FIRST MIDTERM STAT 6869 SPRING 2000

ECONOMICS DEPARTMENT, UNIVERSITY OF UTAH

Problem 1. *5 points Show that* $A \perp B | C$ *is equivalent to* $\Pr[A|B \cap C] = \Pr[A|C]$. *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.

Problem 2. *3 points Using the following definition of a quantile function*

$$F_y^{-1}(p) = \inf\{u : \; F_y(u) \geq p\} \tag{1}$$

prove that

$$p \leq F_y(y) \quad \textit{iff} \quad F_y^{-1}(p) \leq y \tag{2}$$

Date of exam Spring 2000.

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].

□

Problem 3. [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 For this part you will need*

$$\frac{\mathrm{H}[\mathcal{F}]}{\text{bits}} = \sum_{k=1}^{n} p_k \log_2 \frac{1}{p_k} \tag{3}$$

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* $x \neq a$*. 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 4. *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*

$$\operatorname{var}[y|x] = \operatorname{E}[y^2|x] - (\operatorname{E}[y|x])^2. \tag{4}$$

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

$$\operatorname{var}[y] = \operatorname{E}[y^2] - (\operatorname{E}[y])^2 \tag{5}$$

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*

$$\operatorname{var}[y] = \operatorname{var}\big[\operatorname{E}[y|x]\big] + \operatorname{E}\big[\operatorname{var}[y|x]\big], \tag{6}$$

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 $\mathrm{E}[(\mathrm{E}[y|x])^2] - (\mathrm{E}[\mathrm{E}[y|x]])^2$, and the second term, due to (4), becomes $\mathrm{E}[\mathrm{E}[y^2|x]] - \mathrm{E}[(\mathrm{E}[y|x])^2]$. If one adds, the two $\mathrm{E}[(\mathrm{E}[y|x])^2]$ cancel out, and the other two terms can be simplified by the law of iterated expectations to give $\mathrm{E}[y^2] - (\mathrm{E}[y])^2$. □

• **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*

$$\frac{\mathrm{E}\big[\mathrm{var}[y|x]\big]}{\mathrm{var}\big[\mathrm{E}[y|x]\big]} \tag{7}$$

can be used to determine whether the regression function $\mathrm{E}[y|x]$ *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." □

Problem 5. *2 points You want to regress a given data vector $\boldsymbol{y}$ on the matrix of regressors $\boldsymbol{X}$, and your computer gives you a parameter estimate $\hat{\boldsymbol{\beta}}$ and a vector of residuals $\hat{\boldsymbol{\varepsilon}}$. You mistrust the numerical accuracy and want to check independently whether $\hat{\boldsymbol{\beta}}$ is indeed the OLS estimator. First you verify whether $\boldsymbol{y} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\varepsilon}}$ holds, and indeed it does. Which other simple equation do you have to verify? (Hint: it only involves $\boldsymbol{X}$ and $\hat{\boldsymbol{\varepsilon}}$.)*

Answer. We know $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$, i.e., once we know $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\varepsilon}}$ is uniquely determined. Here the claim is that $\hat{\boldsymbol{\beta}}$ is the OLS estimator if and only if $\boldsymbol{X}^\top\hat{\boldsymbol{\varepsilon}} = \boldsymbol{o}$ or, written out, $\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \boldsymbol{o}$, which multiplies out to $\boldsymbol{X}^\top\boldsymbol{y} - \boldsymbol{X}^\top\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{o}$, which is exactly (**??**). □

Problem 6. *Assume $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{z}$ have a joint probability distribution, and the conditional expectation $\mathcal{E}[\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y}] = \boldsymbol{\alpha}^* + \boldsymbol{A}^*\boldsymbol{x} + \boldsymbol{B}^*\boldsymbol{y}$ is linear in $\boldsymbol{x}$ and $\boldsymbol{y}$.*

• **a.** *1 point Show that $\mathcal{E}[\boldsymbol{z}|\boldsymbol{x}] = \boldsymbol{\alpha}^* + \boldsymbol{A}^*\boldsymbol{x} + \boldsymbol{B}^*\mathcal{E}[\boldsymbol{y}|\boldsymbol{x}]$. Hint: you may use the law of iterated expectations in the following form: $\mathcal{E}[\boldsymbol{z}|\boldsymbol{x}] = \mathcal{E}\big[\mathcal{E}[\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y}]\big|\boldsymbol{x}\big]$.*

Answer. With this hint it is trivial: $\mathcal{E}[\boldsymbol{z}|\boldsymbol{x}] = \mathcal{E}\Big[\boldsymbol{\alpha}^* + \boldsymbol{A}^*\boldsymbol{x} + \boldsymbol{B}^*\boldsymbol{y}\Big|\boldsymbol{x}\Big] = \boldsymbol{\alpha}^* + \boldsymbol{A}^*\boldsymbol{x} + \boldsymbol{B}^*\mathcal{E}[\boldsymbol{y}|\boldsymbol{x}]$. □

