SECOND MIDTERM ECON 7800 FALL 2003

ECONOMICS DEPARTMENT, UNIVERSITY OF UTAH

Problem 21. *2 points The conditional density is the joint divided by the marginal:*

$$f_{y|x}(y,x) = \frac{f_{x,y}(x,y)}{f_x(x)}. \tag{1}$$

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,

$$\int_{y=-\infty}^{+\infty} f_{y|x=x}(y,x)\,dy = \frac{\int_{y=-\infty}^{+\infty} f_{x,y}(x,y)\,dy}{f_x(x)} = \frac{f_x(x)}{f_x(x)} = 1 \tag{2}$$

Date of exam Thursday, October 16, 2003, 12:25–1:45 pm in BUC 207.

because of the formula for the marginal:

$$f_x(x) = \int_{y=-\infty}^{+\infty} f_{x,y}(x,y)\,dy \tag{3}$$

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

Problem 22. *Assume the vector* $\boldsymbol{x} = [x_1, \dots x_j]^\top$ *and the scalar* y *are jointly distributed random variables, and assume conditional means exist. Define* $\varepsilon = y - \mathrm{E}[y|\boldsymbol{x}]$.

For this problem you may use (1) the theorem of iterated expectations $\mathrm{E}\big[\mathrm{E}[y|\boldsymbol{x}]\big] = \mathrm{E}[y]$, *(2) the additivity* $\mathrm{E}[g(y)+h(y)|\boldsymbol{x}] = \mathrm{E}[g(y)|\boldsymbol{x}] + \mathrm{E}[h(y)|\boldsymbol{x}]$, *and (3) the fact that* $\mathrm{E}[g(\boldsymbol{x})h(y)|\boldsymbol{x}] = g(\boldsymbol{x})\mathrm{E}[h(y)|\boldsymbol{x}]$.

- **a.** *5 points Demonstrate the following identities:*

$$\mathrm{E}[\varepsilon|\boldsymbol{x}] = 0 \tag{4}$$

$$\mathrm{E}[\varepsilon] = 0 \tag{5}$$

$$\mathrm{E}[x_i\varepsilon|\boldsymbol{x}] = 0 \quad \textit{for all } i,\ 1 \le i \le j \tag{6}$$

$$\mathrm{E}[x_i\varepsilon] = 0 \quad \textit{for all } i,\ 1 \le i \le j \tag{7}$$

$$\mathrm{cov}[x_i, \varepsilon] = 0 \quad \textit{for all } i,\ 1 \le i \le j. \tag{8}$$

Interpretation of (8)*:* ε *is the error in the best prediction of* y *based on* $\boldsymbol{x}$*. 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. (4): $\mathrm{E}[\varepsilon|\boldsymbol{x}] = \mathrm{E}[y|\boldsymbol{x}] - \mathrm{E}\Big[\mathrm{E}[y|\boldsymbol{x}]|\boldsymbol{x}\Big] = 0$ since $\mathrm{E}[y|\boldsymbol{x}]$ is a function of $\boldsymbol{x}$ and therefore equal to its own expectation conditionally on $\boldsymbol{x}$. (This is *not* the law of iterated expectations but the law that the expected value of a constant is a constant.)

(5) follows from (4) (i.e., (4) is stronger than (5)): if an expectation is zero conditionally on every possible outcome of $\boldsymbol{x}$ then it is zero altogether. In formulas, $\mathrm{E}[\varepsilon] = \mathrm{E}\Big[\mathrm{E}[\varepsilon|\boldsymbol{x}]\Big] = \mathrm{E}[0] = 0$. It is also easy to show it in one swoop, without using (4): $\mathrm{E}[\varepsilon] = \mathrm{E}[y - \mathrm{E}[y|\boldsymbol{x}]] = 0$. Either way you need the law of iterated expectations for this.

(6): $\mathrm{E}[x_i\varepsilon|\boldsymbol{x}] = x_i\mathrm{E}[\varepsilon|\boldsymbol{x}] = 0$.

