ECONOMETRICS FIELD EXAM SUMMER 2001, PART I

ECONOMICS DEPARTMENT, UNIVERSITY OF UTAH

Part 1 of this exam (Hans Ehrbar) has three subparts *a*. *b*, and *c*. Part 2 is provided by Peter Philips.

- For part 1*a* (a simple property of conditional probabilities with two novel applications) you have to answer Problem 1 and then either Problem 2 or Problem 3. Problem 3 is a little more difficult than the Problem 2, but you will learn much more doing it, and you will get also more points for it.
- For part 1*b* you have to answer Problem 4 (regression diagnostics).
- For part 1*c* you may answer either Problem 5 or Problem 6 (theory of multivariate estimation problems).

This is a closed book exam but you may bring one sheet with formulas with you; write your name on the formula sheet and submit it together with your exam. If the

Date of exam Aug 14, 2001.

formula sheet is missing, your exam will count as failed. General formulas, please, not the answers to the questions in the class notes.

Problem 1. *2 points A and B are arbitrary events. Prove that the probability of B can be written as:*

$$\Pr[B] = \Pr[B|A]\Pr[A] + \Pr[B|A']\Pr[A'] \tag{1}$$

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 2. [Rao97, pp. 16–17] *You want to know how many people smoke marijuana. In order to get a honest response, you list the following two questions:*

- *Do you smoke marijuana?*
- *Does your telephone number end with an even digit?*

You ask the respondent to toss a coin and answer the first question if head turns up, and answer the second question if tail turns up. Since you do not know which question the respondent is answering, the privacy of the information is maintained.

• **a.** *1 point Assuming the probability that the telephone number ends with an even digit is λ, and the probability that the person smokes marijuana is π, what is the probability that the answer to this combined question is yes?*

Answer. Let's call the probability of a yes answer τ. Then $\tau = \frac{1}{2}\pi + \frac{1}{2}\lambda$. □

• **b.** *1 point Assume you make such a survey and $\frac{1}{3}$ of the respondents answer yes. You also sample 300 numbers from the telephone book and it turns out that exactly $\frac{1}{2}$ of the telephone numbers end in an even digit. What is your estimate of π?*

Answer. From the above formula you get $2\tau = \pi + \lambda$, which gives $\pi = 2\tau - \lambda$. Inserting $\tau = \frac{1}{3}$ and $\lambda = \frac{1}{2}$ the answer is $\pi = \frac{1}{6}$. □

Problem 3. *Our universal set U consists of patients who have a certain disease. We will explore the causal effect of a given treatment with the help of three events, T, C, and S, the first two of which are* counterfactual*, compare* [Hol86]*. These events are defined as follows: T consists of all patients who would recover if given treatment; C consists of all patients who would recover if not given treatment (i.e., if included in the control group). The event S consists of all patients actually receiving treatment. The average causal effect of the treatment is defined as* $\Pr[T] - \Pr[C]$.

• **a.** *2 points Show that*

$$\Pr[T] = \Pr[T|S]\Pr[S] + \Pr[T|S'](1 - \Pr[S]) \tag{2}$$

and that

$$\Pr[C] = \Pr[C|S]\Pr[S] + \Pr[C|S'](1 - \Pr[S]) \tag{3}$$

Which of these probabilities can be estimated as the frequencies of observable outcomes and which cannot?

Answer. This is a direct application of (1). The problem here is that for all $\omega \in C$, i.e., for those patients who do not receive treatment, we do not know whether they would have recovered if given treatment, and for all $\omega \in T$, i.e., for those patients who do receive treatment, we do not know whether they would have recovered if not given treatment. In other words, neither $\Pr[T|S]$ nor $\mathrm{E}[C|S']$ can be estimated as the frequencies of observable outcomes. □

• **b.** *2 points Assume now that S is independent of T and C, because the subjects are assigned randomly to treatment or control. How can this be used to estimate those elements in the equations (2) and (3) which could not be estimated before?*

Answer. In this case, $\Pr[T|S] = \Pr[T|S']$ and $\Pr[C|S'] = \Pr[C|S]$. Therefore, the average causal effect can be simplified as follows:

$$\begin{aligned}\Pr[T] - \Pr[C] &= \Pr[T|S]\Pr[S] + \Pr[T|S'](1 - \Pr[S]) - \Pr[C|S]\Pr[S] + \Pr[C|S'](1 - \Pr[S]) \\ &= \Pr[T|S]\Pr[S] + \Pr[T|S](1 - \Pr[S]) - \Pr[C|S']\Pr[S] + \Pr[C|S'](1 - \Pr[S]) \\ &= \Pr[T|S] - \Pr[C|S'] \qquad (4)\end{aligned}$$

□

• **c.** *2 points Why were all these calculations necessary? Could one not have defined from the beginning that the causal effect of the treatment is $\Pr[T|S] - \Pr[C|S']$?*

Answer. $\Pr[T|S] - \Pr[C|S']$ is only the empirical difference in recovery frequencies between those who receive treatment and those who do not. It is always possible to measure these differences, but these differences are not necessarily due to the treatment but may be due to other reasons. □

Problem 4. *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.

