Chapter 1

Problems

1 Suppose that you are asked to conduct a study to determine whether smaller class sizes lead to improved student performance of those aged 9–10 years old.
   (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 9 and 10 year-olds in a given region of a country. You can obtain the size of their class and a standardised test score taken at the end of the academic year. 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 programmes 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 Birmingham in the UK. 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 characterise the problem as inferring whether study “causes” work or work “causes” study? Explain.

Computer Exercises

C1 Use the data in WAGE1.RAW 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) Assume the wage data are reported in 1976 UK pounds sterling. Go to the UK Office for National Statistics and find the Retail Prices Index (RPI) for the years 1976 and 2003.
   (iv) Use the RPI values from part (iii) to find the average hourly wage in 2003 pounds. Now does the average hourly wage seem reasonable?

C2 Use the data in BWGHT.RAW 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 \( f_{educ} \) in the sample. How come only 1,785 observations are used to compute this average?

(v) Find the most common (modal) value \( n_{pvis} \).

C3 Use the data in MEAP01.RAW to answer the following questions.

(i) Find the largest and smallest values of \( math10 \). Does the range make sense? Explain.

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

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

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

(v) Find the correlation between \( math10 \) and \( read10 \). 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 Assume the data in JTRAIN2.RAW come from a job training experiment conducted for low-income men during one year.

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

(ii) Assume the variable \( re78 \) is earnings from 1978, measured in thousands of 1982 UK pounds. 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 programme was effective? What would make our conclusions more convincing?
Chapter 2

Problems

1 Let \( k \)ids denote the number of children ever born to a woman, and let \( e \)duc denote years of education for the woman. A simple model relating fertility to years of education is

\[
kids = \beta_0 + \beta_1 e\text{duc} + u,
\]

where \( u \) is the unobserved error.

(i) What kinds of factors are contained in \( u \)? Are these likely to be correlated with level of education?

(ii) Will a simple regression analysis uncover the ceteris paribus effect of education on fertility? Explain.

2 In the simple linear regression model \( y = \beta_0 + \beta_1 x + u \), suppose that \( E(u) \neq 0 \). Letting \( \alpha_0 = E(u) \), show that the model can always be rewritten with the same slope, but a new intercept and error, where the new error has a zero expected value.

3 The following table contains the \( ACT \) scores and the \( GPA \) (grade point average) for eight university students. Grade point average is based on a four-point scale and has been rounded to one digit after the decimal.

<table>
<thead>
<tr>
<th>Student</th>
<th>GPA</th>
<th>ACT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.8</td>
<td>21</td>
</tr>
<tr>
<td>2</td>
<td>3.4</td>
<td>24</td>
</tr>
<tr>
<td>3</td>
<td>3.0</td>
<td>26</td>
</tr>
<tr>
<td>4</td>
<td>3.5</td>
<td>27</td>
</tr>
<tr>
<td>5</td>
<td>3.6</td>
<td>29</td>
</tr>
<tr>
<td>6</td>
<td>3.0</td>
<td>25</td>
</tr>
<tr>
<td>7</td>
<td>2.7</td>
<td>25</td>
</tr>
<tr>
<td>8</td>
<td>3.7</td>
<td>30</td>
</tr>
</tbody>
</table>

(i) Estimate the relationship between \( GPA \) and \( ACT \) using OLS; that is, obtain the intercept and slope estimates in the equation

\[
\overline{GPA} = \hat{\beta}_0 + \hat{\beta}_1 ACT.
\]

Comment on the direction of the relationship. Does the intercept have a useful interpretation here? Explain. How much higher is the \( GPA \) predicted to be if the \( ACT \) score is increased by five points?

(ii) Compute the fitted values and residuals for each observation, and verify that the residuals (approximately) sum to zero.

(iii) What is the predicted value of \( GPA \) when \( ACT = 20 \)?