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

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

• **b.** *2 points If* $\boldsymbol{x}$ *and* y *are jointly normal, show that* $\boldsymbol{x}$ *and* ε *are independent, and that the variance of* ε *does not depend on* $\boldsymbol{x}$*. (This is why one can consider it an error term.)*

Answer. If $\boldsymbol{x}$ and y are jointly normal, then $\boldsymbol{x}$ 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., $\mathrm{E}[\varepsilon^2] = \mathrm{E}\left[\mathrm{E}[\varepsilon^2|\boldsymbol{x}]\right]$ = constant (does not depend on $\boldsymbol{x}$). □

Problem 23. *4 points Let $y_1, \dots, y_n$ be an arbitrary vector and α an arbitrary number. As usual, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. Show that*

$$\sum_{i=1}^{n}(y_i - \alpha)^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2 + n(\bar{y} - \alpha)^2 \tag{9}$$

Answer.

$$\sum_{i=1}^{n}(y_i - \alpha)^2 = \sum_{i=1}^{n}\Big((y_i - \bar{y}) + (\bar{y} - \alpha)\Big)^2 \tag{10}$$

$$= \sum_{i=1}^{n}(y_i - \bar{y})^2 + 2\sum_{i=1}^{n}\Big((y_i - \bar{y})(\bar{y} - \alpha)\Big) + \sum_{i=1}^{n}(\bar{y} - \alpha)^2 \tag{11}$$

$$= \sum_{i=1}^{n}(y_i - \bar{y})^2 + 2(\bar{y} - \alpha)\sum_{i=1}^{n}(y_i - \bar{y}) + n(\bar{y} - \alpha)^2 \tag{12}$$

Since the middle term is zero, (9) follows.

□

Problem 24. *4 points Show that*

$$s_u^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{13}$$

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

$$\mathrm{E}[\sum_{i=1}^{n}(y_i - \bar{y})^2] = \sum_{i=1}^{n}\mathrm{E}[(y_i - \mu)^2] - n\,\mathrm{E}[(\bar{y} - \mu)^2] \tag{14}$$

$$= \sum_{i=1}^{n}\sigma^2 - n\frac{\sigma^2}{n} = (n-1)\sigma^2. \tag{15}$$

You do not need Normality for this. □

Problem 25.

- **a.** *2 points Verify that the matrix $\boldsymbol{D} = \boldsymbol{I} - \frac{1}{n}\boldsymbol{\iota}\boldsymbol{\iota}^\top$ is symmetric and idempotent.*
- **b.** *1 point Compute the trace* $\operatorname{tr}\boldsymbol{D}$.

Answer. $\operatorname{tr} \boldsymbol{D} = n - 1$. One can see this either by writing down the matrix element by element, or use the linearity of the trace plus the rule that $\operatorname{tr}(\boldsymbol{AB}) = \operatorname{tr}(\boldsymbol{BA})$. $\operatorname{tr} \boldsymbol{I} = n$ and $\operatorname{tr}(\boldsymbol{\iota}\boldsymbol{\iota}^\top) = \operatorname{tr}(\boldsymbol{\iota}^\top\boldsymbol{\iota}) = \operatorname{tr} n = n$. □

• **c.** *1 point For any vector of observations* $\boldsymbol{y}$ *compute* $\boldsymbol{Dy}$.

Answer. Element by element one can write

$$\boldsymbol{Dy} = \begin{bmatrix} 1-\frac{1}{n} & -\frac{1}{n} & \dots & -\frac{1}{n} \\ -\frac{1}{n} & 1-\frac{1}{n} & \dots & -\frac{1}{n} \\ \vdots & & \ddots & \vdots \\ -\frac{1}{n} & -\frac{1}{n} & \dots & 1-\frac{1}{n} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{bmatrix} \tag{16}$$

There is also a more elegant matrix theoretical proof available □

• **d.** *1 point Is there a vector* $\boldsymbol{a} \neq \boldsymbol{o}$ *for which* $\boldsymbol{Da} = \boldsymbol{o}$*? If so, give an example of such a vector.*

Answer. $\boldsymbol{\iota}$ is, up to a scalar factor, the only nonzero vector with $\boldsymbol{D\iota} = \boldsymbol{o}$. □

• **e.** *1 point Show that the sample variance of a vector of observations* $\boldsymbol{y}$ *can be written in matrix notation as*

$$\textit{The sample variance of } \boldsymbol{y} \textit{ is } \frac{1}{n}\sum (y_i - \bar{y})^2 = \frac{1}{n}\boldsymbol{y}^\top \boldsymbol{Dy} \tag{17}$$