□

-

Problem 5. *We are working in the dummy-variable model for pooled data, which can be written as*

$$\boldsymbol{Y} = \boldsymbol{\iota}\boldsymbol{\alpha}^\top + \begin{bmatrix}\boldsymbol{X}_1\boldsymbol{\beta} & \cdots & \boldsymbol{X}_m\boldsymbol{\beta}\end{bmatrix} + \boldsymbol{E} \tag{5}$$

where $\boldsymbol{Y} = \begin{bmatrix}\boldsymbol{y}_1 & \cdots & \boldsymbol{y}_m\end{bmatrix}$ *is* $t \times m$*, each of the* $\boldsymbol{X}_i$ *is* $t \times k$*,* $\boldsymbol{\iota}$ *is the* t*-vector of ones,* $\boldsymbol{E}$ *is a* $t \times m$ *matrix of identically distributed independent error terms with zero mean, and* $\boldsymbol{\alpha}$ *is a* m*-vector and* $\boldsymbol{\beta}$ *a* k*-vector of unknown nonrandom parameters.*

- **a.** *3 points Describe in words the characteristics of this model and how it can come about.*

Answer. Each of the m units has a different intercept, slope is the same. Equal marginal costs but different fixed costs. □

- **b.** *4 points Describe the issues in estimating this model and how it should be estimated.*

Answer. After vectorization OLS is fine, but design matrix very big. One can derive formulas that are easier to evaluate numerically because they involve smaller matrices, by exploiting the structure of the overall design matrix. First estimate the slope parameters by sweeping out the means, then the intercepts. □

- **c.** *3 points Set up an* F*-test testing whether the individual intercept parameters are indeed different, by running two separate regressions on the restricted and the*

unrestricted model and using the generic formula for the F-test:

$$\frac{(SSE_{constrained} - SSE_{unconstrained})/number\ of\ constraints}{SSE_{\text{unconstrained}}/(numb.\ of\ obs. - numb.\ of\ coeff.\ in\ unconst.\ mod.)} \tag{6}$$

Describe how you would run the restricted and how the unrestricted model. Give the number of constraints, the number of observations, and the number of coefficients in the unrestricted model in terms of m, t, and k.

Answer. The unrestricted regression is the dummy variables regression which was described here: first form $\boldsymbol{DY}$ and all the $\boldsymbol{DX}_i$, then run regression (**??**) without intercept, which is already enough to get the SSE_r.

Number of constraints is $m-1$, number of observations is tm, and number of coefficients in the unrestricted model is $k+m$. The test statistic is given in [JHG+88, (11.4.25) on p. 475]:

$$F = \frac{(SSE_r - SSE_u)/(m-1)}{SSE_u/(mt-m-k)} \tag{7}$$

□

• **d.** *3 points An alternative model specification is the variance components model. Describe it as well as you can, and discuss situations when it would be more appropriate than the model above.*

Answer. If one believes that variances are similar, and if one is not interested in those particular firms in the sample, but in all firms. □

Problem 6. [Gre97, p. 709 ff]. *Here is a demand and supply curve with* $\boldsymbol{q}$ *quantity,* $\boldsymbol{p}$ *price,* $\boldsymbol{y}$ *income, and* $\boldsymbol{\iota}$ *is the vector of ones. All vectors are t-vectors.*

$$\boldsymbol{q} = \alpha_0\boldsymbol{\iota} + \alpha_1\boldsymbol{p} + \alpha_2\boldsymbol{y} + \boldsymbol{\varepsilon}_d \qquad \boldsymbol{\varepsilon}_d \sim (\boldsymbol{o}, \sigma_d^2\boldsymbol{I}) \qquad \textit{(demand)} \tag{8}$$

$$\boldsymbol{q} = \beta_0\boldsymbol{\iota} + \beta_1\boldsymbol{p} + \boldsymbol{\varepsilon}_s \qquad \boldsymbol{\varepsilon}_s \sim (\boldsymbol{o}, \sigma_s^2\boldsymbol{I}) \qquad \textit{(supply)} \tag{9}$$