(iv) How much of the variation in \( GPA \) for these eight students is explained by \( ACT \)? Explain.
The data set BWGHT.RAW contains data on births to women in a country. Two variables of interest are the dependent variable, infant birth weight in kilograms \( (\text{bhwght}) \), and an explanatory variable, average number of cigarettes the mother smoked per day during pregnancy \( (\text{cigs}) \). The following simple regression was estimated using data on \( n = 1388 \) births:

\[
\hat{\text{bhwght}} = 119.77 - 0.514 \text{cigs}
\]

(i) What is the predicted birth weight when \( \text{cigs} = 0 \)? What about when \( \text{cigs} = 20 \) (one pack per day)? Comment on the difference.
(ii) Does this simple regression necessarily capture a causal relationship between the child’s birth weight and the mother’s smoking habits? Explain.
(iii) To predict a birth weight of 3.54kg, what would \( \text{cigs} \) have to be? Comment.
(iv) The proportion of women in the sample who do not smoke while pregnant is about .85. Does this help reconcile your finding from part (iii)?

In the linear consumption function

\[
\hat{\text{cons}} = \hat{\beta}_0 + \hat{\beta}_1 \text{inc},
\]

the (estimated) marginal propensity to consume (MPC) out of income is simply the slope, \( \hat{\beta}_1 \), while the average propensity to consume (APC) is \( \frac{\hat{\text{cons}}}{\text{inc}} = \frac{\hat{\beta}_0}{\text{inc}} + \hat{\beta}_1 \). Using observations for 100 families on annual income and consumption (both measured in euros), the following equation is obtained:

\[
\hat{\text{cons}} = -124.84 + 0.853 \text{inc}
\]

\( n = 100, R^2 = 0.692. \)

(i) Interpret the intercept in this equation, and comment on its sign and magnitude.
(ii) What is the predicted consumption when family income is €30,000?
(iii) With \( \text{inc} \) on the x-axis, draw a graph of the estimated MPC and APC.

Using assumed data from 2012 for houses sold in Andover, Hampshire, UK, the following equation relates housing price \( (\text{price}) \) to the distance from a recently built refuse incinerator \( (\text{dist}) \):

\[
\log(\text{price}) = 9.40 + 0.312 \log(\text{dist})
\]

\( n = 135, R^2 = 0.162. \)

(i) Interpret the coefficient on \( \log(\text{dist}) \). Is the sign of this estimate what you expect it to be?
(ii) Do you think simple regression provides an unbiased estimator of the ceteris paribus elasticity of \( \text{price} \) with respect to \( \text{dist} \)? (Think about the town’s decision on where to put the incinerator.)
(iii) What other factors about a house affect its price? Might these be correlated with distance from the incinerator?

Consider the savings function

\[
\text{sav} = \beta_0 + \beta_1 \text{inc} + u, u = \sqrt{\text{inc}} \cdot e,
\]

where \( e \) is a random variable with \( \text{E}(e) = 0 \) and \( \text{Var}(e) = \sigma_e^2 \). Assume that \( e \) is independent of \( \text{inc} \).
(i) Show that \( E(u^2) = 0 \), so that the key zero conditional mean assumption (Assumption SLR.4) is satisfied. \([\text{Hint: if } e \text{ is independent of } inc, \text{ then } E(e | inc) = E(e).]\)

(ii) Show that \( \text{Var}(u | inc) = \sigma^2inc \), so that the homoskedasticity Assumption SLR.5 is violated. In particular, the variance of \( sav \) increases with \( inc \). \([\text{Hint: } \text{Var}(e | inc) = \text{Var}(e), \text{ if } e \text{ and } inc \text{ are independent.}]\)

(iii) Provide a discussion that supports the assumption that the variance of \( sav \) increases with family income.

8 Consider the standard simple regression model \( y = \beta_0 + \beta_1 x + u \) under the Gauss-Markov Assumptions SLR.1 through SLR.5. The usual OLS estimators \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \) are unbiased for their respective population parameters. Let \( \hat{\beta}_1 \) be the estimator of \( \beta_1 \) obtained by assuming the intercept is zero (see Section 2.6).

(i) Find \( E(\hat{\beta}_1) \) in terms of the \( x_i, \beta_0 \), and \( \beta_1 \). Verify that \( \hat{\beta}_1 \) is unbiased for \( \beta_1 \) when the population intercept \( (\beta_0) \) is zero. Are there other cases where \( \hat{\beta}_1 \) is unbiased?