Answer. Let's get rid of the factor $\frac{1}{n}$ which appears on both sides: we have to show that

$$\sum (y_i - \bar{y})^2 = \boldsymbol{y}^\top \boldsymbol{D}\boldsymbol{y} = \boldsymbol{y}^\top \boldsymbol{D}^\top \boldsymbol{D}\boldsymbol{y} \tag{18}$$

This is the squared length of the vector $\boldsymbol{D}\boldsymbol{y}$ which we computed in part c. □

Problem 26. [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. $\mathrm{E}[t-\theta]$ is called the bias of t, and $\mathrm{E}\big[(t-\theta)^2\big]$ is called the mean squared error of t as an estimator of θ, written $\mathrm{MSE}[t;\theta]$. Show that the MSE is the variance plus the squared bias, i.e., that*

$$\mathrm{MSE}[t;\theta] = \mathrm{var}[t] + \big(\mathrm{E}[t-\theta]\big)^2. \tag{19}$$

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

$$\mathrm{MSE}[t;\theta] = \mathrm{E}\big[(t-\theta)^2\big] = \mathrm{var}[t-\theta] + (\mathrm{E}[t-\theta])^2 = \mathrm{var}[t] + (\mathrm{E}[t-\theta])^2. \tag{20}$$

□

- **b.** *2 points For the rest of this problem assume that t is an unbiased estimator of θ with $\mathrm{var}[t] > 0$. We will investigate whether one can get a better MSE if one*

estimates θ by a constant multiple at instead of t. Show that

$$\text{MSE}[at;\theta] = a^2 \operatorname{var}[t] + (a-1)^2\theta^2. \tag{21}$$

Answer. $\operatorname{var}[at] = a^2 \operatorname{var}[t]$ and the bias of at is $\text{E}[at - \theta] = (a-1)\theta$. Now apply (20). □

• **c.** *1 point Show that, whenever $a > 1$, then $\text{MSE}[at;\theta] > \text{MSE}[t;\theta]$. If one wants to decrease the* MSE*, one should therefore not choose $a > 1$.*

Answer. $\text{MSE}[at;\theta] - \text{MSE}[t;\theta] = (a^2-1)\operatorname{var}[t] + (a-1)^2\theta^2 > 0$ since $a > 1$ and $\operatorname{var}[t] > 0$. □

• **d.** *2 points Show that*

$$\frac{d}{da}\text{MSE}[at;\theta]\bigg|_{a=1} > 0. \tag{22}$$

From this follows that the MSE *of at is smaller than the* MSE *of t, as long as $a < 1$ and close enough to 1.*

Answer. The derivative of (21) is

$$\frac{d}{da}\text{MSE}[at;\theta] = 2a\operatorname{var}[t] + 2(a-1)\theta^2 \tag{23}$$

Plug $a = 1$ into this to get $2\operatorname{var}[t] > 0$. □

• **e.** *2 points By solving the first order condition show that the factor a which gives smallest* MSE *is*

$$a = \frac{\theta^2}{\operatorname{var}[t] + \theta^2}. \tag{24}$$

Answer. Rewrite (23) as $2a(\operatorname{var}[t] + \theta^2) - 2\theta^2$ and set it zero. □

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

$$f_t(t) = \lambda \exp(-\lambda t), \qquad t \geq 0 \quad \textit{and} \quad f_t(t) = 0 \quad \textit{otherwise.} \tag{25}$$

Check that $f_t(t)$ is indeed a density function.

Answer. Since $\lambda > 0$, $f_t(t) > 0$ 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* $\operatorname{var}[t] = 1/\lambda^2$.

Answer. To evaluate $\int_0^\infty \lambda t \exp(-\lambda t)\, dt$, use partial integration $\int uv'\, dt = uv - \int u'v\, dt$ with $u = t$, $u' = 1$, $v = -\exp(-\lambda t)$, $v' = \lambda \exp(-\lambda t)$. Therefore the integral is $-t \exp(-\lambda t)\Big|_0^\infty + \int_0^\infty \exp(-\lambda t)\, dt = 1/\lambda$, since we just saw that $\int_0^\infty \lambda \exp(-\lambda t)\, dt = 1$.

