple regression" is regression with an intercept and one explanatory variable only, i.e.,*

$$\mathsf{y}_t = \alpha + \beta x_t + \varepsilon_t \tag{14.2.15}$$

Here $\boldsymbol{X} = \begin{bmatrix}\boldsymbol{\iota} & \boldsymbol{x}\end{bmatrix}$ and $\boldsymbol{\beta} = \begin{bmatrix}\alpha & \beta\end{bmatrix}^\top$. Evaluate (14.2.4) to get the following formulas for $\hat{\boldsymbol{\beta}} = \begin{bmatrix}\hat{\alpha} & \hat{\beta}\end{bmatrix}^\top$:

$$\hat{\alpha} = \frac{\sum x_t^2 \sum \mathsf{y}_t - \sum x_t \sum x_t\mathsf{y}_t}{n\sum x_t^2 - (\sum x_t)^2} \tag{14.2.16}$$

$$\hat{\beta} = \frac{n\sum x_t\mathsf{y}_t - \sum x_t \sum \mathsf{y}_t}{n\sum x_t^2 - (\sum x_t)^2} \tag{14.2.17}$$

Answer.

$$\boldsymbol{X}^\top\boldsymbol{X} = \begin{bmatrix}\boldsymbol{\iota}^\top\\ \boldsymbol{x}^\top\end{bmatrix}\begin{bmatrix}\boldsymbol{\iota} & \boldsymbol{x}\end{bmatrix} = \begin{bmatrix}\boldsymbol{\iota}^\top\boldsymbol{\iota} & \boldsymbol{\iota}^\top\boldsymbol{x}\\ \boldsymbol{x}^\top\boldsymbol{\iota} & \boldsymbol{x}^\top\boldsymbol{x}\end{bmatrix} = \begin{bmatrix}n & \sum x_t\\ \sum x_t & \sum x_t^2\end{bmatrix} \tag{14.2.18}$$

$$\boldsymbol{X}^\top\boldsymbol{X}^{-1} = \frac{1}{n\sum x_t^2 - (\sum x_t)^2}\begin{bmatrix}\sum x_t^2 & -\sum x_t\\ -\sum x_t & n\end{bmatrix} \tag{14.2.19}$$

$$\boldsymbol{X}^\top\mathsf{y} = \begin{bmatrix}\boldsymbol{\iota}^\top\mathsf{y}\\ \boldsymbol{x}^\top\mathsf{y}\end{bmatrix} = \begin{bmatrix}\sum \mathsf{y}_t\\ \sum x_i\mathsf{y}_t\end{bmatrix} \tag{14.2.20}$$

Therefore $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\mathsf{y}$ gives equations (14.2.16) and (14.2.17). □

Problem 199. *Show that*

$$\sum_{t=1}^n (x_t - \bar{x})(\mathsf{y}_t - \bar{\mathsf{y}}) = \sum_{t=1}^n x_t\mathsf{y}_t - n\bar{x}\bar{\mathsf{y}} \tag{14.2.21}$$

(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 200. *Show that* (14.2.17) *and* (14.2.16) *can also be written as follows:*

$$\hat{\beta} = \frac{\sum(x_t - \bar{x})(\mathsf{y}_t - \bar{\mathsf{y}})}{\sum(x_t - \bar{x})^2} \tag{14.2.22}$$

$$\hat{\alpha} = \bar{\mathsf{y}} - \hat{\beta}\bar{x} \tag{14.2.23}$$

Answer. Using $\sum x_i = n\bar{x}$ and $\sum \mathsf{y}_i = n\bar{\mathsf{y}}$ in (14.2.17), it can be written as

$$\hat{\beta} = \frac{\sum x_t\mathsf{y}_t - n\bar{x}\bar{\mathsf{y}}}{\sum x_t^2 - n\bar{x}^2} \tag{14.2.24}$$

Now apply Problem 199 to the numerator of (14.2.24), and Problem 199 with $\mathsf{y} = \boldsymbol{x}$ to the denominator, to get (14.2.22).

To prove equation (14.2.23) for $\hat{\alpha}$, let us work backwards and plug (14.2.24) into the righthand side of (14.2.23):

$$\bar{\mathsf{y}} - \bar{x}\hat{\beta} = \frac{\bar{\mathsf{y}}\sum x_t^2 - \bar{\mathsf{y}}n\bar{x}^2 - \bar{x}\sum x_t\mathsf{y}_t + n\bar{x}\bar{\mathsf{y}}}{\sum x_t^2 - n\bar{x}^2} \tag{14.2.25}$$

The second and the fourth term in the numerator cancel out, and what remains can be shown to be equal to (14.2.16). □

PROBLEM 201. *3 points Show that in the simple regression model, the fitted regression line can be written in the form*

$$\hat{y}_t = \bar{y} + \hat{\beta}(x_t - \bar{x}). \tag{14.2.26}$$

From this follows in particular that the fitted regression line always goes through the point $\bar{x}, \bar{y}$.

ANSWER. Follows immediately if one plugs (14.2.23) into the defining equation $\hat{y}_t = \hat{\alpha} + \hat{\beta}x_t$. □

Formulas (14.2.22) and (14.2.23) are interesting because they express the regression coefficients in terms of the sample means and covariances. Problem 202 derives the properties of the population equivalents of these formulas:

PROBLEM 202. *Given two random variables* x *and* y *with finite variances, and* $\operatorname{var}[x] > 0$. *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 20.1.1: If you want to predict* y *by an affine expression of the form* $a+bx$, *you will get the lowest mean squared error* MSE *with* $b = \operatorname{cov}[x, y]/\operatorname{var}[x]$ *and* $a = \operatorname{E}[y] - b\operatorname{E}[x]$.

ANSWER. The MSE is variance plus squared bias (see e.g. problem 165), therefore

$$\operatorname{MSE}[a + bx; y] = \operatorname{var}[a + bx - y] + (\operatorname{E}[a + bx - y])^2 = \operatorname{var}[bx - y] + (a - \operatorname{E}[y] + b\operatorname{E}[x])^2. \tag{14.2.27}$$

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

$$\operatorname{var}[bx - y] = b^2\operatorname{var}[x] - 2b\operatorname{cov}[x, y] + \operatorname{var}[y] \tag{14.2.28}$$

the first order condition is

$$2b\operatorname{var}[x] - 2\operatorname{cov}[x, y] = 0 \tag{14.2.29}$$

□

• b. *2 points For the first-order conditions you needed the partial derivatives* $\frac{\partial}{\partial a}\operatorname{E}[(y-a-bx)^2]$ *and* $\frac{\partial}{\partial b}\operatorname{E}[(y-a-bx)^2]$. *It is also possible, and probably shorter, to interchange taking expected value and partial derivative, i.e., to compute* $\operatorname{E}\left[\frac{\partial}{\partial a}(y - a - bx)^2\right]$ *and* $\operatorname{E}\left[\frac{\partial}{\partial b}(y - a - bx)^2\right]$ *and set those zero. Do the above proof in this alternative fashion.*

ANSWER. $\operatorname{E}\left[\frac{\partial}{\partial a}(y-a-bx)^2\right] = -2\operatorname{E}[y-a-bx] = -2(\operatorname{E}[y]-a-b\operatorname{E}[x])$. Setting this zero gives the formula for a. Now $\operatorname{E}\left[\frac{\partial}{\partial b}(y - a - bx)^2\right] = -2\operatorname{E}[x(y - a - bx)] = -2(\operatorname{E}[xy] - a\operatorname{E}[x] - b\operatorname{E}[x^2])$. Setting this zero gives $\operatorname{E}[xy] - a\operatorname{E}[x] - b\operatorname{E}[x^2] = 0$. Plug in formula for a and solve for b:

$$b = \frac{\operatorname{E}[xy] - \operatorname{E}[x]\operatorname{E}[y]}{\operatorname{E}[x^2] - (\operatorname{E}[x])^2} = \frac{\operatorname{cov}[x, y]}{\operatorname{var}[x]}. \tag{14.2.30}$$

□

• c. *2 points Compute the* MSE *of this predictor.*

ANSWER. If one plugs the optimal a into (14.2.27), this just annulls the last term of (14.2.27) so that the MSE is given by (14.2.28). If one plugs the optimal $b = \operatorname{cov}[x, y]/\operatorname{var}[x]$ into (14.2.28), one gets

$$\operatorname{MSE} = \left(\frac{\operatorname{cov}[x, y]}{\operatorname{var}[x]}\right)^2\operatorname{var}[x] - 2\frac{(\operatorname{cov}[x, y])}{\operatorname{var}[x]}\operatorname{cov}[x, y] + \operatorname{var}[x] \tag{14.2.31}$$

$$= \operatorname{var}[y] - \frac{(\operatorname{cov}[x, y])^2}{\operatorname{var}[x]}. \tag{14.2.32}$$

□

• d. *2 points Show that the prediction error is uncorrelated with the observed* x.

ANSWER.

$$\operatorname{cov}[x, y - a - bx] = \operatorname{cov}[x, y] - a\operatorname{cov}[x, x] = 0 \tag{14.2.33}$$

□

• e. *4 points If* $\operatorname{var}[x] = 0$, *the quotient* $\operatorname{cov}[x, y]/\operatorname{var}[x]$ *can no longer be formed, but if you replace the inverse by the g-inverse, so that the above formula becomes*

$$b = \operatorname{cov}[x, y](\operatorname{var}[x])^- \tag{14.2.34}$$

then it always gives the minimum MSE *predictor, whether or not* $\operatorname{var}[x] = 0$, *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* $\operatorname{var}[x] = 0$? *(b) what is the g-inverse of a nonzero scalar? (c) what is the g-inverse of the scalar number 0? (d) if* $\operatorname{var}[x] = 0$, *what do we know about* $\operatorname{cov}[x, y]$?

ANSWER. (a) If $\operatorname{var}[x] = 0$ then $x = \mu$ almost surely, therefore the observation of $\mathbf{x}$ 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 $\operatorname{var}[x] = 0$, then $\operatorname{cov}[x, y] = 0$.

Therefore pick a g-inverse 0, an arbitrary number will do, call it c. Then formula (14.2.34) says $b = 0 \cdot c = 0$. □

PROBLEM 203. *3 points Carefully state the specifications of the random variables involved in the linear regression model. How does the model in Problem 202 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 204. *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* $\bar{\mathbf{y}} = \bar{\boldsymbol{x}}^\top\hat{\boldsymbol{\beta}}$ *(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 205. *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, $\operatorname{cov}[\hat{\mathbf{y}}_i, \hat{\boldsymbol{\varepsilon}}_j] = 0$. *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 246. □

14.3. 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 (14.3.6) 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 161.

PROBLEM 206. *In the* `reggeom` *visualization, see Problem 293, 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:

$$(\boldsymbol{u}^\top\boldsymbol{v})^2 \le (\boldsymbol{u}^\top\boldsymbol{u})(\boldsymbol{v}^\top\boldsymbol{v}) \tag{14.3.1}$$

PROBLEM 207. *If* $\boldsymbol{Q}$ *is any nonnegative definite matrix, show that also*

$$(\boldsymbol{u}^\top\boldsymbol{Q}\boldsymbol{v})^2 \le (\boldsymbol{u}^\top\boldsymbol{Q}\boldsymbol{u})(\boldsymbol{v}^\top\boldsymbol{Q}\boldsymbol{v}). \tag{14.3.2}$$

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 (9.3.17), n times the sample covariance between $\tilde{\boldsymbol{y}}$ and $\boldsymbol{y}$ is

$$n \text{ times sample covariance}(\tilde{\boldsymbol{y}}, \boldsymbol{y}) = \tilde{\boldsymbol{y}}^\top\boldsymbol{D}\boldsymbol{y} = \boldsymbol{c}^\top\boldsymbol{X}^\top\boldsymbol{D}(\hat{\boldsymbol{y}} + \hat{\boldsymbol{\varepsilon}}). \tag{14.3.3}$$

By Problem 208, $\boldsymbol{D}\hat{\boldsymbol{\varepsilon}} = \hat{\boldsymbol{\varepsilon}}$, 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 (14.2.3)), therefore (14.3.3) becomes $\tilde{\boldsymbol{y}}^\top\boldsymbol{D}\boldsymbol{y} = \tilde{\boldsymbol{y}}^\top\boldsymbol{D}\hat{\boldsymbol{y}}$. Together with (14.3.2) this gives

$$\big(n \text{ times sample covariance}(\tilde{\boldsymbol{y}}, \boldsymbol{y})\big)^2 = (\tilde{\boldsymbol{y}}^\top\boldsymbol{D}\hat{\boldsymbol{y}})^2 \le (\tilde{\boldsymbol{y}}^\top\boldsymbol{D}\tilde{\boldsymbol{y}})(\hat{\boldsymbol{y}}^\top\boldsymbol{D}\hat{\boldsymbol{y}}) \tag{14.3.4}$$

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 $\boldsymbol{y}$:

(14.3.5)

$$\big(\text{sample correlation}(\tilde{\boldsymbol{y}}, \boldsymbol{y})\big)^2 = \frac{(\tilde{\boldsymbol{y}}^\top\boldsymbol{D}\boldsymbol{y})^2}{(\tilde{\boldsymbol{y}}^\top\boldsymbol{D}\tilde{\boldsymbol{y}})(\boldsymbol{y}^\top\boldsymbol{D}\boldsymbol{y})} \le \frac{\hat{\boldsymbol{y}}^\top\boldsymbol{D}\hat{\boldsymbol{y}}}{\boldsymbol{y}^\top\boldsymbol{D}\boldsymbol{y}} = \frac{\sum(\hat{y}_j - \bar{\hat{y}})^2}{\sum(y_j - \bar{y})^2} = \frac{\sum(\hat{y}_j - \bar{y})^2}{\sum(y_j - \bar{y})^2}.$$

For the rightmost equal sign in (14.3.5) we need Problem 209.

If $\tilde{\boldsymbol{y}} = \hat{\boldsymbol{y}}$, inequality (14.3.4) becomes an equality, and therefore also (14.3.5) 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

$$R^2 = \frac{\left(\sum(\hat{y}_j - \bar{\hat{y}})(y_j - \bar{y})\right)^2}{\sum(\hat{y}_j - \bar{\hat{y}})^2 \sum(y_j - \bar{y})^2} = \frac{\sum(\hat{y}_j - \bar{y})^2}{\sum(y_j - \bar{y})^2}. \tag{14.3.6}$$

PROBLEM 208. *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* (14.2.3).

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 209. *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 210. *Instead of* (14.3.6) *one often sees the formula*

$$\frac{\left(\sum(\hat{y}_j - \bar{y})(y_j - \bar{y})\right)^2}{\sum(\hat{y}_j - \bar{y})^2 \sum(y_j - \bar{y})^2} = \frac{\sum(\hat{y}_j - \bar{y})^2}{\sum(y_j - \bar{y})^2}. \tag{14.3.7}$$

Prove that they are equivalent. Which equation is better?

The denominator in the righthand side expression of (14.3.6), $\sum(y_j - \bar{y})^2$, 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 $\mathsf{SST} = \mathsf{SSR} + \mathsf{SSE}$.

PROBLEM 211. *In the* `reggeom` *visualization, again with $\boldsymbol{x}_1$ representing the vector of ones, show that $\mathsf{SST} = \mathsf{SSR} + \mathsf{SSE}$, 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 $\mathsf{y} - \boldsymbol{\iota}\bar{\mathsf{y}}$. Sincer $\boldsymbol{\iota}\bar{\mathsf{y}}$ is the projection of 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

$$\boldsymbol{y} = (\boldsymbol{y} - \hat{\boldsymbol{y}}) + (\hat{\boldsymbol{y}} - \boldsymbol{\iota}\bar{y}) + \boldsymbol{\iota}\bar{y} \tag{14.3.8}$$

is an orthogonal decomposition (all three vectors on the righthand side are orthogonal to each other), therefore in particular

$$(\boldsymbol{y} - \hat{\boldsymbol{y}})^\top(\hat{\boldsymbol{y}} - \boldsymbol{\iota}\bar{y}) = 0. \tag{14.3.9}$$

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 212. *Show the decomposition 14.3.8 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 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 (14.3.9) follows (now the same identity written in three different notations):

$$(\boldsymbol{y} - \boldsymbol{\iota}\bar{y})^\top(\boldsymbol{y} - \boldsymbol{\iota}\bar{y}) = (\boldsymbol{y} - \hat{\boldsymbol{y}})^\top(\boldsymbol{y} - \hat{\boldsymbol{y}}) + (\hat{\boldsymbol{y}} - \boldsymbol{\iota}\bar{y})^\top(\hat{\boldsymbol{y}} - \boldsymbol{\iota}\bar{y}) \tag{14.3.10}$$

$$\sum_t (y_t - \bar{y})^2 = \sum_t (y_t - \hat{y}_t)^2 + \sum_t (\hat{y}_t - \bar{y})^2 \tag{14.3.11}$$

$$\mathsf{SST} = \mathsf{SSE} + \mathsf{SSR} \tag{14.3.12}$$

PROBLEM 213. *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:*

$$(\boldsymbol{y} - \boldsymbol{\iota}\bar{y})^\top(\boldsymbol{y} - \boldsymbol{\iota}\bar{y}) = (\boldsymbol{y} - \hat{\boldsymbol{y}})^\top(\boldsymbol{y} - \hat{\boldsymbol{y}}) + (\hat{\boldsymbol{y}} - \boldsymbol{\iota}\bar{y})^\top(\hat{\boldsymbol{y}} - \boldsymbol{\iota}\bar{y}) \tag{14.3.13}$$

$$\sum_t (y_t - \bar{y})^2 = \sum_t (y_t - \hat{y}_t)^2 + \sum_t (\hat{y}_t - \bar{y})^2 \tag{14.3.14}$$

ANSWER. Start with

$$\mathsf{SST} = \sum (y_t - \bar{y})^2 = \sum (y_t - \hat{y}_t + \hat{y}_t - \bar{y})^2 \tag{14.3.15}$$

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 (14.3.12), together with (14.3.6), 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:

$$R^2 = \frac{\left(\sum(\hat{y}_j - \bar{\hat{y}})(y_j - \bar{y})\right)^2}{\sum(\hat{y}_j - \bar{\hat{y}})^2 \sum(y_j - \bar{y})^2} = \frac{\mathsf{SSR}}{\mathsf{SST}} = \frac{\mathsf{SST} - \mathsf{SSE}}{\mathsf{SST}} \tag{14.3.16}$$

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 (14.3.16): $\boldsymbol{\iota}\bar{y}$ 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 214. *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 216). □

PROBLEM 215. *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*

$$R^2 = \hat{\beta}^2 \frac{\sum(x_t - \bar{x})^2}{\sum(y_t - \bar{y})^2} \tag{14.3.17}$$

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

$$R^2 = \frac{\sum(\hat{y}_t - \bar{y})^2}{\sum(y_t - \bar{y})^2} = \hat{\beta}^2 \frac{\sum(x_t - \bar{x})^2}{\sum(y_t - \bar{y})^2} \tag{14.3.18}$$

□

• b. *2 points Furthermore show that R^2 is the sample correlation coefficient between $\boldsymbol{y}$ and $\boldsymbol{x}$, i.e.,*

$$R^2 = \frac{\Big(\sum(x_t - \bar{x})(y_t - \bar{y})\Big)^2}{\sum(x_t - \bar{x})^2 \sum(y_t - \bar{y})^2}. \tag{14.3.19}$$

Hint: you are allowed to use (14.2.22).

ANSWER.

$$R^2 = \hat{\beta}^2 \frac{\sum(x_t - \bar{x})^2}{\sum(y_t - \bar{y})^2} = \frac{\Big(\sum(x_t - \bar{x})(y_t - \bar{y})\Big)^2 \sum(x_t - \bar{x})^2}{\Big(\sum(x_t - \bar{x})^2\Big)^2 \sum(y_t - \bar{y})^2} \tag{14.3.20}$$

which simplifies to (14.3.19). □

• 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 (14.3.16) 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 (14.3.16) 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 (14.3.16) 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.

14.4. 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:

$$\bar{R}^2 = 1 - \frac{\mathsf{SSE}/(n-k)}{\mathsf{SST}/(n-1)} = 1 - (1 - R^2)\frac{n-1}{n-k}. \tag{14.4.1}$$

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

$$\mathsf{F} = \frac{(\mathsf{SSE}_k - \mathsf{SSE}_{k+i})/i}{\mathsf{SSE}_{k+i}/(n-k-i)} = \frac{n-k-i}{i}\Big(\frac{\mathsf{SSE}_k}{\mathsf{SSE}_{k+i}} - 1\Big) \tag{14.4.2}$$

$$= \frac{n-k-i}{i}\Big(\frac{(n-k)\mathsf{s}_k^2}{(n-k-i)\mathsf{s}_{k+i}^2} - 1\Big) \tag{14.4.3}$$

$$= \frac{(n-k)\mathsf{s}_k^2}{i\mathsf{s}_{k+i}^2} - \frac{n-k}{i} + 1 \tag{14.4.4}$$

$$= \frac{(n-k)}{i}\Big(\frac{\mathsf{s}_k^2}{\mathsf{s}_{k+i}^2} - 1\Big) + 1 \tag{14.4.5}$$

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. 143.

CHAPTER 15

Digression about Correlation Coefficients

15.1. 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

$$\rho_{\mathsf{xy}} = \frac{\operatorname{cov}[\mathsf{x},\mathsf{y}]}{\sqrt{\operatorname{var}[\mathsf{x}]}\sqrt{\operatorname{var}[\mathsf{y}]}}. \tag{15.1.1}$$

Because of Cauchy-Schwartz, its value lies between -1 and 1.

Problem 216. *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 $\rho_{12.\boldsymbol{\mathsf{x}}}$ 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 217 makes this point for the *simple* correlation coefficient:

Problem 217. *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 191 and 202. To set notation, call*

the minimum MSE *in the first prediction (Problem 191)* $\operatorname{MSE}[\textit{constant term};\mathsf{y}]$*, and the minimum* MSE *in the second prediction (Problem 202)* $\operatorname{MSE}[\textit{constant term and } \mathsf{x};\mathsf{y}]$*. Show that*

$$\frac{\operatorname{MSE}[\textit{constant term};\mathsf{y}] - \operatorname{MSE}[\textit{constant term and } \mathsf{x};\mathsf{y}]}{\operatorname{MSE}[\textit{constant term};\mathsf{y}]} = \frac{(\operatorname{cov}[\mathsf{y},\mathsf{x}])^2}{\operatorname{var}[\mathsf{y}]\operatorname{var}[\mathsf{x}]} = \rho^2_{\mathsf{yx}}. \tag{15.1.2}$$

Answer. The minimum MSE with only a constant is $\operatorname{var}[\mathsf{y}]$ and (14.2.32) says that $\operatorname{MSE}[\text{constant term and } \mathsf{x};\mathsf{y}] = \operatorname{var}[\mathsf{y}] - (\operatorname{cov}[\mathsf{x},\mathsf{y}])^2/\operatorname{var}[\mathsf{x}]$. 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

$$\begin{bmatrix}\boldsymbol{\mathsf{x}}\\ \mathsf{y}\end{bmatrix} \sim \begin{bmatrix}\boldsymbol{\mu}\\ \nu\end{bmatrix}, \sigma^2\begin{bmatrix}\boldsymbol{\Omega}_{\mathsf{xx}} & \boldsymbol{\omega}_{\mathsf{xy}}\\ \boldsymbol{\omega}^\top_{\mathsf{xy}} & \omega_{\mathsf{yy}}\end{bmatrix}. \tag{15.1.3}$$

By theorem **??**, the best linear predictor of y based on $\boldsymbol{\mathsf{x}}$ has the formula

$$\mathsf{y}^* = \nu + \boldsymbol{\omega}^\top_{\mathsf{xy}}\boldsymbol{\Omega}^-_{\mathsf{xx}}(\boldsymbol{\mathsf{x}} - \boldsymbol{\mu}) \tag{15.1.4}$$

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

$$\rho^2_{\mathsf{y}(\boldsymbol{\mathsf{x}})} = \frac{\boldsymbol{\omega}^\top_{\mathsf{xy}}\boldsymbol{\Omega}^-_{\mathsf{xx}}\boldsymbol{\omega}_{\mathsf{xy}}}{\omega_{\mathsf{yy}}} \tag{15.1.5}$$

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.,

$$\rho^2_{\mathsf{y}(\boldsymbol{\mathsf{x}})} = \frac{\operatorname{MSE}[\text{constant term};\mathsf{y}] - \operatorname{MSE}[\text{constant term and } \boldsymbol{\mathsf{x}};\mathsf{y}]}{\operatorname{MSE}[\text{constant term};\mathsf{y}]} \tag{15.1.6}$$

There are therefore two natural definitions of the multiple correlation coefficient. These two definitions correspond to the two formulas for R^2 in (14.3.6).

Partial Correlation Coefficients. Now assume $\boldsymbol{\mathsf{y}} = \begin{bmatrix}\mathsf{y}_1 & \mathsf{y}_2\end{bmatrix}^\top$ is a vector with two elements and write

$$\begin{bmatrix}\boldsymbol{\mathsf{x}}\\ \mathsf{y}_1\\ \mathsf{y}_2\end{bmatrix} \sim \begin{bmatrix}\boldsymbol{\mu}\\ \nu_1\\ \nu_2\end{bmatrix}, \sigma^2\begin{bmatrix}\boldsymbol{\Omega}_{\mathsf{xx}} & \boldsymbol{\omega}_{\mathsf{y}1} & \boldsymbol{\omega}_{\mathsf{y}2}\\ \boldsymbol{\omega}^\top_{\mathsf{y}1} & \omega_{11} & \omega_{12}\\ \boldsymbol{\omega}^\top_{\mathsf{y}2} & \omega_{21} & \omega_{22}\end{bmatrix}. \tag{15.1.7}$$

Let $\boldsymbol{\mathsf{y}}^*$ be the best linear predictor of $\boldsymbol{\mathsf{y}}$ based on $\boldsymbol{\mathsf{x}}$. The partial correlation coefficient $\rho_{12.\boldsymbol{\mathsf{x}}}$ 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 $\boldsymbol{\mathsf{x}}$. Assume for instance that both y_1 and

y_2 are highly correlated with $\boldsymbol{x}$. Then they will also have a high correlation with each other. Subtracting y_i^* from y_i eliminates this dependency on $\boldsymbol{x}$, 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 $\boldsymbol{x}$ as a predictor of y_1:

(15.1.8)
$$\rho^2_{12.\boldsymbol{x}} = \frac{\text{MSE[constant term and } \boldsymbol{x}; \mathsf{y}_2] - \text{MSE[constant term, } \boldsymbol{x}\text{, and } \mathsf{y}_1; \mathsf{y}_2]}{\text{MSE[constant term and } \boldsymbol{x}; \mathsf{y}_2]}.$$

Problem 218. *Using the definitions in terms of* MSE*'s, show that the following relationship holds between the squares of multiple and partial correlation coefficients:*

$$1 - \rho^2_{2(\boldsymbol{x},1)} = (1 - \rho^2_{21.\boldsymbol{x}})(1 - \rho^2_{2(\boldsymbol{x})}) \tag{15.1.9}$$

Answer. In terms of the MSE, (15.1.9) reads

(15.1.10)
$$\frac{\text{MSE[constant term, } \boldsymbol{x}\text{, and } \mathsf{y}_1; \mathsf{y}_2]}{\text{MSE[constant term; } \mathsf{y}_2]} = \frac{\text{MSE[constant term, } \boldsymbol{x}\text{, and } \mathsf{y}_1; \mathsf{y}_2]}{\text{MSE[constant term and } \boldsymbol{x}; \mathsf{y}_2]} \frac{\text{MSE[constant term and } \boldsymbol{x}; \mathsf{y}}{\text{MSE[constant term; } \mathsf{y}_2]}$$

□

From (15.1.9) follows the following weighted average formula:

$$\rho^2_{2(\boldsymbol{x},1)} = \rho^2_{2(\boldsymbol{x})} + (1 - \rho^2_{2(\boldsymbol{x})})\rho^2_{21.\boldsymbol{x}} \tag{15.1.11}$$

An alternative proof of (15.1.11) 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

(15.1.12)
$$\rho^2_{\mathsf{y}(\boldsymbol{x}).\boldsymbol{z}} = \frac{\text{MSE[constant term and } \boldsymbol{z}; \mathsf{y}] - \text{MSE[constant term, } \boldsymbol{z}\text{, and } \boldsymbol{x}; \mathsf{y}]}{\text{MSE[constant term and } \boldsymbol{z}; \mathsf{y}]}$$

$$= \frac{\boldsymbol{\omega}^\top_{\boldsymbol{x}\mathsf{y}.\boldsymbol{z}} \boldsymbol{\Omega}^-_{\boldsymbol{x}\boldsymbol{x}.\boldsymbol{z}} \boldsymbol{\omega}_{\boldsymbol{x}\mathsf{y}.\boldsymbol{z}}}{\omega_{\mathsf{y}\mathsf{y}.\boldsymbol{z}}} \tag{15.1.13}$$

CHAPTER 16

Specific Datasets

16.1. Cobb Douglas Aggregate Production Function

PROBLEM 219. *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*:*

$$\mathsf{q}_t = \mu \ell_t^{\beta} k_t^{\gamma} \exp(\varepsilon_t). \tag{16.1.1}$$

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*

$$\mathsf{x}_t = \alpha + \beta y_t + \gamma z_t + \varepsilon_t \tag{16.1.2}$$

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*

$$\mathsf{u}_t = \alpha + \gamma v_t + \varepsilon_t \tag{16.1.3}$$

is estimated. How are the regressions (16.1.2) *and* (16.1.3) *related to each other?*

ANSWER. Write (16.1.3) as

$$\mathsf{x}_t - y_t = \alpha + \gamma(z_t - y_t) + \varepsilon_t \tag{16.1.4}$$

and collect terms to get

$$\mathsf{x}_t = \alpha + (1-\gamma) y_t + \gamma z_t + \varepsilon_t \tag{16.1.5}$$

From this follows that running the regression (16.1.3) is equivalent to running the regression (16.1.2) 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 221 below will look for possible alternatives.

PROBLEM 220. *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* (16.1.2) *and* (16.1.3) *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=co` 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=`

□

• 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.*

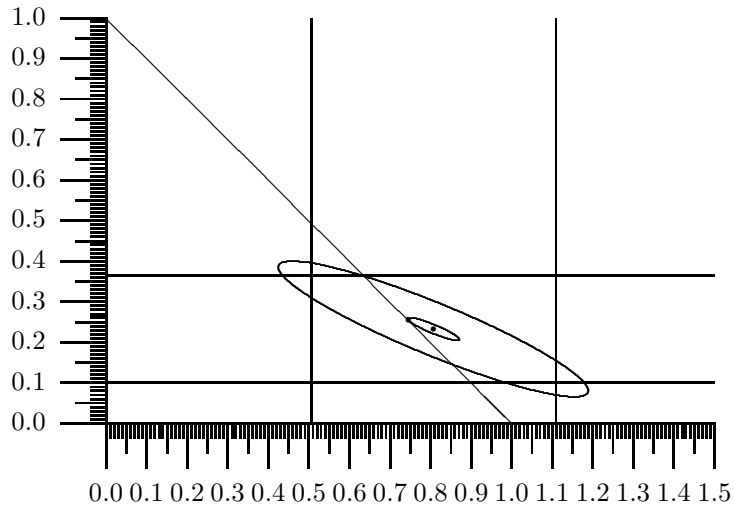


Figure 1. Coefficients of `capital` (vertical) and `labor` (horizontal), dependent variable `output`, unconstrained and constrained, 1899–1922

Answer. `lines(ellipse.lm(cbbfit, which=c(2, 3)))` □

Problem 221. *In this problem we will treat the Cobb-Douglas data as a dataset with errors in all three variables. See chapter* **??** *and problem* **??** *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*

$$\textbf{\textit{output}} = \hat{\alpha}_1 + \hat{\beta}_{12}\,\textbf{\textit{labor}} + \hat{\beta}_{13}\,\textbf{\textit{capital}} + \mathit{residual}_1 \tag{16.1.6}$$

$$\textbf{\textit{labor}} = \hat{\alpha}_2 + \hat{\beta}_{21}\,\textbf{\textit{output}} + \hat{\beta}_{23}\,\textbf{\textit{capital}} + \mathit{residual}_2 \tag{16.1.7}$$

$$\textbf{\textit{capital}} = \hat{\alpha}_3 + \hat{\beta}_{31}\,\textbf{\textit{output}} + \hat{\beta}_{32}\,\textbf{\textit{labor}} + \mathit{residual}_3. \tag{16.1.8}$$

In order to compare the slope parameters in these regressions, first rearrange them in the form

$$-\textbf{\textit{output}} + \hat{\beta}_{12}\,\textbf{\textit{labor}} + \hat{\beta}_{13}\,\textbf{\textit{capital}} + \hat{\alpha}_1 + \mathit{residual}_1 = 0 \tag{16.1.9}$$

$$\hat{\beta}_{21}\,\textbf{\textit{output}} - \textbf{\textit{labor}} + \hat{\beta}_{23}\,\textbf{\textit{capital}} + \hat{\alpha}_2 + \mathit{residual}_2 = 0 \tag{16.1.10}$$

$$\hat{\beta}_{31}\,\textbf{\textit{output}} + \hat{\beta}_{32}\,\textbf{\textit{labor}} - \textbf{\textit{capital}} + \hat{\alpha}_3 + \mathit{residual}_3 = 0 \tag{16.1.11}$$

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.

	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 222. *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 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*

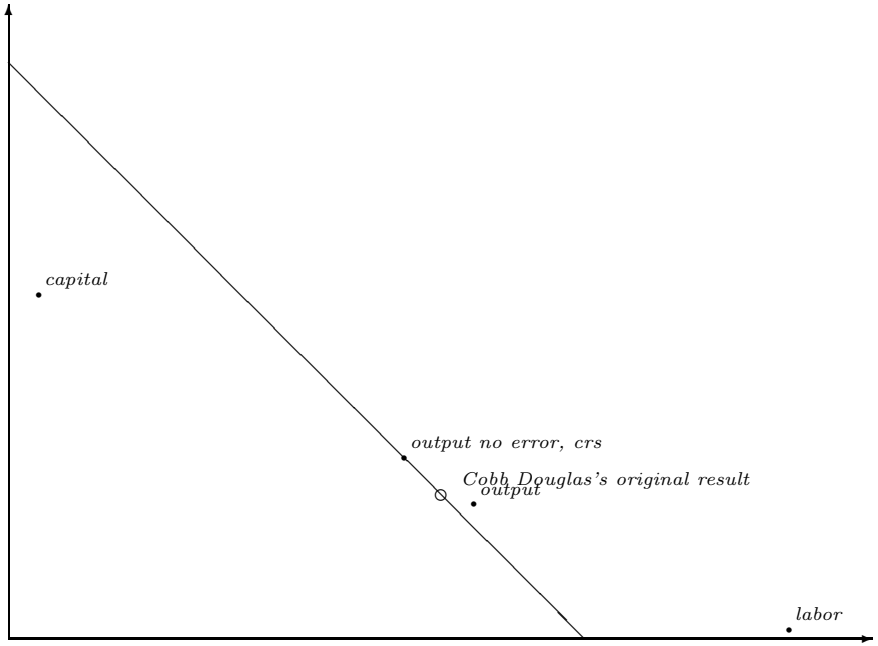


FIGURE 2. Coefficients of `capital` (vertical) and `labor` (horizontal), dependent variable `output`, 1899–1922

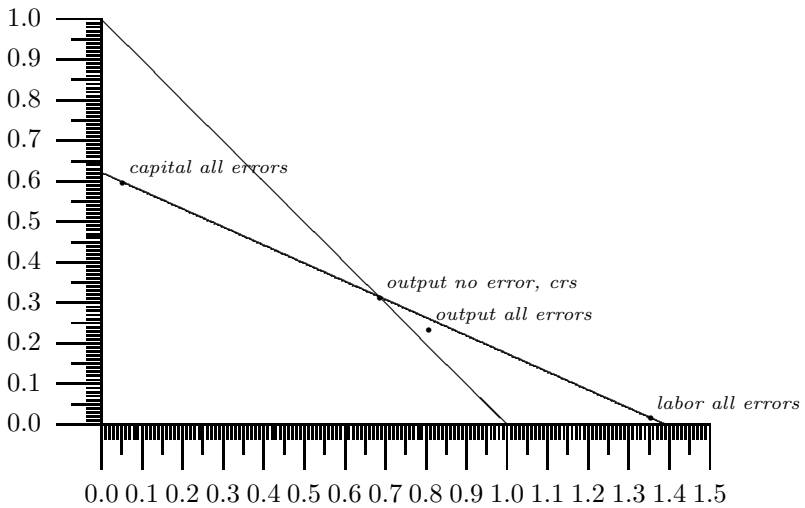


FIGURE 3. Coefficient of `capital` (vertical) and `labor` (horizontal) in the elementary regressions, dependent variable `output`, 1899–1922

$\mathtt{x}$, $\mathtt{y}$, *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* (**??**).

ANSWER.

$$\begin{bmatrix}\mathtt{x}^* & \mathtt{y}^* & \mathtt{z}^*\end{bmatrix}\begin{bmatrix}-1\\ \beta\\ 1-\beta\end{bmatrix}=0 \qquad \text{or} \qquad \begin{aligned}\mathtt{x}^*&=\beta\mathtt{y}^*+(1-\beta)\mathtt{z}^*\\ \mathtt{x}&=\mathtt{x}^*+\mathtt{u}\\ \mathtt{y}&=\mathtt{y}^*+\mathtt{v}\\ \mathtt{z}&=\mathtt{z}^*+\mathtt{w}.\end{aligned} \tag{16.1.12}$$

$$\begin{bmatrix}\mathtt{x} & \mathtt{y} & \mathtt{z}\end{bmatrix}=\begin{bmatrix}\mathtt{x}^* & \mathtt{y}^* & \mathtt{z}^*\end{bmatrix}+\begin{bmatrix}\mathtt{u} & \mathtt{v} & \mathtt{w}\end{bmatrix}$$

□

• 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:

$$\begin{bmatrix}\sigma^2_{\mathtt{x}} & \sigma_{\mathtt{xy}} & \sigma_{\mathtt{xz}}\\ \sigma_{\mathtt{xy}} & \sigma^2_{\mathtt{y}} & \sigma_{\mathtt{yz}}\\ \sigma_{\mathtt{xz}} & \sigma_{\mathtt{yz}} & \sigma^2_{\mathtt{z}}\end{bmatrix}= \tag{16.1.13}$$

$$=\begin{bmatrix}\beta^2\sigma^2_{\mathtt{y}^*}+2\beta(1-\beta)\sigma_{\mathtt{y}^*\mathtt{z}^*}+(1-\beta)^2\sigma^2_{\mathtt{z}^*} & \beta\sigma^2_{\mathtt{y}^*}+(1-\beta)\sigma_{\mathtt{y}^*\mathtt{z}^*} & \beta\sigma_{\mathtt{y}^*\mathtt{z}^*}+(1-\beta)\sigma^2_{\mathtt{z}^*}\\ \beta\sigma^2_{\mathtt{y}^*}+(1-\beta)\sigma_{\mathtt{y}^*\mathtt{z}^*} & \sigma^2_{\mathtt{y}^*} & \sigma_{\mathtt{y}^*\mathtt{z}^*}\\ \beta\sigma_{\mathtt{y}^*\mathtt{z}^*}+(1-\beta)\sigma^2_{\mathtt{z}^*} & \sigma^2_{\mathtt{y}^*} & \sigma^2_{\mathtt{z}^*}\end{bmatrix}+\begin{bmatrix}\sigma^2_{\mathtt{u}} & 0\\ 0 & \sigma^2_{\mathtt{v}}\\ 0 & 0\end{bmatrix} \tag{16.1.14}$$

□

• c. *Show that the unknown parameters are not yet identified. However, if one makes the additional assumption that one of the three error variances* $\sigma^2_{\mathtt{u}}$, $\sigma^2_{\mathtt{v}}$, *or* $\sigma^2_{\mathtt{w}}$ *is zero, then the equations are identified. Since the quantity of output presumably has less error than the other two variables, assume* $\sigma^2_{\mathtt{u}}=0$. *Under this assumption, show that*

$$\beta=\frac{\sigma^2_{\mathtt{x}}-\sigma_{\mathtt{xz}}}{\sigma_{\mathtt{xy}}-\sigma_{\mathtt{xz}}} \tag{16.1.15}$$

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 (16.1.14) one gets from the $\mathtt{yz}$ element of the covariance matrix

$$\sigma_{\mathtt{y}^*\mathtt{z}^*}=\sigma_{\mathtt{yz}} \tag{16.1.16}$$

and from the $\mathtt{xz}$ element

$$\sigma^2_{\mathtt{z}^*}=\frac{\sigma_{\mathtt{xz}}-\beta\sigma_{\mathtt{yz}}}{1-\beta} \tag{16.1.17}$$

Similarly, one gets from the $\mathtt{xy}$ element:

$$\sigma^2_{\mathtt{y}^*}=\frac{\sigma_{\mathtt{xy}}-(1-\beta)\sigma_{\mathtt{yz}}}{\beta} \tag{16.1.18}$$

Now plug (16.1.16), (16.1.17), and (16.1.18) into the equation for the $\mathtt{xx}$ element:

$$\sigma^2_{\mathtt{x}}=\beta(\sigma_{\mathtt{xy}}-(1-\beta)\sigma_{\mathtt{yz}})+2\beta(1-\beta)\sigma_{\mathtt{yz}}+(1-\beta)(\sigma_{\mathtt{xz}}-\beta\sigma_{\mathtt{yz}})+\sigma^2_{\mathtt{u}} \tag{16.1.19}$$

$$=\beta\sigma_{\mathtt{xy}}+(1-\beta)\sigma_{\mathtt{xz}}+\sigma^2_{\mathtt{u}} \tag{16.1.20}$$

Since we are assuming $\sigma_{\mathsf{u}}^2 = 0$ this last equation can be solved for β:

$$\beta = \frac{\sigma_{\mathsf{x}}^2 - \sigma_{\mathsf{xz}}}{\sigma_{\mathsf{xy}} - \sigma_{\mathsf{xz}}} \tag{16.1.21}$$

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

$$\begin{bmatrix}\mathtt{LOGOUT}\\ \mathtt{LOGLAB}\\ \mathtt{LOGCAP}\end{bmatrix} \sim \left(\begin{bmatrix}5.07734\\ 4.96272\\ 5.35648\end{bmatrix}, \begin{bmatrix}0.0724870714 & 0.0522115563 & 0.1169330807\\ 0.0522115563 & 0.0404318579 & 0.0839798588\\ 0.1169330807 & 0.0839798588 & 0.2108441826\end{bmatrix}\right) \tag{16.1.22}$$

Therefore equation (16.1.15) gives

$$\hat\beta = \frac{0.0724870714 - 0.1169330807}{0.0522115563 - 0.1169330807} = 0.686726861149148 \tag{16.1.23}$$

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.,* $\mathsf{y}^* = \delta_1 + \gamma_1\mathsf{x}^*$ *and* $\mathsf{z}^* = \delta_2 + \gamma_2\mathsf{x}^*$. *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

$$\begin{bmatrix}\sigma_{\mathsf{x}}^2 & \sigma_{\mathsf{xy}} & \sigma_{\mathsf{xz}}\\ \sigma_{\mathsf{xy}} & \sigma_{\mathsf{y}}^2 & \sigma_{\mathsf{yz}}\\ \sigma_{\mathsf{xz}} & \sigma_{\mathsf{yz}} & \sigma_{\mathsf{z}}^2\end{bmatrix} = \sigma_{\mathsf{x}^*}^2 \begin{bmatrix}1 & \gamma_1 & \gamma_2\\ \gamma_1 & \gamma_1^2 & \gamma_1\gamma_2\\ \gamma_2 & \gamma_1\gamma_2 & \gamma_2^2\end{bmatrix} + \begin{bmatrix}\sigma_{\mathsf{u}}^2 & 0 & 0\\ 0 & \sigma_{\mathsf{v}}^2 & 0\\ 0 & 0 & \sigma_{\mathsf{w}}^2\end{bmatrix}. \tag{16.1.24}$$

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, $\sigma_{\mathsf{x}^*}^2$ by dividing the 21-element by γ_1, etc.

(16.1.25)

$$\gamma_1 = \frac{\sigma_{\mathsf{yz}}}{\sigma_{\mathsf{xy}}} = \frac{0.0839798588}{0.1169330807} = 0.7181873452513939 \qquad \sigma_{\mathsf{u}}^2 = \sigma_{\mathsf{x}}^2 - \frac{\sigma_{\mathsf{yx}}\sigma_{\mathsf{xz}}}{\sigma_{\mathsf{yz}}} = 0.072487071$$

$$\gamma_2 = \frac{\sigma_{\mathsf{yz}}}{\sigma_{\mathsf{xz}}} = \frac{0.0839798588}{0.0522115563} = 1.608453467992104 \qquad \sigma_{\mathsf{v}}^2 = \sigma_{\mathsf{y}}^2 - \frac{\sigma_{\mathsf{xy}}\sigma_{\mathsf{yz}}}{\sigma_{\mathsf{xz}}}$$

$$\sigma_{\mathsf{x}^*}^2 = \frac{\sigma_{\mathsf{yx}}\sigma_{\mathsf{xz}}}{\sigma_{\mathsf{yz}}} = \frac{0.0522115563 \cdot 0.1169330807}{0.0839798588} = 0.0726990758 \qquad \sigma_{\mathsf{w}}^2 = \sigma_{\mathsf{z}}^2 - \frac{\sigma_{\mathsf{xz}}\sigma_{\mathsf{zy}}}{\sigma_{\mathsf{xy}}}$$

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. □

16.2. 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 223. *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 224. *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 225. *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 226. *2 points Make a plot of the (empirical) density function of `mc6` and `gas6` and interpret the results.*

Problem 227. *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 228. *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 229. *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 230. *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 231. *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 232. *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:/Docu 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 233. *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 234. *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 235. *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* $\hat{\beta}(i)$ *when the* i*th observation is dropped, a vector with all the* $s(i)$, *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.g 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 236. *Which commands do you need to plot the predictive residuals?*

PROBLEM 237. *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 238. *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?*

16.3. Long Term Data about US Economy

The dataset `uslt` is described in [**DL91**]. Home page of the authors is `www.ceprema` `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`×`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(1869:1 cbind(kn2,kns2), type="l")`. If you want the y-axis logarithmic, say `matplot(1869: cbind(gnp/gnp2,kns/kns2,kne/kne2), type="l", log="y")`.

