CHAPTER 1

The Nature of Econometrics and Economic Data

Table of Contents

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 a 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 in teaching the students how econometrics can be used. (And, it is more enjoyable for the students and me.)

I do not use counterfactual notation as in the modern "treatment effects" literature, but I do discuss causality using counterfactual reasoning. The return to education, perhaps focusing on the return to getting a college degree, is a good example of how counterfactual reasoning is easily incorporated into the discussion of causality.

SOLUTIONS TO PROBLEMS

1.1 (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 a larger class size is associated with lower performance. We might find a negative correlation because a 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 might 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 to be placed in 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.

1.2 (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.

1.3 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, which 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.

1.4 (i) Experimental data have to be collected to undertake a statistical analysis.

(ii) Yes, it is feasible to do a controlled experiment. The factors such as consumption, investment, net exports, and so on, would be required for a controlled experiments.

(iii) No, the correlation analysis between GSP growth and tax rates is not likely to be convincing as the tax rates have a significant negative effect on gross state products even after controlling factors like expenditure, fluctuations in the business, control in the supply of money, and so on.

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* in the sample is about $5.90, which seems low.

(iii) Using Table B-60 in the 2004 *Economic Report of the President*, the CPI was 56.9 in 1976 and 233 in 2013.

(iv) To convert 1976 dollars into 2013 dollars, we use the ratio of the CPIs, which is 233 / 56.9 □ 4.09. Therefore, the average hourly wage in 2013 dollars is roughly 4.09($5.90) □ $24.13, 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) 289 out of 1,823, or about 15.85 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.]

(vii) The percentage by which school A outspends school B is

(vii) The percentage by which school A outspends school B is

$$100 \cdot \frac{(6{,}000 - 5{,}500)}{5{,}500} \approx 9.09\%.$$

When we use the approximation based on the difference in the natural logs we get a somewhat smaller number:

$$100 \cdot [\log(6{,}000) - \log(5{,}500)] \approx 8.71\%.$$

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).

C1.5 (i) The smallest and largest values of *children* are 0 and 13, respectively. The average is about 2.27.

(ii) Out of 4,358 women, only 611 have electricity in the home, or about 14.02 percent.

(iii) The average of *children* for women without electricity is about 2.33, and for those with electricity it is about 1.90. So, on average, women with electricity have .43 fewer children than those who do not.

(iv) We cannot infer causality here. There are many confounding factors that may be related to the number of children and the presence of electricity in the home; household income and level of education are two possibilities. For example, it could be that women with more education have fewer children and are more likely to have electricity in the home (the latter due to an income effect).

C1.6 (i) There are 2,197 counties in the dataset. Of these, 1051 counties have zero murders. The percentage of counties having zero executions is 98.6%.

(ii) The largest number of murders is 1403. The largest number of executions is 3. The average number of executions is 0.0159, which is small because most of the counties have zero executions.

(iii) The correlation between *murders* and *execs* is 0.21. There is very low positive relationship

between them.

(iv) No, more executions do not cause more murders to occur. 21% percentage of murders occur by executions that took place of people sentenced to death in the given county.

C1.7 (i) The percentage of men in the sample report abusing alcohol is 9.9. The employment rate is 24.3.

(ii) The employment rate of men who abuse alcohol is 22.6.

(iii) The employment rate who do not abuse alcohol is 24.5.

(iv) The employment rates of men who abuse alcohol and who do not are 22.6 and 24.5, respectively. The difference in these employment rates is very less, which means that alcohol abuse does not cause unemployment.