To evaluate $\int_0^\infty \lambda t^2 \exp(-\lambda t)\, dt$, use partial integration with $u = t^2$, $u' = 2t$, $v = -\exp(-\lambda t)$, $v' = \lambda \exp(-\lambda t)$. Therefore the integral is $-t^2 \exp(-\lambda t)\Big|_0^\infty + 2\int_0^\infty t \exp(-\lambda t)\, dt = \frac{2}{\lambda}\int_0^\infty \lambda t \exp(-\lambda t)\, dt = 2/\lambda^2$. Therefore $\text{var}[t] = \text{E}[t^2] - (\text{E}[t])^2 = 2/\lambda^2 - 1/\lambda^2 = 1/\lambda^2$. □

• **h.** *2 points Which multiple of t has the lowest* MSE *as an estimator of $1/\lambda$?*

Answer. It is $t/2$. Just plug $\theta = 1/\lambda$ into (24).

$$a = \frac{1/\lambda^2}{\text{var}[t] + 1/\lambda^2} = \frac{1/\lambda^2}{1/\lambda^2 + 1/\lambda^2} = \frac{1}{2}. \tag{26}$$

□

• **i.** *2 points Assume $t_1, \ldots, t_n$ are independently distributed, and each of them has the exponential distribution with the same parameter λ. Which multiple of the sample mean $\bar{t} = \frac{1}{n}\sum_{i=1}^n t_i$ has best* MSE *as estimator of $1/\lambda$?*

Answer. $\bar{t}$ has expected value $1/\lambda$ and variance $1/n\lambda^2$. Therefore

$$a = \frac{1/\lambda^2}{\text{var}[t] + 1/\lambda^2} = \frac{1/\lambda^2}{1/n\lambda^2 + 1/\lambda^2} = \frac{n}{n+1}, \tag{27}$$

i.e., for the best estimator $\tilde{t} = \frac{1}{n+1}\sum t_i$ divide the sum by $n+1$ instead of n. □

• **j.** *3 points Assume $q \sim \sigma^2 \chi^2_m$ (in other words, $\frac{1}{\sigma^2} q \sim \chi^2_m$, 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. $\mathrm{E}[q] = m\sigma^2$, therefore q/m is the unbiased estimator. Since $\mathrm{var}[q/m] = 2\sigma^4/m$, it follows from (24) that $a = m/(m+2)$, therefore the minimum MSE multiple of q is $\frac{q}{m}\frac{m}{m+2} = \frac{q}{m+2}$. 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 $\frac{1}{n-1}\sum(y_i - \bar{y})^2$, and the maximum likelihood extimator is $\frac{1}{n}\sum(y_i - \bar{y})^2$. 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 $\sum(y_i - \bar{y})^2$?*

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

• **l.** *3 points We are still in the model defined in* k. *Which multiple of the sample mean $\bar{y}$ 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 $\bar{y}$. Indeed, $\bar{y}$ 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. □

Problem 27. *4 points Assume n independent observations of a variable $y \sim N(\mu, \sigma^2)$ are available, where σ^2 is known. Show that the sample mean $\bar{y}$ attains the Cramer-Rao lower bound for μ.*

Answer. The density function of each y_i is

$$f_{y_i}(y) = (2\pi\sigma^2)^{-1/2} \exp\Big(-\frac{(y-\mu)^2}{2\sigma^2}\Big) \tag{28}$$

therefore the log likelihood function of the whole vector is

$$\ell(y;\mu) = \sum_{i=1}^{n} \log f_{y_i}(y_i) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2 \tag{29}$$

$$\frac{\partial}{\partial\mu}\ell(y;\mu) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\mu) \tag{30}$$

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

$$\mathrm{E}[\Big(\frac{\partial}{\partial\mu}\ell(y;\mu)\Big)^2] = \frac{1}{\sigma^4}\sum \mathrm{E}[(y_i-\mu)^2] = n/\sigma^2 \tag{31}$$

alternatively one may take one more derivative from (30) to get

$$\frac{\partial^2}{\partial \mu^2}\ell(y;\mu) = -\frac{n}{\sigma^2} \tag{32}$$

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

Problem 28. *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 (??) $\binom{12}{9}\theta^9(1-\theta)^3$, and in the second it is by (??) $\binom{11}{8}\theta^9(1-\theta)^3$. They are proportional, the stopping rule therefore does not matter! □

Problem 29. *Mr. Jones is on trial for counterfeiting Picasso paintings, and you are an expert witness who has developed fool-proof statistical significance tests for identifying the painter of a given painting.*

- **a.** *2 points There are two ways you can set up your test.*

a: *You can either say: The null hypothesis is that the painting was done by Picasso, and the alternative hypothesis that it was done by Mr. Jones.*

b: *Alternatively, you might say: The null hypothesis is that the painting was done by Mr. Jones, and the alternative hypothesis that it was done by Picasso.*

Does it matter which way you do the test, and if so, which way is the correct one. Give a reason to your answer, i.e., say what would be the consequences of testing in the incorrect way.