PROBLEM 239. *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**].

16.4. 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.

16.5. 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 sur
> ##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/ecme
> #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 distr
> #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
> 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
> # 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 exper
> ##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 ord
> #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
> ###Exercise 4a (5 points):
> summary(lm(lnwage ~ union + gender + race + ed + ex + I(ex^2), data=cps7

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*e

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=cps7

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, sub

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, subse

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 instanc
> #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 17

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 $\text{MSE}[\hat{\phi};\phi]$ (read it: mean squared error of $\hat{\phi}$ as an estimator or predictor, whatever the case may be, of ϕ):

$$\text{MSE}[\hat{\phi};\phi] = \text{E}[(\hat{\phi}-\phi)^2] \tag{17.0.1}$$

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 $\text{MSE}[\hat{\phi};\phi] \leq \text{MSE}[\tilde{\phi};\phi]$. 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).

17.1. 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 240. *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 $\text{MSE}[\hat{\phi}_i;\phi_i] < \text{MSE}[\tilde{\phi}_i;\phi_i]$ for $i = 1,2$ but $\text{MSE}[\hat{\phi}_1+\hat{\phi}_2;\phi_1+\phi_2] > \text{MSE}[\tilde{\phi}_1+\tilde{\phi}_2;\phi_1+\phi_2]$. 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 17.1.1 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 17.1.1 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

$$\mathcal{MSE}[\hat{\boldsymbol{\phi}};\boldsymbol{\phi}] = \mathcal{E}[(\hat{\boldsymbol{\phi}}-\boldsymbol{\phi})(\hat{\boldsymbol{\phi}}-\boldsymbol{\phi})^\top]\,. \tag{17.1.1}$$

Problem 241. *2 points Let $\boldsymbol{\theta}$ be a vector of possibly random parameters, and $\hat{\boldsymbol{\theta}}$ an estimator of $\boldsymbol{\theta}$. Show that*

$$\mathcal{MSE}[\hat{\boldsymbol{\theta}};\boldsymbol{\theta}] = \mathcal{V}[\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}] + (\mathcal{E}[\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}])(\mathcal{E}[\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}])^\top. \tag{17.1.2}$$

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

$$\begin{aligned}\mathcal{E}[\boldsymbol{x}\boldsymbol{x}^\top] &= \mathcal{E}\big[(\boldsymbol{x}-\mathcal{E}[\boldsymbol{x}]+\mathcal{E}[\boldsymbol{x}])(\boldsymbol{x}-\mathcal{E}[\boldsymbol{x}]+\mathcal{E}[\boldsymbol{x}])^\top\big]\\ &= \mathcal{E}\big[(\boldsymbol{x}-\mathcal{E}[\boldsymbol{x}])(\boldsymbol{x}-\mathcal{E}[\boldsymbol{x}])^\top\big] - \mathcal{E}\big[(\boldsymbol{x}-\mathcal{E}[\boldsymbol{x}])\,\mathcal{E}[\boldsymbol{x}]^\top\big] - \mathcal{E}\big[\mathcal{E}[\boldsymbol{x}](\boldsymbol{x}-\mathcal{E}[\boldsymbol{x}])^\top\big] + \mathcal{E}\big[\mathcal{E}[\boldsymbol{x}]\,\mathcal{E}[\boldsymbol{x}]^\top\big]\\ &= \mathcal{V}[\boldsymbol{x}] - \boldsymbol{O} - \boldsymbol{O} + \mathcal{E}[\boldsymbol{x}]\,\mathcal{E}[\boldsymbol{x}]^\top.\end{aligned}$$

Setting $\boldsymbol{x} = \hat{\boldsymbol{\theta}}-\boldsymbol{\theta}$ the statement follows. □

If $\boldsymbol{\theta}$ is nonrandom, formula (17.1.2) 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 17.1.1. *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* (17.1.3), (17.1.4), *and* (17.1.5) *are equivalent:*

$$\textit{For every constant vector } \boldsymbol{t}, \quad \text{MSE}[\boldsymbol{t}^\top\hat{\boldsymbol{\phi}};\boldsymbol{t}^\top\boldsymbol{\phi}] \leq \text{MSE}[\boldsymbol{t}^\top\tilde{\boldsymbol{\phi}};\boldsymbol{t}^\top\boldsymbol{\phi}] \tag{17.1.3}$$

$$\mathcal{MSE}[\tilde{\boldsymbol{\phi}};\boldsymbol{\phi}] - \mathcal{MSE}[\hat{\boldsymbol{\phi}};\boldsymbol{\phi}] \quad \textit{is a nonnegative definite matrix} \tag{17.1.4}$$

$$\textit{For every nnd } \boldsymbol{\Theta}, \quad \text{E}\big[(\hat{\boldsymbol{\phi}}-\boldsymbol{\phi})^\top\boldsymbol{\Theta}(\hat{\boldsymbol{\phi}}-\boldsymbol{\phi})\big] \leq \text{E}\big[(\tilde{\boldsymbol{\phi}}-\boldsymbol{\phi})^\top\boldsymbol{\Theta}(\tilde{\boldsymbol{\phi}}-\boldsymbol{\phi})\big]. \tag{17.1.5}$$

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 (17.1.3) implies (17.1.4), simply note that $\text{MSE}[\boldsymbol{t}^\top\hat{\boldsymbol{\phi}};\boldsymbol{t}^\top\boldsymbol{\phi}] = \sigma^2\boldsymbol{t}^\top\boldsymbol{\Omega}\boldsymbol{t}$ and likewise $\text{MSE}[\boldsymbol{t}^\top\tilde{\boldsymbol{\phi}};\boldsymbol{t}^\top\boldsymbol{\phi}]$ $\sigma^2\boldsymbol{t}^\top\boldsymbol{\Xi}\boldsymbol{t}$. Therefore (17.1.3) 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 (17.1.4) implies (17.1.5):

$$\mathrm{E}[(\hat{\boldsymbol{\phi}}-\boldsymbol{\phi})^\top\boldsymbol{\Theta}(\hat{\boldsymbol{\phi}}-\boldsymbol{\phi})] = \mathrm{E}[\mathrm{tr}\big((\hat{\boldsymbol{\phi}}-\boldsymbol{\phi})^\top\boldsymbol{\Theta}(\hat{\boldsymbol{\phi}}-\boldsymbol{\phi})\big)] =$$
$$= \mathrm{E}[\mathrm{tr}\big(\boldsymbol{\Theta}(\hat{\boldsymbol{\phi}}-\boldsymbol{\phi})(\hat{\boldsymbol{\phi}}-\boldsymbol{\phi})^\top\big)] = \mathrm{tr}\big(\boldsymbol{\Theta}\,\mathcal{E}[(\hat{\boldsymbol{\phi}}-\boldsymbol{\phi})(\hat{\boldsymbol{\phi}}-\boldsymbol{\phi})^\top]\big) = \sigma^2\,\mathrm{tr}\big(\boldsymbol{\Theta}\boldsymbol{\Omega}\big)$$

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 242 below, this trace is nonnegative.

To complete the proof, (17.1.5) has (17.1.3) as a special case if one sets $\boldsymbol{\Theta} = \boldsymbol{t}\boldsymbol{t}^\top$. □

PROBLEM 242. *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 118 (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 243. *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 $\hat{\boldsymbol{\phi}} = \begin{bmatrix}11\\11\end{bmatrix}$, and the second is $\tilde{\boldsymbol{\phi}} = \begin{bmatrix}12\\8\end{bmatrix}$.*

• a. *2 points Compute the $\mathcal{MSE}$-matrices of these two estimators if the true value of the parameter vector is $\boldsymbol{\phi} = \begin{bmatrix}10\\10\end{bmatrix}$. 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 $\boldsymbol{g} = \begin{bmatrix}g_1\\g_2\end{bmatrix}$ and $\boldsymbol{h} = \begin{bmatrix}h_1\\h_2\end{bmatrix}$ 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 $\boldsymbol{g} = \begin{bmatrix}1\\-1\end{bmatrix}$ and $\boldsymbol{h} = \begin{bmatrix}1\\1\end{bmatrix}$ 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.

$$\mathcal{MSE}[\tilde{\boldsymbol{\phi}};\boldsymbol{\phi}] - \mathcal{MSE}[\hat{\boldsymbol{\phi}};\boldsymbol{\phi}] = \begin{bmatrix}3 & -5\\-5 & 3\end{bmatrix} \tag{17.1.6}$$

Its determinant is negative, and the determinant of its negative is also negative. □

CHAPTER 18

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 (14.1.3), 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:

$$\begin{aligned}\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\mathbf{y} - (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta} = \\ &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}) = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon}\end{aligned} \tag{18.0.7}$$

From (18.0.7) 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 17 gave some reasons why this is a sensible criterion (compare [**DM93**, Chapter 5.5]).

18.1. The Gauss Markov Theorem

Returning to the least squares estimator $\hat{\boldsymbol{\beta}}$, one obtains, using (18.0.7), that

$$\begin{aligned}\mathcal{MSE}[\hat{\boldsymbol{\beta}};\boldsymbol{\beta}] = \mathcal{E}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top] &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\,\mathcal{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top]\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1} = \\ &= \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}.\end{aligned} \tag{18.1.1}$$

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 244. *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 245. *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 246. *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 247. *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 $\bar{\mathbf{y}}$ and $\hat{\boldsymbol{\beta}}$, which is defined as the matrix (here consisting of one row only) that contains all the covariances*

$$\mathcal{C}[\bar{\mathbf{y}}, \hat{\boldsymbol{\beta}}] \equiv \begin{bmatrix}\operatorname{cov}[\bar{\mathbf{y}}, \hat{\beta}_1] & \operatorname{cov}[\bar{\mathbf{y}}, \hat{\beta}_2] & \cdots & \operatorname{cov}[\bar{\mathbf{y}}, \hat{\beta}_k]\end{bmatrix} \tag{18.1.2}$$

has the following form: $\mathcal{C}[\bar{\mathbf{y}}, \hat{\boldsymbol{\beta}}] = \frac{\sigma^2}{n}\begin{bmatrix}1 & 0 & \cdots & 0\end{bmatrix}$ 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 $\bar{\mathbf{y}}$ and $\hat{\boldsymbol{\beta}}$ in terms of $\mathbf{y}$, i.e., $\bar{\mathbf{y}} = \frac{1}{n}\boldsymbol{\iota}^\top\mathbf{y}$ and $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\mathbf{y}$. Therefore

$$\mathcal{C}[\bar{\mathbf{y}}, \hat{\boldsymbol{\beta}}] = \frac{1}{n}\boldsymbol{\iota}^\top\,\mathcal{V}[\mathbf{y}]\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1} = \frac{\sigma^2}{n}\boldsymbol{\iota}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1} = \frac{\sigma^2}{n}\boldsymbol{e}^{(1)\top}\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1} = \frac{\sigma^2}{n}\boldsymbol{e}^{(1)\top}. \tag{18.1.3}$$

□

THEOREM 18.1.1. 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* $\tilde{\phi} = \boldsymbol{a}^\top\mathbf{y}$ *of* $\phi = \boldsymbol{t}^\top\boldsymbol{\beta}$ *has a bigger* MSE *than* $\boldsymbol{t}^\top\hat{\boldsymbol{\beta}}$.

PROOF. Write the alternative linear estimator $\tilde{\phi} = \boldsymbol{a}^\top\mathbf{y}$ in the form

$$\tilde{\phi} = \big(\boldsymbol{t}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top + \boldsymbol{c}^\top\big)\mathbf{y} \tag{18.1.4}$$

then the sampling error is

$$\begin{aligned}\tilde{\phi} - \phi &= \big(\boldsymbol{t}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top + \boldsymbol{c}^\top\big)(\boldsymbol{X}\beta + \boldsymbol{\varepsilon}) - \boldsymbol{t}^\top\beta \\ &= \big(\boldsymbol{t}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top + \boldsymbol{c}^\top\big)\boldsymbol{\varepsilon} + \boldsymbol{c}^\top\boldsymbol{X}\beta. \qquad (18.1.5)\end{aligned}$$

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

$$\begin{aligned}\mathrm{MSE}[\tilde{\phi};\phi] &= \mathrm{E}[(\tilde{\phi}-\phi)^2] = \mathrm{E}[\big(\boldsymbol{t}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top + \boldsymbol{c}^\top\big)\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\big(\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{t} + \boldsymbol{c}\big)] = \\ &= \sigma^2\big(\boldsymbol{t}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top + \boldsymbol{c}^\top\big)\big(\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{t} + \boldsymbol{c}\big) = \sigma^2\boldsymbol{t}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{t} + \sigma^2\boldsymbol{c}^\top\boldsymbol{c},\end{aligned}$$

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 248. *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 $\tilde{\boldsymbol{\beta}} = \big((\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top + \boldsymbol{C}\big)\mathbf{y}$. 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. □

18.2. Digression about Minimax Estimators

Theorem 18.1.1 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 18.2.1. $\hat{\phi}$ 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 $\boldsymbol{\beta}_1$

$$\mathrm{MSE}[\tilde{\phi};\phi|\boldsymbol{\beta} = \boldsymbol{\beta}_0] \geq \mathrm{MSE}[\hat{\phi};\phi|\boldsymbol{\beta} = \boldsymbol{\beta}_1] \tag{18.2.1}$$

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 18.2.2. $\hat{\boldsymbol{\beta}}$ *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* $\tilde{\phi} = \boldsymbol{a}^\top\mathbf{y}$ *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 18.1.1, write the alternative linear estimator $\tilde{\phi}$ in the form $\tilde{\phi} = \big(\boldsymbol{t}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top + \boldsymbol{c}^\top\big)\mathbf{y}$, so that the sampling error is given by (18.1.5). Then it follows

$$\mathrm{MSE}[\tilde{\phi};\phi] = \mathrm{E}[(\tilde{\phi}-\phi)^2] = \mathrm{E}[\Big(\big(\boldsymbol{t}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top + \boldsymbol{c}^\top\big)\boldsymbol{\varepsilon} + \boldsymbol{c}^\top\boldsymbol{X}\beta\Big)\Big(\boldsymbol{\varepsilon}^\top\big(\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{t} + \boldsymbol{c}\big) + \tag{18.2.2}$$

$$= \sigma^2\big(\boldsymbol{t}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top + \boldsymbol{c}^\top\big)\big(\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{t} + \boldsymbol{c}\big) + \boldsymbol{c}^\top\boldsymbol{X}\boldsymbol{\beta}\boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{c} \tag{18.2.3}$$

Now there are two cases: if $\boldsymbol{c}^\top\boldsymbol{X} = \boldsymbol{o}^\top$, then $\mathrm{MSE}[\tilde{\phi};\phi] = \sigma^2\boldsymbol{t}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{t} + \sigma^2\boldsymbol{c}^\top\boldsymbol{c}$. 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 $\mathrm{MSE}[\tilde{\phi};\phi]$ is unbounded, i.e., for any finite number ω one one can always find a $\boldsymbol{\beta}_0$ for which $\mathrm{MSE}[\tilde{\phi};\phi] > \omega$. Since $\mathrm{MSE}[\hat{\phi};\phi]$ is bounded, a $\boldsymbol{\beta}_0$ can be found that satisfies (18.2.1).

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 $\boldsymbol{\beta}$.

18.3. Miscellaneous Properties of the BLUE

PROBLEM 249.

• a. *1 point Instead of* (14.2.22) *one sometimes sees the formula*

$$\hat{\beta} = \frac{\sum(x_t - \bar{x})\mathrm{y}_t}{\sum(x_t - \bar{x})^2}. \tag{18.3.1}$$

for the slope parameter in the simple regression. Show that these formulas are mathematically equivalent.

Answer. Equivalence of (18.3.1) and (14.2.22) follows from $\sum(x_t - \bar{x}) = 0$ and therefore also $\bar{y}\sum(x_t - \bar{x}) = 0$. Alternative proof, using matrix notation and the matrix $\boldsymbol{D}$ defined in Problem 161: (14.2.22) is $\frac{\boldsymbol{x}^\top \boldsymbol{D}^\top \boldsymbol{D}\boldsymbol{y}}{\boldsymbol{x}^\top \boldsymbol{D}^\top \boldsymbol{D}\boldsymbol{x}}$ and (18.3.1) is $\frac{\boldsymbol{x}^\top \boldsymbol{D}\boldsymbol{y}}{\boldsymbol{x}^\top \boldsymbol{D}^\top \boldsymbol{D}\boldsymbol{x}}$. They are equal because $\boldsymbol{D}$ is symmetric and idempotent.

□

• b. *1 point Show that*

$$\text{var}[\hat{\beta}] = \frac{\sigma^2}{\sum(x_i - \bar{x})^2} \tag{18.3.2}$$

Answer. Write (18.3.1) as

$$\hat{\beta} = \frac{1}{\sum(x_t - \bar{x})^2}\sum(x_t - \bar{x})y_t \quad \Rightarrow \quad \text{var}[\hat{\beta}] = \frac{1}{\left(\sum(x_t - \bar{x})^2\right)^2}\sum(x_t - \bar{x})^2\sigma^2 \tag{18.3.3}$$

□

• c. *2 points Show that* $\text{cov}[\hat{\beta}, \bar{y}] = 0$.

Answer. This is a special case of problem 247, but it can be easily shown here separately:

$$\text{cov}[\hat{\beta}, \bar{y}] = \text{cov}\left[\frac{\sum_s(x_s - \bar{x})y_s}{\sum_t(x_t - \bar{x})^2}, \frac{1}{n}\sum_j y_j\right] = \frac{1}{n\sum_t(x_t - \bar{x})^2}\text{cov}\left[\sum_s(x_s - \bar{x})y_s, \sum_j y_j\right] =$$
$$= \frac{1}{n\sum_t(x_t - \bar{x})^2}\sum_s(x_s - \bar{x})\sigma^2 = 0.$$

□

• d. *2 points Using* (14.2.23) *show that*

$$\text{var}[\hat{\alpha}] = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}\right) \tag{18.3.4}$$

Problem 250. *You have two data vectors x_i and y_i $(i = 1, \ldots, n)$, and the true model is*

$$y_i = \beta x_i + \varepsilon_i \tag{18.3.5}$$

where x_i and ε_i satisfy the basic assumptions of the linear regression model. The least squares estimator for this model is

$$\tilde{\beta} = (\boldsymbol{x}^\top\boldsymbol{x})^{-1}\boldsymbol{x}^\top\boldsymbol{y} = \frac{\sum x_i y_i}{\sum x_i^2} \tag{18.3.6}$$

• a. *1 point Is $\tilde{\beta}$ an unbiased estimator of β? (Proof is required.)*

Answer. First derive a nice expression for $\tilde{\beta} - \beta$:

$$\begin{aligned}
\tilde{\beta} - \beta &= \frac{\sum x_i y_i}{\sum x_i^2} - \frac{\sum x_i^2\beta}{\sum x_i^2} \\
&= \frac{\sum x_i(y_i - x_i\beta)}{\sum x_i^2} \\
&= \frac{\sum x_i\varepsilon_i}{\sum x_i^2} \qquad \text{since} \quad y_i = \beta x_i + \varepsilon_i \\
\text{E}[\tilde{\beta} - \beta] &= \text{E}\left[\frac{\sum x_i\varepsilon_i}{\sum x_i^2}\right] \\
&= \frac{\sum \text{E}[x_i\varepsilon_i]}{\sum x_i^2} \\
&= \frac{\sum x_i\,\text{E}[\varepsilon_i]}{\sum x_i^2} = 0 \qquad \text{since} \quad \text{E}\,\varepsilon_i = 0.
\end{aligned}$$

□

• b. *2 points Derive the variance of $\tilde{\beta}$. (Show your work.)*

Answer.

$$\begin{aligned}
\text{var}\,\tilde{\beta} &= \text{E}[\tilde{\beta} - \beta]^2 \\
&= \text{E}\left(\frac{\sum x_i\varepsilon_i}{\sum x_i^2}\right)^2 \\
&= \frac{1}{(\sum x_i^2)^2}\text{E}[\sum x_i\varepsilon_i]^2 \\
&= \frac{1}{(\sum x_i^2)^2}\left(\text{E}\sum(x_i\varepsilon_i)^2 + 2\,\text{E}\sum_{i<j}(x_i\varepsilon_i)(x_j\varepsilon_j)\right) \\
&= \frac{1}{(\sum x_i^2)^2}\sum \text{E}[x_i\varepsilon_i]^2 \qquad \text{since the } \varepsilon_i\text{'s are uncorrelated, i.e., } \text{cov}[\varepsilon_i, \varepsilon_j] = 0 \text{ for } i \neq j \\
&= \frac{1}{(\sum x_i^2)^2}\sigma^2\sum x_i^2 \qquad \text{since all } \varepsilon_i \text{ have equal variance } \sigma^2 \\
&= \frac{\sigma^2}{\sum x_i^2}.
\end{aligned}$$

□

Problem 251. *We still assume* (18.3.5) *is the true model. Consider an alternative estimator:*

$$\hat{\beta} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \tag{18.3.7}$$

i.e., the estimator which would be the best linear unbiased estimator if the true model were (14.2.15).

• a. *2 points Is $\hat{\beta}$ still an unbiased estimator of β if* (18.3.5) *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 (14.2.15) whatever the value of α or β, therefore also when $\alpha = 0$, i.e., when the model is (18.3.5). 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* (18.3.5) *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.

$$\operatorname{var}\hat\beta = \operatorname{var}\left(\beta + \frac{\sum(x_i-\bar x)\varepsilon_i}{\sum(x_i-\bar x)^2}\right) \tag{18.3.8}$$

$$= \operatorname{var}\left(\frac{\sum(x_i-\bar x)\varepsilon_i}{\sum(x_i-\bar x)^2}\right) \tag{18.3.9}$$

$$= \frac{\sum(x_i-\bar x)^2\operatorname{var}\varepsilon_i}{(\sum(x_i-\bar x)^2)^2} \qquad \text{since} \quad \operatorname{cov}[\varepsilon_i\varepsilon_j] = 0 \tag{18.3.10}$$

$$= \frac{\sigma^2}{\sum(x_i-\bar x)^2}. \tag{18.3.11}$$

□

• c. *1 point Still assuming* (18.3.5) *is the true model, would you prefer $\hat\beta$ or the $\tilde\beta$ from Problem 250 as an estimator of β?*

Answer. Since $\tilde\beta$ and $\hat\beta$ are both unbiased estimators, if (18.3.5) 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

$$\operatorname{var}\hat\beta = \frac{\sigma^2}{\sum(x_i-\bar x)^2} \ge \frac{\sigma^2}{\sum x_i^2} = \operatorname{var}\tilde\beta \tag{18.3.12}$$

one must show

$$\sum(x_i-\bar x)^2 \le \sum x_i^2 \tag{18.3.13}$$

which is a simple consequence of (9.1.1). Thus $\operatorname{var}\hat\beta \ge \operatorname{var}\tilde\beta$; the variances are equal only if $\bar x = 0$, i.e., if $\tilde\beta = \hat\beta$. □

Problem 252. *Suppose the true model is* (14.2.15) *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*

$$\mathrm{E}[\tilde\beta - \beta] = \alpha\frac{n\bar x}{\sum x_i^2} \tag{18.3.14}$$

Answer. In situations like this it is always worth while to get a nice simple expression for the sampling error:

$$\tilde\beta - \beta = \frac{\sum x_i y_i}{\sum x_i^2} - \beta \tag{18.3.15}$$

$$= \frac{\sum x_i(\alpha+\beta x_i+\varepsilon_i)}{\sum x_i^2} - \beta \qquad \text{since } y_i = \alpha + \beta x_i + \varepsilon_i \tag{18.3.16}$$

$$= \alpha\frac{\sum x_i}{\sum x_i^2} + \beta\frac{\sum x_i^2}{\sum x_i^2} + \frac{\sum x_i\varepsilon_i}{\sum x_i^2} - \beta \tag{18.3.17}$$

$$= \alpha\frac{\sum x_i}{\sum x_i^2} + \frac{\sum x_i\varepsilon_i}{\sum x_i^2} \tag{18.3.18}$$

$$\mathrm{E}[\tilde\beta - \beta] = \mathrm{E}\,\alpha\frac{\sum x_i}{\sum x_i^2} + \mathrm{E}\,\frac{\sum x_i\varepsilon_i}{\sum x_i^2} \tag{18.3.19}$$

$$= \alpha\frac{\sum x_i}{\sum x_i^2} + \frac{\sum x_i\,\mathrm{E}\,\varepsilon_i}{\sum x_i^2} \tag{18.3.20}$$

$$= \alpha\frac{\sum x_i}{\sum x_i^2} + 0 = \alpha\frac{n\bar x}{\sum x_i^2} \tag{18.3.21}$$

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*

$$\frac{\sigma^2}{\sum(x_i-\bar x)^2} \tag{18.3.22}$$

which is the variance of the OLS estimator in this model?

Answer.

$$\operatorname{var}\tilde\beta = \operatorname{var}\left[\frac{\sum x_i y_i}{\sum x_i^2}\right] \tag{18.3.23}$$

$$= \frac{1}{\left(\sum x_i^2\right)^2}\operatorname{var}[\sum x_i y_i] \tag{18.3.24}$$

$$= \frac{1}{\left(\sum x_i^2\right)^2}\sum x_i^2\operatorname{var}[y_i] \tag{18.3.25}$$

$$= \frac{\sigma^2}{\left(\sum x_i^2\right)^2}\sum x_i^2 \quad \text{since all } y_i \text{ are uncorrelated and have equal variance } \sigma^2 \tag{18.3.26}$$

$$= \frac{\sigma^2}{\sum x_i^2}. \tag{18.3.27}$$

This variance is smaller or equal because $\sum x_i^2 \ge \sum(x_i-\bar x)^2$. □

• 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*

$$\frac{\alpha^2}{\sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{\sum(x_i-\bar{x})^2}\right)}<1 \tag{18.3.28}$$

Answer. This implies some tedious algebra. Here it is important to set it up right.

$$\begin{aligned}
\mathrm{MSE}[\tilde{\beta};\beta] &= \frac{\sigma^2}{\sum x_i^2}+\left(\frac{\alpha n\bar{x}}{\sum x_i^2}\right)^2 \leq \frac{\sigma^2}{\sum(x_i-\bar{x})^2} \\
\left(\frac{\alpha n\bar{x}}{\sum x_i^2}\right)^2 &\leq \frac{\sigma^2}{\sum(x_i-\bar{x})^2}-\frac{\sigma^2}{\sum x_i^2}=\frac{\sigma^2\left(\sum x_i^2-\sum(x_i-\bar{x})^2\right)}{\sum(x_i-\bar{x})^2\sum x_i^2} \\
&= \frac{\sigma^2 n\bar{x}^2}{\sum(x_i-\bar{x})^2\sum x_i^2} \\
\frac{\alpha^2 n}{\sum x_i^2} &= \frac{\alpha^2}{\frac{1}{n}\sum(x_i-\bar{x})^2+\bar{x}^2} \leq \frac{\sigma^2}{\sum(x_i-\bar{x})^2} \\
\frac{\alpha^2}{\sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{\sum(x_i-\bar{x})^2}\right)} &\leq 1
\end{aligned}$$

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 $\hat{\alpha}^2$ 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 $z\sim \mathrm{N}(\mu,1)$ 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 $\boldsymbol{\varepsilon}=\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}$ is the vector of residuals $\hat{\boldsymbol{\varepsilon}}=\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}$.

Problem 253. *Let* $\tilde{\boldsymbol{\varepsilon}}=\boldsymbol{A}\boldsymbol{y}$ *be a linear predictor of the disturbance vector* $\boldsymbol{\varepsilon}$ *in the model* $\boldsymbol{y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$ *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. $\mathcal{E}[\boldsymbol{A}\boldsymbol{y}-\boldsymbol{\varepsilon}]=\mathcal{E}[\boldsymbol{A}\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{A}\boldsymbol{\varepsilon}-\boldsymbol{\varepsilon}]=\boldsymbol{A}\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{o}-\boldsymbol{o}$. 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* $\tilde{\boldsymbol{\varepsilon}}=\boldsymbol{A}\boldsymbol{y}$ *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 $\boldsymbol{A}\boldsymbol{y}-\boldsymbol{\varepsilon}=\boldsymbol{A}\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{A}\boldsymbol{\varepsilon}-\boldsymbol{\varepsilon}=(\boldsymbol{A}-\boldsymbol{I})\boldsymbol{\varepsilon}$; 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 $\hat{\boldsymbol{\varepsilon}}=\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}$. □

Problem 254. *This is a vector generalization of problem 170. 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 $\tilde{\boldsymbol{\beta}}=\tilde{\boldsymbol{B}}\boldsymbol{y}$; unbiasedness means $\tilde{\boldsymbol{B}}\boldsymbol{X}=\boldsymbol{I}$. Therefore

$$\begin{aligned}
\mathcal{C}[\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}},\hat{\boldsymbol{\beta}}] &= \mathcal{C}[\left((\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top-\tilde{\boldsymbol{B}}\right)\boldsymbol{y},(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}] \\
&= \left((\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top-\tilde{\boldsymbol{B}}\right)\mathcal{V}[\boldsymbol{y}]\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1} \\
&= \sigma^2\left((\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top-\tilde{\boldsymbol{B}}\right)\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1} \\
&= \sigma^2\left((\boldsymbol{X}^\top\boldsymbol{X})^{-1}-(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\right)=\boldsymbol{O}.
\end{aligned}$$

□

• 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 255. *3 points Given a simple regression* $y_t=\alpha+\beta x_t+\varepsilon_t$*, 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 $y_t=\alpha+\mu+\beta x_t+\tilde{\varepsilon}_t$, 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 256. *3 points The model is* $\boldsymbol{y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$ *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 257. *This is* [**JHG+88**, 5.3.32]*: Consider the log-linear statistical model*

$$\mathsf{y}_t = \alpha x_t^\beta \exp \varepsilon_t = z_t \exp \varepsilon_t \tag{18.3.29}$$

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:

$$\log \mathsf{y}_t = \log \alpha + \beta \log x_t + \varepsilon_t \tag{18.3.30}$$

□

• b. *1 point Show that the elasticity of the functional relationship between* x_t *and* z_t

$$\eta = \frac{\partial z_t / z_t}{\partial x_t / x_t} \tag{18.3.31}$$

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

$$\frac{\partial \log z_t}{\partial z_t} = \frac{1}{z_t} \tag{18.3.32}$$

which can be rewritten as

$$\frac{\partial z_t}{z_t} = \partial \log z_t; \tag{18.3.33}$$

The same can be done with x_t; therefore

$$\frac{\partial z_t / z_t}{\partial x_t / x_t} = \frac{\partial \log z_t}{\partial \log x_t} = \beta \tag{18.3.34}$$

What we just did was a tricky way to take a derivative. A less tricky way is:

$$\frac{\partial z_t}{\partial x_t} = \alpha \beta x_t^{\beta - 1} = \beta z_t / x_t \tag{18.3.35}$$

Therefore

$$\frac{\partial z_t}{\partial x_t} \frac{x_t}{z_t} = \beta \tag{18.3.36}$$

□

PROBLEM 258.

• a. *2 points What is the elasticity in the simple regression* $\mathsf{y}_t = \alpha + \beta x_t + \varepsilon_t$*?*

ANSWER.

$$\eta_t = \frac{\partial z_t / z_t}{\partial x_t / x_t} = \frac{\partial z_t}{\partial x_t} \frac{x_t}{z_t} = \frac{\beta x_t}{z_t} = \frac{\beta x_t}{\alpha + \beta x_t} \tag{18.3.37}$$

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:

$$\eta_t = \frac{\partial \mathsf{y}_t / \mathsf{y}_t}{\partial x_t / x_t} = \frac{\partial \mathsf{y}_t}{\partial x_t} \frac{x_t}{\mathsf{y}_t} = \frac{\beta x_t}{\mathsf{y}_t} \tag{18.3.38}$$

This is different than (18.3.37), 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*

$$h = \frac{\hat{\beta}\bar{x}}{\bar{\mathsf{y}}} \tag{18.3.39}$$

of the elasticity in the simple regression model has the estimated asymptotic variance

$$\mathsf{s}^2 \begin{bmatrix} \frac{-h}{\bar{\mathsf{y}}} & \frac{\bar{x}(1-h)}{\bar{\mathsf{y}}} \end{bmatrix} \begin{bmatrix} 1 & \bar{x} \\ \bar{x} & \overline{x^2} \end{bmatrix}^{-1} \begin{bmatrix} \frac{-h}{\bar{\mathsf{y}}} \\ \frac{\bar{x}(1-h)}{\bar{\mathsf{y}}} \end{bmatrix} \tag{18.3.40}$$

• d. *Compare* [**Gre97**, example 6.20 on p. 280]*. Assume*

$$\frac{1}{n}(\boldsymbol{X}^\top \boldsymbol{X}) = \begin{bmatrix} 1 & \bar{x} \\ \bar{x} & \overline{x^2} \end{bmatrix} \to \boldsymbol{Q} = \begin{bmatrix} 1 & q \\ q & r \end{bmatrix} \tag{18.3.41}$$

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*

$$\eta = \frac{q\beta}{\alpha + q\beta} \tag{18.3.42}$$

Then

$$\mathsf{h} = \frac{q\hat{\beta}}{\hat{\alpha} + q\hat{\beta}} \tag{18.3.43}$$

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 $\boldsymbol{u} = \boldsymbol{o}$:

$$v = f(\boldsymbol{o}) + \sum u_i \frac{\partial f}{\partial u_i}\Big|_{\boldsymbol{u}=\boldsymbol{o}} + \frac{1}{2} \sum_{i,j} u_i u_j \frac{\partial^2 f}{\partial u_i \partial u_j}\Big|_{\boldsymbol{u}=\boldsymbol{o}} \tag{18.3.44}$$

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:

$$\log(\mathsf{y}) = \alpha + \sum \beta_i \log x_i + \frac{1}{2} \sum_{i,j} \gamma_{ij} \log x_i \log x_j + \varepsilon \tag{18.3.45}$$

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:

$$\log \mathsf{y}_t = \boldsymbol{x}_t^\top \boldsymbol{\beta} + \varepsilon_t \tag{18.3.46}$$

Here usually one of the columns of $\boldsymbol{X}$ is the time subscript t itself; [**Gre97**, p. 227] writes it as

$$\log \mathsf{y}_t = \boldsymbol{x}_t^\top \boldsymbol{\beta} + t\delta + \varepsilon_t \tag{18.3.47}$$

where δ is the autonomous growth rate. The logistic functional form is appropriate for adoption rates $0 \leq \mathsf{y}_t \leq 1$: the rate of adoption is slow at first, then rapid as the innovation gains popularity, then slow again as the market becomes saturated:

$$\mathsf{y}_t = \frac{\exp(\boldsymbol{x}_t^\top \boldsymbol{\beta} + t\delta + \varepsilon_t)}{1 + \exp(\boldsymbol{x}_t^\top \boldsymbol{\beta} + t\delta + \varepsilon_t)} \tag{18.3.48}$$

This can be linearized by the logit transformation:

$$\operatorname{logit}(\mathsf{y}_t) = \log \frac{\mathsf{y}_t}{1 - \mathsf{y}_t} = \boldsymbol{x}_t^\top \boldsymbol{\beta} + t\delta + \varepsilon_t \tag{18.3.49}$$

Problem 259. *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 $\hat{\mu} + \varepsilon_t$. □

18.4. 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

$$\mathrm{E}[\mathsf{SSE}] = \sigma^2(n - q) \tag{18.4.1}$$

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 260.

- a. *2 points Show that*

$$\mathsf{SSE} = \boldsymbol{\varepsilon}^\top \boldsymbol{M}\boldsymbol{\varepsilon} \qquad \textit{where} \qquad \boldsymbol{M} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^- \boldsymbol{X}^\top \tag{18.4.2}$$

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*

$$\mathrm{E}[\mathsf{SSE}] = \sigma^2(n - k) \tag{18.4.3}$$

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 261. *As an alternative proof of* (18.4.3) *show that* $\mathsf{SSE} = \mathsf{y}^\top \boldsymbol{M}\mathsf{y}$ *and use theorem* **??**.

From (18.4.3) 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**]

$$\hat{\sigma}^2_{TS} = \frac{1}{n - q + 2} \mathsf{y}^\top \boldsymbol{M}\mathsf{y} = \frac{1}{n - q + 2} \sum_{i=1}^{n} \hat{\varepsilon}_i^2, \tag{18.4.4}$$

is biased but has lower MSE than s^2. Compare problem 163. 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).

18.5. Mallow's Cp-Statistic as Estimator of the Mean Squared Error

Problem 262. *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*

$$\mathcal{MSE}[\boldsymbol{X}\hat{\boldsymbol{\beta}}; \boldsymbol{\eta}] = \sigma^2 \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^- \boldsymbol{X}^\top + \boldsymbol{M}\boldsymbol{\eta}\boldsymbol{\eta}^\top \boldsymbol{M} \tag{18.5.1}$$

where $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^- \boldsymbol{X}^\top$.

• b. *1 point Formula* (18.5.1) *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*

$$\operatorname{tr}\mathcal{MSE}[\boldsymbol{X}\hat{\boldsymbol{\beta}};\boldsymbol{\eta}] = \sigma^2 q + \boldsymbol{\eta}^\top \boldsymbol{M}\boldsymbol{\eta}. \tag{18.5.2}$$

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* (18.5.2) *can be constructed using this* s^2 *and the* SSE *of the regression* $\mathsf{SSE} = \mathbf{y}^\top \boldsymbol{M}\mathbf{y}$. *Show that*

$$\mathrm{E}[\mathsf{SSE} - (n-2q)\mathsf{s}^2] = \sigma^2 q + \boldsymbol{\eta}^\top \boldsymbol{M}\boldsymbol{\eta}. \tag{18.5.3}$$

Hint: use equation (**??**). *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

$$\mathsf{C}_{\mathrm{p}} = \frac{\mathsf{SSE}}{\mathsf{s}^2} + 2q - n \tag{18.5.4}$$

is called Mallow's 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* (18.5.2) *becomes* $\operatorname{tr}\mathcal{MSE}[\boldsymbol{X}\hat{\boldsymbol{\beta}};\boldsymbol{\eta}] = \sigma^2 q$, *i.e., in this case* C_{p} *should be close to* q. *Therefore the selection rule for regressions should be here to pick that regression for which the* 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 232.

CHAPTER 19

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:

$$\boldsymbol{P}\mathbf{y} = \boldsymbol{P}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{P}\boldsymbol{\varepsilon} \qquad \boldsymbol{P}\boldsymbol{\varepsilon} \sim (\boldsymbol{o}, \sigma^2\boldsymbol{I}) \tag{19.0.5}$$

The OLS estimate of $\boldsymbol{\beta}$ in this transformed model is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{P}^\top\boldsymbol{P}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{P}^\top\boldsymbol{P}\mathbf{y} = (\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\mathbf{y}. \tag{19.0.6}$$

This $\hat{\boldsymbol{\beta}}$ is the BLUE of $\boldsymbol{\beta}$ in model (19.0.5), 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 263. *2 points Show that*

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\varepsilon} \tag{19.0.7}$$

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 (19.0.7) is very similar to proof of (18.0.7). □

The objective function of the associated least squares problem is

$$\boldsymbol{\beta} = \hat{\boldsymbol{\beta}} \qquad \text{minimizes} \qquad (\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta})^\top\boldsymbol{\Psi}^{-1}(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}). \tag{19.0.8}$$

The normal equations are

$$\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\mathbf{y} \tag{19.0.9}$$

If $\boldsymbol{X}$ has full rank, then $\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X}$ is nonsingular, and the unique $\hat{\boldsymbol{\beta}}$ minimizing (19.0.8) is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\mathbf{y} \tag{19.0.10}$$

Problem 264. [**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 $\boldsymbol{a} = \boldsymbol{o}$. □

The least squares objective function of the transformed model, which $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ minimizes, can be written

$$(\boldsymbol{P}\mathbf{y} - \boldsymbol{P}\boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{P}\mathbf{y} - \boldsymbol{P}\boldsymbol{X}\boldsymbol{\beta}) = (\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta})^\top\boldsymbol{\Psi}^{-1}(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}), \tag{19.0.11}$$

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 265. *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 266. *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*

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\mathbf{y}. \tag{19.0.12}$$

We will write $\hat{\boldsymbol{\beta}}_{OLS}$ *for the ordinary least squares estimator*

$$\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \mathbf{y} \tag{19.0.13}$$

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 170.) 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:

$$\mathcal{V}[\hat{\boldsymbol{\beta}}_{OLS}] - \mathcal{V}[\hat{\boldsymbol{\beta}}] = \sigma^2 \Big((\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{\Psi} \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} - (\boldsymbol{X}^\top \boldsymbol{\Psi}^{-1} \boldsymbol{X})^{-1} \Big) =$$
$$= \sigma^2 \Big((\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top - (\boldsymbol{X}^\top \boldsymbol{\Psi}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{\Psi}^{-1} \Big) \boldsymbol{\Psi} \Big(\boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} - \boldsymbol{\Psi}^{-1} \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{\Psi}^{-1} \boldsymbol{X})^{-1} \Big)$$

□

Examples of GLS models are discussed in chapters **??** and **??**.

CHAPTER 20

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.

20.1. 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

$$\begin{bmatrix}\mathbf{y}\\ \boldsymbol{z}\end{bmatrix} \sim \begin{bmatrix}\boldsymbol{\mu}\\ \boldsymbol{\nu}\end{bmatrix}, \sigma^2\begin{bmatrix}\boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}} & \boldsymbol{\Omega}_{\mathbf{y}z}\\ \boldsymbol{\Omega}_{z\mathbf{y}} & \boldsymbol{\Omega}_{zz}\end{bmatrix}, \qquad \sigma^2 > 0. \tag{20.1.1}$$

$\mathbf{y}$ 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.,

$$\mathcal{C}[\boldsymbol{z}^* - \boldsymbol{z}, \mathbf{y}] = \boldsymbol{O}. \tag{20.1.2}$$

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

$$\boldsymbol{z}^* = \boldsymbol{\nu} + \boldsymbol{B}^*(\mathbf{y} - \boldsymbol{\mu}). \tag{20.1.3}$$

Since

$$\boldsymbol{z}^* - \boldsymbol{z} = \boldsymbol{B}^*(\mathbf{y} - \boldsymbol{\mu}) - (\boldsymbol{z} - \boldsymbol{\nu}) = \begin{bmatrix}\boldsymbol{B}^* & -\boldsymbol{I}\end{bmatrix}\begin{bmatrix}\mathbf{y} - \boldsymbol{\mu}\\ \boldsymbol{z} - \boldsymbol{\nu}\end{bmatrix}, \tag{20.1.4}$$

the zero correlation condition (20.1.2) translates into

$$\boldsymbol{B}^*\boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}} = \boldsymbol{\Omega}_{z\mathbf{y}}, \tag{20.1.5}$$

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

$$\boldsymbol{z}^* = \boldsymbol{\nu} + \boldsymbol{\Omega}_{z\mathbf{y}}\boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}}^-(\mathbf{y} - \boldsymbol{\mu}) \tag{20.1.6}$$

satisfies the two requirements.