• **b.** *1 point The next three examples are from* [CW99, pp. 264/5]*: Assume $\mathrm{E}[z|x,y] = 1 + 2x + 3y$, x and y are independent, and $\mathrm{E}[y] = 2$. Compute $\mathrm{E}[z|x]$.*

Answer. According to the formula, $\mathrm{E}[z|x] = 1 + 2x + 3\mathrm{E}[y|x]$, but since x and y are independent, $\mathrm{E}[y|x] = \mathrm{E}[y] = 2$; therefore $\mathrm{E}[z|x] = 7+2x$. I.e., the slope is the same, but the intercept changes. □

• **c.** *1 point Assume again* $\mathrm{E}[z|x,y] = 1 + 2x + 3y$, *but this time* x *and* y *are not independent but* $\mathrm{E}[y|x] = 2 - x$. *Compute* $\mathrm{E}[z|x]$.

Answer. $\mathrm{E}[z|x] = 1 + 2x + 3(2 - x) = 7 - x$. In this situation, both slope and intercept change, but it is still a linear relationship. □

• **d.** *1 point Again* $\mathrm{E}[z|x,y] = 1 + 2x + 3y$, *and this time the relationship between* x *and* y *is nonlinear:* $\mathrm{E}[y|x] = 2 - e^x$. *Compute* $\mathrm{E}[z|x]$.

Answer. $\mathrm{E}[z|x] = 1 + 2x + 3(2 - e^x) = 7 + 2x - 3e^x$. This time the marginal relationship between x and y is no longer linear. This is so despite the fact that, if all the variables are included, i.e., if both x and y are included, then the relationship is linear. □

• **e.** *1 point Assume* $\mathrm{E}[f(z)|x,y] = 1 + 2x + 3y$, *where* f *is a nonlinear function, and* $\mathrm{E}[y|x] = 2 - x$. *Compute* $\mathrm{E}[f(z)|x]$.

Answer. $\mathrm{E}[f(z)|x] = 1 + 2x + 3(2 - x) = 7 - x$. If one plots z against x and z, then the plots should be similar, though not identical, since the same transformation f will straighten them out. This is why the plots in the top row or right column of [CW99, p. 435] are so similar. □

Problem 7. *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* $\boldsymbol{y} = \hat{\hat{\boldsymbol{y}}} + \hat{\hat{\boldsymbol{\varepsilon}}}$ *in the regression of* $\boldsymbol{y} = \begin{bmatrix} 3 \\ 3 \\ 4 \end{bmatrix}$ *on* $\boldsymbol{x}_1 = \begin{bmatrix} 5 \\ 0 \\ 0 \end{bmatrix}$?

Answer. Besides the equation $\boldsymbol{y} = \hat{\hat{y}} + \hat{\hat{\varepsilon}}$ we have to check two things: (1) $\hat{\hat{y}}$ is a linear combination of all the explanatory variables (here: is a multiple of $\boldsymbol{x}_1$), and (2) $\hat{\hat{\varepsilon}}$ is orthogonal to all explanatory variables. Compare Problem 5. □

Problem 8. *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* $\boldsymbol{y} = \hat{\boldsymbol{y}} + \boldsymbol{\varepsilon}$ *in the regression of* $\boldsymbol{y} = \begin{bmatrix} 3 \\ 3 \\ 4 \end{bmatrix}$ *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 $\boldsymbol{y} = \hat{\hat{y}} + \hat{\hat{\varepsilon}}$ we have to check two things: (1) $\hat{\hat{y}}$ 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 $\hat{\boldsymbol{y}}$, which also has the third coordinate zero, is their linear combination. (2) $\hat{\hat{\varepsilon}}$ is orthogonal to both explanatory variables because its only nonzero coordinate is the third. □

Problem 9. *5 points Which inferences about the datasets can you draw from looking at the scatterplot matrix in* [BT99, Exhibit 3.2, p. 14]*?*

Answer. The discussion on [BT99, p. 19?] distinguishes three categories. First the univariate phenomena:

- yield is more concentrated for local genotypes (•) than for imports (◦);
- the converse is true for protein % but not as pronounced;
- oil % and seed size are lower for local genotypes (•); regarding seed size, the heaviest • is ligher than the lightest ◦;
- height and lodging are greater for local genotypes.

Bivariate phenomena are either within-group or between-group phenomena or both.:

- negative relationship of protein % and oil % (both within $\bullet$ and $\circ$);
- positive relationship of oil % and seed size (both within $\bullet$ and $\circ$ and also between these groups);
- negative relationship, between groups, of seed size and height;
- positive relationship of height and lodging (within $\circ$ and between groups);
- negative relationship of oil % and lodging (between groups and possibly within $\bullet$);
- negative relationship of seed size and lodging (between groups);
- positive relationship of height and lodging (between groups).