Answer. The determination of what the null and what the alternative hypothesis is depends on what is considered to be the catastrophic error which is to be guarded against. On a trial, Mr. Jones is considered innocent until proven guilty. Mr. Jones should not be convicted unless he can be proven guilty beyond "reasonable doubt." Therefore the test must be set up in such a way that the hypothesis that the painting is by Picasso will only be rejected if the chance that it is actually by Picasso is very small. The error of type one is that the painting is considered counterfeited although it is really by Picasso. Since the error of type one is always the error to reject the null hypothesis although it is true, solution a. is the correct one. You are not proving, you are testing. □

• **b.** *2 points After the trial a customer calls you who is in the process of acquiring a very expensive alleged Picasso painting, and who wants to be sure that this painting is not one of Jones's falsifications. Would you now set up your test in the same way as in the trial or in the opposite way?*

Answer. It is worse to spend money on a counterfeit painting than to forego purchasing a true Picasso. Therefore the null hypothesis would be that the painting was done by Mr. Jones, i.e., it is the opposite way. □

Problem 30. *1 point Compute the matrix product*

$$\begin{bmatrix} 1 & 2 & 4 \\ 0 & 3 & 3 \end{bmatrix} \begin{bmatrix} 4 & 0 \\ 2 & 1 \\ 3 & 8 \end{bmatrix}$$

Answer.

$$\begin{bmatrix} 1 & 2 & 4 \\ 0 & 3 & 3 \end{bmatrix} \begin{bmatrix} 4 & 0 \\ 2 & 1 \\ 3 & 8 \end{bmatrix} = \begin{bmatrix} 1\cdot 4+2\cdot 2+4\cdot 3 & 1\cdot 0+2\cdot 1+4\cdot 8 \\ 0\cdot 4+3\cdot 2+3\cdot 3 & 0\cdot 0+3\cdot 1+3\cdot 8 \end{bmatrix} = \begin{bmatrix} 20 & 34 \\ 15 & 27 \end{bmatrix}$$

□

Problem 31. *4 points Using the matrix differentiation rules*

$$\partial \boldsymbol{w}^\top \boldsymbol{x}/\partial \boldsymbol{x}^\top = \boldsymbol{w}^\top \tag{33}$$

$$\partial \boldsymbol{x}^\top \boldsymbol{M} \boldsymbol{x}/\partial \boldsymbol{x}^\top = 2\boldsymbol{x}^\top \boldsymbol{M} \tag{34}$$

for symmetric $\boldsymbol{M}$*, compute the least-squares estimate* $\hat{\boldsymbol{\beta}}$ *which minimizes*

$$SSE = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \tag{35}$$

You are allowed to assume that $\boldsymbol{X}^\top \boldsymbol{X}$ has an inverse.

Answer. First you have to multiply out

$$(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{y}^\top \boldsymbol{y} - 2\boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta}. \tag{36}$$

The matrix differentiation rules (33) and (34) allow us to differentiate (36) to get

$$\partial SSE/\partial \boldsymbol{\beta}^\top = -2\boldsymbol{y}^\top \boldsymbol{X} + 2\boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{X}. \tag{37}$$

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"

$$\boldsymbol{X}^\top \boldsymbol{y} = \boldsymbol{X}^\top \boldsymbol{X}\hat{\boldsymbol{\beta}}. \tag{38}$$

Since $\boldsymbol{X}^\top \boldsymbol{X}$ has an inverse, one can premultiply both sides of (38) by $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}. \tag{39}$$

□

Problem 32. *1 point In the decomposition*

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \hat{\varepsilon} \tag{40}$$

which of $\boldsymbol{\varepsilon}$ and $\hat{\varepsilon}$ is called the residual and which is called the disturbance? Which of $\boldsymbol{y}$, $\boldsymbol{X}$, $\boldsymbol{\beta}$, $\boldsymbol{\varepsilon}$, $\hat{\boldsymbol{\beta}}$, and $\hat{\varepsilon}$ is observed (or a function of observed quantities) and which is unobserved?