Unbiasedness and condition (20.1.2) 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 20.1.1 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 20.1.1. *In situation* (20.1.1)*, the predictor* (20.1.6) *has, among all predictors of* $\boldsymbol{z}$ *which are affine functions of* $\mathbf{y}$*, the smallest* $\mathcal{MSE}$ *matrix. Its* $\mathcal{MSE}$ *matrix is*

$$\mathcal{MSE}[\boldsymbol{z}^*; \boldsymbol{z}] = \mathcal{E}[(\boldsymbol{z}^* - \boldsymbol{z})(\boldsymbol{z}^* - \boldsymbol{z})^\top] = \sigma^2(\boldsymbol{\Omega}_{zz} - \boldsymbol{\Omega}_{z\mathbf{y}}\boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}}^-\boldsymbol{\Omega}_{\mathbf{y}z}) = \sigma^2\boldsymbol{\Omega}_{zz.\mathbf{y}}. \tag{20.1.7}$$

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 (17.1.2) one can write

$$\mathcal{E}[(\tilde{\boldsymbol{z}} - \boldsymbol{z})(\tilde{\boldsymbol{z}} - \boldsymbol{z})^\top] = \mathcal{V}[(\tilde{\boldsymbol{z}} - \boldsymbol{z})] + \tilde{\boldsymbol{d}}\tilde{\boldsymbol{d}}^\top \tag{20.1.8}$$

$$= \mathcal{V}\Big[\begin{bmatrix}\tilde{\boldsymbol{B}} & -\boldsymbol{I}\end{bmatrix}\begin{bmatrix}\mathbf{y}\\ \boldsymbol{z}\end{bmatrix}\Big] + \tilde{\boldsymbol{d}}\tilde{\boldsymbol{d}}^\top \tag{20.1.9}$$

$$= \sigma^2\begin{bmatrix}\tilde{\boldsymbol{B}} & -\boldsymbol{I}\end{bmatrix}\begin{bmatrix}\boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}} & \boldsymbol{\Omega}_{\mathbf{y}z}\\ \boldsymbol{\Omega}_{z\mathbf{y}} & \boldsymbol{\Omega}_{zz}\end{bmatrix}\begin{bmatrix}\tilde{\boldsymbol{B}}^\top\\ -\boldsymbol{I}\end{bmatrix} + \tilde{\boldsymbol{d}}\tilde{\boldsymbol{d}}^\top. \tag{20.1.10}$$

This $\mathcal{MSE}$-matrix is minimized if and only if $\boldsymbol{d}^* = \boldsymbol{o}$ and $\boldsymbol{B}^*$ satisfies (20.1.5). To see this, take any solution $\boldsymbol{B}^*$ of (20.1.5), 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

$$\mathcal{MSE}[\tilde{\boldsymbol{z}}; \boldsymbol{z}] = \sigma^2\begin{bmatrix}\boldsymbol{B}^* + \tilde{\boldsymbol{D}} & -\boldsymbol{I}\end{bmatrix}\begin{bmatrix}\boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}} & \boldsymbol{\Omega}_{\mathbf{y}z}\\ \boldsymbol{\Omega}_{z\mathbf{y}} & \boldsymbol{\Omega}_{zz}\end{bmatrix}\begin{bmatrix}\boldsymbol{B}^{*\top} + \tilde{\boldsymbol{D}}^\top\\ -\boldsymbol{I}\end{bmatrix} + \tilde{\boldsymbol{d}}\tilde{\boldsymbol{d}}^\top$$

$$= \sigma^2\begin{bmatrix}\boldsymbol{B}^* + \tilde{\boldsymbol{D}} & -\boldsymbol{I}\end{bmatrix}\begin{bmatrix}\boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}}\tilde{\boldsymbol{D}}^\top\\ -\boldsymbol{\Omega}_{zz.\mathbf{y}} + \boldsymbol{\Omega}_{z\mathbf{y}}\tilde{\boldsymbol{D}}^\top\end{bmatrix} + \tilde{\boldsymbol{d}}\tilde{\boldsymbol{d}}^\top \tag{20.1.11}$$

$$= \sigma^2(\boldsymbol{\Omega}_{zz.\mathbf{y}} + \tilde{\boldsymbol{D}}\boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}}\tilde{\boldsymbol{D}}^\top) + \tilde{\boldsymbol{d}}\tilde{\boldsymbol{d}}^\top. \tag{20.1.12}$$

The $\mathcal{MSE}$ matrix is therefore minimized (with minimum value $\sigma^2\boldsymbol{\Omega}_{zz.\mathbf{y}}$) 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 (20.1.5). □

PROBLEM 267. *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* (20.1.5) *and* $\boldsymbol{y}$ *is any feasible observed value* $\boldsymbol{y}$*, plugged into equations* (20.1.3) *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 $\boldsymbol{B}_i^*\boldsymbol{y} = \boldsymbol{B}_i^*\boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}}\boldsymbol{q} = \boldsymbol{\Omega}_{z\mathbf{y}}\boldsymbol{q}$. □

The matrix $\boldsymbol{B}^*$ is also called the regression matrix of $\boldsymbol{z}$ on $\mathbf{y}$, and the unscaled covariance matrix has the form

$$\boldsymbol{\Omega} = \begin{bmatrix} \boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}} & \boldsymbol{\Omega}_{\mathbf{y}z} \\ \boldsymbol{\Omega}_{z\mathbf{y}} & \boldsymbol{\Omega}_{zz} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}} & \boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}}\boldsymbol{X}^\top \\ \boldsymbol{X}\boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}} & \boldsymbol{X}\boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}}\boldsymbol{X}^\top + \boldsymbol{\Omega}_{zz.\mathbf{y}} \end{bmatrix} \tag{20.1.13}$$

Where we wrote here $\boldsymbol{B}^* = \boldsymbol{X}$ in order to make the analogy with regression clearer. A g-inverse is

$$\boldsymbol{\Omega}^- = \begin{bmatrix} \boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}}^- + \boldsymbol{X}^\top\boldsymbol{\Omega}_{zz.\mathbf{y}}^-\boldsymbol{X} & -\boldsymbol{X}^\top\boldsymbol{\Omega}_{zz.\mathbf{y}}^- \\ -\boldsymbol{X}^\top\boldsymbol{\Omega}_{zz.\mathbf{y}}^- & \boldsymbol{\Omega}_{zz.\mathbf{y}}^- \end{bmatrix} \tag{20.1.14}$$

and every g-inverse of the covariance matrix has a g-inverse of $\boldsymbol{\Omega}_{zz.\mathbf{y}}$ as its zz-partition. (Proof in Problem 392.)

If $\boldsymbol{\Omega} = \begin{bmatrix} \boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}} & \boldsymbol{\Omega}_{\mathbf{y}z} \\ \boldsymbol{\Omega}_{z\mathbf{y}} & \boldsymbol{\Omega}_{zz} \end{bmatrix}$ is nonsingular, 20.1.5 is also solved by $\boldsymbol{B}^* = -(\boldsymbol{\Omega}^{zz})^-\boldsymbol{\Omega}^{z\mathbf{y}}$ where $\boldsymbol{\Omega}^{zz}$ and $\boldsymbol{\Omega}^{z\mathbf{y}}$ are the corresponding partitions of the inverse $\boldsymbol{\Omega}^{-1}$. See Problem 392 for a proof. Therefore instead of 20.1.6 the predictor can also be written

$$z^* = \boldsymbol{\nu} - \left(\boldsymbol{\Omega}^{zz}\right)^{-1}\boldsymbol{\Omega}^{z\mathbf{y}}(\mathbf{y} - \boldsymbol{\mu}) \tag{20.1.15}$$

(note the minus sign) or

$$z^* = \boldsymbol{\nu} - \boldsymbol{\Omega}_{zz.\mathbf{y}}\boldsymbol{\Omega}^{z\mathbf{y}}(\mathbf{y} - \boldsymbol{\mu}). \tag{20.1.16}$$

Problem 268. *This problem utilizes the concept of a bounded risk estimator, which is not yet explained very well in these notes. Assume $\mathbf{y}$, z, $\boldsymbol{\mu}$, and $\boldsymbol{\nu}$ are jointly distributed random vectors. First assume $\boldsymbol{\nu}$ and $\boldsymbol{\mu}$ are observed, but $\mathbf{y}$ and z are not. Assume we know that in this case, the best linear bounded $\mathcal{MSE}$ predictor of $\mathbf{y}$ and z is $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$, with prediction errors distributed as follows:*

$$\begin{bmatrix} \mathbf{y} - \boldsymbol{\mu} \\ z - \boldsymbol{\nu} \end{bmatrix} \sim \begin{bmatrix} \boldsymbol{o} \\ \boldsymbol{o} \end{bmatrix}, \sigma^2 \begin{bmatrix} \boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}} & \boldsymbol{\Omega}_{\mathbf{y}z} \\ \boldsymbol{\Omega}_{z\mathbf{y}} & \boldsymbol{\Omega}_{zz} \end{bmatrix}. \tag{20.1.17}$$

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 z still isn't. Then the predictor

$$z^* = \boldsymbol{\nu} + \boldsymbol{\Omega}_{z\mathbf{y}}\boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}}^-(\mathbf{y} - \boldsymbol{\mu}) \tag{20.1.18}$$

is the best linear bounded $\mathcal{MSE}$ predictor of 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 z random, and one in which $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ and $\mathbf{y}$ are random and z is constant, and one in which $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ are random and $\mathbf{y}$ and 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 $\boldsymbol{\nu} = \mathcal{E}[z]$ and $\sigma^2 \begin{bmatrix} \boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}} & \boldsymbol{\Omega}_{\mathbf{y}z} \\ \boldsymbol{\Omega}_{z\mathbf{y}} & \boldsymbol{\Omega}_{zz} \end{bmatrix} = \mathcal{V}[\begin{bmatrix} \mathbf{y} \\ rx \end{bmatrix}]$ 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 z is constant, is (21.0.26). Another example, in which $\mathbf{y}$ and z both are constants and $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ random, is constrained least squares (22.4.3). □

• b. *Prove equation 20.1.18.*

Answer. In this proof we allow all four $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ and $\mathbf{y}$ and z to be random. A linear predictor based on $\mathbf{y}$, $\boldsymbol{\mu}$, and $\boldsymbol{\nu}$ can be written as $\tilde{z} = \boldsymbol{B}\mathbf{y} + \boldsymbol{C}\boldsymbol{\mu} + \boldsymbol{D}\boldsymbol{\nu} + \boldsymbol{d}$, therefore $\tilde{z} - z = \boldsymbol{B}(\mathbf{y} - \boldsymbol{\mu}) + (\boldsymbol{C} + \boldsymbol{B})\boldsymbol{\mu} + (\boldsymbol{D} - \boldsymbol{I})\boldsymbol{\nu} - (z - \boldsymbol{\nu}) + \boldsymbol{d}$. $\mathcal{E}[\tilde{z} - z] = \boldsymbol{o} + (\boldsymbol{C} + \boldsymbol{B})\,\mathcal{E}[\boldsymbol{\mu}] + (\boldsymbol{D} - \boldsymbol{I})\,\mathcal{E}[\boldsymbol{\nu}] - \boldsymbol{o} + \boldsymbol{d}$. 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 $\tilde{z} = \boldsymbol{\nu} + \boldsymbol{B}(\mathbf{y} - \boldsymbol{\mu}) + \boldsymbol{d}$, and the estimation error is $\tilde{z} - z = \boldsymbol{B}(\mathbf{y} - \boldsymbol{\mu}) - (z - \boldsymbol{\nu}) + \boldsymbol{d}$. Now continue as in the proof of theorem 20.1.1. I must still carry out this proof much more carefully! □

Problem 269. *4 points According to (20.1.2), the prediction error $z^* - z$ 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 z are jointly normal), then z^* as defined in (20.1.6) is the conditional mean $z^* = \mathcal{E}[z|\mathbf{y}]$, and its $\mathcal{MSE}$-matrix as defined in (20.1.7) is the conditional variance $\mathcal{V}[z|\mathbf{y}]$.*

Answer. From independence follows $\mathcal{E}[z^* - z|\mathbf{y}] = \mathcal{E}[z^* - z]$, and by the law of iterated expectations $\mathcal{E}[z^* - z] = \boldsymbol{o}$. Rewrite this as $\mathcal{E}[z|\mathbf{y}] = \mathcal{E}[z^*|\mathbf{y}]$. But since z^* is a function of $\mathbf{y}$, $\mathcal{E}[z^*|\mathbf{y}] = z^*$. Now the proof that the conditional dispersion matrix is the $\mathcal{MSE}$ matrix:

$$\begin{aligned} \mathcal{V}[z|\mathbf{y}] &= \mathcal{E}[(z - \mathcal{E}[z|\mathbf{y}])(z - \mathcal{E}[z|\mathbf{y}])^\top|\mathbf{y}] = \mathcal{E}[(z - z^*)(z - z^*)^\top|\mathbf{y}] \\ &= \mathcal{E}[(z - z^*)(z - z^*)^\top] = \mathcal{MSE}[z^*; z]. \end{aligned} \tag{20.1.19}$$

□

Problem 270. *Assume the expected values of $\mathbf{x}$, $\mathbf{y}$ and z are known, and their joint covariance matrix is known up to an unknown scalar factor $\sigma^2 > 0$.*

$$\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \\ z \end{bmatrix} \sim \begin{bmatrix} \boldsymbol{\lambda} \\ \boldsymbol{\mu} \\ \boldsymbol{\nu} \end{bmatrix}, \sigma^2 \begin{bmatrix} \boldsymbol{\Omega}_{\mathbf{x}\mathbf{x}} & \boldsymbol{\Omega}_{\mathbf{x}\mathbf{y}} & \boldsymbol{\Omega}_{\mathbf{x}z} \\ \boldsymbol{\Omega}_{\mathbf{x}\mathbf{y}}^\top & \boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}} & \boldsymbol{\Omega}_{\mathbf{y}z} \\ \boldsymbol{\Omega}_{\mathbf{x}z}^\top & \boldsymbol{\Omega}_{\mathbf{y}z}^\top & \boldsymbol{\Omega}_{zz} \end{bmatrix}. \tag{20.1.20}$$

$\mathbf{x}$ is the original information, $\mathbf{y}$ is additional information which becomes available, and z is the variable which we want to predict on the basis of this information.

• a. *2 points Show that $\mathbf{y}^* = \boldsymbol{\mu} + \boldsymbol{\Omega}_{\mathbf{x}\mathbf{y}}^\top\boldsymbol{\Omega}_{\mathbf{x}\mathbf{x}}^-(\mathbf{x} - \boldsymbol{\lambda})$ is the best linear predictor of $\mathbf{y}$ and $z^* = \boldsymbol{\nu} + \boldsymbol{\Omega}_{\mathbf{x}z}^\top\boldsymbol{\Omega}_{\mathbf{x}\mathbf{x}}^-(\mathbf{x} - \boldsymbol{\lambda})$ the best linear predictor of z on the basis of the observation of $\mathbf{x}$, and that their joint $\mathcal{MSE}$-matrix is*

$$\mathcal{E}\left[\begin{bmatrix} \mathbf{y}^* - \mathbf{y} \\ z^* - z \end{bmatrix}\left[(\mathbf{y}^* - \mathbf{y})^\top \quad (z^* - z)^\top\right]\right] = \sigma^2 \begin{bmatrix} \boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}} - \boldsymbol{\Omega}_{\mathbf{x}\mathbf{y}}^\top\boldsymbol{\Omega}_{\mathbf{x}\mathbf{x}}^-\boldsymbol{\Omega}_{\mathbf{x}\mathbf{y}} & \boldsymbol{\Omega}_{\mathbf{y}z} - \boldsymbol{\Omega}_{\mathbf{x}\mathbf{y}}^\top\boldsymbol{\Omega}_{\mathbf{x}\mathbf{x}}^-\boldsymbol{\Omega}_{\mathbf{x}z} \\ \boldsymbol{\Omega}_{\mathbf{y}z}^\top - \boldsymbol{\Omega}_{\mathbf{x}z}^\top\boldsymbol{\Omega}_{\mathbf{x}\mathbf{x}}^-\boldsymbol{\Omega}_{\mathbf{x}\mathbf{y}} & \boldsymbol{\Omega}_{zz} - \boldsymbol{\Omega}_{\mathbf{x}z}^\top\boldsymbol{\Omega}_{\mathbf{x}\mathbf{x}}^-\boldsymbol{\Omega}_{\mathbf{x}z} \end{bmatrix}$$

which can also be written

$$= \sigma^2 \begin{bmatrix} \boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}.\mathbf{x}} & \boldsymbol{\Omega}_{\mathbf{y}z.\mathbf{x}} \\ \boldsymbol{\Omega}_{\mathbf{y}z.\mathbf{x}}^\top & \boldsymbol{\Omega}_{zz.\mathbf{x}} \end{bmatrix}.$$

Answer. This part of the question is a simple application of the formulas derived earlier. For the $\mathcal{MSE}$-matrix you first get

$$\sigma^2\left(\begin{bmatrix}\boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}} & \boldsymbol{\Omega}_{\mathbf{y}z}\\ \boldsymbol{\Omega}_{\mathbf{y}z}^\top & \boldsymbol{\Omega}_{zz}\end{bmatrix} - \begin{bmatrix}\boldsymbol{\Omega}_{\mathbf{x}\mathbf{y}}^\top\\ \boldsymbol{\Omega}_{\mathbf{x}z}^\top\end{bmatrix}\boldsymbol{\Omega}_{\mathbf{x}\mathbf{x}}^-\begin{bmatrix}\boldsymbol{\Omega}_{\mathbf{x}\mathbf{y}} & \boldsymbol{\Omega}_{\mathbf{x}z}\end{bmatrix}\right)$$

□

• 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*

$$\mathbf{z}^{**} = \mathbf{z}^* + \boldsymbol{\Omega}_{\mathbf{y}z.\mathbf{x}}^\top\boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}.\mathbf{x}}^-(\mathbf{y}-\mathbf{y}^*) \tag{20.1.21}$$

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 (20.1.20) *do not enter this formula.*

Answer. Follows from

$$z^{**} = \boldsymbol{\nu} + \begin{bmatrix}\boldsymbol{\Omega}_{\mathbf{x}z}^\top & \boldsymbol{\Omega}_{\mathbf{y}z}^\top\end{bmatrix}\begin{bmatrix}\boldsymbol{\Omega}_{\mathbf{x}\mathbf{x}} & \boldsymbol{\Omega}_{\mathbf{x}\mathbf{y}}\\ \boldsymbol{\Omega}_{\mathbf{x}\mathbf{y}}^\top & \boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}}\end{bmatrix}^-\begin{bmatrix}\mathbf{x}-\boldsymbol{\lambda}\\ \mathbf{y}-\boldsymbol{\mu}\end{bmatrix} =$$

Now apply (A.8.2):

$$= \boldsymbol{\nu} + \begin{bmatrix}\boldsymbol{\Omega}_{\mathbf{x}z}^\top & \boldsymbol{\Omega}_{\mathbf{y}z}^\top\end{bmatrix}\begin{bmatrix}\boldsymbol{\Omega}_{\mathbf{x}\mathbf{x}}^- + \boldsymbol{\Omega}_{\mathbf{x}\mathbf{x}}^-\boldsymbol{\Omega}_{\mathbf{x}\mathbf{y}}\boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}.\mathbf{x}}^-\boldsymbol{\Omega}_{\mathbf{x}\mathbf{y}}^\top\boldsymbol{\Omega}_{\mathbf{x}\mathbf{x}}^- & -\boldsymbol{\Omega}_{\mathbf{x}\mathbf{x}}^-\boldsymbol{\Omega}_{\mathbf{x}\mathbf{y}}\boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}.\mathbf{x}}^-\\ -\boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}.\mathbf{x}}^-\boldsymbol{\Omega}_{\mathbf{x}\mathbf{y}}^\top\boldsymbol{\Omega}_{\mathbf{x}\mathbf{x}}^- & \boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}.\mathbf{x}}^-\end{bmatrix}\begin{bmatrix}\mathbf{x}-\boldsymbol{\lambda}\\ \mathbf{y}-\boldsymbol{\mu}\end{bmatrix} =$$

$$= \boldsymbol{\nu} + \begin{bmatrix}\boldsymbol{\Omega}_{\mathbf{x}z}^\top & \boldsymbol{\Omega}_{\mathbf{y}z}^\top\end{bmatrix}\begin{bmatrix}\boldsymbol{\Omega}_{\mathbf{x}\mathbf{x}}^-(\mathbf{x}-\boldsymbol{\lambda}) + \boldsymbol{\Omega}_{\mathbf{x}\mathbf{x}}^-\boldsymbol{\Omega}_{\mathbf{x}\mathbf{y}}\boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}.\mathbf{x}}^-(\mathbf{y}^*-\boldsymbol{\mu}) - \boldsymbol{\Omega}_{\mathbf{x}\mathbf{x}}^-\boldsymbol{\Omega}_{\mathbf{x}\mathbf{y}}\boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}.\mathbf{x}}^-(\mathbf{y}-\boldsymbol{\mu})\\ -\boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}.\mathbf{x}}^-(\mathbf{y}^*-\boldsymbol{\mu}) + \boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}.\mathbf{x}}^-(\mathbf{y}-\boldsymbol{\mu})\end{bmatrix} =$$

$$= \boldsymbol{\nu} + \begin{bmatrix}\boldsymbol{\Omega}_{\mathbf{x}z}^\top & \boldsymbol{\Omega}_{\mathbf{y}z}^\top\end{bmatrix}\begin{bmatrix}\boldsymbol{\Omega}_{\mathbf{x}\mathbf{x}}^-(\mathbf{x}-\boldsymbol{\lambda}) - \boldsymbol{\Omega}_{\mathbf{x}\mathbf{x}}^-\boldsymbol{\Omega}_{\mathbf{x}\mathbf{y}}\boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}.\mathbf{x}}^-(\mathbf{y}-\mathbf{y}^*)\\ +\boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}.\mathbf{x}}^-(\mathbf{y}-\mathbf{y}^*)\end{bmatrix} =$$

$$= \boldsymbol{\nu} + \boldsymbol{\Omega}_{\mathbf{x}z}^\top\boldsymbol{\Omega}_{\mathbf{x}\mathbf{x}}^-(\mathbf{x}-\boldsymbol{\lambda}) - \boldsymbol{\Omega}_{\mathbf{x}z}^\top\boldsymbol{\Omega}_{\mathbf{x}\mathbf{x}}^-\boldsymbol{\Omega}_{\mathbf{x}\mathbf{y}}\boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}.\mathbf{x}}^-(\mathbf{y}-\mathbf{y}^*) + \boldsymbol{\Omega}_{\mathbf{y}z}^\top\boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}.\mathbf{x}}^-(\mathbf{y}-\mathbf{y}^*) =$$

$$= z^* + \left(\boldsymbol{\Omega}_{\mathbf{y}z}^\top - \boldsymbol{\Omega}_{\mathbf{x}z}^\top\boldsymbol{\Omega}_{\mathbf{x}\mathbf{x}}^-\boldsymbol{\Omega}_{\mathbf{x}\mathbf{y}}\right)\boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}.\mathbf{x}}^-(\mathbf{y}-\mathbf{y}^*) = z^* + \boldsymbol{\Omega}_{\mathbf{y}z.\mathbf{x}}^\top\boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}.\mathbf{x}}^-(\mathbf{y}-\mathbf{y}^*)$$

□

Problem 271. *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: $\mathcal{E}[\mathbf{z}|\mathbf{x}] = \mathcal{E}\big[\mathcal{E}[\mathbf{z}|\mathbf{x},\mathbf{y}]\big|\mathbf{x}\big]$.*

Answer. With this hint it is trivial: $\mathcal{E}[\mathbf{z}|\mathbf{x}] = \mathcal{E}\big[\boldsymbol{\alpha}^*+\boldsymbol{A}^*\mathbf{x}+\boldsymbol{B}^*\mathbf{y}\big|\mathbf{x}\big] = \boldsymbol{\alpha}^*+\boldsymbol{A}^*\mathbf{x}+\boldsymbol{B}^*\mathcal{E}[\mathbf{y}|\mathbf{x}]$. □

• b. *1 point The next three examples are from* [**CW99**, pp. 264/5]*: Assume $\mathrm{E}[z|\mathbf{x},\mathbf{y}] = 1+2\mathbf{x}+3\mathbf{y}$, $\mathbf{x}$ and $\mathbf{y}$ are independent, and $\mathrm{E}[\mathbf{y}] = 2$. Compute $\mathrm{E}[z|\mathbf{x}]$.*

Answer. According to the formula, $\mathrm{E}[z|\mathbf{x}] = 1+2\mathbf{x}+3\mathrm{E}[\mathbf{y}|\mathbf{x}]$, but since $\mathbf{x}$ and $\mathbf{y}$ are independent, $\mathrm{E}[\mathbf{y}|\mathbf{x}] = \mathrm{E}[\mathbf{y}] = 2$; therefore $\mathrm{E}[z|\mathbf{x}] = 7+2\mathbf{x}$. I.e., the slope is the same, but the intercept changes. □

• c. *1 point Assume again $\mathrm{E}[z|\mathbf{x},\mathbf{y}] = 1+2\mathbf{x}+3\mathbf{y}$, but this time $\mathbf{x}$ and $\mathbf{y}$ are not independent but $\mathrm{E}[\mathbf{y}|\mathbf{x}] = 2-\mathbf{x}$. Compute $\mathrm{E}[z|\mathbf{x}]$.*

Answer. $\mathrm{E}[z|\mathbf{x}] = 1+2\mathbf{x}+3(2-\mathbf{x}) = 7-\mathbf{x}$. In this situation, both slope and intercept change, but it is still a linear relationship. □

• d. *1 point Again $\mathrm{E}[z|\mathbf{x},\mathbf{y}] = 1+2\mathbf{x}+3\mathbf{y}$, and this time the relationship between $\mathbf{x}$ and $\mathbf{y}$ is nonlinear: $\mathrm{E}[\mathbf{y}|\mathbf{x}] = 2-e^{\mathbf{x}}$. Compute $\mathrm{E}[z|\mathbf{x}]$.*

Answer. $\mathrm{E}[z|\mathbf{x}] = 1+2\mathbf{x}+3(2-e^{\mathbf{x}}) = 7+2\mathbf{x}-3e^{\mathbf{x}}$. This time the marginal relationship between $\mathbf{x}$ and $\mathbf{y}$ is no longer linear. This is so despite the fact that, if all the variables are included, i.e., if both $\mathbf{x}$ and $\mathbf{y}$ are included, then the relationship is linear. □

• e. *1 point Assume $\mathrm{E}[f(z)|\mathbf{x},\mathbf{y}] = 1+2\mathbf{x}+3\mathbf{y}$, where f is a nonlinear function, and $\mathrm{E}[\mathbf{y}|\mathbf{x}] = 2-\mathbf{x}$. Compute $\mathrm{E}[f(z)|\mathbf{x}]$.*

Answer. $\mathrm{E}[f(z)|\mathbf{x}] = 1+2\mathbf{x}+3(2-\mathbf{x}) = 7-\mathbf{x}$. If one plots z against $\mathbf{x}$ 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 $\boldsymbol{B}^* = \boldsymbol{\Omega}_{\mathbf{z}\mathbf{y}}\boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}}^-$. 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 $\boldsymbol{C}^* = \boldsymbol{\Omega}_{\mathbf{y}\mathbf{z}}\boldsymbol{\Omega}_{\mathbf{z}\mathbf{z}}^-$. $\boldsymbol{B}^*$ and $\boldsymbol{C}^*$ are connected by the formula

$$\boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}}\boldsymbol{B}^{*\top} = \boldsymbol{C}^*\boldsymbol{\Omega}_{\mathbf{z}\mathbf{z}}. \tag{20.1.22}$$

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 272. *In order to make relationship* (20.1.22) *more intuitive, assume $\mathbf{x}$ and ε are Normally distributed and independent of each other, and $\mathrm{E}[\varepsilon] = 0$. Define $\mathbf{y} = \alpha + \beta\mathbf{x} + \varepsilon$.*

• a. *Show that $\alpha + \beta\mathbf{x}$ is the best linear predictor of $\mathbf{y}$ based on the observation of $\mathbf{x}$.*

Answer. Follows from the fact that the predictor is unbiased and the prediction error is uncorrelated with $\mathbf{x}$. □

• b. *Express β in terms of the variances and covariances of $\mathbf{x}$ and $\mathbf{y}$.*

Answer. $\mathrm{cov}[\mathbf{x},\mathbf{y}] = \beta\,\mathrm{var}[\mathbf{x}]$, therefore $\beta = \frac{\mathrm{cov}[\mathbf{x},\mathbf{y}]}{\mathrm{var}[\mathbf{x}]}$ □

• c. *Since $\mathbf{x}$ and $\mathbf{y}$ are jointly normal, they can also be written $\mathbf{x} = \gamma + \delta\mathbf{y} + \omega$ where ω is independent of $\mathbf{y}$. Express δ in terms of the variances and covariances of $\mathbf{x}$ and $\mathbf{y}$, and show that $\mathrm{var}[\mathbf{y}]\beta = \gamma\,\mathrm{var}[\mathbf{x}]$.*

Answer. $\delta = \frac{\mathrm{cov}[\mathbf{x},\mathbf{y}]}{\mathrm{var}[\mathbf{y}]}$. □

• d. *Now let us extend the model a little: assume $\mathbf{x}_1$, $\mathbf{x}_2$, and ε are Normally distributed and independent of each other, and $\mathrm{E}[\varepsilon] = 0$. Define $\mathbf{y} = \alpha + \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \varepsilon$. Again express β_1 and β_2 in terms of variances and covariances of $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{y}$.*

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*

$$\begin{bmatrix}\delta_1\\ \delta_2\end{bmatrix}\operatorname{var}[\mathsf{y}] = \begin{bmatrix}\operatorname{var}[\mathsf{x}_1] & 0\\ 0 & \operatorname{var}[\mathsf{x}_2]\end{bmatrix}\begin{bmatrix}\beta_1\\ \beta_2\end{bmatrix} \tag{20.1.23}$$

This is (20.1.22) *in the present situation.*

Answer. $\delta_1 = \frac{\operatorname{cov}[\mathsf{x}_1,\mathsf{y}]}{\operatorname{var}[\mathsf{y}]}$ and $\delta_2 = \frac{\operatorname{cov}[\mathsf{x}_2,\mathsf{y}]}{\operatorname{var}[\mathsf{y}]}$. □

20.2. 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 $\boldsymbol{z}^* = \boldsymbol{\nu} + \boldsymbol{\Omega}_{\mathsf{zy}}\boldsymbol{\Omega}_{\mathsf{yy}}^-(\boldsymbol{y}-\boldsymbol{\mu})$ 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 $\sigma^2\begin{bmatrix}\boldsymbol{\Omega}_{\mathsf{yy}} & \boldsymbol{\Omega}_{\mathsf{yz}}\\ \boldsymbol{\Omega}_{\mathsf{zy}} & \boldsymbol{\Omega}_{\mathsf{zz}}\end{bmatrix}$.

In the case of singular $\boldsymbol{\Omega}_{\mathsf{zz}}$, 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 $\boldsymbol{z} = \boldsymbol{\nu} + \boldsymbol{\Omega}_{\mathsf{zz}}\boldsymbol{q}$ for some $\boldsymbol{q}$. We will also write $r = \operatorname{rank}\begin{bmatrix}\boldsymbol{\Omega}_{\mathsf{yy}} & \boldsymbol{\Omega}_{\mathsf{yz}}\\ \boldsymbol{\Omega}_{\mathsf{zy}} & \boldsymbol{\Omega}_{\mathsf{zz}}\end{bmatrix}$. Therefore, $\boldsymbol{z}^*$ solves the following "least squares problem:"

$$\boxed{\boldsymbol{z} = \boldsymbol{z}^* \quad \text{min.}\ \tfrac{1}{r\sigma^2}\begin{bmatrix}\boldsymbol{y}-\boldsymbol{\mu}\\ \boldsymbol{z}-\boldsymbol{\nu}\end{bmatrix}^\top \begin{bmatrix}\boldsymbol{\Omega}_{\mathsf{yy}} & \boldsymbol{\Omega}_{\mathsf{yz}}\\ \boldsymbol{\Omega}_{\mathsf{zy}} & \boldsymbol{\Omega}_{\mathsf{zz}}\end{bmatrix}^- \begin{bmatrix}\boldsymbol{y}-\boldsymbol{\mu}\\ \boldsymbol{z}-\boldsymbol{\nu}\end{bmatrix} \ \text{s.t.}\ \boldsymbol{z} = \boldsymbol{\nu} + \boldsymbol{\Omega}_{\mathsf{zz}}\boldsymbol{q} \text{ for some } \boldsymbol{q}.} \tag{20.2.1}$$

To prove this, use (A.8.2) to invert the dispersion matrix:

$$\begin{bmatrix}\boldsymbol{\Omega}_{\mathsf{yy}} & \boldsymbol{\Omega}_{\mathsf{yz}}\\ \boldsymbol{\Omega}_{\mathsf{zy}} & \boldsymbol{\Omega}_{\mathsf{zz}}\end{bmatrix}^- = \begin{bmatrix}\boldsymbol{\Omega}_{\mathsf{yy}}^- + \boldsymbol{\Omega}_{\mathsf{yy}}^-\boldsymbol{\Omega}_{\mathsf{yz}}\boldsymbol{\Omega}_{\mathsf{zz.y}}^-\boldsymbol{\Omega}_{\mathsf{zy}}\boldsymbol{\Omega}_{\mathsf{yy}}^- & -\boldsymbol{\Omega}_{\mathsf{yy}}^-\boldsymbol{\Omega}_{\mathsf{yz}}\boldsymbol{\Omega}_{\mathsf{zz.y}}^-\\ -\boldsymbol{\Omega}_{\mathsf{zz.y}}^-\boldsymbol{\Omega}_{\mathsf{zy}}\boldsymbol{\Omega}_{\mathsf{yy}}^- & \boldsymbol{\Omega}_{\mathsf{zz.y}}^-\end{bmatrix}. \tag{20.2.2}$$

If one plugs $\boldsymbol{z} = \boldsymbol{z}^*$ into this objective function, one obtains a very simple expression:

$$(\boldsymbol{y}-\boldsymbol{\mu})^\top\begin{bmatrix}\boldsymbol{I} & \boldsymbol{\Omega}_{\mathsf{yy}}^-\boldsymbol{\Omega}_{\mathsf{yz}}\end{bmatrix}\begin{bmatrix}\boldsymbol{\Omega}_{\mathsf{yy}}^- + \boldsymbol{\Omega}_{\mathsf{yy}}^-\boldsymbol{\Omega}_{\mathsf{yz}}\boldsymbol{\Omega}_{\mathsf{zz.y}}^-\boldsymbol{\Omega}_{\mathsf{zy}}\boldsymbol{\Omega}_{\mathsf{yy}}^- & -\boldsymbol{\Omega}_{\mathsf{yy}}^-\boldsymbol{\Omega}_{\mathsf{yz}}\boldsymbol{\Omega}_{\mathsf{zz.y}}^-\\ -\boldsymbol{\Omega}_{\mathsf{zz.y}}^-\boldsymbol{\Omega}_{\mathsf{zy}}\boldsymbol{\Omega}_{\mathsf{yy}}^- & \boldsymbol{\Omega}_{\mathsf{zz.y}}^-\end{bmatrix}\begin{bmatrix}\boldsymbol{I}\\ \boldsymbol{\Omega}_{\mathsf{zy}}\boldsymbol{\Omega}_{\mathsf{yy}}^-\end{bmatrix}(\boldsymbol{y}-\boldsymbol{\mu}) \tag{20.2.3}$$

$$= (\boldsymbol{y}-\boldsymbol{\mu})^\top\boldsymbol{\Omega}_{\mathsf{yy}}^-(\boldsymbol{y}-\boldsymbol{\mu}). \tag{20.2.4}$$

Now take any $\boldsymbol{z}$ of the form $\boldsymbol{z} = \boldsymbol{\nu} + \boldsymbol{\Omega}_{\mathsf{zz}}\boldsymbol{q}$ for some $\boldsymbol{q}$ and write it in the form $\boldsymbol{z} = \boldsymbol{z}^* + \boldsymbol{\Omega}_{\mathsf{zz}}\boldsymbol{d}$, i.e.,

$$\begin{bmatrix}\mathsf{y}-\boldsymbol{\mu}\\ \boldsymbol{z}-\boldsymbol{\nu}\end{bmatrix} = \begin{bmatrix}\mathsf{y}-\boldsymbol{\mu}\\ \boldsymbol{z}^*-\boldsymbol{\nu}\end{bmatrix} + \begin{bmatrix}\boldsymbol{o}\\ \boldsymbol{\Omega}_{\mathsf{zz}}\boldsymbol{d}\end{bmatrix}.$$

Then the cross product terms in the objective function disappear:

$$\begin{bmatrix}\boldsymbol{o}^\top & \boldsymbol{d}^\top\boldsymbol{\Omega}_{\mathsf{zz}}\end{bmatrix}\begin{bmatrix}\boldsymbol{\Omega}_{\mathsf{yy}}^- + \boldsymbol{\Omega}_{\mathsf{yy}}^-\boldsymbol{\Omega}_{\mathsf{yz}}\boldsymbol{\Omega}_{\mathsf{zz.y}}^-\boldsymbol{\Omega}_{\mathsf{zy}}\boldsymbol{\Omega}_{\mathsf{yy}}^- & -\boldsymbol{\Omega}_{\mathsf{yy}}^-\boldsymbol{\Omega}_{\mathsf{yz}}\boldsymbol{\Omega}_{\mathsf{zz.y}}^-\\ -\boldsymbol{\Omega}_{\mathsf{zz.y}}^-\boldsymbol{\Omega}_{\mathsf{zy}}\boldsymbol{\Omega}_{\mathsf{yy}}^- & \boldsymbol{\Omega}_{\mathsf{zz.y}}^-\end{bmatrix}\begin{bmatrix}\boldsymbol{I}\\ \boldsymbol{\Omega}_{\mathsf{zy}}\boldsymbol{\Omega}_{\mathsf{yy}}^-\end{bmatrix}(\boldsymbol{y}-\boldsymbol{\mu}) = \tag{20.2.5}$$

$$= \begin{bmatrix}\boldsymbol{o}^\top & \boldsymbol{d}^\top\boldsymbol{\Omega}_{\mathsf{zz}}\end{bmatrix}\begin{bmatrix}\boldsymbol{\Omega}_{\mathsf{yy}}^-\\ \boldsymbol{O}\end{bmatrix}(\boldsymbol{y}-\boldsymbol{\mu}) = 0$$

Therefore this gives a larger value of the objective function.

Problem 273. *Use problem 379 for an alternative proof of this.*

From (20.2.1) 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.

20.3. Prediction of Future Observations in the Regression Model

For a moment let us go back to the model $\mathsf{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ 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 $\mathsf{y}_0 = \boldsymbol{X}_0\boldsymbol{\beta} + \boldsymbol{\varepsilon}_0$ have not yet been observed and we want to predict them. The obvious predictor is $\mathsf{y}_0^* = \boldsymbol{X}_0\hat{\boldsymbol{\beta}} = \boldsymbol{X}_0(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\mathsf{y}$.

Since

$$\mathsf{y}_0^* - \mathsf{y}_0 = \boldsymbol{X}_0(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\mathsf{y} - \mathsf{y}_0 = \tag{20.3.1}$$

$$= \boldsymbol{X}_0(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{X}_0(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon} - \boldsymbol{X}_0\boldsymbol{\beta} - \boldsymbol{\varepsilon}_0 = \boldsymbol{X}_0(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}_0$$

one sees that $\mathrm{E}[\mathsf{y}_0^* - \mathsf{y}_0] = \boldsymbol{o}$, i.e., it is an unbiased predictor. And since $\boldsymbol{\varepsilon}$ and $\boldsymbol{\varepsilon}_0$ are uncorrelated, one obtains

$$\mathcal{MSE}[\mathsf{y}_0^*; \mathsf{y}_0] = \mathcal{V}[\mathsf{y}_0^* - \mathsf{y}_0] = \mathcal{V}[\boldsymbol{X}_0(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon}] + \mathcal{V}[\boldsymbol{\varepsilon}_0] \tag{20.3.2}$$

$$= \sigma^2(\boldsymbol{X}_0(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}_0^\top + \boldsymbol{I}). \tag{20.3.3}$$

Problem 274 shows that this is the Best Linear Unbiased Predictor (BLUP) of y_0 on the basis of y.

Problem 274. *The prediction problem in the Ordinary Least Squares model can be formulated as follows:*

$$\begin{bmatrix}\mathsf{y}\\ \mathsf{y}_0\end{bmatrix} = \begin{bmatrix}\boldsymbol{X}\\ \boldsymbol{X}_0\end{bmatrix}\boldsymbol{\beta} + \begin{bmatrix}\boldsymbol{\varepsilon}\\ \boldsymbol{\varepsilon}_0\end{bmatrix} \qquad \mathcal{E}[\begin{bmatrix}\boldsymbol{\varepsilon}\\ \boldsymbol{\varepsilon}_0\end{bmatrix}] = \begin{bmatrix}\boldsymbol{o}\\ \boldsymbol{o}\end{bmatrix} \qquad \mathcal{V}[\begin{bmatrix}\boldsymbol{\varepsilon}\\ \boldsymbol{\varepsilon}_0\end{bmatrix}] = \sigma^2\begin{bmatrix}\boldsymbol{I} & \boldsymbol{O}\\ \boldsymbol{O} & \boldsymbol{I}\end{bmatrix}. \tag{20.3.4}$$

$\boldsymbol{X}$ *and* $\boldsymbol{X}_0$ *are known,* y *is observed,* 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 18.1.1 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

$$\begin{bmatrix}\mathbf{y}\\ \mathbf{y}_0\end{bmatrix} = \begin{bmatrix}\boldsymbol{X}\\ \boldsymbol{X}_0\end{bmatrix}\boldsymbol{\beta} + \begin{bmatrix}\boldsymbol{\varepsilon}\\ \boldsymbol{\varepsilon}_0\end{bmatrix} \qquad \mathcal{E}\begin{bmatrix}\boldsymbol{\varepsilon}\\ \boldsymbol{\varepsilon}_0\end{bmatrix} = \begin{bmatrix}\boldsymbol{o}\\ \boldsymbol{o}\end{bmatrix} \qquad \mathcal{V}\begin{bmatrix}\boldsymbol{\varepsilon}\\ \boldsymbol{\varepsilon}_0\end{bmatrix} = \sigma^2\begin{bmatrix}\boldsymbol{\Psi} & \boldsymbol{C}\\ \boldsymbol{C}^\top & \boldsymbol{\Psi}_0\end{bmatrix}. \tag{20.3.5}$$

$\boldsymbol{X}$ 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

$$\mathbf{y}_0^* = \boldsymbol{X}_0\hat{\boldsymbol{\beta}} + \boldsymbol{C}^\top\boldsymbol{\Psi}^{-1}(\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}). \tag{20.3.6}$$

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

$$\sigma^2\Big(\boldsymbol{\Psi}_0 - \boldsymbol{C}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{C} + (\boldsymbol{X}_0 - \boldsymbol{C}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X}_0^\top - \boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{C})\Big). \tag{20.3.7}$$

Problem 275. *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 (19.0.7),

$$\mathbf{y}_0^* - \mathbf{y}_0 = \boldsymbol{X}_0\hat{\boldsymbol{\beta}} - \boldsymbol{X}_0\boldsymbol{\beta} + \boldsymbol{X}_0\boldsymbol{\beta} - \mathbf{y}_0 + \boldsymbol{C}^\top\boldsymbol{\Psi}^{-1}(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \tag{20.3.8}$$

$$= \boldsymbol{X}_0(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \boldsymbol{\varepsilon}_0 + \boldsymbol{C}^\top\boldsymbol{\Psi}^{-1}(\boldsymbol{\varepsilon} - \boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})) \tag{20.3.9}$$

$$= \boldsymbol{C}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\varepsilon} + (\boldsymbol{X}_0 - \boldsymbol{C}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \boldsymbol{\varepsilon}_0 \tag{20.3.10}$$

$$= \begin{bmatrix}\boldsymbol{C}^\top\boldsymbol{\Psi}^{-1} + (\boldsymbol{X}_0 - \boldsymbol{C}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1} & -\boldsymbol{I}\end{bmatrix}\begin{bmatrix}\boldsymbol{\varepsilon}\\ \boldsymbol{\varepsilon}_0\end{bmatrix} \tag{20.3.11}$$

The $\mathcal{MSE}$-matrix is therefore $\mathcal{E}[(\mathbf{y}_0^* - \mathbf{y}_0)(\mathbf{y}_0^* - \mathbf{y}_0)^\top] =$

$$\begin{aligned}
&= \sigma^2\begin{bmatrix}\boldsymbol{C}^\top\boldsymbol{\Psi}^{-1} + (\boldsymbol{X}_0 - \boldsymbol{C}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1} & -\boldsymbol{I}\end{bmatrix}\\
&\qquad\begin{bmatrix}\boldsymbol{\Psi} & \boldsymbol{C}\\ \boldsymbol{C}^\top & \boldsymbol{\Psi}_0\end{bmatrix}\begin{bmatrix}\boldsymbol{\Psi}^{-1}\boldsymbol{C} + \boldsymbol{\Psi}^{-1}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X}_0^\top - \boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{C})\\ -\boldsymbol{I}\end{bmatrix}
\end{aligned} \tag{20.3.12}$$

and the joint $\mathcal{MSE}$ matrix with the sampling error of the parameter vector $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ is

$$\sigma^2\begin{bmatrix}\boldsymbol{C}^\top\boldsymbol{\Psi}^{-1} + (\boldsymbol{X}_0 - \boldsymbol{C}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1} & -\boldsymbol{I}\\ (\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1} & \boldsymbol{O}\end{bmatrix} \tag{20.3.13}$$

$$\begin{bmatrix}\boldsymbol{\Psi} & \boldsymbol{C}\\ \boldsymbol{C}^\top & \boldsymbol{\Psi}_0\end{bmatrix}\begin{bmatrix}\boldsymbol{\Psi}^{-1}\boldsymbol{C} + \boldsymbol{\Psi}^{-1}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X}_0^\top - \boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{C}) & \boldsymbol{\Psi}^{-1}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\\ -\boldsymbol{I} & \boldsymbol{O}\end{bmatrix} =$$

$$\begin{aligned}
&= \sigma^2\begin{bmatrix}\boldsymbol{C}^\top\boldsymbol{\Psi}^{-1} + (\boldsymbol{X}_0 - \boldsymbol{C}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1} & -\boldsymbol{I}\\ (\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1} & \boldsymbol{O}\end{bmatrix}\\
&\begin{bmatrix}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X}_0^\top - \boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{C}) & \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\\ \boldsymbol{C}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{C} + \boldsymbol{C}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X}_0^\top - \boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{C}) - \boldsymbol{\Psi}_0 & \boldsymbol{C}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\end{bmatrix}
\end{aligned} \tag{20.3.14}$$

If one multiplies this out, one gets

$$\begin{bmatrix}\boldsymbol{\Psi}_0 - \boldsymbol{C}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{C} + (\boldsymbol{X}_0 - \boldsymbol{C}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X}_0^\top - \boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{C}) & (\boldsymbol{X}_0 - \boldsymbol{C}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\\ (\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X}_0^\top - \boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{C}) & (\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\end{bmatrix} \tag{20.3.15}$$

The upper left diagonal element is as claimed in (20.3.7). □

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 $\mathbf{y}$ and $\mathbf{y}_0$ will be transformed, but this transformation must have the following additional property: the transformed $\mathbf{y}$ must be a function of $\mathbf{y}$ alone, not of $\mathbf{y}_0$. Once such a transformation is found, it is easy to predict the transformed $\mathbf{y}_0$ on the basis of the transformed $\mathbf{y}$, and from this one also obtains a prediction of $\mathbf{y}_0$ on the basis of $\mathbf{y}$.

Here is some heuristics in order to understand formula (20.3.6). Assume for a moment that $\boldsymbol{\beta}$ was known. Then you can apply theorem **??** to the model

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{y}_0 \end{bmatrix} \sim \begin{bmatrix} \boldsymbol{X\beta} \\ \boldsymbol{X}_0\boldsymbol{\beta} \end{bmatrix}, \sigma^2 \begin{bmatrix} \boldsymbol{\Psi} & \boldsymbol{C} \\ \boldsymbol{C}^\top & \boldsymbol{\Psi}_0 \end{bmatrix} \tag{20.3.16}$$

to get $\mathbf{y}_0^* = \boldsymbol{X}_0\boldsymbol{\beta} + \boldsymbol{C}^\top\boldsymbol{\Psi}^{-1}(\mathbf{y} - \boldsymbol{X\beta})$ as best linear predictor of $\mathbf{y}_0$ on the basis of $\mathbf{y}$. 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 (20.3.6). This adds $\mathcal{MSE}[\boldsymbol{X}_0\hat{\boldsymbol{\beta}} + \boldsymbol{C}^\top\boldsymbol{\Psi}^{-1}(\mathbf{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}); \boldsymbol{X}_0\boldsymbol{\beta}+\boldsymbol{C}^\top\boldsymbol{\Psi}^{-1}(\mathbf{y}-\boldsymbol{X\beta})]$ to the $\mathcal{MSE}$-matrix, which gives (20.3.7).