$\boldsymbol{\varepsilon}_d$ *and* $\boldsymbol{\varepsilon}_s$ *are independent of* $\boldsymbol{y}$*, but amongst each other they are contemporaneously correlated, with their covariance constant over time:*

$$\operatorname{cov}[\varepsilon_{dt}, \varepsilon_{su}] = \begin{cases} 0 & \text{if } t \neq u \\ \sigma_{ds} & \text{if } t = u \end{cases} \tag{10}$$

- **a.** *1 point Which variables are exogenous and which are endogenous?*

Answer. $\boldsymbol{p}$ and $\boldsymbol{q}$ are called *jointly dependent* or *endogenous.* $\boldsymbol{y}$ is determined outside the system or *exogenous.* □

- **b.** *2 points Assuming* $\alpha_1 \neq \beta_1$*, verify that the reduced-form equations for* $\boldsymbol{p}$ *and* $\boldsymbol{q}$ *are as follows:*

$$\boldsymbol{p} = \frac{\alpha_0 - \beta_0}{\beta_1 - \alpha_1}\boldsymbol{\iota} + \frac{\alpha_2}{\beta_1 - \alpha_1}\boldsymbol{y} + \frac{\boldsymbol{\varepsilon}_d - \boldsymbol{\varepsilon}_s}{\beta_1 - \alpha_1} \tag{11}$$

$$\boldsymbol{q} = \frac{\beta_1\alpha_0 - \beta_0\alpha_1}{\beta_1 - \alpha_1}\boldsymbol{\iota} + \frac{\beta_1\alpha_2}{\beta_1 - \alpha_1}\boldsymbol{y} + \frac{\beta_1\boldsymbol{\varepsilon}_d - \alpha_1\boldsymbol{\varepsilon}_s}{\beta_1 - \alpha_1} \tag{12}$$

Answer. One gets the reduced form equation for $\boldsymbol{p}$ by simply setting the righthand sides equal:

$$\begin{aligned}\beta_0\boldsymbol{\iota} + \beta_1\boldsymbol{p} + \boldsymbol{\varepsilon}_s &= \alpha_0\boldsymbol{\iota} + \alpha_1\boldsymbol{p} + \alpha_2\boldsymbol{y} + \boldsymbol{\varepsilon}_d \\ (\beta_1 - \alpha_1)\boldsymbol{p} &= (\alpha_0 - \beta_0)\boldsymbol{\iota} + \alpha_2\boldsymbol{y} + \boldsymbol{\varepsilon}_d - \boldsymbol{\varepsilon}_s,\end{aligned}$$

hence (11). To get the reduced form equation for $\boldsymbol{q}$, plug that for $\boldsymbol{p}$ into the supply function (one might also plug it into the demand function but the math would be more complicated):

$$\boldsymbol{q} = \beta_0\boldsymbol{\iota} + \beta_1\boldsymbol{p} + \boldsymbol{\varepsilon}_s = \beta_0\boldsymbol{\iota} + \frac{\beta_1(\alpha_0 - \beta_0)}{\beta_1 - \alpha_1}\boldsymbol{\iota} + \frac{\beta_1\alpha_2}{\beta_1 - \alpha_1}\boldsymbol{y} + \frac{\beta_1(\boldsymbol{\varepsilon}_d - \boldsymbol{\varepsilon}_s)}{\beta_1 - \alpha_1} + \boldsymbol{\varepsilon}_s$$

Combining the first two and the last two terms gives (12). □

• **c.** *2 points Show that one will in general not get consistent estimates of the supply equation parameters if one regresses $\boldsymbol{q}$ on $\boldsymbol{p}$ (with an intercept).*

Answer. By (11) (the reduced form equation for $\boldsymbol{p}$), $\operatorname{cov}[\varepsilon_{st}, p_t] = \operatorname{cov}[\varepsilon_{st}, \frac{\varepsilon_{dt} - \varepsilon_{st}}{\beta_1 - \alpha_1}] = \frac{\sigma_{sd} - \sigma_s^2}{\beta_1 - \alpha_1}$. This is generally $\neq 0$, therefore inconsistency. □

• **d.** *2 points If one estimates the supply function by instrumental variables, using $\boldsymbol{y}$ as an instrument for $\boldsymbol{p}$ and $\boldsymbol{\iota}$ as instrument for itself, write down the formula for the resulting estimator $\tilde{\beta}_1$ of β_1 and show that it is consistent. You are allowed to use, without proof, the following equation for the simple instrumental variables estimator:*