(ii) Find the variance of \( \beta_1 \). \( \text{[Hint: The variance does not depend on } \beta_0 \text{.]} \)

(iii) Show that \( \text{Var} (\hat{\beta}_1) = \text{Var} (\beta_1) \). \( \text{[Hint: For any sample of data, } \sum_{i=1}^{n} x_i^2 \geq \sum_{i=1}^{n} (x_i - \bar{x})^2, \text{ with strict inequality unless } \bar{x} = 0.]} \)

(iv) Comment on the tradeoff between bias and variance when choosing between \( \hat{\beta}_1 \) and \( \hat{\beta}_1 \).

9 (i) Let \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \) be the intercept and slope from the regression of \( y_i \) on \( x_i \), using \( n \) observations. Let \( c_1 \) and \( c_2 \), with \( c_2 \neq 0 \), be constants. Let \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \) be the intercept and slope from the regression of \( c_1 y_i \) on \( c_2 x_i \). Show that \( \hat{\beta}_1 = (c_1 / c_2) \hat{\beta}_1 \) and \( \hat{\beta}_0 = c_1 \hat{\beta}_0 \), thereby verifying the claims on units of measurement in Section 2.4. \( \text{[Hint: To obtain } \hat{\beta}_1, \text{ plug the scaled versions of } x \text{ and } y \text{ into (2.19). Then, use (2.17) for } \hat{\beta}_0, \text{ being sure to plug in the scaled } x \text{ and } y \text{ and the correct slope.]} \)

(ii) Now, let \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \) be the regression of \( (c_1 + y_i) \) on \( (c_2 + x_i) \) (with no restriction on \( c_1 \) or \( c_2 \)). Show that \( \hat{\beta}_1 = \hat{\beta}_1 \) and \( \hat{\beta}_0 = \hat{\beta}_0 + c_1 - c_2 \hat{\beta}_1 \).

(iii) Now, let \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \) be the OLS estimates from the regression \( \log(y) \) on \( x \), where we must assume \( y > 0 \) for all \( i \). For \( c_1 > 0 \), let \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \) be the intercept and slope from the regression of \( \log(c_1 y) \) on \( x \). Show that \( \hat{\beta}_1 = \hat{\beta}_1 \) and \( \hat{\beta}_0 = \log(c_1) + \hat{\beta}_0 \).

(iv) Now, assuming that \( x_i > 0 \) for all \( i \), let \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \) be the intercept and slope from the regression of \( y_i \) on \( \log(c_2 x_i) \). How do \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \) compare with the intercept and slope from the regression of \( y_i \) on \( \log(x_i) \)?

10 Let \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \) be the OLS intercept and slope estimators, respectively, and let \( \bar{u} \) be the sample average of the errors (not the residuals!).

(i) Show that \( \hat{\beta}_1 \) can be written as \( \hat{\beta}_1 = \beta_1 + \sum_{i=1}^{n} w_i u, \text{ where } w_i = d / \text{SST}, \text{ and } d_i = x_i - \bar{x}. \)

(ii) Use part (i), along with \( \sum_{i=1}^{n} w_i = 0 \), to show that \( \hat{\beta}_1 \) and \( \bar{u} \) are uncorrelated. \( \text{[Hint: You are being asked to show that } E(\hat{\beta}_1 - \beta_1) \cdot \bar{u} = 0 \text{.]} \)

(iii) Show that \( \hat{\beta}_0 \) can be written as \( \hat{\beta}_0 = \beta_0 + \bar{u} - (\hat{\beta}_1 - \beta_1) \bar{x}. \)

(iv) Use parts (ii) and (iii) to show that \( \text{Var}(\hat{\beta}_0) = \sigma^2/n + \sigma^2(\bar{x})^2 / \text{SST}. \)

(v) Do the algebra to simplify the expression in part (iv) to equation (2.58). \( \text{[Hint: } \text{SST}_x/n = n^{-1} \sum_{i=1}^{n} x_i^2 - (\bar{x})^2. \text{]} \)

11 Suppose you are interested in estimating the effect of hours spent per week studying (\( study \)) on university grade point average (\( gpa \)) for first-year university students.