Problem 276. *Show that*

$$\begin{aligned} \mathcal{MSE}[\boldsymbol{X}_0\hat{\boldsymbol{\beta}} + \boldsymbol{C}^\top\boldsymbol{\Psi}^{-1}(\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}); \boldsymbol{X}_0\boldsymbol{\beta} + \boldsymbol{C}^\top\boldsymbol{\Psi}^{-1}(\mathbf{y} - \boldsymbol{X\beta})] &= \\ = \sigma^2(\boldsymbol{X}_0 - \boldsymbol{C}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X}_0^\top - \boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{C}). \end{aligned} \tag{20.3.17}$$

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 277. *In the following we work with partitioned matrices. Given the model*

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{y}_0 \end{bmatrix} = \begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{X}_0 \end{bmatrix}\boldsymbol{\beta} + \begin{bmatrix} \boldsymbol{\varepsilon} \\ \boldsymbol{\varepsilon}_0 \end{bmatrix} \qquad \mathrm{E}[\begin{bmatrix} \boldsymbol{\varepsilon} \\ \boldsymbol{\varepsilon}_0 \end{bmatrix}] = \begin{bmatrix} \boldsymbol{o} \\ \boldsymbol{o} \end{bmatrix} \qquad \mathcal{V}[\begin{bmatrix} \boldsymbol{\varepsilon} \\ \boldsymbol{\varepsilon}_0 \end{bmatrix}] = \sigma^2 \begin{bmatrix} \boldsymbol{\Psi} & \boldsymbol{C} \\ \boldsymbol{C}^\top & \boldsymbol{\Psi}_0 \end{bmatrix}. \tag{20.3.18}$$

$\boldsymbol{X}$ has full rank. $\mathbf{y}$ is observed, $\mathbf{y}_0$ is not observed. $\boldsymbol{C}$ is not the null matrix.

• a. *Someone predicts $\mathbf{y}_0$ by $\mathbf{y}_0^* = \boldsymbol{X}_0\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\mathbf{y}$ is the BLUE of $\boldsymbol{\beta}$. Is this predictor unbiased?*

Answer. Yes, since $\mathcal{E}[\mathbf{y}_0] = \boldsymbol{X}_0\boldsymbol{\beta}$, and $\mathcal{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$. □

• b. *Compute the $\mathcal{MSE}$ matrix $\mathcal{MSE}[\boldsymbol{X}_0\hat{\boldsymbol{\beta}}; \mathbf{y}_0]$ of this predictor.* Hint: *For any matrix $\boldsymbol{B}$, the difference $\boldsymbol{B}\mathbf{y} - \mathbf{y}_0$ can be written in the form $\begin{bmatrix} \boldsymbol{B} & -\boldsymbol{I} \end{bmatrix}\begin{bmatrix} \mathbf{y} \\ \mathbf{y}_0 \end{bmatrix}$.* Hint: *For an unbiased predictor (or estimator), the $\mathcal{MSE}$ matrix is the covariance matrix of the prediction (or estimation) error.*

Answer.

$$\mathrm{E}[(\boldsymbol{B}\mathbf{y} - \mathbf{y}_0)(\boldsymbol{B}\mathbf{y} - \mathbf{y}_0)^\top] = \mathcal{V}[\boldsymbol{B}\mathbf{y} - \mathbf{y}_0] \tag{20.3.19}$$

$$= \mathcal{V}\left[\begin{bmatrix} \boldsymbol{B} & -\boldsymbol{I} \end{bmatrix}\begin{bmatrix} \mathbf{y} \\ \mathbf{y}_0 \end{bmatrix}\right] \tag{20.3.20}$$

$$= \sigma^2\begin{bmatrix} \boldsymbol{B} & -\boldsymbol{I} \end{bmatrix}\begin{bmatrix} \boldsymbol{\Psi} & \boldsymbol{C} \\ \boldsymbol{C}^\top & \boldsymbol{\Psi}_0 \end{bmatrix}\begin{bmatrix} \boldsymbol{B}^\top \\ -\boldsymbol{I} \end{bmatrix} \tag{20.3.21}$$

$$= \sigma^2\big(\boldsymbol{B\Psi B}^\top - \boldsymbol{C}^\top\boldsymbol{B}^\top - \boldsymbol{CB} + \boldsymbol{\Psi}_0\big). \tag{20.3.22}$$

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

(20.3.23)

$$\mathcal{MSE}[\boldsymbol{X}_0\hat{\boldsymbol{\beta}}; \mathbf{y}_0] = \sigma^2\Big(\boldsymbol{X}_0(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}_0^\top - \boldsymbol{C}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}_0^\top - \boldsymbol{X}_0(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top$$

□

• c. *Compare its $\mathcal{MSE}$-matrix with formula* (20.3.7). *Is the difference nonnegative definite?*

Answer. To compare it with the minimum $\mathcal{MSE}$ matrix, it can also be written as

(20.3.24)

$$\mathcal{MSE}[\boldsymbol{X}_0\hat{\boldsymbol{\beta}}; \mathbf{y}_0] = \sigma^2\Big(\boldsymbol{\Psi}_0 + (\boldsymbol{X}_0 - \boldsymbol{C}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X}_0^\top - \boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{C}) - \boldsymbol{C}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^-$$

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 $\boldsymbol{M} = \boldsymbol{M\Psi M}$, refer here to Problem 265. □

CHAPTER 21

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 21 we will see how estimates must be updated when more observations become available, in chapter 22 how the imposition of a linear constraint affects the parameter estimates, and in chapter 23 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 278. *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 $\mathbf{y}_0 = \boldsymbol{x}_0^\top\boldsymbol{\beta} + \varepsilon_0$ becomes available. Show that the updated estimator has the form*

$$\hat{\hat{\boldsymbol{\beta}}} = \hat{\boldsymbol{\beta}} + (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_0\frac{\mathbf{y}_0 - \boldsymbol{x}_0^\top\hat{\boldsymbol{\beta}}}{1 + \boldsymbol{x}_0^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_0}. \tag{21.0.25}$$

Answer. Set it up as follows:

$$\begin{bmatrix} \mathbf{y}_0 - \boldsymbol{x}_0^\top\hat{\boldsymbol{\beta}} \\ \boldsymbol{\beta} - \hat{\boldsymbol{\beta}} \end{bmatrix} \sim \begin{bmatrix} 0 \\ \boldsymbol{o} \end{bmatrix}, \sigma^2 \begin{bmatrix} \boldsymbol{x}_0^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_0 + 1 & \boldsymbol{x}_0^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1} \\ (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_0 & (\boldsymbol{X}^\top\boldsymbol{X})^{-1} \end{bmatrix} \tag{21.0.26}$$

and use (20.1.18). 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 (20.3.6)

$$\mathbf{y}_0^* = \boldsymbol{x}_0^\top\hat{\boldsymbol{\beta}} + \boldsymbol{c}^\top\boldsymbol{\Psi}^{-1}(\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \tag{21.0.27}$$

and from (20.3.15)

$$\begin{bmatrix} \mathbf{y}_0 - \mathbf{y}_0^* \\ \boldsymbol{\beta} - \hat{\boldsymbol{\beta}} \end{bmatrix} \sim \begin{bmatrix} 0 \\ \boldsymbol{o} \end{bmatrix}, \sigma^2 \tag{21.0.28}$$

$$\begin{bmatrix} \psi_0 - \boldsymbol{c}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{c} + (\boldsymbol{x}_0^\top - \boldsymbol{c}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{x}_0 - \boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{c}) & (\boldsymbol{x}_0^\top - \boldsymbol{c}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1} \\ (\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{x}_0 - \boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{c}) & (\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1} \end{bmatrix}$$

□

• a. *Show that the residual $\hat{\hat{\varepsilon}}_0$ from the full regression is the following nonrandom multiple of the "predictive" residual $\mathbf{y}_0 - \boldsymbol{x}_0^\top\hat{\boldsymbol{\beta}}$:*

$$\hat{\hat{\varepsilon}}_0 = \mathbf{y}_0 - \boldsymbol{x}_0^\top\hat{\hat{\boldsymbol{\beta}}} = \frac{1}{1 + \boldsymbol{x}_0^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_0}(\mathbf{y}_0 - \boldsymbol{x}_0^\top\hat{\boldsymbol{\beta}}) \tag{21.0.29}$$

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 (24.2.9).

Answer. (21.0.29) can either be derived from (21.0.25), 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 $\hat{\mathbf{y}}_0 = \boldsymbol{x}_0^\top\hat{\boldsymbol{\beta}}$. 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":

$$\begin{bmatrix} \hat{\mathbf{y}}_0 \\ \mathbf{y}_0 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}\boldsymbol{x}_0^\top\boldsymbol{\beta} + \begin{bmatrix} \delta_1 \\ \delta_2 \end{bmatrix} \qquad \begin{bmatrix} \delta_1 \\ \delta_2 \end{bmatrix} \sim \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \sigma^2 \begin{bmatrix} \boldsymbol{x}_0^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_0 & 0 \\ 0 & 1 \end{bmatrix}\right) \tag{21.0.30}$$

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 (21.0.30) is

(21.0.31)
$$\hat{\mathsf{y}}_0 = \left(\begin{bmatrix}1 & 1\end{bmatrix}\begin{bmatrix}\boldsymbol{x}_0^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_0 & 0\\ 0 & 1\end{bmatrix}^{-1}\begin{bmatrix}1\\1\end{bmatrix}\right)^{-1}\begin{bmatrix}1 & 1\end{bmatrix}\begin{bmatrix}\boldsymbol{x}_0^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_0 & 0\\ 0 & 1\end{bmatrix}^{-1}\begin{bmatrix}\hat{\mathsf{y}}_0\\ \mathsf{y}_0\end{bmatrix}$$

(21.0.32)
$$= \frac{1}{1+\frac{1}{\boldsymbol{x}_0^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_0}}\left(\frac{\hat{\mathsf{y}}_0}{\boldsymbol{x}_0^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_0}+\mathsf{y}_0\right)$$

(21.0.33)
$$= \frac{1}{1+\boldsymbol{x}_0^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_0}(\hat{\mathsf{y}}_0+\boldsymbol{x}_0^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_0\mathsf{y}_0).$$

Now subtract (21.0.33) from y_0 to get (21.0.29).

□

Using (21.0.29), one can write (21.0.25) as

$$\hat{\hat{\boldsymbol{\beta}}} = \hat{\boldsymbol{\beta}} + (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_0\hat{\hat{\varepsilon}}_0 \tag{21.0.34}$$

Later, in (25.4.1), one will see that it can also be written in the form

$$\hat{\hat{\boldsymbol{\beta}}} = \hat{\boldsymbol{\beta}} + (\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{x}_0(\mathsf{y}_0-\boldsymbol{x}_0^\top\hat{\boldsymbol{\beta}}) \tag{21.0.35}$$

where $\boldsymbol{Z} = \begin{bmatrix}\boldsymbol{X}\\ \boldsymbol{x}_0^\top\end{bmatrix}$.

Problem 279. *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

$$\mathsf{SSE}^* - \mathsf{SSE} = \frac{(\mathsf{y}_0-\boldsymbol{x}_0^\top\hat{\boldsymbol{\beta}})^2}{1+\boldsymbol{x}_0^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_0} \tag{21.0.36}$$

Answer. The sum of squared errors in the old regression is $\mathsf{SSE} = (\boldsymbol{\mathsf{y}}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\boldsymbol{\mathsf{y}}-\boldsymbol{X}\hat{\boldsymbol{\beta}})$; the sum of squared errors in the updated regression is $\mathsf{SSE}^* = (\boldsymbol{\mathsf{y}}-\boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}})^\top(\boldsymbol{\mathsf{y}}-\boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}})+\hat{\hat{\varepsilon}}_0{}^2$. From (21.0.34) follows

$$\boldsymbol{\mathsf{y}}-\boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}} = \boldsymbol{\mathsf{y}}-\boldsymbol{X}\hat{\boldsymbol{\beta}}-\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_0\hat{\hat{\varepsilon}}_0. \tag{21.0.37}$$

If one squares this, the cross product terms fall away: $(\boldsymbol{\mathsf{y}}-\boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}})^\top(\boldsymbol{\mathsf{y}}-\boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}}) = (\boldsymbol{\mathsf{y}}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\boldsymbol{\mathsf{y}}-\boldsymbol{X}\hat{\boldsymbol{\beta}})+\hat{\hat{\varepsilon}}_0\boldsymbol{x}_0^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_0\hat{\hat{\varepsilon}}_0$. 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 (21.0.29) to get (21.0.36). □

CHAPTER 22

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.

22.1. Building the Constraint into the Model

Problem 280. *Given a regression with a constant term and two explanatory variables which we will call x and z, i.e.,*

$$\mathrm{y}_t = \alpha + \beta x_t + \gamma z_t + \varepsilon_t \tag{22.1.1}$$

• a. *1 point How will you estimate β and γ if it is known that $\beta = \gamma$?*

Answer. Write

$$\mathrm{y}_t = \alpha + \beta(x_t + z_t) + \varepsilon_t \tag{22.1.2}$$

□

• 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

$$\mathrm{y}_t - z_t = \alpha + \beta(x_t - z_t) + \varepsilon_t \tag{22.1.3}$$

□

• 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* (22.1.1)*, and it allows you to* test *the constraint.*

Answer. It you add z as additional regressor into (22.1.2), you get $\mathrm{y}_t = \alpha + \beta(x_t + z_t) + \delta z_t + \varepsilon_t$. Now substitute the right hand side from (22.1.1) 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 280 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 219 is another example here.

22.2. Conversion of an Arbitrary Constraint into a Zero Constraint

This section, which is nothing but the matrix version of Problem 280, 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

$$\begin{bmatrix}\boldsymbol{R}_1 & \boldsymbol{R}_2\end{bmatrix}\begin{bmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\end{bmatrix} \equiv \boldsymbol{R}_1\boldsymbol{\beta}_1 + \boldsymbol{R}_2\boldsymbol{\beta}_2 = \boldsymbol{u} \tag{22.2.1}$$

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

$$\mathbf{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon} \tag{22.2.2}$$

Now one can solve (22.2.1) for $\boldsymbol{\beta}_1$ to get

$$\boldsymbol{\beta}_1 = \boldsymbol{R}_1^{-1}\boldsymbol{u} - \boldsymbol{R}_1^{-1}\boldsymbol{R}_2\boldsymbol{\beta}_2 \tag{22.2.3}$$

Plug (22.2.3) into (22.2.2) and rearrange to get a regression which is equivalent to the constrained regression:

$$\boldsymbol{y} - \boldsymbol{X}_1\boldsymbol{R}_1^{-1}\boldsymbol{u} = (\boldsymbol{X}_2 - \boldsymbol{X}_1\boldsymbol{R}_1^{-1}\boldsymbol{R}_2)\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon} \tag{22.2.4}$$

or

$$\boldsymbol{y}^* = \boldsymbol{Z}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon} \tag{22.2.5}$$

One more thing is noteworthy here: if we add $\boldsymbol{X}_1$ as additional regressors into (22.2.5), we get a regression that is equivalent to (22.2.2). To see this, define the difference between the left hand side and right hand side of (22.2.3) 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 (22.2.1) is equivalent to the "zero constraint" $\boldsymbol{\gamma}_1 = \boldsymbol{o}$, and the regression

$$\boldsymbol{y} - \boldsymbol{X}_1\boldsymbol{R}_1^{-1}\boldsymbol{u} = (\boldsymbol{X}_2 - \boldsymbol{X}_1\boldsymbol{R}_1^{-1}\boldsymbol{R}_2)\boldsymbol{\beta}_2 + \boldsymbol{X}_1(\boldsymbol{\beta}_1 - \boldsymbol{R}_1^{-1}\boldsymbol{u} + \boldsymbol{R}_1^{-1}\boldsymbol{R}_2\boldsymbol{\beta}_2) + \boldsymbol{\varepsilon} \tag{22.2.6}$$

is equivalent to the original regression (22.2.2). (22.2.6) can also be written as

$$\boldsymbol{y}^* = \boldsymbol{Z}_2\boldsymbol{\beta}_2 + \boldsymbol{X}_1\boldsymbol{\gamma}_1 + \boldsymbol{\varepsilon} \tag{22.2.7}$$

The coefficient of $\boldsymbol{X}_1$, if it is added back into (22.2.5), is therefore $\boldsymbol{\gamma}_1$.

Problem 281. [**DM93**] *assert on p. 17, middle, that*

$$\mathrm{R}[\boldsymbol{X}_1, \boldsymbol{Z}_2] = \mathrm{R}[\boldsymbol{X}_1, \boldsymbol{X}_2]. \tag{22.2.8}$$

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

$$\{\boldsymbol{z}\colon \boldsymbol{z} = \boldsymbol{X}_1\boldsymbol{\gamma} + \boldsymbol{X}_2\boldsymbol{\delta}\} = \{\boldsymbol{z}\colon \boldsymbol{z} = \boldsymbol{X}_1\boldsymbol{\alpha} + \boldsymbol{Z}_2\boldsymbol{\beta}\} \tag{22.2.9}$$

First $\subset$: given $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$ we need a $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ with

$$\boldsymbol{X}_1\boldsymbol{\gamma} + \boldsymbol{X}_2\boldsymbol{\delta} = \boldsymbol{X}_1\boldsymbol{\alpha} + (\boldsymbol{X}_2 - \boldsymbol{X}_1\boldsymbol{R}_1^{-1}\boldsymbol{R}_2)\boldsymbol{\beta} \tag{22.2.10}$$

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 (22.2.10) 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 $\boldsymbol{\gamma} = \boldsymbol{\alpha} - \boldsymbol{R}_1^{-1}\boldsymbol{R}_2\boldsymbol{\beta}$. □

22.3. Lagrange Approach to Constrained Least Squares

The constrained least squares estimator is that $k \times 1$ vector $\boldsymbol{\beta} = \hat{\hat{\boldsymbol{\beta}}}$ 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":

$$\mathsf{L}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = (\mathsf{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\mathsf{y} - \boldsymbol{X}\boldsymbol{\beta}) + (\boldsymbol{R}\boldsymbol{\beta} - \boldsymbol{u})^\top\boldsymbol{\lambda} \tag{22.3.1}$$

$\boldsymbol{\lambda}$ 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 $\boldsymbol{\beta} = \hat{\hat{\boldsymbol{\beta}}}$ satisfies the constraint. This $\hat{\hat{\boldsymbol{\beta}}}$ is the solution of the constrained minimization problem we are looking for.

Problem 282. *4 points Show the following: If* $\boldsymbol{\beta} = \hat{\hat{\boldsymbol{\beta}}}$ *is the unconstrained minimum argument of the Lagrange function*

$$\mathsf{L}(\boldsymbol{\beta}, \boldsymbol{\lambda}^*) = (\mathsf{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\mathsf{y} - \boldsymbol{X}\boldsymbol{\beta}) + (\boldsymbol{R}\boldsymbol{\beta} - \boldsymbol{u})^\top\boldsymbol{\lambda}^* \tag{22.3.2}$$

for some fixed value $\boldsymbol{\lambda}^*$, *and if at the same time* $\hat{\hat{\boldsymbol{\beta}}}$ *satisfies* $\boldsymbol{R}\hat{\hat{\boldsymbol{\beta}}} = \boldsymbol{u}$, *then* $\boldsymbol{\beta} = \hat{\hat{\boldsymbol{\beta}}}$ *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 $\hat{\hat{\boldsymbol{\beta}}}$ minimizes the Lagrange function, we know that

$$(\mathsf{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}})^\top(\mathsf{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}) + (\boldsymbol{R}\tilde{\boldsymbol{\beta}} - \boldsymbol{u})^\top\boldsymbol{\lambda}^* \geq (\mathsf{y} - \boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}})^\top(\mathsf{y} - \boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}}) + (\boldsymbol{R}\hat{\hat{\boldsymbol{\beta}}} - \boldsymbol{u})^\top\boldsymbol{\lambda}^* \tag{22.3.3}$$

for all $\tilde{\boldsymbol{\beta}}$. Since by assumption, $\hat{\hat{\boldsymbol{\beta}}}$ also satisfies the constraint, this simplifies to:

$$(\mathsf{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}})^\top(\mathsf{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}) + (\boldsymbol{R}\tilde{\boldsymbol{\beta}} - \boldsymbol{u})^\top\boldsymbol{\lambda}^* \geq (\mathsf{y} - \boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}})^\top(\mathsf{y} - \boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}}). \tag{22.3.4}$$

This is still true for all $\tilde{\boldsymbol{\beta}}$. If we only look at those $\tilde{\boldsymbol{\beta}}$ which satisfy the constraint, we get

$$(\mathsf{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}})^\top(\mathsf{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}) \geq (\mathsf{y} - \boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}})^\top(\mathsf{y} - \boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}}). \tag{22.3.5}$$

This means, $\hat{\hat{\boldsymbol{\beta}}}$ 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 $\hat{\hat{\boldsymbol{\beta}}}$ but also shows that there is always a $\boldsymbol{\lambda}^*$ for which $\hat{\hat{\boldsymbol{\beta}}}$ satisfies $\boldsymbol{R}\hat{\hat{\boldsymbol{\beta}}} = \boldsymbol{u}$.

Problem 283. *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*

$$\mathsf{L}(\boldsymbol{\beta}) = (\mathsf{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\mathsf{y} - \boldsymbol{X}\boldsymbol{\beta}) + (\boldsymbol{R}\boldsymbol{\beta} - \boldsymbol{u})^\top\boldsymbol{\lambda} \tag{22.3.6}$$

Answer. Write the objective function as $\mathsf{y}^\top\mathsf{y} - 2\mathsf{y}^\top\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\lambda}^\top\boldsymbol{R}\boldsymbol{\beta} - \boldsymbol{\lambda}^\top\boldsymbol{u}$ to get (22.3.7). □

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

$$\partial\mathsf{L}/\partial\boldsymbol{\beta}^\top = -2\mathsf{y}^\top\boldsymbol{X} + 2\boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{X} + \boldsymbol{\lambda}^{*\top}\boldsymbol{R}, \tag{22.3.7}$$

and (b) $\hat{\hat{\boldsymbol{\beta}}}$ must satisfy the constraint (22.3.9).

(22.3.7) and (22.3.9) are two linear matrix equations which can indeed be solved for $\hat{\hat{\boldsymbol{\beta}}}$ and $\boldsymbol{\lambda}^*$. I wrote (22.3.7) 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 (22.3.7) with its transpose (22.3.8). Our starting point is therefore

$$2\boldsymbol{X}^\top\boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}} = 2\boldsymbol{X}^\top\mathsf{y} - \boldsymbol{R}^\top\boldsymbol{\lambda}^* \tag{22.3.8}$$

$$\boldsymbol{R}\hat{\hat{\boldsymbol{\beta}}} - \boldsymbol{u} = \boldsymbol{o} \tag{22.3.9}$$

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 (22.3.8) for $\hat{\hat{\boldsymbol{\beta}}}$ to get that $\hat{\hat{\boldsymbol{\beta}}}$ which minimizes L for any given $\boldsymbol{\lambda}^*$:

$$\hat{\hat{\boldsymbol{\beta}}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\mathbf{y} - \frac{1}{2}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\boldsymbol{\lambda}^* = \hat{\boldsymbol{\beta}} - \frac{1}{2}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\boldsymbol{\lambda}^* \tag{22.3.10}$$

Here $\hat{\boldsymbol{\beta}}$ on the right hand side is the *unconstrained* OLS estimate. Plug this formula for $\hat{\hat{\boldsymbol{\beta}}}$ into (22.3.9) in order to determine that value of $\boldsymbol{\lambda}^*$ for which the corresponding $\hat{\hat{\boldsymbol{\beta}}}$ satisfies the constraint:

$$\boldsymbol{R}\hat{\boldsymbol{\beta}} - \frac{1}{2}\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\boldsymbol{\lambda}^* - \boldsymbol{u} = \boldsymbol{o}. \tag{22.3.11}$$

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 284). Therefore one can solve for $\boldsymbol{\lambda}^*$:

$$\boldsymbol{\lambda}^* = 2\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{u}) \tag{22.3.12}$$

If one substitutes this $\boldsymbol{\lambda}^*$ back into (22.3.10), one gets the formula for the constrained least squares estimator:

$$\hat{\hat{\boldsymbol{\beta}}} = \hat{\boldsymbol{\beta}} - (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{u}). \tag{22.3.13}$$

Problem 284. *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. $\boldsymbol{b}^\top\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-}$ 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 285. *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* (22.3.13)*:*

$$\hat{\hat{\boldsymbol{\beta}}} = \hat{\boldsymbol{\beta}} - (\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{u}) \tag{22.3.14}$$

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

$$2\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}} = 2\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\mathbf{y} - \boldsymbol{R}^\top\boldsymbol{\lambda}^* \tag{22.3.15}$$

Solve (22.3.15) for $\hat{\hat{\boldsymbol{\beta}}}$:

$$\hat{\hat{\boldsymbol{\beta}}} = (\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\mathbf{y} - \frac{1}{2}(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{R}^\top\boldsymbol{\lambda}^* = \hat{\boldsymbol{\beta}} - \frac{1}{2}(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{R}^\top\boldsymbol{\lambda}^* \tag{22.3.16}$$

Here $\hat{\boldsymbol{\beta}}$ is the unconstrained GLS estimate. Plug $\hat{\hat{\boldsymbol{\beta}}}$ into the constraint (22.3.9):

$$\boldsymbol{R}\hat{\boldsymbol{\beta}} - \frac{1}{2}\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{R}^\top\boldsymbol{\lambda}^* - \boldsymbol{u} = \boldsymbol{o}. \tag{22.3.17}$$

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

$$\boldsymbol{\lambda}^* = 2\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{u}) \tag{22.3.18}$$

Now substitute this $\boldsymbol{\lambda}^*$ back into (22.3.16):

$$\hat{\hat{\boldsymbol{\beta}}} = \hat{\boldsymbol{\beta}} - (\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{u}). \tag{22.3.19}$$

□

22.4. 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, $\boldsymbol{R}\boldsymbol{\beta}$, is observed without error.

Problem 286. *Assume the random $\boldsymbol{\beta} \sim (\hat{\boldsymbol{\beta}}, \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1})$ is unobserved, but one observes $\boldsymbol{R}\boldsymbol{\beta} = \mathbf{u}$.*

• 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.

$$\begin{bmatrix}\mathbf{u}\\ \boldsymbol{\beta}\end{bmatrix} \sim \left(\begin{bmatrix}\boldsymbol{R}\hat{\boldsymbol{\beta}}\\ \hat{\boldsymbol{\beta}}\end{bmatrix}, \sigma^2\begin{bmatrix}\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top & \boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\\ (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top & (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\end{bmatrix}\right). \tag{22.4.1}$$

Therefore application of formula (**??**) gives

$$\boldsymbol{\beta}^* = \hat{\boldsymbol{\beta}} + (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}(\mathbf{u} - \boldsymbol{R}\hat{\boldsymbol{\beta}}). \tag{22.4.2}$$

□

• 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 $\boldsymbol{u} = \boldsymbol{R}\boldsymbol{\beta}$, the best bounded $\mathcal{MSE}$ estimators of $\boldsymbol{u}$ and $\boldsymbol{\beta}$ are $\boldsymbol{R}\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}$, with the sampling errors having the following means and variances:

$$\begin{bmatrix}\boldsymbol{u} - \boldsymbol{R}\hat{\boldsymbol{\beta}}\\ \boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\end{bmatrix} \sim \left(\begin{bmatrix}\boldsymbol{o}\\ \boldsymbol{o}\end{bmatrix}, \sigma^2\begin{bmatrix}\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top & \boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\\ (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top & (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\end{bmatrix}\right) \tag{22.4.3}$$

After the observation of $\boldsymbol{u}$ we can therefore apply (20.1.18) to get exactly equation (22.3.13) 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.

22.5. 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 (14.2.12) which can be used to show optimality of the OLS estimate:

$$(\mathbf{y}-\boldsymbol{X}\boldsymbol{\beta})^\top(\mathbf{y}-\boldsymbol{X}\boldsymbol{\beta}) = (\mathbf{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\mathbf{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}) + (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^\top\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}).$$

Split the second term again, using $\hat{\boldsymbol{\beta}}-\hat{\hat{\boldsymbol{\beta}}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{u})$:

$$\begin{aligned}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^\top\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) &= \big(\boldsymbol{\beta}-\hat{\hat{\boldsymbol{\beta}}}-(\hat{\boldsymbol{\beta}}-\hat{\hat{\boldsymbol{\beta}}})\big)^\top\boldsymbol{X}^\top\boldsymbol{X}\big(\boldsymbol{\beta}-\hat{\hat{\boldsymbol{\beta}}}-(\hat{\boldsymbol{\beta}}-\hat{\hat{\boldsymbol{\beta}}})\big)\\ &= (\boldsymbol{\beta}-\hat{\hat{\boldsymbol{\beta}}})^\top\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{\beta}-\hat{\hat{\boldsymbol{\beta}}})\\ &\quad -2(\boldsymbol{\beta}-\hat{\hat{\boldsymbol{\beta}}})^\top\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\\ &\quad +(\hat{\boldsymbol{\beta}}-\hat{\hat{\boldsymbol{\beta}}})^\top\boldsymbol{X}^\top\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\hat{\hat{\boldsymbol{\beta}}}).\end{aligned}$$

The cross product terms can be simplified to $-2(\boldsymbol{R}\boldsymbol{\beta}-\boldsymbol{u})^\top\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{u})$, and the last term is $(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{u})^\top\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{u})$. Therefore the objective function for an arbitrary $\boldsymbol{\beta}$ can be written as

$$\begin{aligned}(\mathbf{y}-\boldsymbol{X}\boldsymbol{\beta})^\top(\mathbf{y}-\boldsymbol{X}\boldsymbol{\beta}) &= (\mathbf{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\mathbf{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})\\ &\quad +(\boldsymbol{\beta}-\hat{\hat{\boldsymbol{\beta}}})^\top\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{\beta}-\hat{\hat{\boldsymbol{\beta}}})\\ &\quad -2(\boldsymbol{R}\boldsymbol{\beta}-\boldsymbol{u})^\top\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{u})\\ &\quad +(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{u})^\top\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{u})\end{aligned}$$

The first and last terms do not depend on $\boldsymbol{\beta}$ at all; the third term is zero whenever $\boldsymbol{\beta}$ satisfies $\boldsymbol{R}\boldsymbol{\beta}=\boldsymbol{u}$; 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.

22.6. 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 $\boldsymbol{R}\boldsymbol{\beta}=\boldsymbol{u}$, but also in

the case that it does not. For this, let us first get a suitable representation of the sampling error:

$$\begin{aligned}\hat{\hat{\boldsymbol{\beta}}}-\boldsymbol{\beta} &= (\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})+(\hat{\hat{\boldsymbol{\beta}}}-\hat{\boldsymbol{\beta}}) =\\ &= (\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})-(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}\boldsymbol{R}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\\ &\quad -(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\boldsymbol{\beta}-\boldsymbol{u}).\end{aligned} \tag{22.6.1}$$

The last term is zero if $\boldsymbol{\beta}$ satisfies the constraint. Now use (18.0.7) twice to get

$$\hat{\hat{\boldsymbol{\beta}}}-\boldsymbol{\beta} = \boldsymbol{W}\boldsymbol{X}^\top\boldsymbol{\varepsilon}-(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\boldsymbol{\beta}-\boldsymbol{u}) \tag{22.6.2}$$

where

$$\boldsymbol{W} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}-(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}. \tag{22.6.3}$$

If $\boldsymbol{\beta}$ satisfies the constraint, (22.6.2) 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 287). 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 287.

- 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

$$\mathcal{E}[\hat{\hat{\boldsymbol{\beta}}}-\boldsymbol{\beta}] = -(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\boldsymbol{\beta}-\boldsymbol{u}). \tag{22.6.4}$$

Due to the decomposition (17.1.2) of the $\mathcal{MSE}$ matrix into dispersion matrix plus squared bias, it follows

$$\begin{aligned}\mathcal{MSE}[\hat{\hat{\boldsymbol{\beta}}};\boldsymbol{\beta}] &= \sigma^2\boldsymbol{W}+\\ &\quad +(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\boldsymbol{\beta}-\boldsymbol{u})\cdot\\ &\quad \cdot(\boldsymbol{R}\boldsymbol{\beta}-\boldsymbol{u})^\top\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}.\end{aligned} \tag{22.6.5}$$

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

$$(\boldsymbol{R\beta}-\boldsymbol{u})^\top\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R\beta}-\boldsymbol{u})\leq\sigma^2. \tag{22.6.6}$$

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 (22.6.6) is i times the value of the F-test statistic for the hypothesis $\boldsymbol{R\beta}=\boldsymbol{u}$. To test for this, one has to use the *noncentral* F-test with parameters i, $n-k$, and $1/2$.

PROBLEM 288. *2 points This Problem motivates Equation* (22.6.6). *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* $\boldsymbol{R\beta}$ *than* $\boldsymbol{R}\hat{\boldsymbol{\beta}}$. *Show that this latter condition is not only necessary but already sufficient, i.e., if* $\mathcal{MSE}[\boldsymbol{R}\hat{\boldsymbol{\beta}};\boldsymbol{R\beta}]-\mathcal{MSE}[\boldsymbol{u};\boldsymbol{R\beta}]$ *is nonnegative definite then* $\boldsymbol{\beta}$ *and* σ^2 *satisfy* (22.6.6). *You are allowed to use, without proof, theorem A.5.9 in the mathematical Appendix.*

ANSWER. We have to show

$$\sigma^2\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top-(\boldsymbol{R\beta}-\boldsymbol{u})(\boldsymbol{R\beta}-\boldsymbol{u})^\top \tag{22.6.7}$$

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 (22.6.6). □

22.7. 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 $\hat{\hat{\boldsymbol{\varepsilon}}}=\mathbf{y}-\boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}}$, again without necessarily making the assumption that $\boldsymbol{R\beta}=\boldsymbol{u}$:

$$\hat{\hat{\boldsymbol{\varepsilon}}}=\mathbf{y}-\boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}}=\hat{\boldsymbol{\varepsilon}}+\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{u}). \tag{22.7.1}$$

Since $\boldsymbol{X}^\top\hat{\boldsymbol{\varepsilon}}=\boldsymbol{o}$, it follows

$$\hat{\hat{\boldsymbol{\varepsilon}}}^\top\hat{\hat{\boldsymbol{\varepsilon}}}=\hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}}+(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{u})^\top\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{u}). \tag{22.7.2}$$

Now note that $\mathcal{E}[\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{u}]=\boldsymbol{R\beta}-\boldsymbol{u}$ 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 (**??**) in theorem **??** 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

$$\begin{aligned}\mathrm{E}[(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{u})^\top\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{u})]&=\\ &=\sigma^2 i+(\boldsymbol{R\beta}-\boldsymbol{u})^\top\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R\beta}-\boldsymbol{u})\end{aligned} \tag{22.7.3}$$

Since $\mathrm{E}[\hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}}]=\sigma^2(n-k)$, it follows

$$\mathrm{E}[\hat{\hat{\boldsymbol{\varepsilon}}}^\top\hat{\hat{\boldsymbol{\varepsilon}}}]=\sigma^2(n+i-k)+(\boldsymbol{R\beta}-\boldsymbol{u})^\top\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R\beta}-\boldsymbol{u}). \tag{22.7.4}$$

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 22.2), 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

$$\hat{\hat{\sigma}}^2=\hat{\hat{\boldsymbol{\varepsilon}}}^\top\hat{\hat{\boldsymbol{\varepsilon}}}/(n+i-k) \tag{22.7.5}$$

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 $\boldsymbol{X\beta}_0$ where $\boldsymbol{\beta}_0$ satisfies the constraint:

PROBLEM 289. *3 points Assume* $\hat{\hat{\boldsymbol{\beta}}}$ *is the constrained least squares estimate, and* $\boldsymbol{\beta}_0$ *is any vector satisfying* $\boldsymbol{R\beta}_0=\boldsymbol{u}$. *Show that in the decomposition*

$$\mathbf{y}-\boldsymbol{X\beta}_0=\boldsymbol{X}(\hat{\hat{\boldsymbol{\beta}}}-\boldsymbol{\beta}_0)+\hat{\hat{\boldsymbol{\varepsilon}}} \tag{22.7.6}$$

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 (22.3.13),

$$\begin{aligned}(\hat{\hat{\boldsymbol{\beta}}}-\boldsymbol{\beta}_0)^\top\boldsymbol{X}^\top\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\hat{\hat{\boldsymbol{\beta}}})&=(\hat{\hat{\boldsymbol{\beta}}}-\boldsymbol{\beta}_0)^\top\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{u})\\ &=(\boldsymbol{u}-\boldsymbol{u})^\top\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{u})=0.\end{aligned}$$

□

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 290. *Assume* $\hat{\hat{\boldsymbol{\beta}}}$ *is the constrained least squares estimator subject to the constraint* $\boldsymbol{R\beta}=\boldsymbol{o}$, *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*

$$\mathbf{y}=\hat{\hat{\mathbf{y}}}+(\hat{\mathbf{y}}-\hat{\hat{\mathbf{y}}})+\hat{\boldsymbol{\varepsilon}} \tag{22.7.7}$$

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* (22.7.7) *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* (22.3.13)*:*

Answer. One has to verify that the scalar products of the three vectors on the right hand side of (22.7.7) are zero. $\hat{\mathbf{y}}^\top\hat{\boldsymbol{\varepsilon}} = \hat{\hat{\boldsymbol{\beta}}}^\top \boldsymbol{X}^\top\hat{\boldsymbol{\varepsilon}} = 0$ 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 (22.3.13) for $\hat{\boldsymbol{\beta}}-\hat{\hat{\boldsymbol{\beta}}}$,

$$\begin{aligned}\hat{\hat{\mathbf{y}}}^\top(\hat{\mathbf{y}}-\hat{\hat{\mathbf{y}}}) &= \hat{\hat{\boldsymbol{\beta}}}^\top\boldsymbol{X}^\top\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\hat{\hat{\boldsymbol{\beta}}}) = \\ &= \hat{\hat{\boldsymbol{\beta}}}^\top\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}\boldsymbol{R}\hat{\boldsymbol{\beta}} = \\ &= \hat{\hat{\boldsymbol{\beta}}}^\top\boldsymbol{R}^\top\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}\boldsymbol{R}\hat{\boldsymbol{\beta}} = 0\end{aligned}$$

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 291.

• 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 (22.3.13). 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 $\hat{\hat{\boldsymbol{\varepsilon}}} = \boldsymbol{\iota}\bar{\mathbf{y}}$, i.e., $\hat{\hat{\boldsymbol{\varepsilon}}} = [0.5,0.5,0.5,0.5]$. Since one has n observations, n parameters and 1 constraint, the number of degrees of freedom is 1. Therefore $\hat{\hat{\sigma}}^2 = \hat{\hat{\boldsymbol{\varepsilon}}}^\top\hat{\hat{\boldsymbol{\varepsilon}}}/1 = n\bar{\mathbf{y}}^2$ 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 $\mathbf{y}^\top = \begin{bmatrix}-1 & 0 & 1 & 2\end{bmatrix}$. 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 292. *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*

$$\hat{\hat{\boldsymbol{\beta}}} = \hat{\boldsymbol{\beta}} - \boldsymbol{V}\boldsymbol{r}\big(\boldsymbol{r}^\top\boldsymbol{V}\boldsymbol{r}\big)^{-1}(\boldsymbol{r}^\top\hat{\boldsymbol{\beta}} - u). \tag{22.7.8}$$

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*

$$\frac{\boldsymbol{t}^\top(\hat{\boldsymbol{\beta}}-\hat{\hat{\boldsymbol{\beta}}})}{\sigma\sqrt{\boldsymbol{t}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{t}}}. \tag{22.7.9}$$

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 (22.7.8) and equation (25.4.1) one obtains

$$\max_{\boldsymbol{t}}\frac{(\boldsymbol{t}^\top(\hat{\boldsymbol{\beta}}-\hat{\hat{\boldsymbol{\beta}}}))^2}{\sigma^2\boldsymbol{t}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{t}} = \frac{(\hat{\boldsymbol{\beta}}-\hat{\hat{\boldsymbol{\beta}}})^\top\boldsymbol{X}^\top\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\hat{\hat{\boldsymbol{\beta}}})}{\sigma^2} =$$

$$= \frac{(\boldsymbol{r}^\top\hat{\boldsymbol{\beta}}-u)^\top\big(\boldsymbol{r}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{r}\big)^{-1}(\boldsymbol{r}^\top\hat{\boldsymbol{\beta}}-u)}{\sigma^2} = \frac{(\boldsymbol{r}^\top\hat{\boldsymbol{\beta}}-u)^2}{\sigma^2\boldsymbol{r}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{r}}$$

□

22.8. 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.

22.9. 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 23

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

$$\mathbf{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon} = \begin{bmatrix}\boldsymbol{X}_1 & \boldsymbol{X}_2\end{bmatrix}\begin{bmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\end{bmatrix} + \boldsymbol{\varepsilon} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim (\boldsymbol{o}, \sigma^2\boldsymbol{I}) \tag{23.0.1}$$

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 (**??**).
- 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 293. *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 294. *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 $\mathbf{y} = \hat{\mathbf{y}} + \hat{\hat{\boldsymbol{\varepsilon}}}$ 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 $\mathbf{y} = \hat{\mathbf{y}} + \hat{\hat{\boldsymbol{\varepsilon}}}$ we have to check two things: (1) $\hat{\mathbf{y}}$ 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 295. *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 $\mathbf{y} = \hat{\mathbf{y}} + \hat{\hat{\boldsymbol{\varepsilon}}}$ we have to check two things: (1) $\hat{\mathbf{y}}$ 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 293 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

$$\boldsymbol{M}_1\mathbf{y} = \boldsymbol{M}_1\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 + \hat{\boldsymbol{\varepsilon}} \tag{23.0.2}$$

In formulas, the correct $\hat{\boldsymbol{\beta}}_2$ is

$$\hat{\boldsymbol{\beta}}_2 = (\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top\boldsymbol{M}_1\mathbf{y}. \tag{23.0.3}$$

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 296. *2 points Show that in the model* (23.0.1), *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

$$\hat{\boldsymbol{\beta}}_1 = (\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top(\mathbf{y} - \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2). \tag{23.0.4}$$

This regression also gives the right residuals, but not the right estimates of the covariance matrix.

PROBLEM 297. *The three Figures in* [**DM93**, p. 22] *can be seen in `XGobi` if you use the instructions in Problem 293. 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 $\hat{\hat{\varepsilon}}$ is orthogonal to $\boldsymbol{x}_1$, and the green line $\hat{\varepsilon}$ is also orthogonal to $\boldsymbol{x}_1$. Together, $\hat{\varepsilon}$ and $\hat{\hat{\varepsilon}}$ span therefore the plane orthogonal to $\boldsymbol{x}_1$. Since the light blue line lies in the plane spanned by $\hat{\varepsilon}$ and $\hat{\hat{\varepsilon}}$, it is orthogonal to $\boldsymbol{x}_1$. □

Question 297 shows that the decomposition $\mathbf{y} = \hat{\hat{\mathbf{y}}} + (\hat{\mathbf{y}} - \hat{\hat{\mathbf{y}}}) + \hat{\varepsilon}$ is orthogonal, i.e., all 3 vectors $\hat{\hat{\mathbf{y}}}$, $\hat{\mathbf{y}} - \hat{\hat{\mathbf{y}}}$, and $\hat{\varepsilon}$ are orthogonal to each other. This is (22.7.6) 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 $\hat{\hat{\varepsilon}}$.
- 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 $\hat{\varepsilon}$.
- 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 $\hat{\varepsilon}$. The residual is $\hat{\varepsilon}$ because the space spanned by the regressors is the same as in the regression on $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, and $\hat{\varepsilon}$ 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 $\hat{\varepsilon} + \hat{\beta}_1\boldsymbol{x}_1$.
- But in the regression of $\hat{\hat{\varepsilon}} = \boldsymbol{M}_1\mathbf{y}$ on $\boldsymbol{M}_1\boldsymbol{x}_2$, the coefficient is $\hat{\beta}_2$ and the residual $\hat{\varepsilon}$.

This last statement is (23.0.3).

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 (14.2.4) with partitioned matrix inverses. For this we need (23.0.5).

Problem 298. *4 points This is a simplified version of question 393. 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:*

(23.0.5)
$$\begin{bmatrix}\boldsymbol{X}_1^\top\boldsymbol{X}_1 & \boldsymbol{X}_1^\top\boldsymbol{X}_2\\ \boldsymbol{X}_2^\top\boldsymbol{X}_1 & \boldsymbol{X}_2^\top\boldsymbol{X}_2\end{bmatrix}^{-1} = \begin{bmatrix}(\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1} + \boldsymbol{K}_1^\top\boldsymbol{X}_2(\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top\boldsymbol{K}_1 & -\boldsymbol{K}_1^\top\boldsymbol{X}_2(\boldsymbol{X}_2^\top \\ -(\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top\boldsymbol{K}_1 & (\boldsymbol{X}_2^\top\boldsymbol{M}_1\end{bmatrix}$$

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 (23.0.5) one sees that the covariance matrix in regression (23.0.3) is the lower left partition of the covariance matrix in the full regression (23.0.1).

Problem 299. *6 points Use the usual formula* $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\mathbf{y}$ *together with* (23.0.5) *to prove* (23.0.3) *and* (23.0.4).