(13)
$$\tilde{\beta} = \frac{\sum(w_t - \bar{w})(y_t - \bar{y})}{\sum(w_t - \bar{w})(x_t - \bar{x})}$$

Answer. $\tilde{\beta}_1 = \frac{\frac{1}{n}\sum(y_i-\bar{y})(q_i-\bar{q})}{\frac{1}{n}\sum(y_i-\bar{y})(p_i-\bar{p})}$. Its plim is $\frac{\operatorname{cov}[\boldsymbol{y},\boldsymbol{q}]}{\operatorname{cov}[\boldsymbol{y},\boldsymbol{p}]} = \frac{\beta_1\alpha_2 \operatorname{var}[\boldsymbol{y}]/(\beta_1-\alpha_1)}{\alpha_2 \operatorname{var}[\boldsymbol{y}]/(\beta_1-\alpha_1)} = \beta_1$. These covariances were derived from (11) and (12). □

• **e.** *2 points Show that the Indirect Least Squares estimator of β_1 is identical to the instrumental variables estimator.*

Answer. For indirect least squares one estimates the two reduced form equations by OLS:

$$\text{the slope parameter in (11), } \frac{\alpha_2}{\beta_1-\alpha_1}\text{, estimated by } \frac{\sum(y_i-\bar{y})(p_i-\bar{p})}{\sum(y_i-\bar{y})^2};$$

$$\text{the slope parameter in (12), } \frac{\beta_1\alpha_2}{\beta_1-\alpha_1}\text{, estimated by } \frac{\sum(y_i-\bar{y})(q_i-\bar{q})}{\sum(y_i-\bar{y})^2}$$

Divide to get

$$\beta_1 \text{ estimated by } \frac{\sum(y_i-\bar{y})(q_i-\bar{q})}{\sum(y_i-\bar{y})(p_i-\bar{p})}$$

which is the same $\tilde{\beta}_1$ as in part d. □

• **f.** *1 point Since the error terms in the reduced form equations are contemporaneously correlated, wouldn't one get more precise estimates if one estimates the reduced form equations as a seemingly unrelated system, instead of OLS?*

Answer. Not as long as one does not impose any constraints on the reduced form equations, since all regressors are the same. □

• **g.** *2 points We have shown above that the regression of* q *on* p *does not give a consistent estimator of* β_1*. However one does get a consistent estimator of* β_1 *if one regresses* q *on the predicted values of* p *from the reduced form equation. (This is 2SLS.) Show that this estimator is also the same as above.*

Answer. This gives $\tilde{\beta}_1 = \frac{\sum(q_i-\bar{q})(\hat{p}_i-\bar{p})}{\sum(\hat{p}_i-\bar{p})^2}$. Now use $\hat{p}_i-\bar{p} = \hat{\pi}_1(y_i-\bar{y})$ where $\hat{\pi}_1 = \sum \frac{(p_i-\bar{p})(y_i-\bar{y})}{\sum(y_i-\bar{y})^2}$. Therefore $\tilde{\beta} = \hat{\pi}_1 \frac{\sum(q_i-\bar{q})(y_i-\bar{y})}{\hat{\pi}_1^2\sum(y_i-\bar{y})^2} = \frac{\sum(y_i-\bar{y})(q_i-\bar{q})}{\sum(y_i-\bar{y})(p_i-\bar{p})}$ again. □

• **h.** *1 point So far we have only discussed estimators of the parameters in the supply function. How would you estimate the demand function?*

Answer. You can't. The supply function can be estimated because it stays put while the demand function shifts around, therefore the observed intersection points lie on the same supply function but different demand functions. The demand function itself cannot be estimated, it is underidentified in this system. □

References

[BKW80] David A. Belsley, Edwin Kuh, and Roy E. Welsch, *Regression diagnostics: Identifying influential data and sources of collinearity*, Wiley, New York, 1980. 6

[Gre97] William H. Greene, *Econometric analysis*, third ed., Prentice Hall, Upper Saddle River, NJ, 1997. 9

[Hol86] Paul W. Holland, *Statistics and causal inference*, JASA **81** (1986), no. 396, 945–960. 3

[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*, second ed., Wiley, New York, 1988. 8

[Rao97] C. Radhakrishna Rao, *Statistics and truth: Putting chance to work*, second ed., World Scientific, Singapore, 1997. 2

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

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

URL: http://www.econ.utah.edu/ehrbar