(i) What would a controlled experiment entail in this context? Does such an experiment seem feasible?
(ii) Consider the more realistic case where students choose how much time to study per week, and you can only randomly sample the pairs study and gpa (at the end of the first year) from the population. Write the population model as

\[ gpa = \beta_0 + \beta_1 \text{study} + u, \]

where, as usual in a model with an intercept, we can assume \( E(u) = 0 \). List at least two factors contained in \( u \). Are these likely to have positive or negative correlation with study?

(iii) In the equation from part (ii), what do you think should be the sign of \( \beta_1 \) if the equation is interpreted causally?

(iv) In the equation from part (ii), what is the interpretation of \( \beta_0 \)?

**Computer Exercises**

**C1** The data in DCPP.RAW can be used to study the relationship between participation in a defined contribution pension plan and the generosity of the plan. The variable \( \text{prate} \) is the percentage of eligible workers with an active account; this is the variable we would like to explain. The measure of generosity is the plan match rate, \( \text{mrate} \). This variable gives the average amount the firm contributes to each worker’s plan for each €1 contribution by the worker. For example, if \( \text{mrate} = 0.50 \), then a €1 contribution by the worker is matched by a 50 cent contribution by the firm.

(i) Find the average participation rate and the average match rate in the sample of plans.

(ii) Now, estimate the simple regression equation

\[ \hat{\text{prate}} = \hat{\beta}_0 + \hat{\beta}_1 \text{mrate}, \]

and report the results along with the sample size and \( R \)-squared.

(iii) Interpret the intercept in your equation. Interpret the coefficient on \( \text{mrate} \).

(iv) Find the predicted \( \text{prate} \) when \( \text{mrate} = 3.5 \). Is this a reasonable prediction? Explain what is happening here.

(v) How much of the variation in \( \text{prate} \) is explained by \( \text{mrate} \)? Is this a lot in your opinion?

**C2** The data set in CEOSAL2.RAW contains information on chief executive officers for a range of corporations. The variable \( \text{salary} \) is annual compensation, in thousands of euro, and \( \text{ceoterm} \) is prior number of years as company CEO.

(i) Find the average salary and the average time spent in post in the sample.

(ii) How many CEOs are in their first year as CEO (that is, \( \text{ceoterm} = 0 \))? What is the longest time spent in post as a CEO?

(iii) Estimate the simple regression model

\[ \log(\text{salary}) = \beta_0 + \beta_1 \text{ceoterm} + u, \]

and report your results in the usual form. What is the (approximate) predicted percentage increase in salary given one more year as a CEO?

**C3** Use the data in SLEEP75.RAW from Biddle and Hamermesh (1990) to study whether there is a tradeoff between the time spent sleeping per week and the time spent in paid
work. We could use either variable as the dependent variable. For concreteness, estimate
the model

\[ \text{sleep} = \beta_0 + \beta_1 \text{totwrk} + u, \]

where \( \text{sleep} \) is minutes spent sleeping at night per week and \( \text{totwrk} \) is total minutes
worked during the week.

(i) Report your results in equation form along with the number of observations and
\( R^2 \). What does the intercept in this equation mean?

(ii) If \( \text{totwrk} \) increases by 2 hours, by how much is \( \text{sleep} \) estimated to fall? Do you
find this to be a large effect?

C4 Use the data in WAGE2.RAW to estimate a simple regression explaining monthly
salary (\( \text{wage} \)) in terms of IQ score (\( \text{IQ} \)).

(i) Find the average salary and average IQ in the sample. What is the sample standard
deviation of IQ? (IQ scores are standardised so that the average in the population
is 100 with a standard deviation equal to 15.)

(ii) Estimate a simple regression model where a one-point increase in \( \text{IQ} \) changes \( \text{wage} \)
by a constant euros amount. Use this model to find the predicted increase in wage
for an increase in \( \text{IQ} \) of 15 points. Does \( \text{IQ} \) explain most of the variation in \( \text{wage} \)?

(iii) Now, estimate a model where each one-point increase in \( \text{IQ} \) has the same percent-
age effect on \( \text{wage} \). If \( \text{IQ} \) increases by 15 points, what is the approximate percent-
age increase in predicted \( \text{wage} \)?