Answer. (14.2.4) reads here

(23.0.6)
$$\begin{bmatrix}\hat{\boldsymbol{\beta}}_1\\ \hat{\boldsymbol{\beta}}_2\end{bmatrix} = \begin{bmatrix}(\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1} + \boldsymbol{K}_1^\top\boldsymbol{X}_2(\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top\boldsymbol{K}_1 & -\boldsymbol{K}_1^\top\boldsymbol{X}_2(\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\\ -(\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top\boldsymbol{K}_1 & (\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\end{bmatrix}\begin{bmatrix}\boldsymbol{X}_1^\top\mathbf{y}\\ \boldsymbol{X}_2^\top\mathbf{y}\end{bmatrix}$$

Since $\boldsymbol{M}_1 = \boldsymbol{I} - \boldsymbol{K}_1\boldsymbol{X}_1^\top$, one can simplify

(23.0.7)
$$\hat{\boldsymbol{\beta}}_2 = -(\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top\boldsymbol{K}_1\boldsymbol{X}_1^\top\mathbf{y} + (\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top\mathbf{y}$$

(23.0.8)
$$= (\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top\boldsymbol{M}\mathbf{y}$$

(23.0.9)
$$\hat{\boldsymbol{\beta}}_1 = (\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top\mathbf{y} + \boldsymbol{K}_1^\top\boldsymbol{X}_2(\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top\boldsymbol{K}_1\boldsymbol{X}_1^\top\mathbf{y} - \boldsymbol{K}_1^\top\boldsymbol{X}_2(\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top\mathbf{y}$$

(23.0.10)
$$= \boldsymbol{K}_1^\top\mathbf{y} - \boldsymbol{K}_1^\top\boldsymbol{X}_2(\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top(\boldsymbol{I} - \boldsymbol{K}_1\boldsymbol{X}_1^\top)\mathbf{y}$$

(23.0.11)
$$= \boldsymbol{K}_1^\top\mathbf{y} - \boldsymbol{K}_1^\top\boldsymbol{X}_2(\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top\boldsymbol{M}_1\mathbf{y}$$

(23.0.12)
$$= \boldsymbol{K}_1^\top(\mathbf{y} - \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2)$$

□

[**Gre97**, pp. 245–7] follow a different proof strategy: he solves the partitioned normal equations

(23.0.13)
$$\begin{bmatrix}\boldsymbol{X}_1^\top\boldsymbol{X}_1 & \boldsymbol{X}_1^\top\boldsymbol{X}_2\\ \boldsymbol{X}_2^\top\boldsymbol{X}_1 & \boldsymbol{X}_2^\top\boldsymbol{X}_2\end{bmatrix}\begin{bmatrix}\hat{\boldsymbol{\beta}}_1\\ \hat{\boldsymbol{\beta}}_2\end{bmatrix}\begin{bmatrix}\boldsymbol{X}_1^\top\mathbf{y}\\ \boldsymbol{X}_2^\top\mathbf{y}\end{bmatrix}$$

directly, without going through the inverse. A third proof strategy, used by [**Seb77**, pp. 65–72], is followed in Problems 301 and 302.

Problem 300. *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:*

(23.0.14)

$\boldsymbol{y}$	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	
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.

(23.0.15)
$$\begin{bmatrix}\boldsymbol{x}_1^\top\boldsymbol{x}_1 & \boldsymbol{x}_1^\top\boldsymbol{x}_2\\ \boldsymbol{x}_2^\top\boldsymbol{x}_1 & \boldsymbol{x}_2^\top\boldsymbol{x}_2\end{bmatrix}\begin{bmatrix}\hat{\beta}_1\\ \hat{\beta}_2\end{bmatrix} = \begin{bmatrix}\boldsymbol{x}_1^\top\mathbf{y}\\ \boldsymbol{x}_2^\top\mathbf{y}\end{bmatrix}$$

The first row reads

$$\hat\beta_1 + (\boldsymbol{x}_1^\top \boldsymbol{x}_1)^{-1}\boldsymbol{x}_1^\top \boldsymbol{x}_2\hat\beta_2 = (\boldsymbol{x}_1^\top \boldsymbol{x}_1)^{-1}\boldsymbol{x}_1^\top \boldsymbol{y} \tag{23.0.16}$$

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

$$(\boldsymbol{x}_2^\top \boldsymbol{x}_2)^{-1}\boldsymbol{x}_2^\top \boldsymbol{x}_1\hat\beta_1 + \hat\beta_2 = (\boldsymbol{x}_2^\top \boldsymbol{x}_2)^{-1}\boldsymbol{x}_2^\top \boldsymbol{y} \tag{23.0.17}$$

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 301. *Derive* (23.0.3) *and* (23.0.4) *from the first order conditions for minimizing*

$$(\mathbf{y} - \boldsymbol{X}_1\boldsymbol{\beta}_1 - \boldsymbol{X}_2\boldsymbol{\beta}_2)^\top(\mathbf{y} - \boldsymbol{X}_1\boldsymbol{\beta}_1 - \boldsymbol{X}_2\boldsymbol{\beta}_2). \tag{23.0.18}$$

Answer. Start by writing down the OLS objective function for the full model. Perhaps we can use the more sophisticated matrix differentiation rules?

$$(\mathbf{y}-\boldsymbol{X}_1\boldsymbol{\beta}_1-\boldsymbol{X}_2\boldsymbol{\beta}_2)^\top(\mathbf{y}-\boldsymbol{X}_1\boldsymbol{\beta}_1-\boldsymbol{X}_2\boldsymbol{\beta}_2) = \mathbf{y}^\top\mathbf{y}+\boldsymbol{\beta}_1^\top\boldsymbol{X}_1^\top\boldsymbol{X}_1\boldsymbol{\beta}_1+\boldsymbol{\beta}_2^\top\boldsymbol{X}_2^\top\boldsymbol{X}_2\boldsymbol{\beta}_2-2\mathbf{y}^\top\boldsymbol{X}_1\boldsymbol{\beta}_1-2\mathbf{y}^\top\boldsymbol{X}_2\boldsymbol{\beta}_2+ \tag{23.0.19}$$

Taking partial derivatives with respect to $\boldsymbol{\beta}_1^\top$ and $\boldsymbol{\beta}_2^\top$ gives

$$2\boldsymbol{\beta}_1^\top\boldsymbol{X}_1^\top\boldsymbol{X}_1 - 2\mathbf{y}^\top\boldsymbol{X}_1 + 2\boldsymbol{\beta}_2^\top\boldsymbol{X}_2^\top\boldsymbol{X}_1 \quad \text{or, transposed} \quad 2\boldsymbol{X}_1^\top\boldsymbol{X}_1\boldsymbol{\beta}_1 - 2\boldsymbol{X}_1^\top\mathbf{y} + 2\boldsymbol{X}_1^\top\boldsymbol{X}_2\boldsymbol{\beta}_2 \tag{23.0.20}$$

$$2\boldsymbol{\beta}_2^\top\boldsymbol{X}_2^\top\boldsymbol{X}_2 - 2\mathbf{y}^\top\boldsymbol{X}_2^\top + 2\boldsymbol{\beta}_1^\top\boldsymbol{X}_1^\top\boldsymbol{X}_2 \quad \text{or, transposed} \quad 2\boldsymbol{X}_2^\top\boldsymbol{X}_2\boldsymbol{\beta}_2 - 2\boldsymbol{X}_2^\top\mathbf{y} + 2\boldsymbol{X}_2^\top\boldsymbol{X}_1\boldsymbol{\beta}_1 \tag{23.0.21}$$

Setting them zero and replacing $\boldsymbol{\beta}_1$ by $\hat\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ by $\hat\boldsymbol{\beta}_2$ gives

$$\boldsymbol{X}_1^\top\boldsymbol{X}_1\hat\boldsymbol{\beta}_1 = \boldsymbol{X}_1^\top(\mathbf{y} - \boldsymbol{X}_2\hat\boldsymbol{\beta}_2) \tag{23.0.22}$$

$$\boldsymbol{X}_2^\top\boldsymbol{X}_2\hat\boldsymbol{\beta}_2 = \boldsymbol{X}_2^\top(\mathbf{y} - \boldsymbol{X}_1\hat\boldsymbol{\beta}_1). \tag{23.0.23}$$

Premultiply (23.0.22) by $\boldsymbol{X}_1(\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1}$:

$$\boldsymbol{X}_1\hat\boldsymbol{\beta}_1 = \boldsymbol{X}_1(\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top(\mathbf{y} - \boldsymbol{X}_2\hat\boldsymbol{\beta}_2). \tag{23.0.24}$$

Plug this into (23.0.23):

$$\boldsymbol{X}_2^\top\boldsymbol{X}_2\hat\boldsymbol{\beta}_2 = \boldsymbol{X}_2^\top\Big(\mathbf{y} - \boldsymbol{X}_1(\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top\mathbf{y} + \boldsymbol{X}_1(\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top\boldsymbol{X}_2\hat\boldsymbol{\beta}_2\Big) \tag{23.0.25}$$

$$\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2\hat\boldsymbol{\beta}_2 = \boldsymbol{X}_2^\top\boldsymbol{M}_1\mathbf{y}. \tag{23.0.26}$$

(23.0.26) is the normal equation of the regression of $\boldsymbol{M}_1\mathbf{y}$ on $\boldsymbol{M}_1\boldsymbol{X}_2$; it immediately implies (23.0.3). Once $\hat\boldsymbol{\beta}_2$ is known, (23.0.22) is the normal equation of the regression of $\mathbf{y} - \boldsymbol{X}_2\hat\boldsymbol{\beta}_2$ on $\boldsymbol{X}_1$, which gives (23.0.4). □

Problem 302. *Using* (23.0.3) *and* (23.0.4) *show that the residuals in regression* (23.0.1) *are identical to those in the regression of* $\boldsymbol{M}_1\mathbf{y}$ *on* $\boldsymbol{M}_1\boldsymbol{X}_2$.

Answer.

$$\hat\varepsilon = \mathbf{y} - \boldsymbol{X}_1\hat\boldsymbol{\beta}_1 - \boldsymbol{X}_2\hat\boldsymbol{\beta}_2 \tag{23.0.27}$$

$$= \mathbf{y} - \boldsymbol{X}_1(\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top(\mathbf{y} - \boldsymbol{X}_2\hat\boldsymbol{\beta}_2) - \boldsymbol{X}_2\hat\boldsymbol{\beta}_2 \tag{23.0.28}$$

$$= \boldsymbol{M}_1\mathbf{y} - \boldsymbol{M}_1\boldsymbol{X}_2\hat\boldsymbol{\beta}_2. \tag{23.0.29}$$

□

Problem 303. *The following problem derives one of the main formulas for adding regressors, following* [**DM93**, pp. 19–24]. *We are working in model* (23.0.1).

• 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 $\boldsymbol{M}_1\boldsymbol{M} = (\boldsymbol{I} - \boldsymbol{X}\boldsymbol{A}(\boldsymbol{A}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{A})^{-1}\boldsymbol{A}^\top\boldsymbol{X}^\top)\boldsymbol{M}$ $\boldsymbol{M}$ 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 196. 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*

$$\mathbf{y} = \boldsymbol{X}_1\hat\boldsymbol{\beta}_1 + \boldsymbol{X}_2\hat\boldsymbol{\beta}_2 + \hat\varepsilon \tag{23.0.30}$$

The difference between (23.0.1) *and* (23.0.30) *is that* (23.0.30) *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* (23.0.30) *by* $\boldsymbol{M}_1$ *gives*

$$\boldsymbol{M}_1\mathbf{y} = \boldsymbol{M}_1\boldsymbol{X}_2\hat\boldsymbol{\beta}_2 + \hat\varepsilon \tag{23.0.31}$$

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* (23.0.31) *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* (23.0.31), *which are the same* $\hat{\boldsymbol{\beta}}_2$ *and* $\hat{\boldsymbol{\varepsilon}}$ *as in the complete regression* (23.0.30).

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* (23.0.1) *by* $\boldsymbol{M}_1$, *one obtains*

$$\boldsymbol{M}_1\mathbf{y} = \boldsymbol{M}_1\boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{M}_1\boldsymbol{\varepsilon}, \qquad \boldsymbol{M}_1\boldsymbol{\varepsilon} \sim (\boldsymbol{o}, \sigma^2\boldsymbol{M}_1) \tag{23.0.32}$$

Although the covariance matrix of the disturbance $\boldsymbol{M}_1\boldsymbol{\varepsilon}$ *in* (23.0.32) *is no longer spherical, show that nevertheless the* $\hat{\boldsymbol{\beta}}_2$ *obtained by running OLS on* (23.0.32) *is the BLUE of* $\boldsymbol{\beta}_2$ *based on the information given in* (23.0.32) *(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 $\boldsymbol{M}_1\mathbf{y}$ will give the null vector, therefore the *values* of the BLUE for those values of $\mathbf{y}$ 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* (23.0.30) *to get*

$$\mathbf{y} - \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 = \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 + \hat{\boldsymbol{\varepsilon}} \tag{23.0.33}$$

Prove that one gets the right values of $\hat{\boldsymbol{\beta}}_1$ *and of* $\hat{\boldsymbol{\varepsilon}}$ *if one regresses* $\mathbf{y} - \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2$ *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 $\boldsymbol{X}_1^\top\mathbf{y} - \boldsymbol{X}_1^\top\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 = \boldsymbol{X}_1^\top\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1$, which holds due to the normal equation for the full model. □

• k. *1 point Does* (23.0.33) *also give the right covariance matrix for* $\hat{\boldsymbol{\beta}}_1$*?*

Answer. No, since $\mathbf{y} - \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2$ has a different covariance matrix than $\sigma^2\boldsymbol{I}$. □

This following Problems gives some applications of the results in Problem 303. You are allowed to use the results of Problem 303 without proof.

Problem 304. *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*

$$\mathbf{y} = \boldsymbol{\iota}\alpha + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{23.0.34}$$

• a. *1 point Show that the OLS estimate of the slope parameters* $\boldsymbol{\beta}$ *can be obtained by regressing* $\underline{\mathbf{y}}$ *on* $\underline{\boldsymbol{X}}$ *without intercept, where* $\underline{\mathbf{y}}$ *and* $\underline{\boldsymbol{X}}$ *are the variables with their means taken out, i.e.,* $\underline{\mathbf{y}} = \boldsymbol{D}\mathbf{y}$ *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 (23.0.3). This is the usual procedure to do regression with a constant term: in simple regression $y_i = \alpha + \beta x_i + \varepsilon_i$, (23.0.3) is equation (14.2.22):

$$\hat{\beta} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}. \tag{23.0.35}$$

□

• b. *Show that the OLS estimate of the intercept is* $\hat{\alpha} = \bar{y} - \bar{\boldsymbol{x}}^\top\hat{\boldsymbol{\beta}}$ *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 (23.0.4). Here is a more specific argument: The intercept $\hat{\alpha}$ is obtained by regressing $\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ on $\boldsymbol{\iota}$. The normal equation for this second regression is $\boldsymbol{\iota}^\top\mathbf{y} - \boldsymbol{\iota}^\top\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{\iota}^\top\boldsymbol{\iota}\hat{\alpha}$. If $\bar{y}$ is the mean of $\mathbf{y}$, and $\bar{\boldsymbol{x}}^\top$ the row vector consisting of means of the colums of $\boldsymbol{X}$, then this gives $\bar{y} = \bar{\boldsymbol{x}}^\top\hat{\boldsymbol{\beta}} + \hat{\alpha}$. In the case of simple regression, this was derived earlier as formula (14.2.23). □

• 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

$$\begin{bmatrix}\boldsymbol{\iota}^\top \\ \boldsymbol{X}^\top\end{bmatrix}\begin{bmatrix}\boldsymbol{\iota} & \boldsymbol{X}\end{bmatrix} = \begin{bmatrix} n & n\bar{\boldsymbol{x}}^\top \\ \bar{\boldsymbol{x}}n & \boldsymbol{X}^\top\boldsymbol{X}\end{bmatrix}, \tag{23.0.36}$$

it follows by Problem 393

$$\left(\begin{bmatrix}\boldsymbol{\iota}^\top \\ \boldsymbol{X}^\top\end{bmatrix}\begin{bmatrix}\boldsymbol{\iota} & \boldsymbol{X}\end{bmatrix}\right)^{-1} = \begin{bmatrix} 1/n + \bar{\boldsymbol{x}}^\top(\underline{\boldsymbol{X}}^\top\underline{\boldsymbol{X}})^{-1}\bar{\boldsymbol{x}} & -\bar{\boldsymbol{x}}^\top(\underline{\boldsymbol{X}}^\top\underline{\boldsymbol{X}})^{-1} \\ -(\underline{\boldsymbol{X}}^\top\underline{\boldsymbol{X}})^{-1}\bar{\boldsymbol{x}} & (\underline{\boldsymbol{X}}^\top\underline{\boldsymbol{X}})^{-1}\end{bmatrix} \tag{23.0.37}$$

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* $\hat{\mathbf{y}} - \boldsymbol{\iota}\bar{y} = \underline{\boldsymbol{X}}\hat{\boldsymbol{\beta}}$.

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, $\hat{\mathbf{y}} = \boldsymbol{\iota}\hat{\alpha} + \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{\iota}\hat{\alpha} + \underline{\boldsymbol{X}}\hat{\boldsymbol{\beta}} + \boldsymbol{\iota}\bar{\boldsymbol{x}}^\top\hat{\boldsymbol{\beta}} = \boldsymbol{\iota}(\hat{\alpha} + \bar{\boldsymbol{x}}^\top\hat{\boldsymbol{\beta}}) + \underline{\boldsymbol{X}}\hat{\boldsymbol{\beta}} = \boldsymbol{\iota}\bar{y} + \underline{\boldsymbol{X}}\hat{\boldsymbol{\beta}}$. □

• 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.

$$R^2 = \frac{(\hat{\mathsf{y}} - \bar{\mathsf{y}}\iota)^\top(\hat{\mathsf{y}} - \bar{\mathsf{y}}\iota)}{\underline{\mathsf{y}}^\top \underline{\mathsf{y}}} = \frac{\hat{\boldsymbol{\beta}}^\top \underline{\boldsymbol{X}}^\top \underline{\boldsymbol{X}}\hat{\boldsymbol{\beta}}}{\underline{\mathsf{y}}^\top \underline{\mathsf{y}}} \tag{23.0.38}$$

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*

$$\operatorname{var}[\hat{\beta}_2] = \frac{\sigma^2}{\underline{\boldsymbol{x}}^\top \underline{\boldsymbol{x}}}\frac{1}{(1 - R^2_{2.})} \tag{23.0.39}$$

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* (23.0.39) *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* $1/(1 - R^2_{2.})$ *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.

$$\underline{\boldsymbol{x}}_2^\top \underline{\boldsymbol{M}}_1 \underline{\boldsymbol{x}}_2 = \underline{\boldsymbol{x}}_2^\top \underline{\boldsymbol{x}}_2 + \underline{\boldsymbol{x}}_2^\top \underline{\boldsymbol{X}}_1(\underline{\boldsymbol{X}}_1^\top \underline{\boldsymbol{X}}_1)^{-1}\underline{\boldsymbol{X}}_1^\top \underline{\boldsymbol{x}}_2 = \underline{\boldsymbol{x}}_2^\top \underline{\boldsymbol{x}}_2\left(1 + \frac{\underline{\boldsymbol{x}}_2^\top \underline{\boldsymbol{X}}_1(\underline{\boldsymbol{X}}_1^\top \underline{\boldsymbol{X}}_1)^{-1}\underline{\boldsymbol{X}}_1^\top \underline{\boldsymbol{x}}_2}{\underline{\boldsymbol{x}}_2^\top \underline{\boldsymbol{x}}_2}\right) \tag{23.0.40}$$

and since the fraction is $R^2_{2.}$, 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. □

CHAPTER 24

Residuals: Standardized, Predictive, "Studentized"

24.1. 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

$$\hat{\varepsilon}_i = y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}} \tag{24.1.1}$$

($\boldsymbol{x}_i^\top$ 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

$$\hat{\varepsilon}_i(i) = y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}(i). \tag{24.1.2}$$

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 $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{M}\boldsymbol{y}$, the variance of the ith ordinary residual is

$$\operatorname{var}[\hat{\varepsilon}_i] = \sigma^2 m_{ii} = \sigma^2(1 - h_{ii}), \tag{24.1.3}$$

and regarding the predictive residuals it will be shown below, see (24.2.9), that

$$\operatorname{var}[\hat{\varepsilon}_i(i)] = \frac{\sigma^2}{m_{ii}} = \frac{\sigma^2}{1 - h_{ii}}. \tag{24.1.4}$$

Here

$$h_{ii} = \boldsymbol{x}_i^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_i. \tag{24.1.5}$$

(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 $\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$, i.e., $\boldsymbol{H}$ puts the "hat" on $\boldsymbol{y}$.

Problem 305. *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 306. *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* (24.2.9) *without proof.*

Answer. Here is only the math part of the answer: $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{M}\boldsymbol{y}$. Since $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{H}$ is idempotent and symmetric, we get $\mathcal{V}[\boldsymbol{M}\boldsymbol{y}] = \sigma^2\boldsymbol{M}$, 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 (24.2.9) it becomes

$$\hat{\varepsilon}_i(i) = \frac{1}{m_{ii}}\hat{\varepsilon}_i \qquad \text{therefore} \qquad \operatorname{var}[\hat{\varepsilon}_i(i)] = \frac{1}{m_{ii}^2}\sigma^2 m_{ii} = \frac{\sigma^2}{m_{ii}} \tag{24.1.6}$$

which is bigger than σ^2. □

Problem 307. *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 (23.0.2). 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 (23.0.4), 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.

24.2. Relationship between Ordinary and Predictive Residuals

In equation (24.1.2), 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 (24.2.9). (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 308. *2 points Show that*

$$\boldsymbol{X}(i)^\top \boldsymbol{X}(i) = \boldsymbol{X}^\top \boldsymbol{X} - \boldsymbol{x}_i \boldsymbol{x}_i^\top \tag{24.2.1}$$

$$\boldsymbol{X}(i)^\top \mathbf{y}(i) = \boldsymbol{X}^\top \mathbf{y} - \boldsymbol{x}_i \mathrm{y}_i. \tag{24.2.2}$$

ANSWER. Write (24.2.2) 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

$$\boldsymbol{X}^\top \mathbf{y} = \begin{bmatrix} \boldsymbol{x}_1 & \dots & \boldsymbol{x}_n \end{bmatrix} \begin{bmatrix} \mathrm{y}_1 \\ \vdots \\ \mathrm{y}_n \end{bmatrix} = \boldsymbol{x}_1 \mathrm{y}_1 + \cdots + \boldsymbol{x}_n \mathrm{y}_n. \tag{24.2.3}$$

□

An important stepping stone towards the proof of (24.2.9) is equation (24.2.8), which gives a relationship between h_{ii} and

$$h_{ii}(i) = \boldsymbol{x}_i^\top (\boldsymbol{X}(i)^\top \boldsymbol{X}(i))^{-1} \boldsymbol{x}_i. \tag{24.2.4}$$

$\hat{\mathrm{y}}_i(i) = \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}(i)$ has variance $\sigma^2 h_{ii}(i)$. The following problems give the steps necessary to prove (24.2.8). We begin with a simplified version of theorem A.8.2 in the Mathematical Appendix:

THEOREM 24.2.1. *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*

$$\left(\boldsymbol{A} + \frac{\boldsymbol{b}\boldsymbol{b}^\top}{\delta} \right)^{-1} = \boldsymbol{A}^{-1} - \frac{\boldsymbol{A}^{-1}\boldsymbol{b}\boldsymbol{b}^\top \boldsymbol{A}^{-1}}{\delta + \boldsymbol{b}^\top \boldsymbol{A}^{-1} \boldsymbol{b}}. \tag{24.2.5}$$

PROBLEM 309. *Prove* (24.2.5) *by showing that the product of the matrix with its alleged inverse is the unit matrix.*

PROBLEM 310. *As an application of* (24.2.5) *show that*

$$(\boldsymbol{X}^\top \boldsymbol{X})^{-1} + \frac{(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_i \boldsymbol{x}_i^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1}}{1 - h_{ii}} \quad \textit{is the inverse of} \quad \boldsymbol{X}(i)^\top \boldsymbol{X}(i). \tag{24.2.6}$$

ANSWER. This is (24.2.5), or (A.8.20), with $\boldsymbol{A} = \boldsymbol{X}^\top \boldsymbol{X}$, $\boldsymbol{b} = \boldsymbol{x}_i$, and $\delta = -1$. □

PROBLEM 311. *Using* (24.2.6) *show that*

$$(\boldsymbol{X}(i)^\top \boldsymbol{X}(i))^{-1} \boldsymbol{x}_i = \frac{1}{1 - h_{ii}} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_i, \tag{24.2.7}$$

and using (24.2.7) *show that $h_{ii}(i)$ is related to h_{ii} by the equation*

$$1 + h_{ii}(i) = \frac{1}{1 - h_{ii}} \tag{24.2.8}$$

[**Gre97**, (9-37) on p. 445] *was apparently not aware of this relationship.*

PROBLEM 312. *Prove the following mathematical relationship between predictive residuals and ordinary residuals:*

$$\hat{\varepsilon}_i(i) = \frac{1}{1 - h_{ii}} \hat{\varepsilon}_i \tag{24.2.9}$$

which is the same as (21.0.29)*, only in a different notation.*

Answer. For this we have to apply the above mathematical tools. With the help of (24.2.7) (transpose it!) and (24.2.2), (24.1.2) 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 (24.2.9) 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

$$\hat{\mathsf{y}}_i(i) = \eta_i + \delta_i \qquad \delta_i \sim (0, \sigma^2 h_{ii}(i)) \tag{24.2.10}$$

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

$$\begin{bmatrix}\hat{\mathsf{y}}_i(i)\\ \mathsf{y}_i\end{bmatrix} = \begin{bmatrix}1\\1\end{bmatrix}\eta_i + \begin{bmatrix}\delta_i\\ \varepsilon_i\end{bmatrix} \qquad \begin{bmatrix}\delta_i\\ \varepsilon_i\end{bmatrix} \sim \left(\begin{bmatrix}0\\0\end{bmatrix}, \sigma^2\begin{bmatrix}h_{ii}(i) & 0\\ 0 & 1\end{bmatrix}\right) \tag{24.2.11}$$

This is a regression model similar to model (14.1.1), but this time with a nonspherical covariance matrix.

Problem 313. *Show that the BLUE of η_i in model* (24.2.11) *is*

$$\hat{\mathsf{y}}_i = (1-h_{ii})\hat{\mathsf{y}}_i(i) + h_{ii}\mathsf{y}_i = \hat{\mathsf{y}}_i(i) + h_{ii}\hat{\varepsilon}_i(i) \tag{24.2.12}$$

Hint: apply (24.2.8). *Use this to prove* (24.2.9).

Answer. As shown in problem 178, 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

$$\frac{1/h_{ii}(i)}{1/h_{ii}(i)+1} = \frac{1}{1+h_{ii}(i)} = 1-h_{ii}. \tag{24.2.13}$$

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 (19.0.6) for the BLUE, which reads here

$$\hat{\mathsf{y}}_i = \left(\begin{bmatrix}1 & 1\end{bmatrix}\begin{bmatrix}\frac{h_{ii}}{1-h_{ii}} & 0\\ 0 & 1\end{bmatrix}^{-1}\begin{bmatrix}1\\1\end{bmatrix}\right)^{-1}\begin{bmatrix}1 & 1\end{bmatrix}\begin{bmatrix}\frac{h_{ii}}{1-h_{ii}} & 0\\ 0 & 1\end{bmatrix}^{-1}\begin{bmatrix}\hat{\mathsf{y}}_i(i)\\ \mathsf{y}_i\end{bmatrix} =$$

$$= h_{ii}\begin{bmatrix}1 & 1\end{bmatrix}\begin{bmatrix}\frac{1-h_{ii}}{h_{ii}} & 0\\ 0 & 1\end{bmatrix}\begin{bmatrix}\hat{\mathsf{y}}_i(i)\\ \mathsf{y}_i\end{bmatrix} = (1-h_{ii})\hat{\mathsf{y}}_i(i) + h_{ii}\mathsf{y}_i.$$

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 (24.2.9). □

24.3. 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,

$$\text{the } i\text{th standardized ordinary residual} = \frac{\hat{\varepsilon}_i}{\mathsf{s}\sqrt{1-h_{ii}}}. \tag{24.3.1}$$

Standardization of the ith *predictive* residual has the same two steps: first divide the predictive residual (24.2.9) by the relative standard deviation, and then divide by $\mathsf{s}(i)$. But a look at formula (24.2.9) 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:

$$\text{the } i\text{th standardized predictive residual} = \frac{\hat{\varepsilon}_i}{\mathsf{s}(i)\sqrt{1-h_{ii}}}. \tag{24.3.2}$$

Note that equation (24.3.2) 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 314. *3 points The ith predictive residual has the formula*

$$\hat{\varepsilon}_i(i) = \frac{1}{1-h_{ii}}\hat{\varepsilon}_i \tag{24.3.3}$$

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 315. *Following* [**DM93**, p. 34], *we will use* (23.0.3) *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* i*th observation is the last observation, i.e.,* $i = n$, *and put the dummy variable first in the regression:*

$$\begin{bmatrix} \mathbf{y}(n) \\ \mathbf{y}_n \end{bmatrix} = \begin{bmatrix} \boldsymbol{o} & \boldsymbol{X}(n) \\ 1 & \boldsymbol{x}_n^\top \end{bmatrix} \begin{bmatrix} \alpha \\ \boldsymbol{\beta} \end{bmatrix} + \begin{bmatrix} \hat{\boldsymbol{\varepsilon}}(i) \\ \hat{\varepsilon}_n \end{bmatrix} \qquad \text{or} \quad \mathbf{y} = \begin{bmatrix} \boldsymbol{e}_n & \boldsymbol{X} \end{bmatrix} \begin{bmatrix} \alpha \\ \boldsymbol{\beta} \end{bmatrix} + \boldsymbol{\varepsilon} \tag{24.3.4}$$

• 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.

$$\boldsymbol{M}_1 = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{o} \\ \boldsymbol{o}^\top & 1 \end{bmatrix} - \begin{bmatrix} \boldsymbol{o} \\ 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{o}^\top & 1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{o} \\ \boldsymbol{o}^\top & 0 \end{bmatrix}; \qquad \begin{bmatrix} \boldsymbol{I} & \boldsymbol{o} \\ \boldsymbol{o}^\top & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{z}(i) \\ z_i \end{bmatrix} = \begin{bmatrix} \boldsymbol{z}(i) \\ 0 \end{bmatrix} \tag{24.3.5}$$

i.e., $\boldsymbol{M}_1$ simply annulls the last element. □

• b. *2 points Either show mathematically, perhaps by evaluating* $(\boldsymbol{X}_2^\top \boldsymbol{M}_1 \boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top$ *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* n*th observation dropped.*

Answer. (23.0.2) reads here

$$\begin{bmatrix} \mathbf{y}(n) \\ 0 \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}(n) \\ \boldsymbol{o}^\top \end{bmatrix} \hat{\boldsymbol{\beta}}(i) + \begin{bmatrix} \hat{\boldsymbol{\varepsilon}}(i) \\ 0 \end{bmatrix} \tag{24.3.6}$$

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 (23.0.3)

$$\begin{aligned} \hat{\boldsymbol{\beta}}_2 &= (\boldsymbol{X}_2^\top \boldsymbol{M}_1 \boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top \boldsymbol{M}_1 \mathbf{y} \\ &= \left(\begin{bmatrix} \boldsymbol{X}(n)^\top & \boldsymbol{x}_n \end{bmatrix} \begin{bmatrix} \boldsymbol{I} & \boldsymbol{o} \\ \boldsymbol{o}^\top & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{X}(n) \\ \boldsymbol{x}_n^\top \end{bmatrix} \right)^{-1} \begin{bmatrix} \boldsymbol{X}(n)^\top & \boldsymbol{x}_n \end{bmatrix} \begin{bmatrix} \boldsymbol{I} & \boldsymbol{o} \\ \boldsymbol{o}^\top & 0 \end{bmatrix} \begin{bmatrix} \mathbf{y}(n) \\ \mathbf{y}_n \end{bmatrix} \\ &= (\boldsymbol{X}(n)^\top \boldsymbol{X}(n))^{-1}\boldsymbol{X}(n)^\top \mathbf{y}(n) = \hat{\boldsymbol{\beta}}(n) \end{aligned}$$

□

• 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* (24.3.4) *to show that* $\hat{\alpha}$ *is the* n*th 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* (23.0.3) *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* (23.0.4) *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*

$$\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}(i) = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{x}_i \hat{\varepsilon}_i \frac{1}{1-h_{ii}} \tag{24.3.7}$$

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 (25.4.1). □

CHAPTER 25

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 316. *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?*

25.1. 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.

25.2. 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.

25.3. 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.)

25.3.1. 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

$$\hat{\mathsf{y}}_i = (1 - h_{ii})\hat{\mathsf{y}}_i(i) + h_{ii}\mathsf{y}_i \tag{25.3.1}$$

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 317. *3 points Explain the meanings of all the terms in equation* (25.3.1) *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 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 $\hat{\boldsymbol{\beta}}$.

$$h_{ii} = \frac{\det(\boldsymbol{X}^\top\boldsymbol{X}) - \det(\boldsymbol{X}(i)^\top\boldsymbol{X}(i))}{\det(\boldsymbol{X}^\top\boldsymbol{X})}, \tag{25.3.2}$$

where $\boldsymbol{X}(i)$ is the $\boldsymbol{X}$-matrix without the ith observation.

PROBLEM 318. *Prove equation* (25.3.2).

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 319. *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* $\boldsymbol{H}\mathsf{y} = \hat{\mathsf{y}}$, *i.e., it puts the hat on* 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*

$$0 \leq h_{ii} \leq 1 \tag{25.3.3}$$

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* (25.3.3) *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 (24.2.12). Effect on any other linear combination of $\hat{\boldsymbol{\beta}}$ is less than the effect on $\hat{\mathsf{y}}_i$. Distinguish from influence. Leverage depends only on $\boldsymbol{X}$ matrix, not on 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.

25.4. 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 $\hat{\varepsilon}_i$.

25.4.1. 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

$$\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}(i) = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_i\hat{\varepsilon}_i(i) \tag{25.4.1}$$

PROBLEM 320. *Show* (25.4.1) *by methods very similar to the proof of* (24.2.9)

ANSWER. Here is this brute-force proof, I think from [**BKW80**]: Let $\mathsf{y}(i)$ be the y vector with the ith observation deleted. As shown in Problem 308, $\boldsymbol{X}^\top(i)\mathsf{y}(i) = \boldsymbol{X}^\top\mathsf{y} - \boldsymbol{x}_i\mathsf{y}_i$. Therefore by (24.2.6)

$$\begin{aligned}\hat{\boldsymbol{\beta}}(i) = (\boldsymbol{X}^\top(i)\boldsymbol{X}(i))^{-1}\boldsymbol{X}^\top(i)\mathsf{y}(i) &= \left((\boldsymbol{X}^\top\boldsymbol{X})^{-1} + \frac{(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}}{1-h_{ii}}\right)\left(\boldsymbol{X}^\top\mathsf{y} - \boldsymbol{x}_i\mathsf{y}_i\right) = \\ &= \hat{\boldsymbol{\beta}} - (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_i\mathsf{y}_i + \frac{1}{1-h_{ii}}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}} - \frac{h_{ii}}{1-h_{ii}}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_i\mathsf{y}_i \\ &= \hat{\boldsymbol{\beta}} - \frac{1}{1-h_{ii}}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_i\mathsf{y}_i + \frac{1}{1-h_{ii}}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - \frac{1}{1-h_{ii}}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_i\hat{\varepsilon}_i\end{aligned}$$

□

To understand (25.4.1), 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 $\hat{\mathsf{y}}_i(i)$. Writing $\mathsf{y}((i))$ for the vector y whose ith observation has been replaced in this way, one obtains

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\mathsf{y}; \qquad \hat{\boldsymbol{\beta}}(i) = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\mathsf{y}((i)). \tag{25.4.2}$$

Since $\mathsf{y} - \mathsf{y}((i)) = \boldsymbol{e}_i\hat{\varepsilon}_i(i)$ and $\boldsymbol{x}_i = \boldsymbol{X}^\top\boldsymbol{e}_i$ (25.4.1) follows.

The quantities h_{ii}, $\hat{\boldsymbol{\beta}}(i)-\hat{\boldsymbol{\beta}}$, and $s^2(i)$ are computed by the R-function `lm.influence`. Compare [**CH93**, pp. 129–131].

25.4.2. 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

$$\frac{\boldsymbol{t}^\top(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}(i))}{\sigma\sqrt{\boldsymbol{t}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{t}}}. \tag{25.4.3}$$

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 25.4.1. *If $\boldsymbol{\Omega}$ is positive definite symmetric, then*

$$\max_{\boldsymbol{g}} \frac{(\boldsymbol{g}^\top\boldsymbol{x})^2}{\boldsymbol{g}^\top\boldsymbol{\Omega}\boldsymbol{g}} = \boldsymbol{x}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{x}. \tag{25.4.4}$$

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 (25.4.4) follows from the Cauchy-Schwartz inequality $(\boldsymbol{h}^\top\boldsymbol{y})^2 \leq (\boldsymbol{h}^\top\boldsymbol{h})(\boldsymbol{y}^\top\boldsymbol{y})$.

Using Theorem 25.4.1 and equation (25.4.1) one obtains

$$\begin{aligned}\max_{\boldsymbol{t}} \frac{(\boldsymbol{t}^\top(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}(i)))^2}{\sigma^2\boldsymbol{t}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{t}} &= \frac{1}{\sigma^2}(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}(i))^\top\boldsymbol{X}^\top\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}(i)) = \\ &= \frac{1}{\sigma^2}\boldsymbol{x}_i^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_i\hat{\varepsilon}_i^2(i) = \frac{h_{ii}}{\sigma^2}\hat{\varepsilon}_i^2(i)\end{aligned} \tag{25.4.5}$$

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 (25.4.1) by $\boldsymbol{x}_i^\top$ one obtains

$$\hat{\mathfrak{y}}_i - \hat{\mathfrak{y}}_i(i) = \boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}} - \boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}(i) = \frac{h_{ii}}{1-h_{ii}}\hat{\varepsilon}_i = h_{ii}\hat{\varepsilon}_i(i) \tag{25.4.6}$$

If one divides (25.4.6) by the standard deviation of $\hat{\mathfrak{y}}_i$, i.e., if one applies the construction (25.4.3), one obtains

$$\frac{\hat{\mathfrak{y}}_i - \hat{\mathfrak{y}}_i(i)}{\sigma\sqrt{h_{ii}}} = \frac{\sqrt{h_{ii}}}{\sigma}\hat{\varepsilon}_i(i) = \frac{\sqrt{h_{ii}}}{\sigma(1-h_{ii})}\hat{\varepsilon}_i \tag{25.4.7}$$

If $\hat{\mathfrak{y}}_i$ changes only little (compared with the standard deviation of $\hat{\mathfrak{y}}_i$) 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 (25.4.7), with σ estimated by $s(i)$, 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 $s^2(i)$, 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**]. (25.4.5) gives an equation for Cook's distance in terms of $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}(i)$:

$$\text{Cook's distance} = \frac{(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}(i))^\top\boldsymbol{X}^\top\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}(i))}{ks^2} = \frac{h_{ii}}{ks^2}\hat{\varepsilon}_i^2(i) = \frac{h_{ii}}{ks^2(1-h_{ii})^2}\hat{\varepsilon}_i^2 \tag{25.4.8}$$

PROBLEM 321. *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 (25.3.2), i.e., in this case only the h_{ii} count. □

PROBLEM 322. *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.For + 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(lon 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 Longle 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* (24.2.9)*. Also plot the standardized predictive residuals, and compare them.*

Answer. The predictive residuals are `plot(Year, residuals(longley.fit)/(1-hatvalues(longl 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.*

25.4.3. 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:

$$\mathsf{SSE} - \mathsf{SSE}(i) = \frac{\hat{\varepsilon}_i^2}{1-h_{ii}} \tag{25.4.9}$$

Problem 323. *Use* (25.4.9) *to derive the following formula for* $\mathsf{s}^2(i)$*:*

$$\mathsf{s}^2(i) = \frac{1}{n-k-1}\Big((n-k)\mathsf{s}^2 - \frac{\hat{\varepsilon}_i^2}{1-h_{ii}}\Big) \tag{25.4.10}$$

Answer. This merely involves re-writing SSE and $\mathsf{SSE}(i)$ in terms of s^2 and $\mathsf{s}^2(i)$.

$$\mathsf{s}^2(i) = \frac{\mathsf{SSE}(i)}{n-1-k} = \frac{1}{n-k-1}\Big(\mathsf{SSE} - \frac{\hat{\varepsilon}_i^2}{1-h_{ii}}\Big) \tag{25.4.11}$$

□

Proof of equation (25.4.9)*:*

$$\begin{aligned}
\mathsf{SSE}(i) &= \sum_{j:\ j\neq i} (\mathsf{y}_j - \boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}(i))^2 = \sum_{j:\ j\neq i} \big(\mathsf{y}_j - \boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}} - \boldsymbol{x}_j^\top(\hat{\boldsymbol{\beta}}(i) - \hat{\boldsymbol{\beta}})\big)^2 \\
&= \sum_{j:\ j\neq i} \Big(\hat{\varepsilon}_j + \frac{h_{ji}}{1-h_{ii}}\hat{\varepsilon}_i\Big)^2 \\
&= \sum_j \Big(\hat{\varepsilon}_j + \frac{h_{ji}}{1-h_{ii}}\hat{\varepsilon}_i\Big)^2 - \Big(\frac{1}{1-h_{ii}}\hat{\varepsilon}_i\Big)^2 \\
&= \sum_j \hat{\varepsilon}_j^2 + \frac{2\hat{\varepsilon}_i}{1-h_{ii}}\sum_j h_{ij}\hat{\varepsilon}_j + \Big(\frac{\hat{\varepsilon}_i}{1-h_{ii}}\Big)^2 \sum_j h_{ji}^2 - \Big(\frac{\hat{\varepsilon}_i}{1-h_{ii}}\Big)^2
\end{aligned}$$

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 324. *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 $\mathbf{y}$ at all. □

PROBLEM 325. *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 $\mathbf{y}$. 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 326. *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 26

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:

$$\boldsymbol{Q} = \lim_{n\to\infty} \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X} \quad \text{exists and is nonsingular.} \tag{26.0.12}$$

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 (26.0.12) 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

$$\boldsymbol{X}^\top\boldsymbol{X} = \begin{bmatrix} 1+1+\cdots+1 & \lambda+\lambda^2+\cdots+\lambda^n \\ \lambda+\lambda^2+\cdots+\lambda^n & \lambda^2+\lambda^4+\cdots+\lambda^{2n} \end{bmatrix} = \\ = \begin{bmatrix} n & (\lambda-\lambda^{n+1})/(1-\lambda) \\ (\lambda-\lambda^{n+1})/(1-\lambda) & (\lambda^2-\lambda^{2n+2})/(1-\lambda^2) \end{bmatrix}.$$

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 (26.0.12), $\hat{\boldsymbol{\beta}}$ 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,

$$\operatorname{plim} g(\boldsymbol{z}) = g(\operatorname{plim}(\boldsymbol{z})). \tag{26.0.13}$$

(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.