Answer. $\boldsymbol{\varepsilon}$ is called the disturbance and $\hat{\varepsilon}$ the residual. $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$ are unobserved, all others are observed. □

Problem 33. *2 points Assume that $\boldsymbol{X}$ has full column rank. Show that $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{M}\boldsymbol{y}$ where $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$. Show that $\boldsymbol{M}$ is symmetric and idempotent.*

Answer. By definition, $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{y} = \big(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}\big)\boldsymbol{y}$. Idempotent, i.e. $\boldsymbol{M}\boldsymbol{M} = \boldsymbol{M}$:

(41)
$$\boldsymbol{M}\boldsymbol{M} = \big(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\big)\big(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\big) \quad = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$$

□

Problem 34. *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 35. *The fitted values $\hat{\boldsymbol{y}}$ and the residuals $\hat{\boldsymbol{\varepsilon}}$ are "orthogonal" in two different ways.*

• **a.** *2 points Show that the inner product $\hat{\boldsymbol{y}}^\top \hat{\boldsymbol{\varepsilon}} = 0$. Why should you expect this from the geometric intuition of Least Squares?*

Answer. Use $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{M}\boldsymbol{y}$ and $\hat{\boldsymbol{y}} = (\boldsymbol{I} - \boldsymbol{M})\boldsymbol{y}$: $\hat{\boldsymbol{y}}^\top \hat{\boldsymbol{\varepsilon}} = \boldsymbol{y}^\top (\boldsymbol{I} - \boldsymbol{M})\boldsymbol{M}\boldsymbol{y} = 0$ 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 $\hat{\boldsymbol{y}}$ and $\hat{\boldsymbol{\varepsilon}}$ are orthogonal also in this sense, i.e., show that for every i and j, $\operatorname{cov}[\hat{y}_i, \hat{\varepsilon}_j] = 0$. In matrix notation this can also be written $\mathcal{C}[\hat{\boldsymbol{y}}, \hat{\boldsymbol{\varepsilon}}] = \boldsymbol{O}$. Here the covariance matrix $\mathcal{C}[\boldsymbol{x}, \boldsymbol{z}]$ is that matrix whose (i, j) element is $\operatorname{cov}[x_i, z_j]$. The covariance matrix satisfies the rules $\mathcal{C}[\boldsymbol{B}\boldsymbol{y}, \boldsymbol{T}\boldsymbol{z}] = \boldsymbol{B}\,\mathcal{C}[\boldsymbol{y}, \boldsymbol{z}]\boldsymbol{T}^\top$, $\mathcal{C}[\boldsymbol{y}, \boldsymbol{y}] = \mathcal{V}[\boldsymbol{y}]$, $\mathcal{C}[\boldsymbol{z}, \boldsymbol{y}] = (\mathcal{C}[\boldsymbol{y}, \boldsymbol{z}])^\top$, $\mathcal{C}[\boldsymbol{x} + \boldsymbol{y}, \boldsymbol{z}] = \mathcal{C}[\boldsymbol{x}, \boldsymbol{z}] + \mathcal{C}[\boldsymbol{y}, \boldsymbol{z}]$, and $\mathcal{C}[\boldsymbol{x}, \boldsymbol{c}] = \boldsymbol{O}$ if $\boldsymbol{c}$ is a vector of constants.*

Answer. $\mathcal{C}[\hat{\boldsymbol{y}}, \hat{\boldsymbol{\varepsilon}}] = \mathcal{C}[(\boldsymbol{I} - \boldsymbol{M})\boldsymbol{y}, \boldsymbol{M}\boldsymbol{y}] = (\boldsymbol{I} - \boldsymbol{M})\,\mathcal{V}[\boldsymbol{y}]\boldsymbol{M}^\top = (\boldsymbol{I} - \boldsymbol{M})(\sigma^2 \boldsymbol{I})\boldsymbol{M} = \sigma^2(\boldsymbol{I} - \boldsymbol{M})\boldsymbol{M} = \boldsymbol{O}$. This is a consequence of the more general result given in question **??**. □

-

Maximum number of points: 75.

References

[KS79] Sir Maurice Kendall and Alan Stuart, *The advanced theory of statistics*, fourth ed., vol. 2, Griffin, London, 1979. 7

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