C5 For the population of firms in the chemical industry, let \( \text{rd} \) denote annual expenditures
on research and development, and let \( \text{sales} \) denote annual sales (both are in millions
of euros).

(i) Write down a model (not an estimated equation) that implies a constant elasticity
between \( \text{rd} \) and \( \text{sales} \). Which parameter is the elasticity?

(ii) Now, estimate the model using the data in RDCHEM.RAW. Write out the estimated
equation in the usual form. What is the estimated elasticity of \( \text{rd} \) with respect to
\( \text{sales} \)? Explain in words what this elasticity means.

C6 Use the data in ATTEND.RAW for this exercise. We want to study the relationship
between attendance rate (\( \text{atndrte} \)), which is measured as a percent, and score on an
achievement test (\( \text{ACT} \)), which has a maximum possible value of 32.

(i) Find the smallest and largest values of \( \text{atndrte} \) and \( \text{ACT} \) in the sample.

(ii) In the population model

\[ \text{atndrte} = \beta_0 + \beta_1 \text{ACT} + u, \]

interpret the coefficient \( \beta_1 \). Is the sign of \( \beta_1 \) obvious? Explain.

(iii) Use the data in ATTEND.RAW to estimate the model from part (ii). Report the
estimated equation in the usual way, including the sample size and \( R \)-squared.
Does \( \text{ACT} \) explain a lot of the variation in the attendance rate?

(iv) Find the predicted difference in attendance rates for two students whose \( \text{ACT} \)
scores differ by 10. What do you make of your answer?

(v) Before estimating the model, one might worry that the regression analysis can
produce fitted values for \( \text{atndrte} \) that are less than zero or greater than 100. What
is the range of fitted values for this sample?
Chapter 3

Problems

1. Using the data in GPA2.RAW on 4,137 university students, the following equation was estimated by OLS:

\[
\hat{\text{unipa}} = 1.392 - .0135 \text{psperc} + .00148 \text{sat}
\]

\[n = 4,137, R^2 = .273,\]

where \(\text{unipa}\) is measured on a four-point scale, \(\text{psperc}\) is the percentile in the post-sixteen class (defined so that, for example, \(\text{psperc} = 5\) means the top 5% of the class), and \(\text{sat}\) is the combined maths and verbal scores on the student achievement test.

(i) Why does it make sense for the coefficient on \(\text{psperc}\) to be negative?

(ii) What is the predicted university GPA when \(\text{psperc} = 20\) and \(\text{sat} = 1,050\)?

(iii) Suppose that two post-sixteen students, A and B, both end their studies in the same percentile, but Student A’s SAT score was 140 points higher (about one standard deviation in the sample). What is the predicted difference in university GPA for these two students? Is the difference large?

(iv) Holding \(\text{psperc}\) fixed, what difference in SAT scores leads to a predicted \(\text{unipa}\) difference of .50, or one-half of a grade point? Comment on your answer.

2. The data in WAGE2.RAW on working men was used to estimate the following equation:

\[
\hat{\text{educ}} = 10.36 - .094 \text{sibs} + .131 \text{meduc} + .210 \text{feduc}
\]

\[n = 722, R^2 = .214,\]

where \(\text{educ}\) is years of schooling, \(\text{sibs}\) is number of siblings, \(\text{meduc}\) is mother’s years of schooling, and \(\text{feduc}\) is father’s years of schooling.

(i) Does \(\text{sibs}\) have the expected effect? Explain. Holding \(\text{meduc}\) and \(\text{feduc}\) fixed, by how much does \(\text{sibs}\) have to increase to reduce predicted years of education by one year? (A noninteger answer is acceptable here.)

(ii) Discuss the interpretation of the coefficient on \(\text{meduc}\).

(iii) Suppose that Man A has no siblings, and his mother and father each have 12 years of education. Man B has no siblings, and his mother and father each have 16 years of education. What is the predicted difference in years of education between B and A?

