CHAPTER 1

TEACHING NOTES

You have substantial latitude about what to emphasize in Chapter 1. I find it useful to talk about the economics of crime example (Example 1.1) and the wage example (Example 1.2) so that students see, at the outset, that econometrics is linked to economic reasoning, even if the economics is not complicated theory.

I like to familiarize students with the important data structures that empirical economists use, focusing primarily on cross-sectional and time series data sets, as these are what I cover in a first-semester course. It is probably a good idea to mention the growing importance of data sets that have both a cross-sectional and time dimension.

I spend almost an entire lecture talking about the problems inherent in drawing causal inferences in the social sciences. I do this mostly through the agricultural yield, return to education, and crime examples. These examples also contrast experimental and nonexperimental (observational) data. Students studying business and finance tend to find the term structure of interest rates example more relevant, although the issue there is testing the implication of a simple theory, as opposed to inferring causality. I have found that spending time talking about these examples, in place of a formal review of probability and statistics, is more successful (and more enjoyable for the students and me).

SOLUTIONS TO PROBLEMS

1.1 It does not make sense to pose the question in terms of causality. Economists would assume that students choose a mix of studying and working (and other activities, such as attending class, leisure, and sleeping) based on rational behavior, such as maximizing utility subject to the constraint that there are only 168 hours in a week. We can then use statistical methods to measure the association between studying and working, including regression analysis that we cover starting in Chapter 2. But we would not be claiming that one variable "causes" the other. They are both choice variables of the student.

1.2 (i) Ideally, we could randomly assign students to classes of different sizes. That is, each student is assigned a different class size without regard to any student characteristics such as ability and family background. For reasons we will see in Chapter 2, we would like substantial variation in class sizes (subject, of course, to ethical considerations and resource constraints).

(ii) A negative correlation means that larger class size is associated with lower performance. We might find a negative correlation because larger class size actually hurts performance. However, with observational data, there are other reasons we might find a negative relationship. For example, children from more affluent families might be more likely to attend schools with smaller class sizes, and affluent children generally score better on standardized tests. Another possibility is that, within a school, a principal might assign the better students to smaller classes. Or, some parents might insist their children are in the smaller classes, and these same parents tend to be more involved in their children's education.

(iii) Given the potential for confounding factors – some of which are listed in (ii) – finding a negative correlation would not be strong evidence that smaller class sizes actually lead to better performance. Some way of controlling for the confounding factors is needed, and this is the subject of multiple regression analysis.

1.3 (i) Here is one way to pose the question: If two firms, say *A* and *B*, are identical in all respects except that firm *A* supplies job training one hour per worker more than firm *B*, by how much would firm *A*'s output differ from firm *B*'s?

(ii) Firms are likely to choose job training depending on the characteristics of workers. Some observed characteristics are years of schooling, years in the workforce, and experience in a particular job. Firms might even discriminate based on age, gender, or race. Perhaps firms choose to offer training to more or less able workers, where "ability" might be difficult to quantify but where a manager has some idea about the relative abilities of different employees. Moreover, different kinds of workers might be attracted to firms that offer more job training on average, and this might not be evident to employers.

(iii) The amount of capital and technology available to workers would also affect output. So, two firms with exactly the same kinds of employees would generally have different outputs if they use different amounts of capital or technology. The quality of managers would also have an effect.

(iv) No, unless the amount of training is randomly assigned. The many factors listed in parts (ii) and (iii) can contribute to finding a positive correlation between *output* and *training* even if job training does not improve worker productivity.

SOLUTIONS TO COMPUTER EXERCISES

C1.1 (i) The average of *educ* is about 12.6 years. There are two people reporting zero years of education, and 19 people reporting 18 years of education.

(ii) The average of *wage* is about \$5.90, which seems low in the year 2008.

(iii) Using Table B-60 in the 2004 *Economic Report of the President*, the CPI was 56.9 in 1976 and 184.0 in 2003.