26.1. 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:

$$\operatorname{plim}\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{\varepsilon} = \boldsymbol{o}. \tag{26.1.1}$$

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

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon} = \boldsymbol{\beta} + \Big(\frac{\boldsymbol{X}^\top\boldsymbol{X}}{n}\Big)^{-1}\frac{\boldsymbol{X}^\top\boldsymbol{\varepsilon}}{n} \tag{26.1.2}$$

$$\operatorname{plim}\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \lim\Big(\frac{\boldsymbol{X}^\top\boldsymbol{X}}{n}\Big)^{-1}\operatorname{plim}\frac{\boldsymbol{X}^\top\boldsymbol{\varepsilon}}{n} \tag{26.1.3}$$

$$= \boldsymbol{\beta} + \boldsymbol{Q}^{-1}\boldsymbol{o} = \boldsymbol{\beta}. \tag{26.1.4}$$

Let's look at the geometry of this when there is only one explanatory variable. The specification is therefore $\mathsf{y} = \boldsymbol{x}\beta + \boldsymbol{\varepsilon}$. 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 (26.0.12) 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 (26.1.1) $\operatorname{plim}\frac{1}{n}\boldsymbol{x}^\top\boldsymbol{\varepsilon}=0$ 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

$$\begin{aligned}\frac{\hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}}}{n-k}&=\frac{n}{n-k}\boldsymbol{\varepsilon}^\top\Big(\frac{\boldsymbol{I}}{n}-\frac{\boldsymbol{X}}{n}\Big(\frac{\boldsymbol{X}^\top\boldsymbol{X}}{n}\Big)^{-1}\frac{\boldsymbol{X}^\top}{n}\Big)\boldsymbol{\varepsilon}=\\&=\frac{n}{n-k}\Big(\frac{\boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon}}{n}-\frac{\boldsymbol{\varepsilon}^\top\boldsymbol{X}}{n}\Big(\frac{\boldsymbol{X}^\top\boldsymbol{X}}{n}\Big)^{-1}\frac{\boldsymbol{X}^\top\boldsymbol{\varepsilon}}{n}\Big)\to 1\cdot\Big(\sigma^2-\boldsymbol{o}^\top\boldsymbol{Q}^{-1}\boldsymbol{o}\Big).\end{aligned}$$

26.2. 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

$$n(\boldsymbol{R}\hat{\boldsymbol{\beta}}_n-\boldsymbol{R}\boldsymbol{\beta})\big(\boldsymbol{R}\boldsymbol{Q}^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}_n-\boldsymbol{R}\boldsymbol{\beta})\to\sigma^2\chi^2_i. \tag{26.2.1}$$

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

$$\frac{(\boldsymbol{R}\hat{\boldsymbol{\beta}}_n-\boldsymbol{R}\boldsymbol{\beta})\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}_n-\boldsymbol{R}\boldsymbol{\beta})}{\mathsf{s}^2}\to\chi^2_i \tag{26.2.2}$$

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 327. *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 (26.0.12) 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 328. *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 (18.0.7), 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 27

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 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

$$f_{\mathsf{y}_t}(y_t) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y_t - \boldsymbol{x}_t^\top \boldsymbol{\beta})^2/2\sigma^2}. \tag{27.0.3}$$

Since the y_i are stochastically independent, their joint density function is the product, which can be written as

$$f_{\mathsf{y}}(\boldsymbol{y}) = (2\pi\sigma^2)^{-n/2} \exp\Big(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\Big). \tag{27.0.4}$$

To compute the maximum likelihood estimator, it is advantageous to start with the log likelihood function:

$$\log f_{\mathsf{y}}(\mathsf{y}; \boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(\mathsf{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\mathsf{y} - \boldsymbol{X}\boldsymbol{\beta}). \tag{27.0.5}$$

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 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 (27.0.5) by $\hat{\boldsymbol{\beta}}$, we get what is called the log likelihood function with $\boldsymbol{\beta}$ "concentrated out."

$$\log f_{\mathsf{y}}(\mathsf{y}; \boldsymbol{\beta} = \hat{\boldsymbol{\beta}}, \sigma^2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(\mathsf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top (\mathsf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}). \tag{27.0.6}$$

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

$$\frac{\partial}{\partial\sigma^2}\log f_{\mathsf{y}}(\mathsf{y}; \hat{\boldsymbol{\beta}}) = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}(\mathsf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top (\mathsf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \tag{27.0.7}$$

Setting this zero gives

$$\tilde{\sigma}^2 = \frac{(\mathsf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top (\mathsf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})}{n} = \frac{\hat{\boldsymbol{\varepsilon}}^\top \hat{\boldsymbol{\varepsilon}}}{n}. \tag{27.0.8}$$

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 7.4.9 characterizing all the quadratic forms of multivariate normal variables that are χ^2's. Here it is again: Assume 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 $(\mathsf{y} - \boldsymbol{\mu})^\top \boldsymbol{\Omega}(\mathsf{y} - \boldsymbol{\mu}) \sim \sigma^2\chi^2_k$ iff

$$\boldsymbol{\Psi}\boldsymbol{\Omega}\boldsymbol{\Psi}\boldsymbol{\Omega}\boldsymbol{\Psi} = \boldsymbol{\Psi}\boldsymbol{\Omega}\boldsymbol{\Psi}, \tag{27.0.9}$$

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.

$$\hat{\sigma}^2 = \frac{(\mathsf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top (\mathsf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})}{n-k} = \frac{\boldsymbol{\varepsilon}^\top(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top)\boldsymbol{\varepsilon}}{n-k} = \frac{\boldsymbol{\varepsilon}^\top\boldsymbol{M}\boldsymbol{\varepsilon}}{n-k} \tag{27.0.10}$$

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 329. *4 points Show that, if y is normally distributed, s^2 and $\hat{\boldsymbol{\beta}}$ are independent.*

ANSWER. We showed in question 246 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 330. *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-Sc 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 $\mathbf{y} = \boldsymbol{X\beta}+\boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathsf{N}(\boldsymbol{o}, \sigma^2\boldsymbol{\Psi})$. The density function is

$$f_{\mathbf{y}}(\boldsymbol{y}) = (2\pi\sigma^2)^{-n/2}\,|\det\boldsymbol{\Psi}|^{-1/2}\exp\Big(-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X\beta})^\top\boldsymbol{\Psi}^{-1}(\boldsymbol{y}-\boldsymbol{X\beta})\Big). \tag{27.0.11}$$

Problem 331. *Derive* (27.0.11) *as follows: Take a matrix* $\boldsymbol{P}$ *with the property that* $\boldsymbol{P\varepsilon}$ *has covariance matrix* $\sigma^2\boldsymbol{I}$. *Write down the joint density function of* $\boldsymbol{P\varepsilon}$. *Since* $\mathbf{y}$ *is a linear transformation of* $\boldsymbol{\varepsilon}$, *one can apply the rule for the density function of a transformed random variable.*

Answer. Write $\boldsymbol{\Psi} = \boldsymbol{QQ}^\top$ with $\boldsymbol{Q}$ nonsingular and define $\boldsymbol{P} = \boldsymbol{Q}^{-1}$ and $\boldsymbol{\nu} = \boldsymbol{P\varepsilon}$. Then $\mathcal{V}[\boldsymbol{\nu}] = \sigma^2\boldsymbol{PQQ}^\top\boldsymbol{P}^\top = \sigma^2\boldsymbol{I}$, therefore

$$f_{\boldsymbol{\nu}}(\boldsymbol{v}) = (2\pi\sigma^2)^{-n/2}\exp\Big(-\frac{1}{2\sigma^2}\boldsymbol{v}^\top\boldsymbol{v}\Big). \tag{27.0.12}$$

For the transformation rule, write $\boldsymbol{\nu}$, whose density function you know, as a function of $\mathbf{y}$, whose density function you want to know. $\boldsymbol{\nu} = \boldsymbol{P}(\mathbf{y}-\boldsymbol{X\beta})$; therefore the Jacobian matrix is $\partial\boldsymbol{\nu}/\partial\mathbf{y}^\top = \partial(\boldsymbol{P}\mathbf{y}-\boldsymbol{PX\beta})/\partial\mathbf{y}^\top = \boldsymbol{P}$, or one can see it also element by element

$$\begin{bmatrix} \frac{\partial\nu_1}{\partial y_1} & \cdots & \frac{\partial\nu_1}{\partial y_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial\nu_n}{\partial y_1} & \cdots & \frac{\partial\nu_n}{\partial y_n} \end{bmatrix} = \boldsymbol{P}, \tag{27.0.13}$$

therefore one has to do two things: first, substitute $\boldsymbol{P}(\boldsymbol{y}-\boldsymbol{X\beta})$ for $\boldsymbol{v}$ in formula (27.0.12), 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 $\boldsymbol{\Psi}^{-1} = (\boldsymbol{QQ}^\top)^{-1} = (\boldsymbol{Q}^\top)^{-1}\boldsymbol{Q}^{-1} = (\boldsymbol{Q}^{-1})^\top\boldsymbol{Q}^{-1} = \boldsymbol{P}^\top\boldsymbol{P}$ follows $(\det\boldsymbol{P})^2 = (\det\boldsymbol{\Psi})^{-1}$, hence $|\det\boldsymbol{P}| = \sqrt{\det\boldsymbol{\Psi}}$. □

From (27.0.11) one obtains the following log likelihood function:

$$\log f_{\mathbf{y}}(\mathbf{y}) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2}\ln\det[\boldsymbol{\Psi}] - \frac{1}{2\sigma^2}(\mathbf{y}-\boldsymbol{X\beta})^\top\boldsymbol{\Psi}^{-1}(\mathbf{y}-\boldsymbol{X\beta}). \tag{27.0.14}$$

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:

$$\frac{\partial}{\partial\sigma^2}\log f_{\mathbf{y}}(\mathbf{y}) = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{(\mathbf{y}-\boldsymbol{X\beta})^\top\boldsymbol{\Psi}^{-1}(\mathbf{y}-\boldsymbol{X\beta})}{2\sigma^4} \tag{27.0.15}$$

$$\tilde\sigma^2 = \frac{(\mathbf{y}-\boldsymbol{X\beta})^\top\boldsymbol{\Psi}^{-1}(\mathbf{y}-\boldsymbol{X\beta})}{n}. \tag{27.0.16}$$

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 $\boldsymbol{\Psi}$:

$$\log f_{\mathbf{y}}(\mathbf{y};\tilde\sigma^2) = -\frac{n}{2}(1+\ln 2\pi-\ln n) - \frac{n}{2}\ln(\mathbf{y}-\boldsymbol{X\beta})^\top\boldsymbol{\Psi}^{-1}(\mathbf{y}-\boldsymbol{X\beta}) - \frac{1}{2}\ln\det[\boldsymbol{\Psi}] \tag{27.0.17}$$

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 (19.0.11), 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 $(\mathbf{y}-\boldsymbol{X\beta})^\top\boldsymbol{\Psi}^{-1}(\mathbf{y}-\boldsymbol{X\beta})$, 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

$$\begin{aligned}(\boldsymbol{y}-\boldsymbol{X\beta})^\top(\boldsymbol{y}-\boldsymbol{X\beta}) &= (\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}) + (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^\top\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) \\ &= (n-k)\mathrm{s}^2 + (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^\top\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}).\end{aligned} \tag{27.0.18}$$

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 332. *12 points The log likelihood function in the linear model is given by* (27.0.5). *Show that the inverse of the information matrix is*

$$\begin{bmatrix} \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1} & \boldsymbol{o} \\ \boldsymbol{o}^\top & 2\sigma^4/n \end{bmatrix} \tag{27.0.19}$$

The information matrix can be obtained in two different ways. Its typical element has the following two forms:

$$\mathrm{E}[\frac{\partial\ln\ell}{\partial\theta_i}\frac{\partial\ln\ell}{\partial\theta_k}] = -\mathrm{E}[\frac{\partial^2\ln\ell}{\partial\theta_i\partial\theta_k}], \tag{27.0.20}$$

or written as matrix derivatives

$$\mathcal{E}[\frac{\partial\ln\ell}{\partial\boldsymbol{\theta}}\frac{\partial\ln\ell}{\partial\boldsymbol{\theta}^\top}] = -\mathcal{E}[\frac{\partial^2\ln\ell}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}]. \tag{27.0.21}$$

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

$$\ln\ell = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}(\mathbf{y}^\top\mathbf{y} - 2\mathbf{y}^\top\boldsymbol{X\beta} + \boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{X\beta}). \tag{27.0.22}$$

The first derivatives were already computed for the maximum likelihood estimators:

$$\frac{\partial}{\partial \boldsymbol{\beta}^\top} \ln \ell = -\frac{1}{2\sigma^2}(2\mathbf{y}^\top \boldsymbol{X} + 2\boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{X}) = \frac{1}{\sigma^2}(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta})^\top \boldsymbol{X} = \frac{1}{\sigma^2}\boldsymbol{\varepsilon}^\top \boldsymbol{X} \tag{27.0.23}$$

$$\frac{\partial}{\partial \sigma^2} \ln \ell = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon} \tag{27.0.24}$$

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

$$\mathcal{E}[\frac{1}{\sigma^4}\boldsymbol{X}^\top \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top \boldsymbol{X}] = \frac{1}{\sigma^2}\boldsymbol{X}^\top \boldsymbol{X} \tag{27.0.25}$$

and

$$\mathrm{E}[\Big(-\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon}\Big)^2] = \mathrm{var}[-\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon}] = \mathrm{var}[\frac{1}{2\sigma^4}\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon}] = \frac{1}{4\sigma^8}2n\sigma^4 = \frac{n}{2\sigma^4} \tag{27.0.26}$$

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:

$$\frac{\partial^2}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^\top} \ln \ell = -\frac{1}{\sigma^2}\boldsymbol{X}^\top \boldsymbol{X} \tag{27.0.27}$$

$$\frac{\partial^2}{\partial \boldsymbol{\beta}\partial \sigma^2} \ln \ell = -\frac{1}{\sigma^4}\boldsymbol{X}^\top (\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}) = -\frac{1}{\sigma^4}\boldsymbol{X}^\top \boldsymbol{\varepsilon} \tag{27.0.28}$$

$$\frac{\partial^2}{(\partial \sigma^2)^2} \ln \ell = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}) = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon} \tag{27.0.29}$$

This gives the top matrix in [**JHG+88**, (6.1.24b)]:

$$\begin{bmatrix} -\frac{1}{\sigma^2}\boldsymbol{X}^\top \boldsymbol{X} & -\frac{1}{\sigma^4}(\boldsymbol{X}^\top \mathbf{y} - \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta}) \\ -\frac{1}{\sigma^4}(\boldsymbol{X}^\top \mathbf{y} - \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta})^\top & \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}) \end{bmatrix} \tag{27.0.30}$$

Now assume that $\boldsymbol{\beta}$ and σ^2 are the true values, take expected values, and reverse the sign. This gives the information matrix

$$\begin{bmatrix} \sigma^{-2}\boldsymbol{X}^\top \boldsymbol{X} & \boldsymbol{o} \\ \boldsymbol{o}^\top & n/(2\sigma^4) \end{bmatrix} \tag{27.0.31}$$

For the lower righthand side corner we need that $\mathrm{E}[(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta})] = \mathrm{E}[\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon}] = n\sigma^2$.

Taking inverses gives (27.0.19), which is a lower bound for the covariance matrix; we see that s^2 with $\mathrm{var}[\mathrm{s}^2] = 2\sigma^4/(n-k)$ does not attain the bound. However one can show with other means that it is nevertheless efficient. □

CHAPTER 28

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 $\mathbf{X}$.

28.1. 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 (18.0.7):

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\varepsilon} \tag{28.1.1}$$

$$\mathcal{E}[\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|\mathbf{X}] = \mathcal{E}[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\varepsilon}|\mathbf{X}] = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\,\mathcal{E}[\boldsymbol{\varepsilon}|\mathbf{X}] = \boldsymbol{o}. \tag{28.1.2}$$

$$\mathcal{E}[\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}] = \mathcal{E}\big[\mathcal{E}[\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|\mathbf{X}]\big] = \boldsymbol{o}. \tag{28.1.3}$$

Problem 333. *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 334. *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 $\hat{\boldsymbol{\beta}}$:

$$\mathcal{MSE}[\hat{\boldsymbol{\beta}};\boldsymbol{\beta}] = \mathcal{E}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top] \tag{28.1.4}$$

$$= \mathcal{E}\big[\mathcal{E}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top|\mathbf{X}]\big] \tag{28.1.5}$$

$$= \mathcal{E}[\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}] \tag{28.1.6}$$

$$= \sigma^2\,\mathcal{E}[(\mathbf{X}^\top\mathbf{X})^{-1}]. \tag{28.1.7}$$

Problem 335. *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.

$$\mathcal{E}[\mathrm{s}^2(\mathbf{X}^\top\mathbf{X})^{-1}] = \mathcal{E}\big[\mathcal{E}[\mathrm{s}^2(\mathbf{X}^\top\mathbf{X})^{-1}|\mathbf{X}]\big] \tag{28.1.8}$$

$$= \mathcal{E}[\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}] \tag{28.1.9}$$

$$= \sigma^2\,\mathcal{E}[(\mathbf{X}^\top\mathbf{X})^{-1}] \tag{28.1.10}$$

$$= \mathcal{MSE}[\hat{\boldsymbol{\beta}};\boldsymbol{\beta}] \quad \text{by (28.1.7).} \tag{28.1.11}$$

□

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 **??**; the generalized linear model).

28.2. 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 336. *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! □

28.3. 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 **??**), 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 337. *3 points What does one have to watch out for if some of the regressors are random?*

CHAPTER 29

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

$$\begin{bmatrix}\mathbf{y}\\ \boldsymbol{z}\end{bmatrix} \sim \begin{bmatrix}\boldsymbol{X}\\ \boldsymbol{W}\end{bmatrix}\boldsymbol{\beta},\ \sigma^2\begin{bmatrix}\boldsymbol{\Omega}_{\mathbf{y}\mathbf{y}} & \boldsymbol{\Omega}_{\mathbf{y}\boldsymbol{z}}\\ \boldsymbol{\Omega}_{\boldsymbol{z}\mathbf{y}} & \boldsymbol{\Omega}_{\boldsymbol{z}\boldsymbol{z}}\end{bmatrix}, \qquad (29.0.1)$$

$\sigma^2 > 0$, otherwise u
$\boldsymbol{\beta}$ 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.

29.1. 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 29.1.1. 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

$$\mathrm{MHD}[\boldsymbol{y};\mathbf{y}] = \frac{1}{n}\max_{\boldsymbol{g}}\frac{\left(\boldsymbol{g}^\top\boldsymbol{y} - \mathrm{E}[\boldsymbol{g}^\top\mathbf{y}]\right)^2}{\mathrm{var}[\boldsymbol{g}^\top\mathbf{y}]}. \qquad (29.1.1)$$

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 29.1.2. *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*

$$\mathrm{MHD}[\boldsymbol{y};\mathbf{y}] = \frac{1}{n\sigma^2}(\boldsymbol{y}-\boldsymbol{\mu})^\top\boldsymbol{\Omega}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}) \qquad (29.1.2)$$

Proof. (29.1.2) is a simple consequence of (25.4.4). It is also somewhat intuitive since the righthand side of (29.1.2) 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 338. *Let $\mathbf{y}$ 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 $\mathbf{y}$ from the distribution of $\mathbf{y}$ is 1, i.e.,*

$$\mathrm{E}\big[\mathrm{MHD}[\mathbf{y};\mathbf{y}]\big] = 1 \qquad (29.1.3)$$

Answer.

(29.1.4)
$$\mathrm{E}[\frac{1}{n\sigma^2}(\mathbf{y}-\boldsymbol{\mu})^\top\boldsymbol{\Omega}^{-1}(\mathbf{y}-\boldsymbol{\mu})] = \mathrm{E}[\mathrm{tr}\,\frac{1}{n\sigma^2}\boldsymbol{\Omega}^{-1}(\mathbf{y}-\boldsymbol{\mu})(\mathbf{y}-\boldsymbol{\mu})^\top]\,\mathrm{tr}(\frac{1}{n\sigma^2}\boldsymbol{\Omega}^{-1}\sigma^2\boldsymbol{\Omega}) = \frac{1}{n}\,\mathrm{tr}(\boldsymbol{I}) = 1.$$

□

(29.1.2) is, up to a constant factor, the quadratic form in the exponent of the normal density function of $\mathbf{y}$. For a normally distributed $\mathbf{y}$, 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 $\mathbf{y}$ is singular. In this case, certain nonzero linear combinations of the elements of $\mathbf{y}$ are known with certainty. Certain vectors can therefore not possibly be realizations of $\mathbf{y}$, i.e., the set of realizations of $\mathbf{y}$ does not fill the whole $\mathbb{R}^n$.

Problem 339. *2 points The random vector* $\mathbf{y} = \begin{bmatrix} \mathrm{y}_1 \\ \mathrm{y}_2 \\ \mathrm{y}_3 \end{bmatrix}$ *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* $\mathbf{y}$ *which is known with certainty. And give a value which can never be a realization of* $\mathbf{y}$. *Prove everything you state.*

Answer. Yes, it is singular;

(29.1.5)
$$\begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

I.e., $\mathrm{y}_1+\mathrm{y}_2+\mathrm{y}_3 = 0$ 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 29.1.3. Given a vector random variable $\mathbf{y}$ 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 $\mathrm{var}[\boldsymbol{g}^\top\mathbf{y}] = 0$ but $\boldsymbol{g}^\top\boldsymbol{y} \neq \boldsymbol{g}^\top\,\mathcal{E}[\mathbf{y}]$. If such a $\boldsymbol{g}$ does not exist, then the squared Mahalanobis distance of $\boldsymbol{y}$ from $\mathbf{y}$ is defined as in (29.1.1), with n replaced by $\mathrm{rank}[\boldsymbol{\Omega}]$. If the denominator in (29.1.1) 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 (29.1.2) 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 340. *a is a scalar. What is its g-inverse* a^-*?*

Theorem 29.1.4. *Let* $\mathbf{y}$ *be a random variable with* $\mathcal{E}[\mathbf{y}] = \boldsymbol{\mu}$ *and* $\mathcal{V}[\mathbf{y}] = \sigma^2\boldsymbol{\Omega}$, $\sigma^2 > 0$. *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* $\mathbf{y}$ *is infinite, i.e.,* $\mathrm{MHD}[\boldsymbol{y};\mathbf{y}] = \infty$*; otherwise*

(29.1.6)
$$\mathrm{MHD}[\boldsymbol{y};\mathbf{y}] = \frac{1}{\sigma^2\,\mathrm{rank}[\boldsymbol{\Omega}]}(\boldsymbol{y}-\boldsymbol{\mu})^\top\boldsymbol{\Omega}^-(\boldsymbol{y}-\boldsymbol{\mu})$$

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.

CHAPTER 30

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.

30.1. 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,

$$\tilde{\boldsymbol{\beta}} \in B_{\boldsymbol{\beta};\pi}(\mathbf{y}) \iff \frac{(\mathbf{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}})^\top(\mathbf{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}})}{(\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})} \leq c_{\pi;n-k,k} \tag{30.1.1}$$

It will be shown below, in the discussion following (30.2.1), 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 (30.1.1) (with the parameters adjusted to reflect the dimensionality of $\tilde{\boldsymbol{u}}$) which satisfies $\boldsymbol{R}\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{u}}$:

$$\tilde{\boldsymbol{u}} \in B_{\boldsymbol{R}\boldsymbol{\beta};\pi}(\mathbf{y}) \iff \text{exist } \tilde{\boldsymbol{\beta}} \text{ with } \tilde{\boldsymbol{u}} = \boldsymbol{R}\tilde{\boldsymbol{\beta}} \text{ and } \frac{(\mathbf{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}})^\top(\mathbf{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}})}{(\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})} \leq c_{\pi;n-k,i} \tag{30.1.2}$$

Problem 341. *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 $\boldsymbol{R}\boldsymbol{\beta}$ 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 (30.1.1), this best $\hat{\hat{\boldsymbol{\beta}}}$ being, of course, the constrained least squares estimate subject to the constraint $\boldsymbol{R}\boldsymbol{\beta} = \tilde{\boldsymbol{u}}$, whose formula is given by (22.3.13). The confidence region for $\boldsymbol{R}\boldsymbol{\beta}$ 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 (30.1.1):

$$\tilde{\boldsymbol{u}} \in B_{\boldsymbol{R\beta}}(\mathsf{y}) \iff \frac{(\mathsf{y}-\boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}})^\top(\mathsf{y}-\boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}})}{(\mathsf{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\mathsf{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})} \leq c_{\pi;n-k,i} \tag{30.1.3}$$

One can also write it as

$$\tilde{\boldsymbol{u}} \in B_{\boldsymbol{R\beta}}(\mathsf{y}) \iff \frac{\mathsf{SSE}_{\text{constrained}}}{\mathsf{SSE}_{\text{unconstrained}}} \leq c_{\pi;n-k,i} \tag{30.1.4}$$

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.

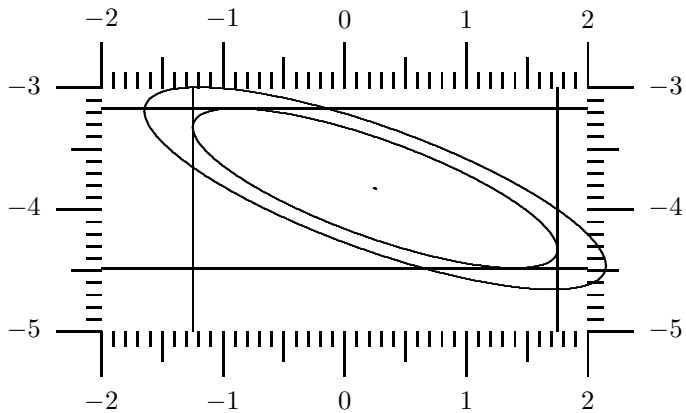


FIGURE 1. Confidence Ellipse with "Shadows"

In order to transform (30.1.3) 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 (22.7.2) to get

$$\tilde{\boldsymbol{u}} \in B_{\boldsymbol{R\beta};\pi}(\mathsf{y}) \iff \frac{(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{u}})^\top\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{u}})}{(\mathsf{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\mathsf{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})} \leq c_{\pi;n-k,i}-1 \tag{30.1.5}$$

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 342. *Using* (14.2.12), *show that* (30.1.1) *can be rewritten as*

$$\tilde{\boldsymbol{\beta}} \in B_{\boldsymbol{\beta};\pi}(\mathsf{y}) \iff \frac{(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})^\top\boldsymbol{X}^\top\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})}{(\mathsf{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\mathsf{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})} \leq c_{\pi;n-k,k}-1 \tag{30.1.6}$$

Verify that this is the same as (30.1.5) *in the special case* $\boldsymbol{R}=\boldsymbol{I}$.

FIGURE 2. Confidence Band for Regression Line

PROBLEM 343. *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 304, show that this confidence region has the form*

$$\tilde{\boldsymbol{\beta}} \in B_{\boldsymbol{\beta};\pi}(\mathsf{y}) \iff \frac{(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})^\top\underline{\boldsymbol{X}}^\top\underline{\boldsymbol{X}}(\hat{\boldsymbol{\beta}}-\tilde{\boldsymbol{\beta}})}{(\underline{\mathsf{y}}-\underline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})^\top(\underline{\mathsf{y}}-\underline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})} \leq c_{\pi;n-k,k-1}-1 \tag{30.1.7}$$

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 (30.1.6) *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 (30.1.5) but instead of $\tilde{\boldsymbol{u}}$ write $\tilde{\boldsymbol{\beta}}$. The only tricky part is the following which uses (23.0.37):

$$\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top=\begin{bmatrix}\boldsymbol{o} & \boldsymbol{I}\end{bmatrix}\begin{bmatrix}1/n+\bar{\boldsymbol{x}}^\top(\underline{\boldsymbol{X}}^\top\underline{\boldsymbol{X}})^{-1}\bar{\boldsymbol{x}} & -\bar{\boldsymbol{x}}^\top(\underline{\boldsymbol{X}}^\top\underline{\boldsymbol{X}})^{-1}\\ -(\underline{\boldsymbol{X}}^\top\underline{\boldsymbol{X}})^{-1}\bar{\boldsymbol{x}} & (\underline{\boldsymbol{X}}^\top\underline{\boldsymbol{X}})^{-1}\end{bmatrix}\begin{bmatrix}\boldsymbol{o}^\top\\ \boldsymbol{I}\end{bmatrix}=(\underline{\boldsymbol{X}}^\top\underline{\boldsymbol{X}})^{-1} \tag{30.1.8}$$

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 204, 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 344. *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.*

30.2. Coverage Probability of the Confidence Regions

The probability that any *given known value* $\tilde{\boldsymbol{u}}$ lies in the confidence region (30.1.3) 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 (30.1.5). This gives

(30.2.1)

$$\boldsymbol{R\beta} \in B_{\boldsymbol{R\beta};\pi}(\mathsf{y}) \iff \frac{(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{R\beta})^\top \big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{R\beta})}{(\mathsf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\mathsf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})} \leq c_{\pi;n-k,i} - 1$$

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 (7.4.9), the distribution of the numerator of (30.2.1) is

$$(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{R\beta})^\top \big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{R\beta}) \sim \sigma^2\chi^2_i. \tag{30.2.2}$$

This probability distribution only depends on one unknown parameter, namely, σ^2. Regarding the denominator, remember that, by (18.4.2), $(\mathsf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\mathsf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \boldsymbol{\varepsilon}^\top\boldsymbol{M}\boldsymbol{\varepsilon}$, and if we apply (7.4.9) to this we can see that

$$(\mathsf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\mathsf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \sim \sigma^2\chi^2_{n-k} \tag{30.2.3}$$

Furthermore, numerator and denominator are independent. To see this, look first at $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\varepsilon}}$. By Problem 246 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 (30.2.1) 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 (30.2.1) 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 (30.1.5) 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.

30.3. 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 (30.1.5) 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:

(30.3.1)

$$\frac{(\mathsf{SSE}_{\text{constrained}} - \mathsf{SSE}_{\text{unconstrained}})/\text{number of constraints}}{\mathsf{SSE}_{\text{unconstr.}}/(\text{numb. of obs.} - \text{numb. of coeff. in unconstr. model})} \leq F_{(i,n-k;\alpha)}$$

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 $\Pr[\mathsf{F} \geq c] = \alpha$.

30.4. 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 29. If one divides the denominator by its degrees of freedom, one gets an unbiased estimate of σ^2

$$\mathsf{s}^2 = \frac{1}{n-k}(\mathsf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\mathsf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}). \tag{30.4.1}$$

Therefore from (30.1.5) one gets the following alternative formula for the joint confidence region $B(\mathsf{y})$ for the vector parameter $\boldsymbol{u} = \boldsymbol{R\beta}$ for confidence level $\pi = 1-\alpha$:

(30.4.2)

$$\tilde{\boldsymbol{u}} \in B_{\boldsymbol{R\beta};1-\alpha}(\mathsf{y}) \iff \frac{1}{\mathsf{s}^2}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{u}})^\top \big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{u}}) \leq iF_{(i,n-k;\alpha)}$$

Here $\hat{\boldsymbol{\beta}}$ is the least squares estimator of $\boldsymbol{\beta}$, and $\mathsf{s}^2 = (\mathsf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\mathsf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})/(n-k)$ the unbiased estimator of σ^2. Therefore $\hat{\boldsymbol{\Sigma}} = \mathsf{s}^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$ is the estimated covariance matrix as available in the regression printout. Therefore $\hat{\mathbf{V}} = \mathsf{s}^2\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top$ is the estimate of the covariance matrix of $\boldsymbol{R}\hat{\boldsymbol{\beta}}$. Another way to write (30.4.2) is therefore

$$B(\mathsf{y}) = \{\tilde{\boldsymbol{u}} \in \mathbb{R}^i\colon\ (\boldsymbol{R}\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{u}})^\top\hat{\mathbf{V}}^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{u}}) \leq iF_{(i,n-k;\alpha)}\}. \tag{30.4.3}$$

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 (7.3.22) 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 146.

PROBLEM 345. *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:*

(30.4.4)
$$\mathsf{B}(\mathbf{y}) = \{\boldsymbol{u} \in \mathbb{R}^i\colon\ (\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{u})^\top(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{u}) \leq i\mathsf{s}^2 F_{(i,n-k;\alpha)}\}.$$

Compute the probability that B *contains the true* $\boldsymbol{R\beta}$*.*

ANSWER.

$$\Pr[B(\mathbf{y}) \ni \boldsymbol{R\beta}] = \Pr[(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{R\beta})^\top(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{R\beta}) \leq iF_{(i,n-k;\alpha)}\mathsf{s}^2] = \tag{30.4.5}$$

$$= \Pr[\frac{(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{R\beta})^\top(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{R\beta})/i}{\mathsf{s}^2} \leq F_{(i,n-k;\alpha)}] = 1 - \alpha \tag{30.4.6}$$

□

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

$$\mathsf{B}(\mathbf{y}) = \{u \in \mathbb{R}\colon\ |u - \boldsymbol{r}^\top\hat{\boldsymbol{\beta}}| \leq t_{(n-k;\alpha/2)}\mathsf{s}_{\boldsymbol{r}^\top\hat{\boldsymbol{\beta}}}\}. \tag{30.4.7}$$

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 $\Pr[\mathsf{t} \geq c] = \alpha/2$. Since by symmetry $\Pr[\mathsf{t} \leq -c] = \alpha/2$ as well, one obtains the inequality relevant for a two-sided test:

$$\Pr[|\mathsf{t}| \geq t_{(n-k;\alpha/2)}] = \alpha. \tag{30.4.8}$$

Finally, $\mathsf{s}_{\boldsymbol{r}^\top\hat{\boldsymbol{\beta}}}$ 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 $\boldsymbol{r}^\top\hat{\boldsymbol{\beta}}$:

$$\operatorname{var}[\boldsymbol{r}^\top\hat{\boldsymbol{\beta}}] = \sigma^2\boldsymbol{r}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{r}. \tag{30.4.9}$$

Secondly, replace σ^2 by its unbiased estimator $\mathsf{s}^2 = (\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})/(n-k)$, 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 346. *Which element(s) on the right hand side of* (30.4.7) *depend(s) on* $\mathbf{y}$*?*

ANSWER. $\hat{\boldsymbol{\beta}}$ depends on $\mathbf{y}$, and also $\mathsf{s}_{\boldsymbol{r}^\top\hat{\boldsymbol{\beta}}}$ depends on $\mathbf{y}$ through s^2. □

Let us verify that the coverage probability, i.e., the probability that the confidence interval constructed using formula (30.4.7) contains the true value $\boldsymbol{r}^\top\boldsymbol{\beta}$, is, as claimed, $1 - \alpha$:

(30.4.10)

$$\Pr[\mathsf{B}(\mathbf{y}) \ni \boldsymbol{r}^\top\boldsymbol{\beta}] = \Pr[\left|\boldsymbol{r}^\top\boldsymbol{\beta} - \boldsymbol{r}^\top\hat{\boldsymbol{\beta}}\right| \leq t_{(n-k;\alpha/2)}\mathsf{s}_{\boldsymbol{r}^\top\hat{\boldsymbol{\beta}}}]$$

$$= \Pr\left[\left|\boldsymbol{r}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon}\right| \leq t_{(n-k;\alpha/2)}\mathsf{s}\sqrt{\boldsymbol{r}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{r}}\right] \tag{30.4.11}$$

$$= \Pr[\left|\frac{\boldsymbol{r}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon}}{\mathsf{s}\sqrt{\boldsymbol{r}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{r}}}\right| \leq t_{(n-k;\alpha/2)}] \tag{30.4.12}$$

$$= \Pr[\left|\frac{\boldsymbol{r}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon}}{\sigma\sqrt{\boldsymbol{r}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{r}}}\Big/\frac{\mathsf{s}}{\sigma}\right| \leq t_{(n-k;\alpha/2)}] = 1 - \alpha, \tag{30.4.13}$$

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 (30.4.13) follows from (30.4.8).

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 $\leq$ 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 31 is therefore a discussion of the same subject under a slightly different angle:

CHAPTER 31

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.

31.1. 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

$$\frac{1}{i}\frac{(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{u})^\top\big(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{u})}{\mathsf{s}^2}. \tag{31.1.1}$$

(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

$$\frac{\mathsf{SSE}_r/(n+i-k)}{\mathsf{SSE}_u/(n-k)} \tag{31.1.2}$$

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.

$$\frac{(\mathsf{SSE}_r - \mathsf{SSE}_u)/i}{\mathsf{SSE}_u/(n-k)} \tag{31.1.3}$$

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 (31.1.3) can also be written as

$$\frac{(\mathsf{SSE}_{\text{constrained}} - \mathsf{SSE}_{\text{unconstrained}})/\text{number of constraints}}{\mathsf{SSE}_{\text{unconstrained}}/(\text{numb. of observations} - \text{numb. of coefficients in unconstr. model})} \tag{31.1.4}$$

The equivalence of formulas (31.1.1) and (31.1.4) is a simple consequence of (22.7.2).

(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

$$\frac{\partial}{\partial\boldsymbol{\beta}^\top}(\boldsymbol{y} - \boldsymbol{X\beta})^\top(\boldsymbol{y} - \boldsymbol{X\beta}) = -2(\boldsymbol{y} - \boldsymbol{X\beta})^\top\boldsymbol{X}. \tag{31.1.5}$$

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

$$-2\boldsymbol{X}^\top(\mathsf{y}-\boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}}) \tag{31.1.6}$$

has if the true $\boldsymbol{\beta}$ satisfies $\boldsymbol{R\beta}=\boldsymbol{u}$. If this constraint is satisfied, $\hat{\hat{\boldsymbol{\beta}}}$ is unbiased, therefore (31.1.6) has expected value zero. Furthermore, if one premultiplies (22.7.1) 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

$$\frac{1}{\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}})^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{y}-\boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}}) \tag{31.1.7}$$

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 (22.3.13).

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 (22.3.12), 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

$$\boldsymbol{\lambda}^\top(\mathcal{V}[\boldsymbol{\lambda}])^{-1}\boldsymbol{\lambda}=\frac{1}{4\sigma^2}\boldsymbol{\lambda}^\top\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\boldsymbol{\lambda} \tag{31.1.8}$$

Using (22.7.1) one can verify that this is the same as (31.1.7).

Problem 347. *Show that* (31.1.7) *is equal to the righthand side of* (31.1.8).

Problem 348. *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:*

$$\hat{\hat{\boldsymbol{\varepsilon}}}^\top\hat{\hat{\boldsymbol{\varepsilon}}}-\hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}}=(\hat{\boldsymbol{\beta}}-\hat{\hat{\boldsymbol{\beta}}})^\top\boldsymbol{X}^\top\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\hat{\hat{\boldsymbol{\beta}}}) \tag{31.1.9}$$

$$=(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{u})^\top(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{u}) \tag{31.1.10}$$

$$=\frac{1}{4}\boldsymbol{\lambda}^\top\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top\boldsymbol{\lambda} \tag{31.1.11}$$

$$=\hat{\hat{\boldsymbol{\varepsilon}}}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\hat{\hat{\boldsymbol{\varepsilon}}} \tag{31.1.12}$$

$$=(\hat{\hat{\boldsymbol{\varepsilon}}}-\hat{\boldsymbol{\varepsilon}})^\top(\hat{\hat{\boldsymbol{\varepsilon}}}-\hat{\boldsymbol{\varepsilon}}) \tag{31.1.13}$$

Furthermore show that

$$\boldsymbol{X}^\top\boldsymbol{X} \quad \textit{is } \sigma^2 \textit{ times a g-inverse of} \quad \mathcal{V}[\hat{\boldsymbol{\beta}}-\hat{\hat{\boldsymbol{\beta}}}] \tag{31.1.14}$$

$$(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top)^{-1} \quad \textit{is } \sigma^2 \textit{ times the inverse of} \quad \mathcal{V}[\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{u}] \tag{31.1.15}$$

$$\frac{1}{4}\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top \quad \textit{is } \sigma^2 \textit{ times the inverse of} \quad \mathcal{V}[\boldsymbol{\lambda}] \tag{31.1.16}$$

$$(\boldsymbol{X}^\top\boldsymbol{X})^{-1} \quad \textit{is } \sigma^2 \textit{ times a g-inverse of} \quad \mathcal{V}[\boldsymbol{X}^\top(\mathsf{y}-\boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}})] \tag{31.1.17}$$

$$\boldsymbol{I} \quad \textit{is } \sigma^2 \textit{ times a g-inverse of} \quad \mathcal{V}[\hat{\hat{\boldsymbol{\varepsilon}}}-\hat{\boldsymbol{\varepsilon}}] \tag{31.1.18}$$

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* (31.1.12) *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.

$$\hat{\hat{\boldsymbol{\varepsilon}}}=\boldsymbol{y}-\boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}}=\boldsymbol{X}\hat{\boldsymbol{\beta}}+\hat{\boldsymbol{\varepsilon}}-\boldsymbol{X}\hat{\hat{\boldsymbol{\beta}}}=\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\hat{\hat{\boldsymbol{\beta}}})+\hat{\boldsymbol{\varepsilon}}, \tag{31.1.19}$$

and since $\boldsymbol{X}^\top\hat{\boldsymbol{\varepsilon}}=\boldsymbol{o}$, the righthand decomposition is an orthogonal decomposition. This gives (31.1.9) above:

$$\hat{\hat{\boldsymbol{\varepsilon}}}^\top\hat{\hat{\boldsymbol{\varepsilon}}}=(\hat{\boldsymbol{\beta}}-\hat{\hat{\boldsymbol{\beta}}})^\top\boldsymbol{X}^\top\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\hat{\hat{\boldsymbol{\beta}}})+\hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}}, \tag{31.1.20}$$

Using (22.3.13) 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 (31.1.10), which is (22.7.2), one has to plug (22.3.13) into (31.1.20). Clearly, $\mathcal{V}[\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{u}]=\sigma^2\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top$.

For (31.1.11) one needs the formula for the Lagrange multiplier (22.3.12).

□

The test statistic defined alternatively either by (31.1.1) or (31.1.4) or (31.1.7) or (31.1.8) has the following nice properties:

- $\mathrm{E}(\mathsf{SSE}_u)=\mathrm{E}(\hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}})=\sigma^2(n-k)$, which holds whether or not the constraint is true. Furthermore it was shown earlier that

$$\mathrm{E}(\mathsf{SSE}_r-\mathsf{SSE}_u)=\sigma^2 i+(\boldsymbol{R\beta}-\boldsymbol{u})^\top(\boldsymbol{R}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{R}^\top)^{-1}(\boldsymbol{R\beta}-\boldsymbol{u}), \tag{31.1.21}$$

  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 (31.1.4), 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 $\boldsymbol{\varepsilon}$ is normal, then numerator and denominator are independent. The numerator is a function of $\hat{\boldsymbol{\beta}}$ and the denominator one of $\hat{\boldsymbol{\varepsilon}}$, and $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\varepsilon}}$ 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).

31.2. 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 $\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$, 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 (31.1.1) becomes

$$\frac{(\hat\beta_j - u)^2}{\mathsf{s}^2 d_{jj}} \sim F_{1,n-k} \quad \text{when } H \text{ is true.} \tag{31.2.1}$$

This is the square of a random variable which has a t-distribution:

$$\frac{\hat\beta_j - u}{\mathsf{s}\sqrt{d_{jj}}} \sim t_{n-k} \quad \text{when } H \text{ is true.} \tag{31.2.2}$$

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 $\boldsymbol{r}^\top\hat{\boldsymbol{\beta}}$:

$$\frac{\boldsymbol{r}^\top\hat{\boldsymbol{\beta}} - u}{\mathsf{s}\sqrt{\boldsymbol{r}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{r}}} \sim t_{n-k} \quad \text{when } H \text{ is true.} \tag{31.2.3}$$

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 349. *4 points In the model $\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with intercept, show that the test statistic for testing whether all the slope parameters are zero is*

$$\frac{(\mathbf{y}^\top\boldsymbol{X}\hat{\boldsymbol{\beta}} - n\bar{\mathbf{y}}^2)/(k-1)}{(\mathbf{y}^\top\mathbf{y} - \mathbf{y}^\top\boldsymbol{X}\hat{\boldsymbol{\beta}})/(n-k)} \tag{31.2.4}$$

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 $\sim \mathsf{F}_{k-1,n-k}$. (31.2.4) is most conveniently derived from (31.1.4). In the constrained model, which has only a constant term and no other explanatory variables, i.e., $\mathbf{y} = \boldsymbol{\iota}\mu + \boldsymbol{\varepsilon}$, the BLUE is $\hat\mu = \bar{\mathbf{y}}$. 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):

$$\mathsf{SSE}_{\text{const.}} = \mathsf{SST} = (\mathbf{y} - \boldsymbol{\iota}\bar{\mathbf{y}})^\top(\mathbf{y} - \boldsymbol{\iota}\bar{\mathbf{y}}) = \mathbf{y}^\top(\mathbf{y} - \boldsymbol{\iota}\bar{\mathbf{y}}) = \mathbf{y}^\top\mathbf{y} - n\bar{\mathbf{y}}^2 \tag{31.2.5}$$

while the unconstrained residual sum of squares is what is usually called SSE:

$$\mathsf{SSE}_{\text{unconst.}} = \mathsf{SSE} = (\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \mathbf{y}^\top(\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \mathbf{y}^\top\mathbf{y} - \mathbf{y}^\top\boldsymbol{X}\hat{\boldsymbol{\beta}}. \tag{31.2.6}$$

This last equation because $\boldsymbol{X}^\top(\mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \boldsymbol{X}^\top\hat{\boldsymbol{\varepsilon}} = \boldsymbol{o}$. A more elegant way is perhaps

$$\mathsf{SSE}_{\text{unconst.}} = \mathsf{SSE} = \hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}} = \mathbf{y}^\top\boldsymbol{M}^\top\boldsymbol{M}\mathbf{y} = \mathbf{y}^\top\boldsymbol{M}\mathbf{y} = \mathbf{y}^\top\mathbf{y} - \mathbf{y}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\mathbf{y} = \mathbf{y}^\top\mathbf{y} - \mathbf{y}^\top\boldsymbol{X}\hat{\boldsymbol{\beta}} \tag{31.2.7}$$

According to (14.3.12) we can write $\mathsf{SSR} = \mathsf{SST} - \mathsf{SSE}$, therefore the F-statistic is

$$\frac{\mathsf{SSR}/(k-1)}{\mathsf{SSE}/(n-k)} = \frac{(\mathbf{y}^\top\boldsymbol{X}\hat{\boldsymbol{\beta}} - n\bar{\mathbf{y}}^2)/(k-1)}{(\mathbf{y}^\top\mathbf{y} - \mathbf{y}^\top\boldsymbol{X}\hat{\boldsymbol{\beta}})/(n-k)} \sim \mathsf{F}_{k-1,n-k} \quad \text{if } H_0 \text{ is true.} \tag{31.2.8}$$

□

Problem 350. *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.

$$F = \frac{\mathsf{SSR}/(k-1)}{\mathsf{SSE}/(n-k)} = \frac{n-k}{k-1}\frac{\mathsf{SSR}}{\mathsf{SST}-\mathsf{SSR}} = \frac{n-k}{k-1}\frac{R^2}{1-R^2}. \tag{31.2.9}$$

□

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 351. *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 291, 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 $\boldsymbol{R\beta} = \begin{bmatrix}1 & 0 & -1 & 0 & \cdots & 0\end{bmatrix}\begin{bmatrix}\beta_1\\ \vdots \\ \beta_k\end{bmatrix} = 0$. In Problem 291 we computed the "unconstrained" estimates $\hat{\boldsymbol{\beta}} = \mathsf{y} - \boldsymbol{\iota}\bar{\mathsf{y}}$ and $\mathsf{s}^2 = n\bar{\mathsf{y}}^2 = (\mathsf{y}_1 + \cdots + \mathsf{y}_n)^2/n$. 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

$$\frac{(\mathsf{SSE}_r - \mathsf{SSE}_1 - \mathsf{SSE}_2)/k}{(\mathsf{SSE}_1 + \mathsf{SSE}_2)/(n_1 + n_2 - 2k)}. \tag{31.2.10}$$

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

$$\frac{(\mathsf{SSE}_r - \mathsf{SSE}_1)/n_2}{\mathsf{SSE}_1/(n_1 - k)}. \tag{31.2.11}$$

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 352. *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 $\mathsf{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$.*

Answer.

$$\begin{bmatrix}\mathsf{u}\\ \mathsf{v}\end{bmatrix} = \begin{bmatrix}\boldsymbol{\iota}_1\mu_1 + \boldsymbol{\varepsilon}_1\\ \boldsymbol{\iota}_2\mu_2 + \boldsymbol{\varepsilon}_2\end{bmatrix} = \begin{bmatrix}\boldsymbol{\iota}_1 & \boldsymbol{o}\\ \boldsymbol{o} & \boldsymbol{\iota}_2\end{bmatrix}\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix} + \begin{bmatrix}\boldsymbol{\varepsilon}_1\\ \boldsymbol{\varepsilon}_2\end{bmatrix}. \tag{31.2.12}$$

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.

$$\boldsymbol{X}^\top\boldsymbol{X} = \begin{bmatrix}\boldsymbol{\iota}_1^\top & \boldsymbol{o}^\top\\ \boldsymbol{o}^\top & \boldsymbol{\iota}_2^\top\end{bmatrix}\begin{bmatrix}\boldsymbol{\iota}_1 & \boldsymbol{o}\\ \boldsymbol{o} & \boldsymbol{\iota}_2\end{bmatrix} = \begin{bmatrix}n_1 & 0\\ 0 & n_2\end{bmatrix} \tag{31.2.13}$$

$$(\boldsymbol{X}^\top\boldsymbol{X})^{-1} = \begin{bmatrix}\frac{1}{n_1} & 0\\ 0 & \frac{1}{n_2}\end{bmatrix} \tag{31.2.14}$$

□

• c. *2 points Compute $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\mathsf{y}$ in this case.*

Answer.

$$\boldsymbol{X}^\top\mathsf{y} = \begin{bmatrix}\boldsymbol{\iota}_1^\top & \boldsymbol{o}^\top\\ \boldsymbol{o}^\top & \boldsymbol{\iota}_2^\top\end{bmatrix}\begin{bmatrix}\mathsf{u}\\ \mathsf{v}\end{bmatrix} = \begin{bmatrix}\sum_{i=1}^{n_1}\mathsf{u}_i\\ \sum_{j=1}^{n_2}\mathsf{v}_j\end{bmatrix} \tag{31.2.15}$$