3. The following model is a simplified version of the multiple regression model used by Biddle and Hamermesh (1990) to study the tradeoff between time spent sleeping and working and to look at other factors affecting sleep:

\[
\text{sleep} = \beta_0 + \beta_1 \text{totwrk} + \beta_2 \text{educ} + \beta_3 \text{age} + u,
\]

where \(\text{sleep}\) and \(\text{totwrk}\) (total work) are measured in minutes per week and \(\text{educ}\) and \(\text{age}\) are measured in years. (See also Computer Exercise C3 in Chapter 2 of the online end of chapter questions.)

(i) If adults trade off sleep for work, what is the sign of \(\beta_1\)?

(ii) What signs do you think \(\beta_2\) and \(\beta_3\) will have?
(iii) Using the data in SLEEP75.RAW, the estimated equation is

\[ \text{sleep} = 3,638.25 - .148 \text{totwrk} - 11.13 \text{educ} + 2.20 \text{age} \]

\[ n = 706, R^2 = .113. \]

If someone works five more hours per week, by how many minutes is sleep predicted to fall? Is this a large tradeoff?

(iv) Discuss the sign and magnitude of the estimated coefficient on \( \text{educ} \).

(v) Would you say \( \text{totwrk} \), \( \text{educ} \), and \( \text{age} \) explain much of the variation in \( \text{sleep} \)? What other factors might affect the time spent sleeping? Are these likely to be correlated with \( \text{totwrk} \)?

4 The median starting salary for new law graduates is determined by

\[ \log(\text{salary}) = \beta_0 + \beta_1 \text{LSAT} + \beta_2 \text{GPA} + \beta_3 \log(\text{libvol}) + \beta_4 \log(\text{cost}) \]
\[ + \beta_5 \text{rank} + u, \]

where \( \text{LSAT} \) is the median LSAT score for the graduating students, \( \text{GPA} \) is the median university GPA for the class, \( \text{libvol} \) is the number of volumes in the institution law library, \( \text{cost} \) is the annual cost of attending university to read law, and \( \text{rank} \) is a ranking for the institution offering the law degree (with \( \text{rank} = 1 \) being the best).

(i) Explain why we expect \( \beta_5 \approx 0 \).

(ii) What signs do you expect for the other slope parameters? Justify your answers.

(iii) Using the data in LAWUNI85.RAW, the estimated equation is

\[ \log(\text{salary}) = 8.34 + .0047 \text{LSAT} + .248 \text{GPA} + .095 \log(\text{libvol}) \]
\[ + .038 \log(\text{cost}) + .0033 \text{rank} \]

\[ n = 136, R^2 = .842. \]

What is the predicted ceteris paribus difference in salary for institutions with a median GPA different by one point? (Report your answer as a percentage.)

(iv) Interpret the coefficient on the variable \( \log(\text{libvol}) \).

(v) Would you say it is better to attend a higher ranked institution? How much is a difference in ranking of 20 worth in terms of predicted starting salary?

5 In a study relating college grade point average to time spent in various activities, you distribute a survey to several students. The students are asked how many hours they spend each week in four activities: studying, sleeping, working, and leisure. Any activity is put into one of the four categories, so that for each student, the sum of hours in the four activities must be 168.

(i) In the model

\[ \text{GPA} = \beta_0 + \beta_1 \text{study} + \beta_2 \text{sleep} + \beta_3 \text{work} + \beta_4 \text{leisure} + u, \]

does it make sense to hold \( \text{sleep} \), \( \text{work} \), and \( \text{leisure} \) fixed, while changing \( \text{study} \)?

(ii) Explain why this model violates Assumption MLR.3.

(iii) How could you reformulate the model so that its parameters have a useful interpretation and it satisfies Assumption MLR.3?
Consider the multiple regression model containing three independent variables, under Assumptions MLR.1 through MLR.4:

\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u. \]

You are interested in estimating the sum of the parameters on \( x_1 \) and \( x_2 \); call this \( \theta_1 = \beta_1 + \beta_2 \).

(i) Show that \( \hat{\theta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \) is an unbiased estimator of \( \theta_1 \).

(ii) Find \( \text{Var}(\hat{\theta}_1) \) in terms of \( \text{Var}(\hat{\beta}_1) \), \( \text{Var}(\hat{\beta}_2) \), and \( \text{Corr}(\hat{\beta}_1, \hat{\beta}_2) \).