(iv) To convert 1976 dollars into 2003 dollars, we use the ratio of the CPIs, which is $184/56.9 \approx 3.23$. Therefore, the average hourly wage in 2003 dollars is roughly $3.23(\$5.90) \approx \$19.06$, which is a reasonable figure.

(v) The sample contains 252 women (the number of observations with *female* = 1) and 274 men.

C1.2 (i) There are 1,388 observations in the sample. Tabulating the variable *cigs* shows that 212 women have *cigs* > 0.

(ii) The average of *cigs* is about 2.09, but this includes the 1,176 women who did not smoke. Reporting just the average masks the fact that almost 85 percent of the women did not smoke. It makes more sense to say that the "typical" woman does not smoke during pregnancy; indeed, the median number of cigarettes smoked is zero.

(iii) The average of *cigs* over the women with *cigs* > 0 is about 13.7. Of course this is much higher than the average over the entire sample because we are excluding 1,176 zeros.

(iv) The average of *fatheduc* is about 13.2. There are 196 observations with a missing value for *fatheduc*, and those observations are necessarily excluded in computing the average.

(v) The average and standard deviation of *faminc* are about 29.027 and 18.739, respectively, but *faminc* is measured in thousands of dollars. So, in dollars, the average and standard deviation are \$29,027 and \$18,739.

C1.3 (i) The largest is 100, the smallest is 0.

(ii) 38 out of 1,823, or about 2.1 percent of the sample.

(iii) 17

(iv) The average of *math4* is about 71.9 and the average of *read4* is about 60.1. So, at least in 2001, the reading test was harder to pass.

(v) The sample correlation between *math4* and *read4* is about .843, which is a very high degree of (linear) association. Not surprisingly, schools that have high pass rates on one test have a strong tendency to have high pass rates on the other test.

(vi) The average of *exppp* is about $5,194.87. The standard deviation is $1,091.89, which shows rather wide variation in spending per pupil. [The minimum is $1,206.88 and the maximum is $11,957.64.]

C1.4 (i) 185/445 ≈ .416 is the fraction of men receiving job training, or about 41.6%.

(ii) For men receiving job training, the average of *re78* is about 6.35, or $6,350. For men not receiving job training, the average of *re78* is about 4.55, or $4,550. The difference is $1,800, which is very large. On average, the men receiving the job training had earnings about 40% higher than those not receiving training.

(iii) About 24.3% of the men who received training were unemployed in 1978; the figure is 35.4% for men not receiving training. This, too, is a big difference.

(iv) The differences in earnings and unemployment rates suggest the training program had strong, positive effects. Our conclusions about economic significance would be stronger if we could also establish statistical significance (which is done in Computer Exercise C9.10 in Chapter 9).

CHAPTER 2

TEACHING NOTES

This is the chapter where I expect students to follow most, if not all, of the algebraic derivations. In class I like to derive at least the unbiasedness of the OLS slope coefficient, and usually I derive the variance. At a minimum, I talk about the factors affecting the variance. To simplify the notation, after I emphasize the assumptions in the population model, and assume random sampling, I just condition on the values of the explanatory variables in the sample. Technically, this is justified by random sampling because, for example, $E(u_i|x_1,x_2,\ldots,x_n) = E(u_i|x_i)$ by independent sampling. I find that students are able to focus on the key assumption SLR.4 and subsequently take my word about how conditioning on the independent variables in the sample is harmless. (If you prefer, the appendix to Chapter 3 does the conditioning argument carefully.) Because statistical inference is no more difficult in multiple regression than in simple regression, I postpone inference until Chapter 4. (This reduces redundancy and allows you to focus on the interpretive differences between simple and multiple regression.)

You might notice how, compared with most other texts, I use relatively few assumptions to derive the unbiasedness of the OLS slope estimator, followed by the formula for its variance. This is because I do not introduce redundant or unnecessary assumptions. For example, once SLR.4 is assumed, nothing further about the relationship between u and x is needed to obtain the unbiasedness of OLS under random sampling.