The between group pehnomena are, of course, not due to an interaction between the groups, but they are the consequence of univariate phenomena. As a third category, the authors point out unusual individual points:

- 1 high $\circ$ for yield;
- 1 high $\bullet$ (still lower than all the $\circ$s) for seed size;
- 1 low $\circ$ for lodging;
- 1 low $\bullet$ for protein % and oil % in combination.

$\square$

Problem 10. *3 points In the mussel data set, M is the "response" (according to [Coo98]). Is it justified to call this variable the "response" and the other variables the explanatory variables, and if so, how would you argue for it?*

Answer. This is one of the issues which is not sufficiently discussed in the literature. It would be justified if the dimensions and weight of the shell were exogenous to the weight of the edible part of the mussel. I.e., if the mussel first grows the shell, and then it fills this shell wish muscle, and the dimensions of the shell affect how big the muscle can grow, but the muscle itself does not have an influence on the dimensions of the shell. If this is the case, then it makes sense to look at the distribution of M conditionally on the other variables, i.e., ask the question: given certain weights and dimensions of the shell, what is the nature of the mechanism by which the muscle grows inside this shell. But if muscle and shell grow together, both affected by the same variables (temperature, nutrition, daylight, etc.), then the conditional distribution is not informative. In this case, the joint distribution is of interest. □

Problem 11. *2 points Why can the scatter plot of the dependent variable against one of the independent variables be so misleading?*

Answer. Because the included independent variable becomes a proxy for the excluded variable. The effect of the excluded variable is mistaken to come from the included variable. Now if the included and the excluded variable are independent of each other, then the omission of the excluded variable increases the noise, but does not have a systematic effect. But if there is an empirical relationship between the included and the excluded variable, then this translates into a spurious relationship between included and dependent variables. The mathematics of this is discussed in Problem 6. □

Problem 12. *If $u \mapsto k(u)$ is the kernel, and $\boldsymbol{x} = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}^\top$ the data vector, then $\hat{f}(u) = \frac{1}{n}\sum_{i=1}^{n} k(u - x_i)$ is the kernel estimate of the density at u.*

• **a.** *3 points Compute the mean of the kernel estimator at u.*

Answer. $\mathrm{E}[\hat{f}(u)] = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}[k(u - x_i)]$ but since all x_i are assumed to come from the same distribution, it follows $\mathrm{E}[\hat{f}(u)] = \mathrm{E}[k(u - x)] = \int_{x=-\infty}^{+\infty} k(u - x) f(x)\, dx$. □

- **b.** *4 points Assuming the x_i are independent, show that*

$$\operatorname{var}[\hat{f}(u)] = \frac{1}{n}\Big(\int_{x=-\infty}^{+\infty} k^2(u-x)f(x)\,dx - \big(\int_{x=-\infty}^{+\infty} k(u-x)f(x)\,dx\big)^2\Big). \tag{8}$$

Answer.

$$\operatorname{var}[\hat{f}(u)] = \frac{1}{n^2}\sum_{i=1}^{n}\operatorname{var}[k(u-x_i)] \tag{9}$$

$$= \frac{1}{n}\operatorname{var}[k(u-x)] \tag{10}$$

$$= \frac{1}{n}\Big(\mathrm{E}\Big[\big(k(u-x)\big)^2\Big] - \big(\mathrm{E}[k(u-x)]\big)^2\Big) \tag{11}$$

$$= \frac{1}{n}\Big(\int_{x=-\infty}^{+\infty} k^2(u-x)f(x)\,dx - \big(\int_{x=-\infty}^{+\infty} k(u-x)f(x)\,dx\big)^2\Big). \tag{12}$$

□

Maximum number of points: 57.

References

[Ame94] Takeshi Amemiya, *Introduction to statistics and econometrics*, Harvard University Press, Cambridge, MA, 1994. 4

[BT99] Kaye E. Basford and J. W. Tukey, *Graphical analysis of multiresponse data: Illustrated with a plant breeding trial*, Interdisciplinary Statistics, Chapman & Hall/CRC, Boca Raton, Fla., 1999. 7

[Coo98] R. Dennis Cook, *Regression graphics: Ideas for studying regressions through graphics*, Series in Probability and Statistics, Wiley, New York, 1998. 4, 8

[CT91] Thomas M. Cover and Joy A. Thomas, *Elements of information theory*, Series in Telecommunications, Wiley, New York, 1991. 2

[CW99] R. Dennis Cook and Sanford Weisberg, *Applied regression including computing and graphics*, Wiley, 1999. 5, 6

[Rao73] C. Radhakrishna Rao, *Linear statistical inference and its applications*, second ed., Wiley, New York, 1973. 4

[Rei89] Rolf-Dieter Reiss, *Approximate distributions of order statistics*, Springer-Verlag, New York, 1989. 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`