Which of the following can cause OLS estimators to be biased?

(i) Heteroskedasticity.

(ii) Omitting an important variable.

(iii) A sample correlation coefficient of .95 between two independent variables both included in the model.

Suppose that average worker productivity at manufacturing firms (\( \text{avgprod} \)) depends on two factors, average hours of training (\( \text{avgtrain} \)) and average worker ability (\( \text{avgabil} \)):

\[ \text{avgprod} = \beta_0 + \beta_1 \text{avgtrain} + \beta_2 \text{avgabil} + u. \]

Assume that this equation satisfies the Gauss-Markov assumptions. If grants have been given to firms whose workers have less than average ability, so that \( \text{avgtrain} \) and \( \text{avgabil} \) are negatively correlated, what is the likely bias in \( \hat{\beta}_1 \) obtained from the simple regression of \( \text{avgprod} \) on \( \text{avgtrain} \)?

The following equation describes the relationship between pass rates for students aged 9–10 on a maths test, measured as a percent, spending per student (\( \text{exppp} \), in euros), and the percentage of students eligible for free and reduced-price lunches (\( \text{lunch} \)):

\[ \text{math10} = \beta_0 + \beta_1 \log(\text{exppp}) + \beta_2 \text{lunch} + u. \]

(i) Argue that \( \frac{\beta_1}{100} \) is the (ceteris paribus) percentage point change in \( \text{math10} \) when \( \text{exppp} \) increases by 10 percent.

(ii) If expenditure per student is higher at poor schools, are \( \log(\text{exppp}) \) and \( \text{lunch} \) positively or negatively correlated?

(iii) Using the data in MEAP01.RAW, the following equations were estimated:

\[ \hat{\text{math10}} = 84.84 - 1.52 \log(\text{exppp}), \quad n = 1,823, \quad R^2 = .0003. \]

\[ \hat{\text{math10}} = 84.84 + 11.38 \log(\text{exppp}) - .471 \text{lunch}, \quad n = 1,823, \quad R^2 = .370. \]

From these simple and multiple regression results, determine whether, in this sample, \( \log(\text{exppp}) \) and \( \text{lunch} \) are positively or negatively correlated.

Suppose that you are interested in estimating the ceteris paribus relationship between \( y \) and \( x_1 \). For this purpose, you can collect data on two control variables, \( x_2 \) and \( x_3 \). (For concreteness, you might think of \( y \) as final exam score, \( x_1 \) as lecture attendance, \( x_2 \) as GPA up through the previous semester, and \( x_3 \) as SAT or ACT score.) Let \( \hat{\beta}_1 \) be the simple regression estimate from \( y \) on \( x_1 \) and let \( \hat{\beta}_1 \) be the multiple regression estimate from \( y \) on \( x_1, x_2, x_3 \).
(i) If \( x_1 \) is highly correlated with \( x_2 \) and \( x_3 \) in the sample, and \( x_2 \) and \( x_3 \) have large partial effects on \( y \), would you expect \( \hat{\beta}_1 \) and \( \hat{\beta}_1 \) to be similar or very different? Explain.

(ii) If \( x_1 \) is almost uncorrelated with \( x_2 \) and \( x_3 \), but \( x_2 \) and \( x_3 \) are highly correlated, will \( \hat{\beta}_1 \) and \( \hat{\beta}_1 \) tend to be similar or very different? Explain.

(iii) If \( x_1 \) is highly correlated with \( x_2 \) and \( x_3 \), and \( x_2 \) and \( x_3 \) have small partial effects on \( y \), would you expect \( \text{se}(\hat{\beta}_1) \) or \( \text{se}(\hat{\beta}_1) \) to be smaller? Explain.

(iv) If \( x_1 \) is almost uncorrelated with \( x_2 \) and \( x_3 \), \( x_2 \) and \( x_3 \) have large partial effects on \( y \), and \( x_2 \) and \( x_3 \) are highly correlated, would you expect \( \text{se}(\hat{\beta}_1) \) or \( \text{se}(\hat{\beta}_1) \) to be smaller? Explain.