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\mathsf{y} = \begin{bmatrix}\frac{1}{n_1} & 0\\ 0 & \frac{1}{n_2}\end{bmatrix}\begin{bmatrix}\sum_{i=1}^{n_1}\mathsf{u}_i\\ \sum_{j=1}^{n_2}\mathsf{v}_j\end{bmatrix} = \begin{bmatrix}\bar{\mathsf{u}}\\ \bar{\mathsf{v}}\end{bmatrix} \tag{31.2.16}$$

□

• d. *3 points Compute $\mathsf{SSE} = (\mathsf{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\mathsf{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})$ and s^2, the unbiased estimator of σ^2, in this case.*

Answer.

$$\mathsf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = \begin{bmatrix}\mathsf{u}\\ \mathsf{v}\end{bmatrix} - \begin{bmatrix}\boldsymbol{\iota}_1 & \boldsymbol{o}\\ \boldsymbol{o} & \boldsymbol{\iota}_2\end{bmatrix}\begin{bmatrix}\bar{\mathsf{u}}\\ \bar{\mathsf{v}}\end{bmatrix} = \begin{bmatrix}\mathsf{u} - \boldsymbol{\iota}_1\bar{\mathsf{u}}\\ \mathsf{v} - \boldsymbol{\iota}_2\bar{\mathsf{u}}\end{bmatrix} \tag{31.2.17}$$

$$\mathsf{SSE} = \sum_{i=1}^{n_1}(\mathsf{u}_i - \bar{\mathsf{u}})^2 + \sum_{j=1}^{n_2}(\mathsf{v}_j - \bar{\mathsf{v}})^2 \tag{31.2.18}$$

$$\mathsf{s}^2 = \frac{\sum_{i=1}^{n_1}(\mathsf{u}_i - \bar{\mathsf{u}})^2 + \sum_{j=1}^{n_2}(\mathsf{v}_j - \bar{\mathsf{v}})^2}{n_1 + n_2 - 2} \tag{31.2.19}$$

□

• e. *1 point Next, the hypothesis $\mu_1 = \mu_2$ must be written in the form $\boldsymbol{R\beta} = u$. 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

$$\begin{bmatrix}1 & -1\end{bmatrix}\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix} = 0 \qquad \text{i.e.,} \qquad \boldsymbol{r} = \begin{bmatrix}1\\ -1\end{bmatrix} \qquad \text{and} \qquad u = 0 \tag{31.2.20}$$

□

• 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.

$$\operatorname{var}[\boldsymbol{r}^\top\hat{\boldsymbol{\beta}}] = \sigma^2\boldsymbol{r}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{r} = \sigma^2\begin{bmatrix}1 & -1\end{bmatrix}\begin{bmatrix}\frac{1}{n_1} & 0\\ 0 & \frac{1}{n_2}\end{bmatrix}\begin{bmatrix}1\\ -1\end{bmatrix} = \sigma^2\Big(\frac{1}{n_1} + \frac{1}{n_2}\Big) \tag{31.2.21}$$

One can also see this without matrix algebra. $\text{var}[\bar{\mathsf{u}} = \sigma^2 \frac{1}{n_1}$, $\text{var}[\bar{\mathsf{v}} = \sigma^2 \frac{1}{n_2}$, 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* (31.2.3) *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.,

$$\frac{\bar{\mathsf{u}} - \bar{\mathsf{v}}}{\mathsf{s}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2} \quad \text{when } H \text{ is true.} \tag{31.2.22}$$

□

Problem 353. [**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 $\text{var}[\bar{\mathsf{y}} - \mathsf{y}_{n+1}] = \text{var}[\bar{\mathsf{y}}] + \text{var}[\mathsf{y}_{n+1}] = \sigma^2((1/n) + 1) = \sigma^2(n+1)/n$. 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

$$\mathsf{s}^2 = \frac{1}{n-1}\sum_{j=1}^{n}(\mathsf{y}_j - \bar{\mathsf{y}})^2 \tag{31.2.23}$$

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

$$\frac{\bar{\mathsf{y}} - \mathsf{y}_{n+1}}{\mathsf{s}\sqrt{1 + \frac{1}{n}}} \tag{31.2.24}$$

□

• 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 $\text{MSE}[\hat{\hat{\mathsf{y}}}(n+1)_{n+1}; \mathsf{y}_{n+1}]$, 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, $\text{MSE}[\bar{\mathsf{y}}; \mathsf{y}_{n+1}] = \text{var}[\bar{\mathsf{y}} - \mathsf{y}_{n+1}] = \sigma^2(n+1)/n$. 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:

$$\mathsf{SSE}_r = \sum_{j=1}^{n+1}(y_j - \bar{\bar{y}})^2 = \sum_{j=1}^{n+1} y_j^2 - (n+1)\bar{\bar{y}}^2 = \sum_{j=1}^{n+1} y_j^2 - \frac{1}{n+1}(n\bar{y} + y_{n+1})^2$$

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

$$\frac{\mathsf{SSE}_r - \mathsf{SSE}_u}{\mathsf{SSE}_u/(n+1-2)} = \frac{\frac{n}{n+1}(y_{n+1} - \bar{y})^2}{\sum_1^n (y_j - \bar{y})^2/(n-1)} = \frac{(y_{n+1} - \bar{y})^2\, n(n-1)}{\sum_1^n (y_j - \bar{y})^2\,(n+1)} \sim F_{1,n-1} \tag{31.2.25}$$

This is the square of the t statistic (31.2.24). □

31.2.1. Goodness of Fit Test.

Problem 354. [**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*

$$\mathsf{y}_{mq} = \sum_{j=1}^{k} x_{mj}\beta_j + \varepsilon_{mq} \qquad \text{or} \quad \mathsf{y}_{mq} = \boldsymbol{x}_m^\top \boldsymbol{\beta} + \varepsilon_{mq}, \tag{31.2.26}$$

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*

$$\mathsf{y}_{mq} = \sum_{j=1}^{k} x_{mj}\beta_{mj} + \varepsilon_{mq} \qquad \text{or} \quad \mathsf{y}_{mq} = \boldsymbol{x}_m^\top \boldsymbol{\beta}_m + \varepsilon_{mq}. \tag{31.2.27}$$

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

$$\mathsf{SSE}_u = \sum_{m,q} (\mathsf{y}_{mq} - \tilde{\eta}_m)^2 \tag{31.2.28}$$

First order conditions are

$$\frac{\partial}{\partial \tilde{\eta}_h} \sum_{m,q} (\mathsf{y}_{mq} - \tilde{\eta}_m)^2 = \frac{\partial}{\partial \tilde{\eta}_h} \sum_{q} (\mathsf{y}_{hq} - \tilde{\eta}_h)^2 = -2 \sum_{q} (\mathsf{y}_{hq} - \tilde{\eta}_h) = 0 \tag{31.2.29}$$

□

• 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* (31.2.26). *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*

$$\mathsf{z}_m = \sum_{j=1}^{k} x_{mj}\beta_j + \varepsilon_m, \tag{31.2.30}$$

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* (31.2.26) *and SSE_b the SSE in* (31.2.30), *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

$$\frac{(\mathsf{SSE}_c - \mathsf{SSE}_u)/(n-k)}{\mathsf{SSE}_u/n(r-1)} = \frac{r\mathsf{SSE}_b/(n-k)}{\mathsf{SSE}_u/n(r-1)} \tag{31.2.31}$$

□

31.3. The F-Test Statistic is a Function of the Likelihood Ratio

Problem 355. *The critical region of the generalized likelihood ratio test can be written as*

$$C = \{y_1, \ldots, y_n : \frac{\sup_{\boldsymbol{\theta} \in \Omega} \ell(y_1, \ldots, y_n; \theta_1, \ldots, \theta_k)}{\sup_{\boldsymbol{\theta} \in \omega} \ell(y_1, \ldots, y_n; \theta_1, \ldots, \theta_k)} \geq k\}, \tag{31.3.1}$$

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

$$LR = 2(\log \ell(\mathbf{y}, \hat{\boldsymbol{\theta}}) - \log \ell(\mathbf{y}, \hat{\hat{\boldsymbol{\theta}}})) \to \chi_i^2 \tag{31.3.2}$$

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.

$$\text{Concentrated } \log \ell(\mathbf{y}; \boldsymbol{\beta}) = -\frac{n}{2}\Big(1 + \log 2\pi + \log \frac{1}{n}(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta})\Big) \tag{31.3.3}$$

□

• b. *2 points In the case of a linear restriction, show that LR is connected with the F-statistic F as follows:*

$$LR = n \log\Big(1 + \frac{i}{n-k} F\Big) \tag{31.3.4}$$

Answer. $LR = -n\left(\log \frac{1}{n}\hat{\varepsilon}^\top\hat{\varepsilon} - \log \frac{1}{n}\hat{\hat{\varepsilon}}^\top\hat{\hat{\varepsilon}}\right) = n \log \frac{\hat{\hat{\varepsilon}}^\top\hat{\hat{\varepsilon}}}{\hat{\varepsilon}^\top\hat{\varepsilon}}$ In order to connect this with the F statistic note that

$$F = \frac{n-k}{i}\left(\frac{\hat{\hat{\varepsilon}}^\top\hat{\hat{\varepsilon}}}{\hat{\varepsilon}^\top\hat{\varepsilon}} - 1\right) \tag{31.3.5}$$

□

31.4. Tests of Nonlinear Hypotheses

Make linear approximation, need Jacobian for this. Here is an example where a nonlinear hypothesis arises naturally:

Problem 356. [**Gre97**, Example 7.14 on p. 361]*: The model*

$$C_t = \alpha + \beta Y_t + \gamma C_{t-1} + \varepsilon_t \tag{31.4.1}$$

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 $\delta = \beta/(1-\gamma)$.

□

31.5. 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 32

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 328). 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:

$$\operatorname{plim}\frac{1}{n}\boldsymbol{W}_n^\top\boldsymbol{\varepsilon}_n = \boldsymbol{o} \tag{32.0.1}$$

$$\operatorname{plim}\frac{1}{n}\boldsymbol{W}_n^\top\boldsymbol{W}_n = \boldsymbol{Q} \quad \text{exists, is nonrandom and nonsingular} \tag{32.0.2}$$

$$\operatorname{plim}\frac{1}{n}\boldsymbol{W}_n^\top\mathbf{X}_n = \boldsymbol{D} \quad \text{exists, is nonrandom and has full column rank} \tag{32.0.3}$$

Full column rank in (32.0.3) 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 $\mathbf{x}_i$ is the ith column vector of $\mathbf{X}$, then $\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top\mathbf{x}_i$ is the projection of $\mathbf{x}_i$ 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"

$$\tilde{\boldsymbol{\beta}} = \Big(\mathbf{X}^\top\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top\mathbf{X}\Big)^{-1}\mathbf{X}^\top\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top\mathbf{y} \tag{32.0.4}$$

Problem 357. *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

$$\tilde{\boldsymbol{\beta}} = \Big(\mathbf{X}^\top\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top\mathbf{X}\Big)^{-1}\mathbf{X}^\top\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top\mathbf{y} \tag{32.0.5}$$

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 358. *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 359. *In the situation of Problem 357, 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 359 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 (32.0.4) 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 $\hat{\mathbf{X}} = \mathbf{W}\hat{\mathbf{\Pi}} = \mathbf{W}(\mathbf{W}^\top\mathbf{W})^{-1}\mathbf{W}^\top\mathbf{X}$. The second regression, which regresses $\mathbf{y}$ on $\hat{\mathbf{X}}$, gives the coefficient vector

$$\tilde{\boldsymbol{\beta}} = (\hat{\mathbf{X}}^\top\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}^\top\mathbf{y}. \tag{32.0.6}$$

If you plug this in you see this is exactly (32.0.4) again.

Now let's look at the geometry of instrumental variable regression of one variable $\mathbf{y}$ on one other variable $\mathbf{x}$ with $\mathbf{w}$ as an instrument. The specification is $\mathbf{y} = \mathbf{x}\beta + \boldsymbol{\varepsilon}$. On p. 316 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 $\mathbf{w}$ which is orthogonal to the errors. (Since $\boldsymbol{\varepsilon}$ is vertical, $\mathbf{w}$ is on the horizontal axis.) Call D the projection of $\mathbf{y}$ on $\mathbf{w}$, which is the prolongation of the vector $\boldsymbol{\varepsilon}$, and call E the end of the $\mathbf{w}$-vector, and call F the projection of $\mathbf{x}$ on $\mathbf{w}$. Then $\mathbf{w}^\top\mathbf{y} = \bar{OE}\,\bar{OD}$, and $\mathbf{w}^\top\mathbf{x} = \bar{OE}\,\bar{OF}$. Therefore $\mathbf{w}^\top\mathbf{y}/\mathbf{w}^\top\mathbf{x} = (\bar{OE}\,\bar{OD})(\bar{OE}\,\bar{OF}) = \bar{OD}/\bar{OF} = \bar{OA}/\bar{OC} = \beta$. Or geometrically it is obvious that the regression of $\mathbf{y}$ on the projection of $\mathbf{x}$ on $\mathbf{w}$ 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 $\mathbf{w}$ 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 $\mathbf{w}$ which goes through point C. The projection of $\mathbf{x}$ on $\mathbf{w}$ is the intersection of the ray generated by $\mathbf{w}$ with this plane. Now move this plane parallel until it intersects point A. Then the intersection with the $\mathbf{w}$-ray is the projection of $\mathbf{y}$ on $\mathbf{w}$. But this latter plane contains $\boldsymbol{\varepsilon}$, since $\boldsymbol{\varepsilon}$ is orthogonal to $\mathbf{w}$. This makes sure that the regression gives the right results.

Problem 360. *4 points The asymptotic $\mathcal{MSE}$ matrix of the instrumental variables estimator with $\mathbf{W}$ as matrix of instruments is* $\sigma^2 \operatorname{plim}\Big(\mathbf{X}^\top\mathbf{W}(\mathbf{W}^\top\mathbf{W})^{-1}\mathbf{W}^\top\mathbf{X}\Big)^{-1}$ *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*

$$\Big(\mathbf{X}^\top\mathbf{U}(\mathbf{U}^\top\mathbf{U})^{-1}\mathbf{U}^\top\mathbf{X}\Big)^{-1} - \Big(\mathbf{X}^\top\mathbf{W}(\mathbf{W}^\top\mathbf{W})^{-1}\mathbf{W}^\top\mathbf{X}\Big)^{-1} \tag{32.0.7}$$

is nonnegative definite. Hints: (1) Use theorem A.5.5 in the Appendix (proof is not required). (2) Note that $\mathbf{U} = \mathbf{W}\boldsymbol{G}$ for some $\boldsymbol{G}$. Can you write this $\boldsymbol{G}$ in partitioned matrix form? (3) Show that, whatever $\mathbf{W}$ and $\boldsymbol{G}$, $\mathbf{W}(\mathbf{W}^\top\mathbf{W})^{-1}\mathbf{W}^\top - \mathbf{W}\boldsymbol{G}(\boldsymbol{G}^\top\mathbf{W}^\top\mathbf{W}\boldsymbol{G})^{-1}\boldsymbol{G}^\top\mathbf{W}^\top$ is idempotent.

Answer.

$$\mathbf{U} = \begin{bmatrix}\mathbf{U} & \mathbf{V}\end{bmatrix}\begin{bmatrix}\boldsymbol{I}\\ \boldsymbol{O}\end{bmatrix} = \mathbf{W}\boldsymbol{G} \quad \text{where} \quad \boldsymbol{G} = \begin{bmatrix}\boldsymbol{I}\\ \boldsymbol{O}\end{bmatrix}. \tag{32.0.8}$$

□

Problem 361. *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: $\boldsymbol{L} = \boldsymbol{L}(\boldsymbol{A}\boldsymbol{R}) = (\boldsymbol{L}\boldsymbol{A})\boldsymbol{R} = \boldsymbol{R}$. □

Problem 362. *2 points If $\mathbf{W}^\top\mathbf{X}$ is square and has full column rank, then it is nonsingular. Show that in this case* (32.0.4) *simplifies to the "simple" instrumental variables estimator:*

$$\tilde{\boldsymbol{\beta}} = (\mathbf{W}^\top\mathbf{X})^{-1}\mathbf{W}^\top\mathbf{y} \tag{32.0.9}$$

Answer. In this case the big inverse can be split into three:

$$\tilde{\beta} = \Big(\mathbf{X}^\top\mathbf{W}(\mathbf{W}^\top\mathbf{W})^{-1}\mathbf{W}^\top\mathbf{X}\Big)^{-1}\mathbf{X}^\top\mathbf{W}(\mathbf{W}^\top\mathbf{W})^{-1}\mathbf{W}^\top\mathbf{y} = \tag{32.0.10}$$

$$= (\mathbf{W}^\top\mathbf{X})^{-1}\mathbf{W}^\top\mathbf{W}(\mathbf{X}^\top\mathbf{W})^{-1}\mathbf{X}^\top\mathbf{W}(\mathbf{W}^\top\mathbf{W})^{-1}\mathbf{W}^\top\mathbf{y} \tag{32.0.11}$$

□

Problem 363. *We only have one regressor with intercept, i.e., $\mathbf{X} = \begin{bmatrix}\boldsymbol{\iota} & \mathbf{x}\end{bmatrix}$, and we have one instrument $\mathbf{w}$ for $\mathbf{x}$ (while the constant term is its own instrument), i.e., $\mathbf{W} = \begin{bmatrix}\boldsymbol{\iota} & \mathbf{w}\end{bmatrix}$. Show that the instrumental variables estimators for slope and intercept are*

$$\tilde{\beta} = \frac{\sum(w_t - \bar{w})(y_t - \bar{y})}{\sum(w_t - \bar{w})(x_t - \bar{x})} \tag{32.0.12}$$

$$\tilde{\alpha} = \bar{y} - \tilde{\beta}\bar{x} \tag{32.0.13}$$

Hint: the math is identical to that in question 200.

Problem 364. *2 points Show that, if there are as many instruments as there are observations, then the instrumental variables estimator* (32.0.4) *becomes identical to OLS.*

Answer. In this case $\mathbf{W}$ has an inverse, therefore the projection on $\mathcal{R}[\mathbf{W}]$ is the identity. Staying in the algebraic paradigm, $(\mathbf{W}^\top\mathbf{W})^{-1} = \mathbf{W}^{-1}(\mathbf{W}^\top)^{-1}$. □

An implication of Problem 364 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.

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 365. *Prove theorem A.1.2.*

PROBLEM 366. *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 367. *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 368. *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 369. *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 $\boldsymbol{R}\boldsymbol{L}$ 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 $\boldsymbol{B} = \boldsymbol{A}\boldsymbol{A}^{-}\boldsymbol{B}$ holds for one g-inverse $\boldsymbol{A}^{-}$ of $\boldsymbol{A}$, then it holds for all g-inverses. If $\boldsymbol{A}$ is symmetric and $\boldsymbol{B} = \boldsymbol{A}\boldsymbol{A}^{-}\boldsymbol{B}$, then also $\boldsymbol{B}^{\top} = \boldsymbol{B}^{\top}\boldsymbol{A}^{-}\boldsymbol{A}$. If $\boldsymbol{B} = \boldsymbol{B}\boldsymbol{A}^{-}\boldsymbol{A}$ and $\boldsymbol{C} = \boldsymbol{A}\boldsymbol{A}^{-}\boldsymbol{C}$ then $\boldsymbol{B}\boldsymbol{A}^{-}\boldsymbol{C}$ is independent of the choice of g-inverses.*

PROOF. Assume the identity $\boldsymbol{B} = \boldsymbol{A}\boldsymbol{A}^{+}\boldsymbol{B}$ 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 $\boldsymbol{A}\boldsymbol{A}^{-}\boldsymbol{B} = \boldsymbol{A}\boldsymbol{A}^{-}\boldsymbol{A}\boldsymbol{A}^{+}\boldsymbol{B} = \boldsymbol{A}\boldsymbol{A}^{+}\boldsymbol{B} = \boldsymbol{B}$. 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: $\boldsymbol{B}\boldsymbol{A}^{+}\boldsymbol{C} = \boldsymbol{B}\boldsymbol{A}^{-}\boldsymbol{A}\boldsymbol{A}^{+}\boldsymbol{A}\boldsymbol{A}^{-}\boldsymbol{C} = \boldsymbol{B}\boldsymbol{A}^{-}\boldsymbol{A}\boldsymbol{A}^{-}\boldsymbol{C} = \boldsymbol{B}\boldsymbol{A}^{-}\boldsymbol{C}$. Here $^{+}$ signifies a different g-inverse; again, it is not necessarily the Moore-Penrose one. □

PROBLEM 370. *Show that $\boldsymbol{x}$ satisfies $\boldsymbol{x} = \boldsymbol{B}\boldsymbol{a}$ for some $\boldsymbol{a}$ if and only if $\boldsymbol{x} = \boldsymbol{B}\boldsymbol{B}^{-}\boldsymbol{x}$.*

THEOREM A.3.2. *Both $\boldsymbol{A}^{\top}(\boldsymbol{A}\boldsymbol{A}^{\top})^{-}$ and $(\boldsymbol{A}^{\top}\boldsymbol{A})^{-}\boldsymbol{A}$ are g-inverses of $\boldsymbol{A}$.*

PROOF. We have to show

$$\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^{\top}(\boldsymbol{A}\boldsymbol{A}^{\top})^{-}\boldsymbol{A} \tag{A.3.3}$$

which is [**Rao73**, (1b.5.5) on p. 26]. Define $\boldsymbol{D} = \boldsymbol{A} - \boldsymbol{A}\boldsymbol{A}^{\top}(\boldsymbol{A}\boldsymbol{A}^{\top})^{-}\boldsymbol{A}$ and show, by multiplying out, that $\boldsymbol{D}\boldsymbol{D}^{\top} = \boldsymbol{O}$. □

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 $\boldsymbol{G}\boldsymbol{S} = \boldsymbol{O}$, and for all $\boldsymbol{Q}$ with $\boldsymbol{Q}\boldsymbol{S} = \boldsymbol{O}$ there is an $\boldsymbol{X}$ with $\boldsymbol{Q} = \boldsymbol{X}\boldsymbol{G}$. 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: $\boldsymbol{G}\boldsymbol{S} = \boldsymbol{O}$, and for all $\boldsymbol{R}$ with $\boldsymbol{G}\boldsymbol{R} = \boldsymbol{O}$ there is a $\boldsymbol{Y}$ with $\boldsymbol{R} = \boldsymbol{S}\boldsymbol{Y}$. 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 $\boldsymbol{T}\boldsymbol{U} = \boldsymbol{O}$ and $\boldsymbol{T}^{\top}\boldsymbol{T} + \boldsymbol{U}\boldsymbol{U}^{\top}$ 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 $\boldsymbol{\gamma}^{\top}\boldsymbol{T}^{\top}\boldsymbol{T}\boldsymbol{\gamma} + \boldsymbol{\gamma}^{\top}\boldsymbol{U}\boldsymbol{U}^{\top}\boldsymbol{\gamma} = 0$, 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 $\boldsymbol{T}\boldsymbol{U} = \boldsymbol{O}$ and $\boldsymbol{T}^{\top}\boldsymbol{T} + \boldsymbol{U}\boldsymbol{U}^{\top}$ is nonsingular. To show that $\boldsymbol{T} \perp \boldsymbol{U}$ take any $\boldsymbol{B}$ with $\boldsymbol{B}\boldsymbol{U} = \boldsymbol{O}$. Then $\boldsymbol{B} = \boldsymbol{B}(\boldsymbol{T}^{\top}\boldsymbol{T} + \boldsymbol{U}\boldsymbol{U}^{\top})(\boldsymbol{T}^{\top}\boldsymbol{T} + \boldsymbol{U}\boldsymbol{U}^{\top})^{-1} = \boldsymbol{B}\boldsymbol{T}^{\top}\boldsymbol{T}(\boldsymbol{T}^{\top}\boldsymbol{T} + \boldsymbol{U}\boldsymbol{U}^{\top})^{-1}$. In the same way one gets $\boldsymbol{T} = \boldsymbol{T}\boldsymbol{T}^{\top}\boldsymbol{T}(\boldsymbol{T}^{\top}\boldsymbol{T} + \boldsymbol{U}\boldsymbol{U}^{\top})^{-1}$. Premultiply this last equation by $\boldsymbol{T}^{\top}\boldsymbol{T}(\boldsymbol{T}^{\top}\boldsymbol{T}\boldsymbol{T}^{\top}\boldsymbol{T})^{-}\boldsymbol{T}^{\top}$ and use theorem A.3.2 to get $\boldsymbol{T}^{\top}\boldsymbol{T}(\boldsymbol{T}^{\top}\boldsymbol{T}\boldsymbol{T}^{\top}\boldsymbol{T})^{-}\boldsymbol{T}^{\top}\boldsymbol{T} = \boldsymbol{T}^{\top}\boldsymbol{T}(\boldsymbol{T}^{\top}\boldsymbol{T} + \boldsymbol{U}\boldsymbol{U}^{\top})^{-1}$. Inserting this into the equation for $\boldsymbol{B}$ gives $\boldsymbol{B} = \boldsymbol{B}\boldsymbol{T}^{\top}\boldsymbol{T}(\boldsymbol{T}^{\top}\boldsymbol{T}\boldsymbol{T}^{\top}\boldsymbol{T})^{-}\boldsymbol{T}^{\top}\boldsymbol{T}$, 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 $\boldsymbol{Y}$, $\boldsymbol{B}\boldsymbol{Y} = \boldsymbol{O}$ implies $\boldsymbol{A}\boldsymbol{Y} = \boldsymbol{O}$, then a $\boldsymbol{X}$ exists with $\boldsymbol{A} = \boldsymbol{X}\boldsymbol{B}$.*

PROBLEM 371. *Prove theorem A.4.2.*

ANSWER. Let $\boldsymbol{B} \perp \boldsymbol{C}$. Choosing $\boldsymbol{Y} = \boldsymbol{B}$ follows $\boldsymbol{A}\boldsymbol{B} = \boldsymbol{O}$, hence $\boldsymbol{X}$ exists. □

PROBLEM 372. *Show that $\boldsymbol{I} - \boldsymbol{S}\boldsymbol{S}^{-} \perp \boldsymbol{S}$.*

ANSWER. Clearly, $(\boldsymbol{I} - \boldsymbol{S}\boldsymbol{S}^{-})\boldsymbol{S} = \boldsymbol{O}$. Now if $\boldsymbol{Q}\boldsymbol{S} = \boldsymbol{O}$, then $\boldsymbol{Q} = \boldsymbol{Q}(\boldsymbol{I} - \boldsymbol{S}\boldsymbol{S}^{-})$, i.e., the $\boldsymbol{X}$ whose existence is postulated in the definition of a deficiency matrix is $\boldsymbol{Q}$ itself. □

PROBLEM 373. *Show that $\boldsymbol{S} \perp \boldsymbol{U}$ if and only if $\boldsymbol{S}$ is a matrix with maximal rank which satisfies $\boldsymbol{S}\boldsymbol{U} = \boldsymbol{O}$. In other words, one cannot add linearly independent rows to $\boldsymbol{S}$ in such a way that the new matrix still satisfies $\boldsymbol{T}\boldsymbol{U} = \boldsymbol{O}$.*

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., $\boldsymbol{S}\boldsymbol{Q} = \boldsymbol{S}$, 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 $\boldsymbol{P}\boldsymbol{U} = \boldsymbol{O}$. 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 $\boldsymbol{P} = \boldsymbol{S}\boldsymbol{S}$ 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 374. *Given $\boldsymbol{X}$ and a symmetric nonnegative definite $\boldsymbol{\Omega}$ such that $\boldsymbol{X} = \boldsymbol{\Omega}\boldsymbol{W}$ 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 $\boldsymbol{X}\boldsymbol{Y} = \boldsymbol{O}$ is equivalent to $\boldsymbol{X}^{\top}\boldsymbol{\Omega}^{-}\boldsymbol{X}\boldsymbol{Y} = \boldsymbol{O}$. ⇒ clear; for ⇐ note that $\boldsymbol{X}^{\top}\boldsymbol{\Omega}^{-}\boldsymbol{X} = \boldsymbol{W}^{\top}\boldsymbol{\Omega}\boldsymbol{W}$, therefore $\boldsymbol{X}\boldsymbol{Y} = \boldsymbol{\Omega}\boldsymbol{W}\boldsymbol{Y} = \boldsymbol{\Omega}\boldsymbol{W}(\boldsymbol{W}^{\top}\boldsymbol{\Omega}\boldsymbol{W})^{-}\boldsymbol{W}^{\top}\boldsymbol{\Omega}\boldsymbol{W}\boldsymbol{Y} = \boldsymbol{\Omega}\boldsymbol{W}(\boldsymbol{W}^{\top}\boldsymbol{\Omega}\boldsymbol{W})^{-}\boldsymbol{X}^{\top}\boldsymbol{\Omega}^{-}\boldsymbol{X}\boldsymbol{Y} = \boldsymbol{O}$.

□

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 375. *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

$$\text{For all } \boldsymbol{C} \text{ which satisfy } \boldsymbol{CU} = \boldsymbol{O} \text{ follows } \boldsymbol{CD} = \boldsymbol{O}. \tag{A.4.7}$$

A.5. Nonnegative Definite Symmetric Matrices

By definition, a symmetric matrix $\boldsymbol{\Omega}$ is nonnegative definite if $\boldsymbol{a}^\top\boldsymbol{\Omega}\boldsymbol{a} \geq 0$ for all vectors $\boldsymbol{a}$. It is positive definite if $\boldsymbol{a}^\top\boldsymbol{\Omega}\boldsymbol{a} > 0$ 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 $\boldsymbol{a}^\top\boldsymbol{\Omega}\boldsymbol{a} = 0$, then already $\boldsymbol{\Omega}\boldsymbol{a} = \boldsymbol{o}$.*

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 $\operatorname{tr}(\boldsymbol{\Omega}\boldsymbol{\Sigma}) \geq 0$.*

PROBLEM 376. *Prove theorem A.5.6.*

ANSWER. Find any factorization $\boldsymbol{\Sigma} = \boldsymbol{P}\boldsymbol{P}^\top$. Then $\operatorname{tr}(\boldsymbol{\Omega}\boldsymbol{\Sigma}) = \operatorname{tr}(\boldsymbol{P}^\top\boldsymbol{\Omega}\boldsymbol{P}) \geq 0$. □

THEOREM A.5.7. *If $\boldsymbol{\Omega}$ is nonnegative definite symmetric, then*

$$(\boldsymbol{g}^\top\boldsymbol{\Omega}\boldsymbol{a})^2 \leq \boldsymbol{g}^\top\boldsymbol{\Omega}\boldsymbol{g}\,\boldsymbol{a}^\top\boldsymbol{\Omega}\boldsymbol{a}, \tag{A.5.1}$$

for arbitrary vectors $\boldsymbol{a}$ and $\boldsymbol{g}$. Equality holds if and only if $\boldsymbol{\Omega}\boldsymbol{g}$ and $\boldsymbol{\Omega}\boldsymbol{a}$ are linearly dependent, i.e., α and β exist, not both zero, such that $\boldsymbol{\Omega}\boldsymbol{g}\alpha + \boldsymbol{\Omega}\boldsymbol{a}\beta = \boldsymbol{o}$.

Proof: First we will show that the condition for equality is sufficient. Therefore assume $\boldsymbol{\Omega}\boldsymbol{g}\alpha + \boldsymbol{\Omega}\boldsymbol{a}\beta = 0$ 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 $\boldsymbol{a}^\top\boldsymbol{\Omega}\boldsymbol{g}\alpha + \boldsymbol{a}^\top\boldsymbol{\Omega}\boldsymbol{a}\beta = 0$ to get $\boldsymbol{a}^\top\boldsymbol{\Omega}\boldsymbol{g} = -(\beta/\alpha)\boldsymbol{a}^\top\boldsymbol{\Omega}\boldsymbol{a}$, therefore the lefthand side of (A.5.1) is $(\beta/\alpha)^2(\boldsymbol{a}^\top\boldsymbol{\Omega}\boldsymbol{a})^2$. Furthermore we can solve $\boldsymbol{g}^\top\boldsymbol{\Omega}\boldsymbol{g}\alpha + \boldsymbol{g}^\top\boldsymbol{\Omega}\boldsymbol{a}\beta = 0$ to get $\boldsymbol{g}^\top\boldsymbol{\Omega}\boldsymbol{g} = -(\beta/\alpha)\boldsymbol{g}^\top\boldsymbol{\Omega}\boldsymbol{a} = (\beta/\alpha)^2\boldsymbol{a}^\top\boldsymbol{\Omega}\boldsymbol{a}$, therefore the righthand side of (A.5.1) is $(\beta/\alpha)^2(\boldsymbol{a}^\top\boldsymbol{\Omega}\boldsymbol{a})^2$ 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, $\boldsymbol{\Omega}\boldsymbol{g}$ and $\boldsymbol{\Omega}\boldsymbol{a}$ are linearly dependent. We will split this second half of the proof into two substeps. First verify that (A.5.1) holds if $\boldsymbol{g}^\top\boldsymbol{\Omega}\boldsymbol{g} = 0$. If this is the case, then already $\boldsymbol{\Omega}\boldsymbol{g} = \boldsymbol{o}$, therefore the $\boldsymbol{\Omega}\boldsymbol{g}$ and $\boldsymbol{\Omega}\boldsymbol{a}$ 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 $\boldsymbol{g}^\top\boldsymbol{\Omega}\boldsymbol{g} \neq 0$. Since $\boldsymbol{\Omega}$ is nonnegative definite, it follows

$$0 \leq \left(\boldsymbol{a}-\boldsymbol{g}\frac{\boldsymbol{g}^\top\boldsymbol{\Omega}\boldsymbol{a}}{\boldsymbol{g}^\top\boldsymbol{\Omega}\boldsymbol{g}}\right)^\top\boldsymbol{\Omega}\left(\boldsymbol{a}-\boldsymbol{g}\frac{\boldsymbol{g}^\top\boldsymbol{\Omega}\boldsymbol{a}}{\boldsymbol{g}^\top\boldsymbol{\Omega}\boldsymbol{g}}\right) = \boldsymbol{a}^\top\boldsymbol{\Omega}\boldsymbol{a} - 2\frac{(\boldsymbol{g}^\top\boldsymbol{\Omega}\boldsymbol{a})^2}{\boldsymbol{g}^\top\boldsymbol{\Omega}\boldsymbol{g}} + \frac{(\boldsymbol{g}^\top\boldsymbol{\Omega}\boldsymbol{a})^2}{\boldsymbol{g}^\top\boldsymbol{\Omega}\boldsymbol{g}} = \boldsymbol{a}^\top\boldsymbol{\Omega}\boldsymbol{a} - \frac{(\boldsymbol{g}^\top\boldsymbol{\Omega}\boldsymbol{a})}{\boldsymbol{g}^\top\boldsymbol{\Omega}\boldsymbol{g}} \tag{A.5.2}$$

From this follows (A.5.1). If (A.5.2) is an equality, then already $\boldsymbol{\Omega}\left(\boldsymbol{a}-\boldsymbol{g}\frac{\boldsymbol{g}^\top\boldsymbol{\Omega}\boldsymbol{a}}{\boldsymbol{g}^\top\boldsymbol{\Omega}\boldsymbol{g}}\right) = \boldsymbol{o}$, which means that $\boldsymbol{\Omega}\boldsymbol{g}$ and $\boldsymbol{\Omega}\boldsymbol{a}$ 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*

$$\frac{(\boldsymbol{g}^\top\boldsymbol{\Omega}\boldsymbol{a})^2}{\boldsymbol{g}^\top\boldsymbol{\Omega}\boldsymbol{g}} \leq \boldsymbol{a}^\top\boldsymbol{\Omega}\boldsymbol{a}. \tag{A.5.3}$$

Equality holds if and only if a $\gamma \neq 0$ exists with $\boldsymbol{\Omega}\boldsymbol{g} = \boldsymbol{\Omega}\boldsymbol{a}\gamma$.

PROBLEM 377. *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 378. *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 379. *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 380. *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 $\boldsymbol{g} = \boldsymbol{\Omega a}$ for some $\boldsymbol{a}$, maximize over $\boldsymbol{a}$ instead of $\boldsymbol{g}$. This reduces it to theorem A.5.8:

$$\max_{\boldsymbol{g}:\ \boldsymbol{g}=\boldsymbol{\Omega a} \text{ for some } \boldsymbol{a}} \frac{(\boldsymbol{g}^\top \boldsymbol{x})^2}{\boldsymbol{g}^\top \boldsymbol{\Omega}^- \boldsymbol{g}} = \max_{\boldsymbol{a}} \frac{(\boldsymbol{a}^\top \boldsymbol{\Omega x})^2}{\boldsymbol{a}^\top \boldsymbol{\Omega a}} = \boldsymbol{x}^\top \boldsymbol{\Omega x} \tag{A.5.10}$$

□

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 $\boldsymbol{x} = \boldsymbol{\Omega a}$ 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 381. *Prove theorem A.5.9.*

Answer. Assume $\boldsymbol{x} = \boldsymbol{\Omega a}$ and $\boldsymbol{x}^\top \boldsymbol{\Omega}^- \boldsymbol{x} = \boldsymbol{a}^\top \boldsymbol{\Omega a} \leq 1$; then for any $\boldsymbol{g}$, $\boldsymbol{g}^\top(\boldsymbol{\Omega} - \boldsymbol{x}\boldsymbol{x}^\top)\boldsymbol{g}^\top = \boldsymbol{g}^\top \boldsymbol{\Omega g} - \boldsymbol{g}^\top \boldsymbol{\Omega a}\boldsymbol{a}^\top \boldsymbol{\Omega g} \geq \boldsymbol{a}^\top \boldsymbol{\Omega a}\boldsymbol{g}^\top \boldsymbol{\Omega g} - \boldsymbol{g}^\top \boldsymbol{\Omega a}\boldsymbol{a}^\top \boldsymbol{\Omega g} \geq 0$ by theorem A.5.7.

Conversely, assume $\boldsymbol{x}$ cannot be written in the form $\boldsymbol{x} = \boldsymbol{\Omega a}$ 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 $\boldsymbol{x}^\top \boldsymbol{\Omega}^- \boldsymbol{x} = \boldsymbol{a}^\top \boldsymbol{\Omega a} > 1$; then $\boldsymbol{a}^\top(\boldsymbol{\Omega} - \boldsymbol{x}\boldsymbol{x}^\top)\boldsymbol{a} = \boldsymbol{a}^\top \boldsymbol{\Omega a} - (\boldsymbol{a}^\top \boldsymbol{\Omega a})^2 < 0$, 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 $\boldsymbol{\Sigma K \Omega}$ is defined, then*

$$\boldsymbol{K\Omega} = (\boldsymbol{K\Omega K}^\top + \boldsymbol{\Sigma})(\boldsymbol{K\Omega K}^\top + \boldsymbol{\Sigma})^- \boldsymbol{K\Omega}. \tag{A.5.11}$$

Furthermore, $\boldsymbol{\Omega K}^\top(\boldsymbol{K\Omega K}^\top + \boldsymbol{\Sigma})^- \boldsymbol{K\Omega}$ is independent of the choice of g-inverses.

Problem 382. *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 $\boldsymbol{\Omega} = \boldsymbol{QQ}^\top$ and $\boldsymbol{P}$ with $\boldsymbol{\Sigma} = \boldsymbol{PP}^\top$ and define $\boldsymbol{A} = \begin{bmatrix} \boldsymbol{KQ} & \boldsymbol{P} \end{bmatrix}$. 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 $\boldsymbol{\Omega} = \begin{bmatrix} \boldsymbol{\Omega}_{\mathtt{yy}} & \boldsymbol{\Omega}_{\mathtt{yz}} \\ \boldsymbol{\Omega}_{\mathtt{yz}}^\top & \boldsymbol{\Omega}_{\mathtt{zz}} \end{bmatrix}$ is nonnegative definite if and only if the following conditions hold:*

(A.5.12)

$$\boldsymbol{\Omega}_{\mathtt{yy}} \quad \textit{and} \quad \boldsymbol{\Omega}_{\mathtt{zz.y}} := \boldsymbol{\Omega}_{\mathtt{zz}} - \boldsymbol{\Omega}_{\mathtt{yz}}^\top \boldsymbol{\Omega}_{\mathtt{yy}}^- \boldsymbol{\Omega}_{\mathtt{yz}} \quad \textit{are both nonnegative definite, and}$$

$$\boldsymbol{\Omega}_{\mathtt{yz}} = \boldsymbol{\Omega}_{\mathtt{yy}} \boldsymbol{\Omega}_{\mathtt{yy}}^- \boldsymbol{\Omega}_{\mathtt{yz}} \tag{A.5.13}$$

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, $\boldsymbol{\Omega}_{\mathtt{zz.y}}$ 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 $\begin{bmatrix} \boldsymbol{\Omega}_{\mathtt{yy}} & \boldsymbol{\Omega}_{\mathtt{yz}} \\ \boldsymbol{\Omega}_{\mathtt{yz}}^\top & \boldsymbol{\Omega}_{\mathtt{zz}} \end{bmatrix} =$

$\begin{bmatrix} \boldsymbol{R}_{\mathtt{y}}^\top \boldsymbol{R}_{\mathtt{y}} & \boldsymbol{R}_{\mathtt{y}}^\top \boldsymbol{R}_{\mathtt{z}} \\ \boldsymbol{R}_{\mathtt{z}}^\top \boldsymbol{R}_{\mathtt{y}} & \boldsymbol{R}_{\mathtt{z}}^\top \boldsymbol{R}_{\mathtt{z}} \end{bmatrix}$. $\boldsymbol{\Omega}_{\mathtt{yy}}$ 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): $\boldsymbol{\Omega}_{\mathtt{yy}} \boldsymbol{\Omega}_{\mathtt{yy}}^- \boldsymbol{\Omega}_{\mathtt{yz}} = \boldsymbol{R}_{\mathtt{y}}^\top \boldsymbol{R}_{\mathtt{y}} (\boldsymbol{R}_{\mathtt{y}}^\top \boldsymbol{R}_{\mathtt{y}})^- \boldsymbol{R}_{\mathtt{y}}^\top \boldsymbol{R}_{\mathtt{z}} = \boldsymbol{R}_{\mathtt{y}}^\top \boldsymbol{R}_{\mathtt{z}} = \boldsymbol{\Omega}_{\mathtt{yz}}$. To show that $\boldsymbol{\Omega}_{\mathtt{zz.y}}$ is nonnegative definite, define $\boldsymbol{S} = (\boldsymbol{I} - \boldsymbol{R}_{\mathtt{y}}(\boldsymbol{R}_{\mathtt{y}}^\top \boldsymbol{R}_{\mathtt{y}})^- \boldsymbol{R}_{\mathtt{y}}^\top)\boldsymbol{R}_{\mathtt{z}}$. Then $\boldsymbol{S}^\top \boldsymbol{S} = \boldsymbol{R}_{\mathtt{z}}^\top \big(\boldsymbol{I} - \boldsymbol{R}_{\mathtt{y}}(\boldsymbol{R}_{\mathtt{y}}^\top \boldsymbol{R}_{\mathtt{y}})^- \boldsymbol{R}_{\mathtt{y}}^\top\big)\boldsymbol{R}_{\mathtt{z}} = \boldsymbol{\Omega}_{\mathtt{zz.y}}$.

To show sufficiency of the three conditions of theorem A.5.11, assume the symmetric $\begin{bmatrix} \boldsymbol{\Omega}_{\mathtt{yy}} & \boldsymbol{\Omega}_{\mathtt{yz}} \\ \boldsymbol{\Omega}_{\mathtt{yz}}^\top & \boldsymbol{\Omega}_{\mathtt{zz}} \end{bmatrix}$ satisfies them. Pick two matrices $\boldsymbol{Q}$ and $\boldsymbol{S}$ so that $\boldsymbol{\Omega}_{\mathtt{yy}} = \boldsymbol{Q}^\top \boldsymbol{Q}$ and $\boldsymbol{\Omega}_{\mathtt{zz.y}} = \boldsymbol{S}^\top \boldsymbol{S}$. Then

$$\begin{bmatrix} \boldsymbol{\Omega}_{\mathtt{yy}} & \boldsymbol{\Omega}_{\mathtt{yz}} \\ \boldsymbol{\Omega}_{\mathtt{yz}}^\top & \boldsymbol{\Omega}_{\mathtt{zz}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{Q}^\top & \boldsymbol{O} \\ \boldsymbol{\Omega}_{\mathtt{yz}}^\top \boldsymbol{\Omega}_{\mathtt{yy}}^{-\ \top} \boldsymbol{Q}^\top & \boldsymbol{S}^\top \end{bmatrix} \begin{bmatrix} \boldsymbol{Q} & \boldsymbol{Q}\boldsymbol{\Omega}_{\mathtt{yy}}^- \boldsymbol{\Omega}_{\mathtt{yz}} \\ \boldsymbol{O} & \boldsymbol{S} \end{bmatrix},$$

therefore nonnegative definite.

