Problems

1 Suppose that you are asked to conduct a study to determine whether smaller class sizes lead to improved student performance of fourth graders.

(i) If you could conduct any experiment you want, what would you do? Be specific.

(ii) More realistically, suppose you can collect observational data on several thousand fourth graders in a given state. You can obtain the size of their fourth-grade class and a standardized test score taken at the end of fourth grade. Why might you expect a negative correlation between class size and test score?

(iii) Would a negative correlation necessarily show that smaller class sizes cause better performance? Explain.

2 A justification for job training programs is that they improve worker productivity. Suppose that you are asked to evaluate whether more job training makes workers more productive. However, rather than having data on individual workers, you have access to data on manufacturing firms in Ohio. In particular, for each firm, you have information on hours of job training per worker (*training*) and number of nondefective items produced per worker hour (*output*).

(i) Carefully state the ceteris paribus thought experiment underlying this policy question.

(ii) Does it seem likely that a firm's decision to train its workers will be independent of worker characteristics? What are some of those measurable and unmeasurable worker characteristics?

(iii) Name a factor other than worker characteristics that can affect worker productivity.

(iv) If you find a positive correlation between *output* and *training*, would you have convincingly established that job training makes workers more productive? Explain.

3 Suppose at your university you are asked to find the relationship between weekly hours spent studying (*study*) and weekly hours spent working (*work*). Does it make sense to characterize the problem as inferring whether *study* "causes" *work* or *work* "causes" *study*? Explain.

4 States (and provinces) that have control over taxation sometimes reduce taxes in an attempt to spur economic growth. Suppose that you are hired by a state to estimate the effect of corporate tax rates on, say, the growth in per capita gross state product (GSP).

(i) What kind of data would you need to collect to undertake a statistical analysis?

(ii) Is it feasible to do a controlled experiment? What would be required?

(iii) Is a correlation analysis between GSP growth and tax rates likely to be convincing? Explain.

Computer Exercises

C1 Use the data in WAGE1 for this exercise.

(i) Find the average education level in the sample. What are the lowest and highest years of education?

(ii) Find the average hourly wage in the sample. Does it seem high or low?

(iii) The wage data are reported in 1976 dollars. Using the Internet or a printed source, find the Consumer Price Index (CPI) for the years 1976 and 2013.

(iv) Use the CPI values from part (iii) to find the average hourly wage in 2013 dollars. Now does the average hourly wage seem reasonable?

(v) How many women are in the sample? How many men?

C2 Use the data in BWGHT to answer this question.

(i) How many women are in the sample, and how many report smoking during pregnancy?

(ii) What is the average number of cigarettes smoked per day? Is the average a good measure of the "typical" woman in this case? Explain.

(iii) Among women who smoked during pregnancy, what is the average number of cigarettes smoked per day? How does this compare with your answer from part (ii), and why?

(iv) Find the average of *fatheduc* in the sample. Why are only 1,192 observations used to compute this average?

(v) Report the average family income and its standard deviation in dollars.

C3 The data in MEAP01 are for the state of Michigan in the year 2001. Use these data to answer the following questions.

(i) Find the largest and smallest values of *math4*. Does the range make sense? Explain.

(ii) How many schools have a perfect pass rate on the math test? What percentage is this of the total sample?

(iii) How many schools have math pass rates of exactly 50%?

(iv) Compare the average pass rates for the math and reading scores. Which test is harder to pass?

(v) Find the correlation between *math4* and *read4*. What do you conclude?

(vi) The variable *exppp* is expenditure per pupil. Find the average of *exppp* along with its standard deviation. Would you say there is wide variation in per pupil spending?

(vii) Suppose School A spends $6,000 per student and School B spends $5,500 per student. By what percentage does School A's spending exceed School B's? Compare this to $100 \cdot [\log(6,000) - \log(5,500)]$, which is the approximation percentage difference based on the difference in the natural logs. (See Section A.4 in Appendix A.)

C4 The data in JTRAIN2 come from a job training experiment conducted for low-income men during 1976–1977; see Lalonde (1986).

(i) Use the indicator variable *train* to determine the fraction of men receiving job training.

(ii) The variable *re78* is earnings from 1978, measured in thousands of 1982 dollars. Find the averages of *re78* for the sample of men receiving job training and the sample not receiving job training. Is the difference economically large?

(iii) The variable *unem78* is an indicator of whether a man is unemployed or not in 1978. What fraction of the men who received job training are unemployed? What about for men who did not receive job training? Comment on the difference.

(iv) From parts (ii) and (iii), does it appear that the job training program was effective? What would make our conclusions more convincing?

C5 The data in FERTIL2 were collected on women living in the Republic of Botswana in 1988. The variable *children* refers to the number of living children. The variable *electric* is a binary indicator equal to one if the woman's home has electricity, and zero if not.

(i) Find the smallest and largest values of *children* in the sample. What is the average of *children*?

(ii) What percentage of women have electricity in the home?

(iii) Compute the average of *children* for those without electricity and do the same for those with electricity. Comment on what you find.

(iv) From part (iii), can you infer that having electricity "causes" women to have fewer children? Explain.

C6 Use the data in COUNTYMURDERS to answer this question. Use only the year 1996. The variable *murders* is the number of murders reported in the county. The variable *execs* is the number of executions that took place of people sentenced to death in the given county. Most states in the United States have the death penalty, but several do not.

(i) How many counties are there in the data set? Of these, how many have zero murders? What percentage of counties have zero executions? (Remember, use only the 1996 data.)

(ii) What is the largest number of murders? What is the largest number of executions? Why is the average number of executions so small?

(iii) Compute the correlation coefficient between *murders* and *execs* and describe what you find.

(iv) You should have computed a positive correlation in part (iii). Do you think that more executions *cause* more murders to occur? What might explain the positive correlation?

C7 The data set in ALCOHOL contains information on a sample of men in the United States. Two key variables are self-reported employment status and alcohol abuse (along with many other variables). The variables *employ* and *abuse* are both binary, or indicator, variables: they take on only the values zero and one.

(i) What is percentage of the men in the sample report abusing alcohol? What is the employment rate?

(ii) Consider the group of men who abuse alcohol. What is the employment rate?

(iii) What is the employment rate for the group of men who do not abuse alcohol?

(iv) Discuss the difference in your answers to parts (ii) and (iii). Does this allow you to conclude that alcohol abuse causes unemployment?