SOLUTIONS TO PROBLEMS

2.1 In the equation $y = \beta_0 + \beta_1 x + u$, add and subtract α_0 from the right hand side to get $y = (\alpha_0 + \beta_0) + \beta_1 x + (u - \alpha_0)$. Call the new error $e = u - \alpha_0$, so that $E(e) = 0$. The new intercept is $\alpha_0 + \beta_0$, but the slope is still β_1.

2.2 (i) Let $y_i = GPA_i$, $x_i = ACT_i$, and $n = 8$. Then $\bar{x} = 25.875$, $\bar{y} = 3.2125$, $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = 5.8125$, and $\sum_{i=1}^{n}(x_i - \bar{x})^2 = 56.875$. From equation (2.9), we obtain the slope as $\hat{\beta}_1 = 5.8125/56.875 \approx .1022$, rounded to four places after the decimal. From (2.17), $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} \approx 3.2125 - (.1022)25.875 \approx .5681$. So we can write

$$\widehat{GPA} = .5681 + .1022\, ACT$$
$$n = 8.$$

The intercept does not have a useful interpretation because *ACT* is not close to zero for the population of interest. If *ACT* is 5 points higher, $\widehat{GPA}$ increases by .1022(5) = .511.

(ii) The fitted values and residuals — rounded to four decimal places — are given along with the observation number *i* and *GPA* in the following table:

i	GPA	$\widehat{GPA}$	$\hat{u}$
1	2.8	2.7143	.0857
2	3.4	3.0209	.3791
3	3.0	3.2253	–.2253
4	3.5	3.3275	.1725
5	3.6	3.5319	.0681
6	3.0	3.1231	–.1231
7	2.7	3.1231	–.4231
8	3.7	3.6341	.0659

You can verify that the residuals, as reported in the table, sum to –.0002, which is pretty close to zero given the inherent rounding error.

(iii) When $ACT = 20$, $\widehat{GPA} = .5681 + .1022(20) \approx 2.61$.

(iv) The sum of squared residuals, $\sum_{i=1}^{n} \hat{u}_i^2$, is about .4347 (rounded to four decimal places), and the total sum of squares, $\sum_{i=1}^{n} (y_i - \bar{y})^2$, is about 1.0288. So the *R*-squared from the regression is

$$R^2 = 1 - \text{SSR/SST} \approx 1 - (.4347/1.0288) \approx .577.$$

Therefore, about 57.7% of the variation in *GPA* is explained by *ACT* in this small sample of students.

2.3 (i) Income, age, and family background (such as number of siblings) are just a few possibilities. It seems that each of these could be correlated with years of education. (Income and education are probably positively correlated; age and education may be negatively correlated because women in more recent cohorts have, on average, more education; and number of siblings and education are probably negatively correlated.)

(ii) Not if the factors we listed in part (i) are correlated with *educ*. Because we would like to hold these factors fixed, they are part of the error term. But if *u* is correlated with *educ* then $E(u|educ) \neq 0$, and so SLR.4 fails.

2.4 (i) We would want to randomly assign the number of hours in the preparation course so that *hours* is independent of other factors that affect performance on the SAT. Then, we would collect information on SAT score for each student in the experiment, yielding a data set $\{(sat_i, hours_i) : i = 1, \ldots, n\}$, where *n* is the number of students we can afford to have in the study. From equation (2.7), we should try to get as much variation in $hours_i$ as is feasible.

(ii) Here are three factors: innate ability, family income, and general health on the day of the exam. If we think students with higher native intelligence think they do not need to prepare for the SAT, then ability and *hours* will be negatively correlated. Family income would probably be positively correlated with *hours*, because higher income families can more easily afford preparation courses. Ruling out chronic health problems, health on the day of the exam should be roughly uncorrelated with hours spent in a preparation course.

(iii) If preparation courses are effective, β_1 should be positive: other factors equal, an increase in *hours* should increase *sat*.

(iv) The intercept, β_0, has a useful interpretation in this example: because E(*u*) = 0, β_0 is the average SAT score for students in the population with *hours* = 0.