Problem 383. [**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 384. *Given the vector $\boldsymbol{h} \neq \boldsymbol{o}$. For which values of the scalar γ is the matrix $\boldsymbol{I} - \frac{\boldsymbol{hh}^\top}{\gamma}$ 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 $(\boldsymbol{I} - \frac{\boldsymbol{hh}^\top}{\gamma})\boldsymbol{x} = \boldsymbol{o}$. It follows $\boldsymbol{x} = \boldsymbol{h}\alpha$ where $\alpha = \boldsymbol{h}^\top \boldsymbol{x}/\gamma$, therefore $(\boldsymbol{I} - \frac{\boldsymbol{hh}^\top}{\gamma})\boldsymbol{x} = \boldsymbol{h}\alpha(1 - \boldsymbol{h}^\top \boldsymbol{h}/\gamma)$. 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 385. *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 386. *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 $\boldsymbol{Q}\boldsymbol{c} = \boldsymbol{o}$ implies $\boldsymbol{P}\boldsymbol{c} = \boldsymbol{o}$. Proof: from $\boldsymbol{Q}-\boldsymbol{P}$ nnd and $\boldsymbol{Q}\boldsymbol{c} = \boldsymbol{o}$ follows $0 \leq \boldsymbol{c}^\top(\boldsymbol{Q}-\boldsymbol{P})\boldsymbol{c} = -\boldsymbol{c}^\top\boldsymbol{P}\boldsymbol{c} \leq 0$, therefore equality throughout, i.e., $0 = \boldsymbol{c}^\top\boldsymbol{P}\boldsymbol{c} = \boldsymbol{c}^\top\boldsymbol{P}^\top\boldsymbol{P}\boldsymbol{c} = \|\boldsymbol{P}\boldsymbol{c}\|^2$ and therefore $\boldsymbol{P}\boldsymbol{c} = \boldsymbol{o}$. Secondly: this is also true for matrices:

$\boldsymbol{Q}\boldsymbol{C} = \boldsymbol{O}$ implies $\boldsymbol{P}\boldsymbol{C} = \boldsymbol{O}$, 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 387. *If $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{A}$ 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. $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{A}$ means that every column of $\boldsymbol{Y}$ is a linear combination of columns of $\boldsymbol{A}$:

$$\begin{bmatrix}\boldsymbol{y}_1 & \cdots & \boldsymbol{y}_m\end{bmatrix} = \boldsymbol{X}\begin{bmatrix}\boldsymbol{a}_1 & \cdots & \boldsymbol{a}_m\end{bmatrix} = \begin{bmatrix}\boldsymbol{X}\boldsymbol{a}_1 & \cdots & \boldsymbol{X}\boldsymbol{a}_m\end{bmatrix}. \tag{A.6.6}$$

Therefore geometrically the statement follows from the fact shown in Problem 385 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 388. *(Not eligible for in-class exams) Let $\boldsymbol{Q}$ be a projection matrix (i.e., a symmetric and idempotent matrix) with the property that $\boldsymbol{Q} = \boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^\top$ 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

$$\boldsymbol{X}^\top(\boldsymbol{I}-\boldsymbol{Q})\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^-\boldsymbol{X}^\top(\boldsymbol{I}-\boldsymbol{Q})\boldsymbol{X} = \boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^-\boldsymbol{X}^\top\boldsymbol{X} - \boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^-\boldsymbol{X}^\top\boldsymbol{Q}\boldsymbol{X} - \boldsymbol{X}^\top\boldsymbol{Q}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^- \tag{A.6.8}$$

$$= \boldsymbol{X}^\top\boldsymbol{X} - \boldsymbol{X}^\top\boldsymbol{Q}\boldsymbol{X} - \boldsymbol{X}^\top\boldsymbol{Q}\boldsymbol{X} + \boldsymbol{X}^\top\boldsymbol{Q}\boldsymbol{X} = \boldsymbol{X}^\top(\boldsymbol{I}-\boldsymbol{Q})\boldsymbol{X}. \tag{A.6.9}$$

Only for the fourth term did we need the condition $\boldsymbol{Q} = \boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^\top$:

$$\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^-\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^\top\boldsymbol{X} = \boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^\top\boldsymbol{X} = \boldsymbol{X}^\top\boldsymbol{Q}\boldsymbol{Q}\boldsymbol{X} = \boldsymbol{X}^\top\boldsymbol{X}. \tag{A.6.10}$$

Using this g-inverse we have

$$\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^-\boldsymbol{X}^\top - \tilde{\boldsymbol{X}}(\boldsymbol{X}^\top\boldsymbol{X})^-\tilde{\boldsymbol{X}}^\top = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^-\boldsymbol{X}^\top - (\boldsymbol{I}-\boldsymbol{Q})\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^-\boldsymbol{X}^\top(\boldsymbol{I}-\boldsymbol{Q}) = \tag{A.6.11}$$

$$= \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^-\boldsymbol{X}^\top - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^-\boldsymbol{X}^\top + \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^-\boldsymbol{X}^\top\boldsymbol{Q} + \boldsymbol{Q}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^-\boldsymbol{X}^\top - \boldsymbol{Q}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^-\boldsymbol{X}^\top\boldsymbol{Q} = \boldsymbol{X}(\boldsymbol{X}^- \tag{A.6.12}$$

□

PROBLEM 389. *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 $\boldsymbol{P} = \boldsymbol{P}\boldsymbol{P}$ 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 $\pi\colon\ \{1,\dots,n\} \to \{1,\dots,n\}$ 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

$$\det(\boldsymbol{A}) = \sum_{\pi:\ \pi \text{ even}} a_{1\pi(1)}\cdots a_{n\pi(n)} - \sum_{\pi:\ \pi \text{ odd}} a_{1\pi(1)}\cdots a_{n\pi(n)} \tag{A.7.2}$$

Now assume $\boldsymbol{A}$ is $m \times m$, $1 \le m < n$. If a $j \le m$ exists with $\pi(j) > m$ then not all $i \le m$ can be images of other points $j \le m$, i.e., there must be at least one $j > m$ with $\pi(j) \le m$. Therefore, in a block triangular matrix in which all $a_{ij} = 0$ for $i \le m$, $j > m$, only those permutations give a nonzero product which remain in the two submatrices straddling the diagonal.

Theorem A.7.2. *If $\boldsymbol{B} = \boldsymbol{A}\boldsymbol{A}^-\boldsymbol{B}$, then the following identity is valid between determinants:*

$$\begin{vmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{vmatrix} = |\boldsymbol{A}|\ |\boldsymbol{E}| \quad \textit{where} \quad \boldsymbol{E} = \boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^-\boldsymbol{B}. \tag{A.7.3}$$

Proof: Postmultiply by a matrix whose determinant, by lemma A.7.1, is one, and then apply lemma A.7.1 once more:

(A.7.4)
$$\begin{vmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{vmatrix} = \begin{vmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{vmatrix} \begin{vmatrix} \boldsymbol{I} & -\boldsymbol{A}^-\boldsymbol{B} \\ \boldsymbol{O} & \boldsymbol{I} \end{vmatrix} = \begin{vmatrix} \boldsymbol{A} & \boldsymbol{O} \\ \boldsymbol{C} & \boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^-\boldsymbol{B} \end{vmatrix} = |\boldsymbol{A}|\ \left|\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^-\boldsymbol{B}\right| .$$

Problem 390. *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 391. *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 392. *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* $\begin{bmatrix} \boldsymbol{A} & \boldsymbol{AF} \\ \boldsymbol{GA} & \boldsymbol{E}+\boldsymbol{GAF} \end{bmatrix}$, *show that* $\boldsymbol{X}$ *is a g-inverse of* $\boldsymbol{E}$.

Answer. The g-inverse condition means

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{AF} \\ \boldsymbol{GA} & \boldsymbol{E}+\boldsymbol{GAF} \end{bmatrix} \begin{bmatrix} \boldsymbol{U} & \boldsymbol{V} \\ \boldsymbol{W} & \boldsymbol{X} \end{bmatrix} \begin{bmatrix} \boldsymbol{A} & \boldsymbol{AF} \\ \boldsymbol{GA} & \boldsymbol{E}+\boldsymbol{GAF} \end{bmatrix} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{AF} \\ \boldsymbol{GA} & \boldsymbol{E}+\boldsymbol{GAF} \end{bmatrix} \tag{A.8.5}$$

The first matrix product evaluated is

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{AF} \\ \boldsymbol{GA} & \boldsymbol{E}+\boldsymbol{GAF} \end{bmatrix} \begin{bmatrix} \boldsymbol{U} & \boldsymbol{V} \\ \boldsymbol{W} & \boldsymbol{X} \end{bmatrix} = \begin{bmatrix} \boldsymbol{AU}+\boldsymbol{AFW} & \boldsymbol{AV}+\boldsymbol{AFX} \\ \boldsymbol{GAU}+\boldsymbol{EW}+\boldsymbol{GAFW} & \boldsymbol{GAV}+\boldsymbol{EX}+\boldsymbol{GAFX} \end{bmatrix}. \tag{A.8.6}$$

The g-inverse condition means therefore

$$\begin{bmatrix} \boldsymbol{AU}+\boldsymbol{AFW} & \boldsymbol{AV}+\boldsymbol{AFX} \\ \boldsymbol{GAU}+\boldsymbol{EW}+\boldsymbol{GAFW} & \boldsymbol{GAV}+\boldsymbol{EX}+\boldsymbol{GAFX} \end{bmatrix} \begin{bmatrix} \boldsymbol{A} & \boldsymbol{AF} \\ \boldsymbol{GA} & \boldsymbol{E}+\boldsymbol{GAF} \end{bmatrix} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{AF} \\ \boldsymbol{GA} & \boldsymbol{E}+\boldsymbol{GAF} \end{bmatrix} \tag{A.8.7}$$

For the upper left partition this means $\boldsymbol{AUA}+\boldsymbol{AFWA}+\boldsymbol{AVGA}+\boldsymbol{AFXGA}=\boldsymbol{A}$, and for the upper right partition it means $\boldsymbol{AUAF}+\boldsymbol{AFWAF}+\boldsymbol{AVE}+\boldsymbol{AVGAF}+\boldsymbol{AFXE}+\boldsymbol{AFXGAF}=\boldsymbol{AF}$. Postmultiply the upper left equation by $\boldsymbol{F}$ and subtract from the upper right to get $\boldsymbol{AVE}+\boldsymbol{AFXE}=\boldsymbol{O}$. For the lower left we get $\boldsymbol{GAUA}+\boldsymbol{EWA}+\boldsymbol{GAFWA}+\boldsymbol{GAVGA}+\boldsymbol{EXGA}+\boldsymbol{GAFXGA}=\boldsymbol{GA}$. Premultiplication of the upper left equation by $\boldsymbol{G}$ and subtraction gives $\boldsymbol{EWA}+\boldsymbol{EXGA}=\boldsymbol{O}$. For the lower right corner we get $\boldsymbol{GAUAF}+\boldsymbol{EWAF}+\boldsymbol{GAFWAF}+\boldsymbol{GAVE}+\boldsymbol{EXE}+\boldsymbol{GAFXE}+\boldsymbol{GAVGAF}+\boldsymbol{EXGAF}+\boldsymbol{GAFXGAF}=\boldsymbol{E}+\boldsymbol{GAF}$. Since $\boldsymbol{AVE}+\boldsymbol{AFXE}=\boldsymbol{O}$ and $\boldsymbol{EWA}+\boldsymbol{EXGA}=\boldsymbol{O}$, this simplifies to $\boldsymbol{GAUAF}+\boldsymbol{GAFWAF}+\boldsymbol{EXE}+\boldsymbol{GAVGAF}+\boldsymbol{GAFXGAF}=\boldsymbol{E}+\boldsymbol{GAF}$. 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 393. *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 392. □

Problem 394. *Show that the following are g-inverses:*

$$\begin{bmatrix} \boldsymbol{I} & \boldsymbol{X} \\ \boldsymbol{X}^\top & \boldsymbol{X}^\top\boldsymbol{X} \end{bmatrix}^- = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} \end{bmatrix} \qquad \begin{bmatrix} \boldsymbol{X}^\top\boldsymbol{X} & \boldsymbol{X}^\top \\ \boldsymbol{X} & \boldsymbol{I} \end{bmatrix}^- = \begin{bmatrix} (\boldsymbol{X}^\top\boldsymbol{X})^- & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{I}-\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^- \end{bmatrix} \tag{A.8.10}$$

Answer. Either do it by multiplying it out, or apply problem 392. □

Problem 395. *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 396. *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 397. *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 398. *Prove Theorem A.8.3. Hint: Let $\lambda_1, \dots, \lambda_r$ 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 399. *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 $\boldsymbol{t} \ne \boldsymbol{o}$, $\lambda^2 = \lambda$. 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 400 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 400. *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 401. *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 402. *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. □

APPENDIX B

Arrays of Higher Rank

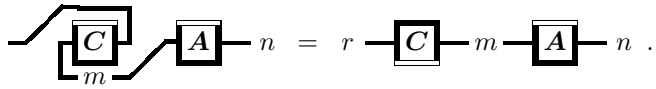
This chapter was presented at the Array Programming Languages Conference in Berlin, on July 24, 2000.

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 $\boldsymbol{ab}^\top$ 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

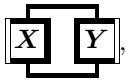
(1) Juxtapose the tiles for $\boldsymbol{X}$ and $\boldsymbol{Y}$, i.e., form their outer product, which is an array of rank 4 with typical element $x_{mp} y_{qn}$.
(2) Connect the East arm of $\boldsymbol{X}$ with the West arm of $\boldsymbol{Y}$. This is a contraction, resulting in an array of rank 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

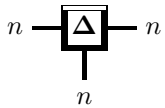
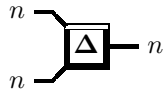
(1) Juxtapose the tiles for $\boldsymbol{X}$ and $\boldsymbol{Y}$.
(2) Connect the West arm of $\boldsymbol{X}$ with the East arm of $\boldsymbol{Y}$ to get the matrix product $\boldsymbol{YX}$.
(3) Now connect the East arm of $\boldsymbol{X}$ with the West arm of $\boldsymbol{Y}$ to get $\operatorname{tr} \boldsymbol{YX}$.

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.

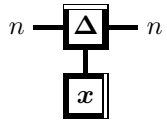
If the user specifies $\operatorname{tr}(\boldsymbol{XY})$, the computer is locked into one evaluation path: it first has to compute the matrix product $\boldsymbol{XY}$, even if $\boldsymbol{X}$ is a column vector and $\boldsymbol{Y}$ a row vector and it would be much more efficient to compute it as $\operatorname{tr}(\boldsymbol{YX})$, 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 

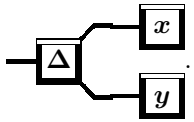
or similar configurations. It has 1's down the main diagonal and 0's elsewhere. It can be used to construct the diagonal matrix diag($\boldsymbol{x}$) of a vector (the square matrix with the vector in the diagonal and zeros elsewhere) as

(B.1.3) 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.

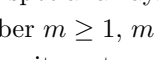
Each such graph can be evaluated in a number of different ways, and all these evaluations give the same result. In principle, each graph can be evaluated as follows: form the outer product of all arrays involved, and then contract along all those pairs of arms which are connected. For practical implementations 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`

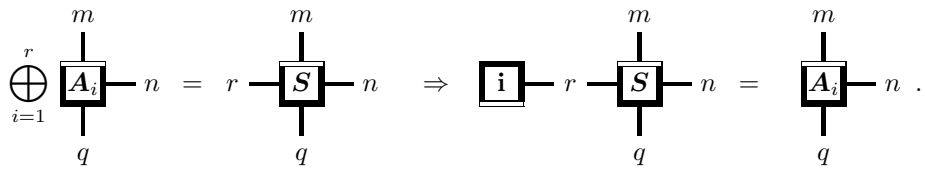
There are two kinds of special arrays: unit vectors and diagonal arrays.

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:

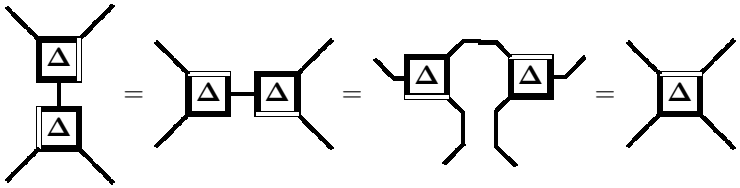


It is impossible to tell which is the first summand and which the second, direct sum is an operation defined on finite sets of arrays (where different elements of a set may be equal to each other in every respect but still have different identities).

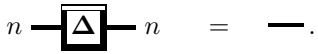
There is a broad rule of associativity: the order in which outer products and contractions are performed does not matter, as long as the at the end, the right arms are connected with each other. And there are distributive rules involving (contracted) outer products and direct sums.

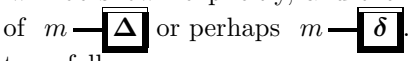
Additional rules apply for the special arrays. If two different diagonal arrays join arms, the result is again a diagonal array. For instance, the following three concatenations of diagonal three-way arrays are identical, and they all evaluate to

the (for a given dimension) unique diagonal array or rank 4:

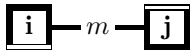
(B.2.1) 

The diagonal array of rank 2 is neutral under concatenation, i.e., it can be written as

(B.2.2) 

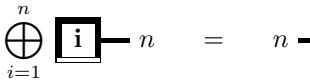
because attaching it to any array will not change this array. (B.2.1) and (B.2.2) make it possible to represent diagonal arrays simply as the branching points of several arms. This will make the array notation even simpler. However in the present introductory article, all diagonal arrays will be shown explicitly, and the vector of ones will be denoted  instead of or perhaps .

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:

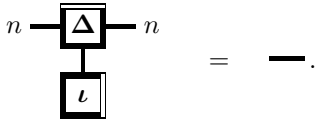
(B.2.4) 

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 403. *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 404. *Show that the diagonal matrix of the vector of ones is the identity matrix, i.e.,*

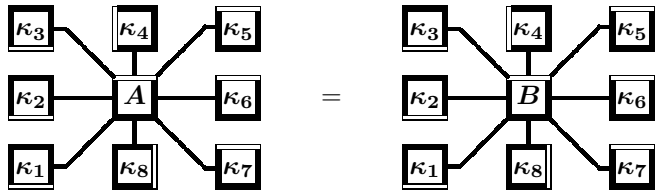
(B.2.5) 

Answer. In view of (B.2.2), this is a special case of Problem 403. □

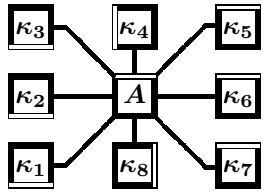
Problem 405. *A trivial array operation is the addition of an arm of dimension 1; for instance, this is how a n-vector can be turned into a $n \times 1$ matrix. Is this operation contained in the above system of axioms?*

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

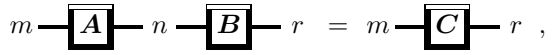
(B.2.6) 

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) 

$$= a_{\kappa_1\kappa_2\kappa_3\kappa_4\kappa_5\kappa_6\kappa_7\kappa_8}.$$

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).

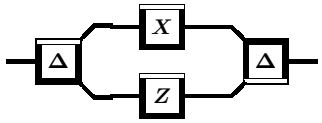
Maybe the following examples may clarify these points. If you specify a set of colors as $\{\text{red}, \text{green}, \text{blue}\}$, then this representation has an ordering built in: red comes first, then green, then blue. However this ordering is not part of the definition of the set; $\{\text{green}, \text{red}, \text{blue}\}$ 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 406. *The trace of the product of two matrices can be written as*

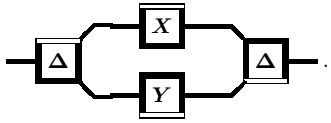
$$\operatorname{tr}(\boldsymbol{XY}) = \boldsymbol{\iota}^\top(\boldsymbol{X} * \boldsymbol{Y}^\top)\boldsymbol{\iota}. \tag{B.2.8}$$

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* $\operatorname{tr}(\boldsymbol{XY})$.

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

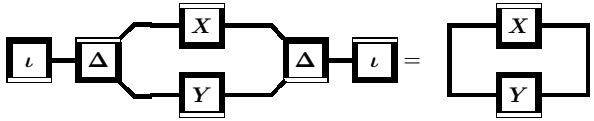
$\boldsymbol{X} * \boldsymbol{Z} =$ 

If $\boldsymbol{Z} = \boldsymbol{Y}^\top$, one gets

$\boldsymbol{X} * \boldsymbol{Y}^\top =$ 

.

Therefore one gets, using (B.2.5):

$\boldsymbol{\iota}^\top(\boldsymbol{X} * \boldsymbol{Y}^\top)\boldsymbol{\iota} =$ 

$= \operatorname{tr}(\boldsymbol{XY})$

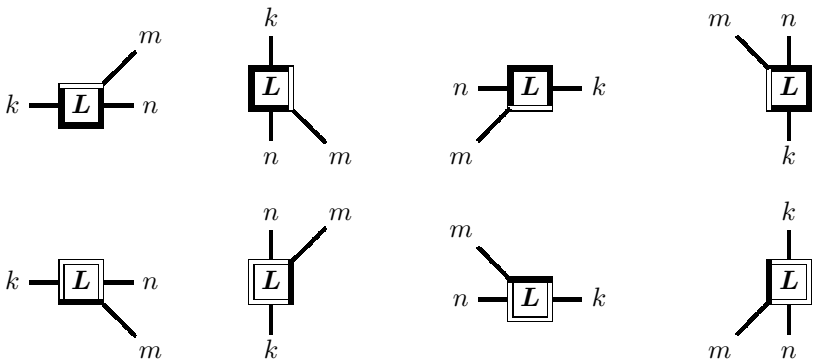
□

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

and the matrix

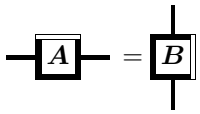
. Since sometimes one and sometimes the other notation seems more natural, both will be used. For higher arrays, flipping over arranges the arms in a different fashion, which is sometimes convenient in order to keep the graphs uncluttered. It will be especially useful for differentiation. If one allows 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:

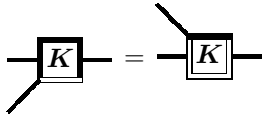


The black-and-white pattern at the edge of the tile indicates whether and how much the tile has been turned and/or flipped over, so that one can keep track which arm is which. In the above example, the arm with dimension k will always be called the West arm, whatever position the tile is in.

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

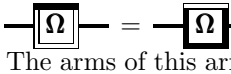


or


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

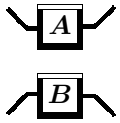
above expressions would therefore be

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 is symmetric if and only if  . 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 satisfying 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 407. *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. □

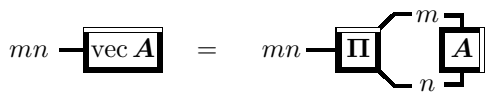
By the way, a better protocol for vectorizing would have been to assemble all rows into one long row vector and then converting it into a column vector. In other words

$$\text{if} \quad \boldsymbol{B} = \begin{bmatrix} \boldsymbol{b}_1^\top \\ \vdots \\ \boldsymbol{b}_m^\top \end{bmatrix} \quad \text{then } \operatorname{vec}(\boldsymbol{B}) \text{ should have been defined as} \quad \begin{bmatrix} \boldsymbol{b}_1 \\ \vdots \\ \boldsymbol{b}_m \end{bmatrix}.$$

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 408. [**The71**, pp. 303–306] *Prove the following simple properties of the Kronecker product:*

$$(\boldsymbol{A}\otimes\boldsymbol{B})^\top = \boldsymbol{A}^\top\otimes\boldsymbol{B}^\top \tag{B.5.4}$$

$$(\boldsymbol{A}\otimes\boldsymbol{B})\otimes\boldsymbol{C} = \boldsymbol{A}\otimes(\boldsymbol{B}\otimes\boldsymbol{C}) \tag{B.5.5}$$

$$\boldsymbol{I}\otimes\boldsymbol{I} = \boldsymbol{I} \tag{B.5.6}$$

$$(\boldsymbol{A}\otimes\boldsymbol{B})(\boldsymbol{C}\otimes\boldsymbol{D}) = \boldsymbol{AC}\otimes\boldsymbol{BD} \tag{B.5.7}$$

$$(\boldsymbol{A}\otimes\boldsymbol{B})^{-1} = \boldsymbol{A}^{-1}\otimes\boldsymbol{B}^{-1} \tag{B.5.8}$$

$$(\boldsymbol{A}\otimes\boldsymbol{B})^{-} = \boldsymbol{A}^{-}\otimes\boldsymbol{B}^{-} \tag{B.5.9}$$

$$\boldsymbol{A}\otimes(\boldsymbol{B}+\boldsymbol{C}) = \boldsymbol{A}\otimes\boldsymbol{B}+\boldsymbol{A}\otimes\boldsymbol{C} \tag{B.5.10}$$

$$(\boldsymbol{A}+\boldsymbol{B})\otimes\boldsymbol{C} = \boldsymbol{A}\otimes\boldsymbol{C}+\boldsymbol{B}\otimes\boldsymbol{C} \tag{B.5.11}$$

$$(c\boldsymbol{A})\otimes\boldsymbol{B} = \boldsymbol{A}\otimes(c\boldsymbol{B}) = c(\boldsymbol{A}\otimes\boldsymbol{B}) \tag{B.5.12}$$

$$\begin{bmatrix}\boldsymbol{A}_{11} & \boldsymbol{A}_{12}\\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22}\end{bmatrix}\otimes\boldsymbol{B} = \begin{bmatrix}\boldsymbol{A}_{11}\otimes\boldsymbol{B} & \boldsymbol{A}_{12}\otimes\boldsymbol{B}\\ \boldsymbol{A}_{21}\otimes\boldsymbol{B} & \boldsymbol{A}_{22}\otimes\boldsymbol{B}\end{bmatrix} \tag{B.5.13}$$

$$\operatorname{rank}(\boldsymbol{A}\otimes\boldsymbol{B}) = (\operatorname{rank}\boldsymbol{A})(\operatorname{rank}\boldsymbol{B}) \tag{B.5.14}$$

$$\operatorname{tr}(\boldsymbol{A}\otimes\boldsymbol{B}) = (\operatorname{tr}\boldsymbol{A})(\operatorname{tr}\boldsymbol{B}) \tag{B.5.15}$$

If a is a 1×1 matrix, then

$$a\otimes\boldsymbol{B} = \boldsymbol{B}\otimes a = a\boldsymbol{B} \tag{B.5.16}$$

$$\det(\boldsymbol{A}\otimes\boldsymbol{B}) = (\det(\boldsymbol{A}))^n(\det(\boldsymbol{B}))^k \tag{B.5.17}$$

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 $\operatorname{rank}(\boldsymbol{A}) = \operatorname{tr}(\boldsymbol{AA}^-)$. compare Problem 366. □

PROBLEM 409. *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 410. [**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 $\boldsymbol{Y}$ is often given as the matrix $\mathcal{V}[\operatorname{vec}\boldsymbol{Y}]$, where the vectorization is usually not made explicit, i.e., this matrix is denoted $\mathcal{V}[\boldsymbol{Y}]$.

PROBLEM 411. *If $\mathcal{V}[\operatorname{vec}\boldsymbol{Y}] = \boldsymbol{\Sigma}\otimes\boldsymbol{\Omega}$ 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): $\mathcal{V}[\operatorname{vec}\boldsymbol{PYQ}] = \mathcal{V}[(\boldsymbol{Q}^\top\otimes\boldsymbol{P})\operatorname{vec}\boldsymbol{Y}] = (\boldsymbol{Q}^\top\otimes\boldsymbol{P})(\boldsymbol{\Sigma}\otimes\boldsymbol{\Omega})(\boldsymbol{Q}\otimes\boldsymbol{P}^\top)$. Now apply (B.5.7). □

PROBLEM 412. *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 $\boldsymbol{C}=\boldsymbol{\gamma}^\top$. □

PROBLEM 413. *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.

$$\boldsymbol{\delta}\otimes\boldsymbol{\alpha} = \begin{bmatrix}\boldsymbol{\alpha}\delta_1\\ \vdots\\ \boldsymbol{\alpha}\delta_n\end{bmatrix} \qquad \mathcal{V}[\boldsymbol{\delta}\otimes\boldsymbol{\alpha}] = \begin{bmatrix}\boldsymbol{\alpha}\operatorname{var}[\delta_1]\boldsymbol{\alpha}^\top & \boldsymbol{\alpha}\operatorname{cov}[\delta_1,\delta_2]\boldsymbol{\alpha}^\top & \cdots & \boldsymbol{\alpha}\operatorname{cov}[\delta_1,\delta_n]\boldsymbol{\alpha}^\top\\ \boldsymbol{\alpha}\operatorname{cov}[\delta_2,\delta_1]\boldsymbol{\alpha}^\top & \boldsymbol{\alpha}\operatorname{var}[\delta_2]\boldsymbol{\alpha}^\top & \cdots & \boldsymbol{\alpha}\operatorname{cov}[\delta_2,\delta_n]\boldsymbol{\alpha}^\top\\ \vdots & \vdots & \ddots & \vdots\\ \boldsymbol{\alpha}\operatorname{cov}[\delta_n,\delta_1]\boldsymbol{\alpha}^\top & \boldsymbol{\alpha}\operatorname{cov}[\delta_n,\delta_2]\boldsymbol{\alpha}^\top & \cdots & \boldsymbol{\alpha}\operatorname{cov}[\delta_n,\delta_n]\boldsymbol{\alpha}^\top\end{bmatrix} =$$

$$= \begin{bmatrix}\operatorname{var}[\delta_1]\boldsymbol{\alpha}\boldsymbol{\alpha}^\top & \operatorname{cov}[\delta_1,\delta_2]\boldsymbol{\alpha}\boldsymbol{\alpha}^\top & \cdots & \operatorname{cov}[\delta_1,\delta_n]\boldsymbol{\alpha}\boldsymbol{\alpha}^\top\\ \operatorname{cov}[\delta_2,\delta_1]\boldsymbol{\alpha}\boldsymbol{\alpha}^\top & \operatorname{var}[\delta_2]\boldsymbol{\alpha}\boldsymbol{\alpha}^\top & \cdots & \operatorname{cov}[\delta_2,\delta_n]\boldsymbol{\alpha}\boldsymbol{\alpha}^\top\\ \vdots & \vdots & \ddots & \vdots\\ \operatorname{cov}[\delta_n,\delta_1]\boldsymbol{\alpha}\boldsymbol{\alpha}^\top & \operatorname{cov}[\delta_n,\delta_2]\boldsymbol{\alpha}\boldsymbol{\alpha}^\top & \cdots & \operatorname{cov}[\delta_n,\delta_n]\boldsymbol{\alpha}\boldsymbol{\alpha}^\top\end{bmatrix} = \mathcal{V}[\boldsymbol{\delta}]\otimes\boldsymbol{\alpha}\boldsymbol{\alpha}^\top$$

□

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 414. *Use* (B.5.20) *to compute* $\boldsymbol{K}^{(2,3)}$.

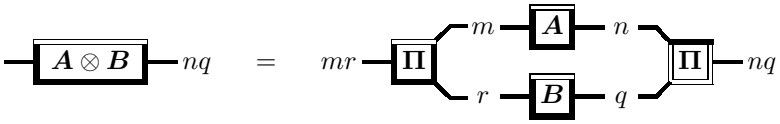
ANSWER.

$$\boldsymbol{K}^{(2,3)} = \begin{bmatrix}1&0&0&0&0&0\\0&0&1&0&0&0\\0&0&0&0&1&0\\0&1&0&0&0&0\\0&0&0&1&0&0\\0&0&0&0&0&1\end{bmatrix} \tag{B.5.22}$$

□

B.5.4. Kronecker Product and Vectorization in Tile Notation. The Kronecker product of

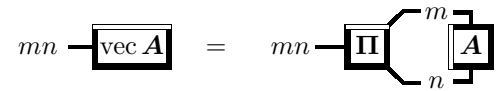
and is the following concatenation of $\boldsymbol{A}$ and $\boldsymbol{B}$ with members of a certain family of three-way arrays $\boldsymbol{\Pi}^{(i,j)}$:

(B.5.23) 

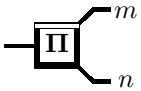
Strictly speaking we should have written $\boldsymbol{\Pi}^{(m,r)}$ and $\boldsymbol{\Pi}^{(n,q)}$ for the two $\boldsymbol{\Pi}$-arrays in (B.5.23), but the superscripts can be inferred from the context: the first superscript is the dimension of the Northeast arm, and the second that of the Southeast arm.

Vectorization uses a member of the same family $\boldsymbol{\Pi}^{(m,n)}$ to convert the matrix

into the vector

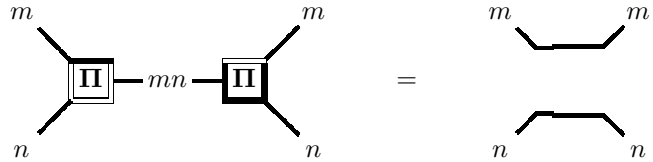
(B.5.24) 

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

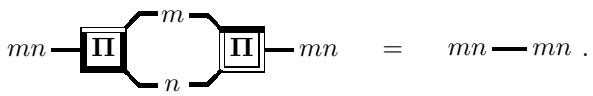
(B.5.25) 

.

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 $\mathbf{\Pi}^{(m,n)}$ 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.

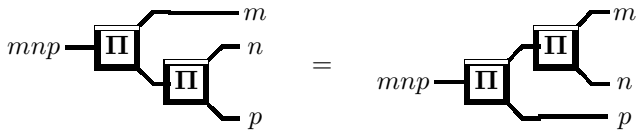
The second identity satisfied by $\mathbf{\Pi}^{(m,n)}$ is

(B.5.26)

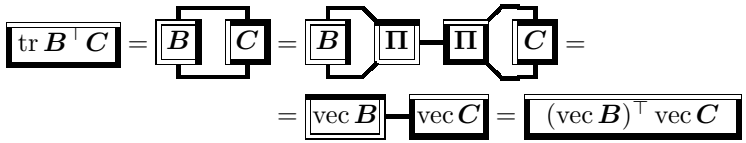


Finally, there is also associativity:

(B.5.27)

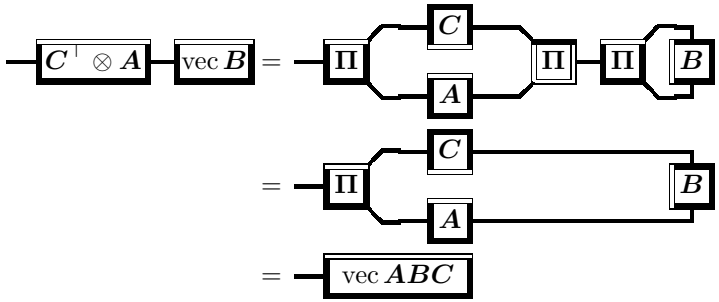


Here is the answer to Problem 407 in tile notation:



(B.5.28)

Equation (B.5.25) was central for obtaining the result. The answer to Problem 410 also relies on equation (B.5.25):

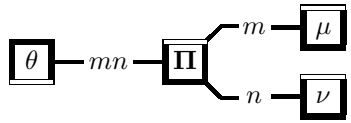


(B.5.29)

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 $\boxed{\boldsymbol{\theta}}$—$\boxed{\text{vec}\,\boldsymbol{A}} = c_\theta$, (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 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.

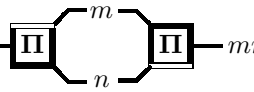
The proof of (B.5.25) uses the fact that for every θ there is exactly one μ and one ν such that $\pi^{(m,n)}_{\theta\mu\nu} \neq 0$:

$$\sum_{\theta=1}^{\theta=mn} \pi^{(m,n)}_{\theta\mu\nu} \pi^{(m,n)}_{\theta\omega\sigma} = \begin{cases} 1 & \text{if } \mu = \omega \text{ and } \nu = \sigma \\ 0 & \text{otherwise} \end{cases} \tag{B.5.33}$$

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

(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 $\boldsymbol{K}^{(m,n)}$:

(B.5.35) $\boldsymbol{K}^{(m,n)} =$

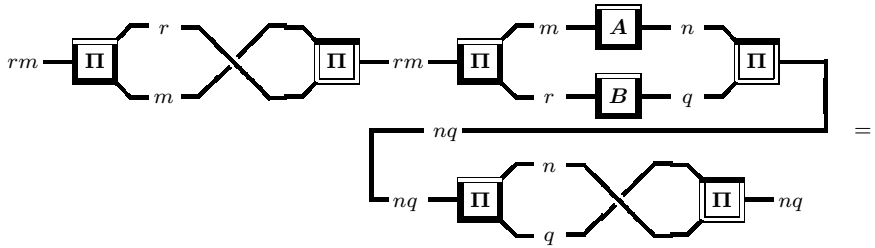
.

PROBLEM 415. *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). □

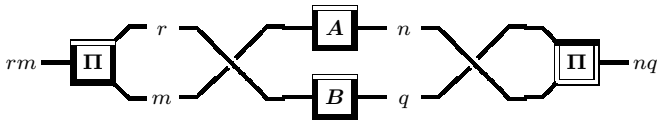
PROBLEM 416. *Prove* (B.5.21) *in tile notation.*

ANSWER. Start with a tile representation of $\boldsymbol{K}^{(r,m)}(\boldsymbol{A} \otimes \boldsymbol{B})\boldsymbol{K}^{(n,q)}$:

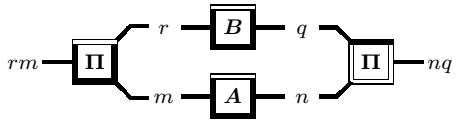


=

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

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} f_1(x_1, \dots, x_m) \\ \vdots \\ f_n(x_1, \dots, x_m) \end{bmatrix} \tag{C.1.2}$$

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 $\boldsymbol{A} = \boldsymbol{Df}$.

Problem 417. *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

$$\begin{array}{lcccccccccc} \boldsymbol{x}^\top\boldsymbol{M}\boldsymbol{x} & = & x_1m_{11}x_1 & + & x_1m_{12}x_2 & + & \cdots & + & x_1m_{1n}x_n & + \\ & + & x_2m_{21}x_1 & + & x_2m_{22}x_2 & + & \cdots & + & x_2m_{2n}x_n & + \\ & & \vdots & & \vdots & & & & \vdots & \\ & + & x_nm_{n1}x_1 & + & x_nm_{n2}x_2 & + & \cdots & + & x_nm_{nn}x_n & \end{array}$$

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

$$\begin{array}{lcccccccc} \partial\boldsymbol{x}^\top\boldsymbol{M}\boldsymbol{x}/\partial x_1 & = & 2m_{11}x_1 & + & m_{12}x_2 & + & \cdots & + & m_{1n}x_n & + \\ & + & x_2m_{21} & + & & & & & & \\ & & \vdots & & & & & & & \\ & + & x_nm_{n1} & & & & & & & \end{array}$$

Now split the upper diagonal element, writing it as $m_{11}x_1 + x_1 m_{11}$, to get

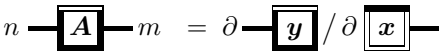
$$\begin{array}{lllllllll} = & m_{11}x_1 & + & m_{12}x_2 & + & \cdots & + & m_{1n}x_n & + \\ + & x_1 m_{11} & + & & & & & & \\ + & x_2 m_{21} & + & & & & & & \\ & \vdots & & & & & & & \\ + & x_n m_{n1} & & & & & & & \end{array}$$

The sum of the elements in the first row is the first element of the column vector $\boldsymbol{Mx}$, 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

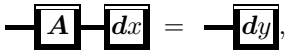
$$\partial \boldsymbol{x}^\top \boldsymbol{Mx} / \partial \boldsymbol{x}^\top = \boldsymbol{x}^\top (\boldsymbol{M} + \boldsymbol{M}^\top). \tag{C.1.8}$$

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 $(\boldsymbol{M} + \boldsymbol{M}^\top)/2$.

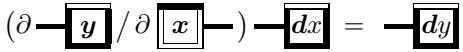
Here is the tile notation for matrix differentiation: If n — $\boldsymbol{y}$ depends on m — $\boldsymbol{x}$, then

(C.1.9) 

is that array which satisfies

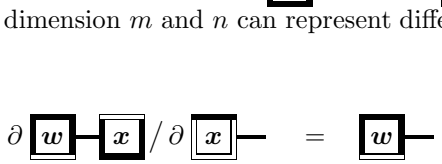
(C.1.10) 

i.e.,

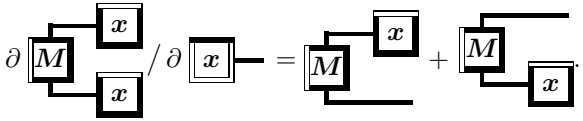
(C.1.11) 

Extended substitutability applies here: n — $\boldsymbol{y}$ and m — $\boldsymbol{x}$ 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

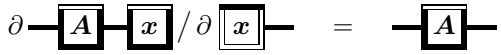
(C.1.13) 

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 $\boldsymbol{y} = \boldsymbol{Ax}$ is a linear vector function, then its derivative is that same linear vector function:

$$\partial \boldsymbol{Ax} / \partial \boldsymbol{x}^\top = \boldsymbol{A}, \tag{C.1.14}$$

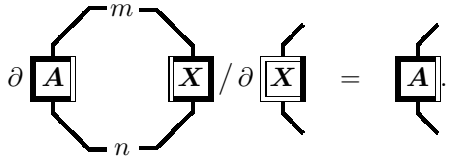
or in tiles

(C.1.15) 

Problem 418. *Show that*

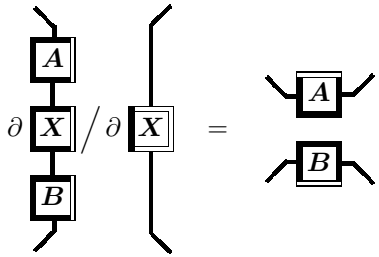
$$\frac{\partial \operatorname{tr} \boldsymbol{AX}}{\partial \boldsymbol{X}^\top} = \boldsymbol{A}. \tag{C.1.16}$$

In tiles it reads

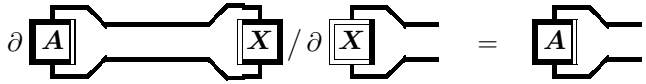
(C.1.17) 

Answer. $\operatorname{tr}(\boldsymbol{AX}) = \sum_{i,j} a_{ij} x_{ji}$ 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

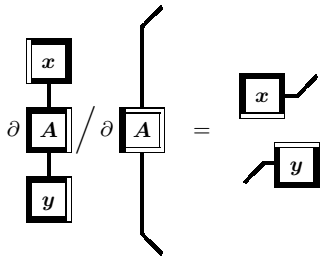


and this is a special case of (C.1.12), since the two parallel arms can be treated as one arm. With a better development of the logic underlying this notation, it will not be necessary to formulate them as separate theorems; all matrix differentiation rules given so far are trivial applications of (C.1.15).

Problem 419. *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.

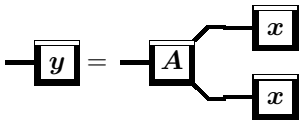
(C.1.19)



□

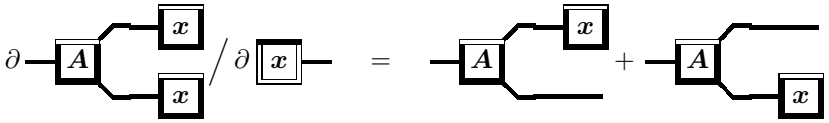
Here is a basic differentiation rule for *bilinear* array concatenations: if

(C.1.20)



then one gets the following simple generalization of (C.1.13):

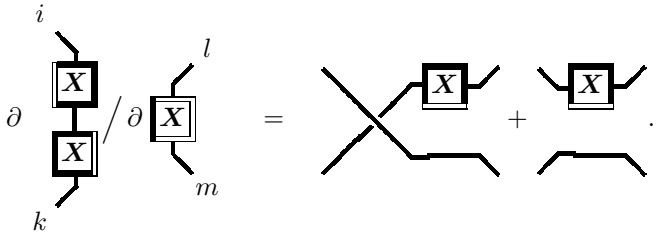
(C.1.21)



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) $\dfrac{\partial \boldsymbol{X}^\top \boldsymbol{X}}{\partial \boldsymbol{X}^\top} =$



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 (14.2.2).

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.

[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.

[Bar82] Vic Barnett. *Comparative Statistical Inference*. Wiley, New York, 1982.

[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.

[BKW80] David A. Belsley, Edwin Kuh, and Roy E. Welsch. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley, New York, 1980.

[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.

[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.

[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.

[Cra43] A. T. Craig. Note on the independence of certain quadratic forms. *Annals of Mathematical Statistics*, 14:195, 1943.

[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.

[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.

[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.

[Dou92] Christopher Dougherty. *Introduction to Econometrics*. Oxford University Press, Oxford, 1992.

[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.

[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.

[FS91] Milton Friedman and Anna J. Schwarz. Alternative approaches to analyzing economic data. *American Economic Review*, 81(1):39–49, March 1991.

[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.

[Gre97] William H. Greene. *Econometric Analysis*. Prentice Hall, Upper Saddle River, NJ, third edition, 1997.

[Gre03] William H. Greene. *Econometric Analysis*. Prentice Hall, Upper Saddle River, New Jersey 07458, fifth edition, 2003.

[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.

[Hen95] David F. Hendry. *Dynamic Econometrics*. Oxford University Press, Oxford, New York, 1995.

[HK79] James M. Henle and Eugene M. Kleinberg. *Infinitesimal Calculus*. MIT Press, 1979.

[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.

[HT83] Robert V. Hogg and Elliot A. Tanis. *Probability and Statistical Inference*. Macmillan, second edition, 1983.

[HVdP02] Ben J. Hejdra and Frederick Van der Ploeg. *Foundations of Modern Macroeconomics*. Oxford University Press, 2002.

[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.

[KA69] J. Koerts and A. P. J. Abramanse. *On the Theory and Application of the General Linear Model*. Rotterdam University Press, Rotterdam, 1969.

[Kap89] Jagat Narain Kapur. *Maximum Entropy Models in Science and Engineering*. Wiley, 1989.

[Ken98] Peter Kennedy. *A Guide to Econometrics*. MIT Press, Cambridge, MA, fourth edition, 1998.

[Khi57] R. T. Khinchin. *Mathematical Foundations of Information Theory*. Dover Publications, New York, 1957.

[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.

[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.

[Lea75] Edward E. Leamer. A result on the sign of the restricted least squares estimator. *Journal of Econometrics*, 3:387–390, 1975.

[Mag88] Jan R. Magnus. *Linear Structures*. Oxford University Press, New York, 1988.

[Mal80] E. Malinvaud. *Statistical Methods of Econometrics*. North-Holland, Amsterdam, third edition, 1980.

[MN88] Jan R. Magnus and Heinz Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Wiley, Chichester, 1988.

[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.

[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.

[Rén70] Alfred Rényi. *Foundations of Probability*. Holden-Day, San Francisco, 1970.

[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.

[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.

[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.

[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.

[SM86] Hans Schneeweiß and Hans-Joachim Mittag. *Lineare Modelle mit fehlerbehafteten Daten*. Physica Verlag, Heidelberg, Wien, 1986.

[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.

[SW76] Thomes J. Sargent and Neil Wallace. Rational expectations and the theory of economic policy. *Journal of Monetary Economics*, 2:169–183, 1976.

[The71] Henri Theil. *Principles of Econometrics*. Wiley, New York, 1971.

[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.

[Wit85] Uli Wittman. *Das Konzept rationaler Preiserwartungen*, volume 241 of *Lecture Notes in Economics and Mathematical Systems*. Springer, 1985.

[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.