2.5 (i) When we condition on *inc* in computing an expectation, $\sqrt{inc}$ becomes a constant. So $E(u|inc) = E(\sqrt{inc} \cdot e|inc) = \sqrt{inc} \cdot E(e|inc) = \sqrt{inc} \cdot 0$ because $E(e|inc) = E(e) = 0$.

(ii) Again, when we condition on *inc* in computing a variance, $\sqrt{inc}$ becomes a constant. So $\text{Var}(u|inc) = \text{Var}(\sqrt{inc} \cdot e|inc) = (\sqrt{inc})^2 \text{Var}(e|inc) = \sigma_e^2\, inc$ because $\text{Var}(e|inc) = \sigma_e^2$.

(iii) Families with low incomes do not have much discretion about spending; typically, a low-income family must spend on food, clothing, housing, and other necessities. Higher income people have more discretion, and some might choose more consumption while others more saving. This discretion suggests wider variability in saving among higher income families.

2.6 (i) This derivation is essentially done in equation (2.52), once $(1/\text{SST}_x)$ is brought inside the summation (which is valid because SST_x does not depend on *i*). Then, just define $w_i = d_i/\text{SST}_x$.

(ii) Because $\text{Cov}(\hat{\beta}_1, \bar{u}) = \text{E}[(\hat{\beta}_1 - \beta_1)\bar{u}]$, we show that the latter is zero. But, from part (i), $\text{E}[(\hat{\beta}_1 - \beta_1)\bar{u}] = \text{E}\left[\left(\sum_{i=1}^n w_i u_i\right)\bar{u}\right] = \sum_{i=1}^n w_i \text{E}(u_i\bar{u}).$ Because the u_i are pairwise uncorrelated (they are independent), $\text{E}(u_i\bar{u}) = \text{E}(u_i^2/n) = \sigma^2/n$ (because $\text{E}(u_i u_h) = 0,\ i \neq h$). Therefore, $\sum_{i=1}^n w_i \text{E}(u_i\bar{u}) = \sum_{i=1}^n w_i(\sigma^2/n) = (\sigma^2/n)\sum_{i=1}^n w_i = 0.$

(iii) The formula for the OLS intercept is $\hat{\beta}_0 = \bar{y} - \hat{\beta}\bar{x}$ and, plugging in $\bar{y} = \beta_0 + \beta_1\bar{x} + \bar{u}$ gives $\hat{\beta}_0 = (\beta_0 + \beta_1\bar{x} + \bar{u}) - \hat{\beta}_1\bar{x} = \beta_0 + \bar{u} - (\hat{\beta}_1 - \beta_1)\bar{x}.$

(iv) Because $\hat{\beta}_1$ and $\bar{u}$ are uncorrelated,

$$\text{Var}(\hat{\beta}_0) = \text{Var}(\bar{u}) + \text{Var}(\hat{\beta}_1)\bar{x}^2 = \sigma^2/n + (\sigma^2/\text{SST}_x)\bar{x}^2 = \sigma^2/n + \sigma^2\bar{x}^2/\text{SST}_x,$$

which is what we wanted to show.

(v) Using the hint and substitution gives $\text{Var}(\hat{\beta}_0) = \sigma^2[(\text{SST}_x/n) + \bar{x}^2]/\text{SST}_x$

$$= \sigma^2\left[\left(n^{-1}\sum_{i=1}^n x_i^2 - \bar{x}^2\right) + \bar{x}^2\right]/\text{SST}_x = \sigma^2\left(n^{-1}\sum_{i=1}^n x_i^2\right)/\text{SST}_x.$$

2.7 (i) Yes. If living closer to an incinerator depresses housing prices, then being farther away increases housing prices.

(ii) If the city chose to locate the incinerator in an area away from more expensive neighborhoods, then log(*dist*) is positively correlated with housing quality. This would violate SLR.4, and OLS estimation is biased.

(iii) Size of the house, number of bathrooms, size of the lot, age of the home, and quality of the neighborhood (including school quality), are just a handful of factors. As mentioned in part (ii), these could certainly be correlated with *dist* [and log(*dist*)].