11 Suppose that the population model determining \( y \) is

\[
y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u,
\]

and this model satisfies Assumptions MLR.1 through MLR.4. However, we estimate the model that omits \( x_3 \). Let \( \hat{\beta}_0, \hat{\beta}_1, \) and \( \hat{\beta}_2 \) be the OLS estimators from the regression of \( y \) on \( x_1 \) and \( x_2 \). Show that the expected value of \( \hat{\beta}_1 \) (given the values of the independent variables in the sample) is

\[
E(\hat{\beta}_1) = \beta_1 + \beta_3 \frac{\sum_{i=1}^{n} \hat{r}_{i1} x_{i3}}{\sum_{i=1}^{n} \hat{r}_{i1}^2},
\]

where the \( \hat{r}_{i1} \) are the OLS residuals from the regression of \( x_1 \) on \( x_2 \). [Hint: The formula for \( \hat{\beta}_1 \) comes from equation (3.22). Plug \( y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + u_i \) into this equation. After some algebra, take the expectation treating \( x_{i3} \) and \( \hat{r}_{i1} \) as nonrandom.]

12 The following equation represents the effects of tax revenue mix on subsequent employment growth for the population of regions in a country:

\[
growth = \beta_0 + \beta_p \text{share}_p + \beta_i \text{share}_i + \beta_s \text{share}_s + \text{other factors},
\]

where \( \text{growth} \) is the percentage change in employment over a ten-year period, \( \text{share}_p \) is the share of property taxes in total tax revenue, \( \text{share}_i \) is the share of income tax revenues, and \( \text{share}_s \) is the share of sales tax revenues. All of these variables are measured in 1980. The omitted share, \( \text{share}_p \), includes fees and miscellaneous taxes. By definition, the four shares add up to one. Other factors would include expenditures on education, infrastructure, and so on (all measured in 1980).

(i) Why must we omit one of the tax share variables from the equation?

(ii) Give a careful interpretation of \( \beta_i \).

13 (i) Consider the simple regression model \( y = \beta_0 + \beta x + u \) under the first four Gauss-Markov assumptions. For some function \( g(x) \), for example \( g(x) = x^2 \) or \( g(x) = \log(1 + x^2) \), define \( z_i = g(x_i) \). Define a slope estimator as

\[
\tilde{\beta}_1 = \frac{\sum_{i=1}^{n} (z_i - \bar{z}) y_i}{\sum_{i=1}^{n} (z_i - \bar{z}) x_i}.
\]

Show that \( \tilde{\beta}_1 \) is linear and unbiased. Remember, because \( E(u|x) = 0 \), you can treat both \( x_i \) and \( z_i \) as nonrandom in your derivation.
(ii) Add the homoskedasticity assumption, MLR.5. Show that

\[
\text{Var}(\hat{\beta}_1) = \sigma^2 \left[ \sum_{i=1}^{n} (z_i - \bar{z})^2 \right] \left[ \sum_{i=1}^{n} (z_i - \bar{z})x_i \right]^2.
\]

(iii) Show directly that, under the Gauss-Markov assumptions, \( \text{Var}(\hat{\beta}_1) \leq \text{Var}(\tilde{\beta}_1) \), where \( \hat{\beta}_1 \) is the OLS estimator. [Hint: The Cauchy-Schwartz inequality in Appendix B implies that

\[
\left( n^{-1} \sum_{i=1}^{n} (z_i - \bar{z})(x_i - \bar{x}) \right)^2 \leq \left( n^{-1} \sum_{i=1}^{n} (z_i - \bar{z})^2 \right) \left( n^{-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right);
\]

notice that we can drop \( \bar{x} \) from the sample covariance.]

**Computer Exercises**

**C1** A problem of interest to health officials (and others) is to determine the effects of smoking during pregnancy on infant health. One measure of infant health is birth weight; a birth weight that is too low can put an infant at risk for contracting various illnesses. Since factors other than cigarette smoking that affect birth weight are likely to be correlated with smoking, we should take those factors into account. For example, higher income generally results in access to better prenatal care, as well as better nutrition for the mother. An equation that recognises this is

\[
bwght = \beta_0 + \beta_1 \text{cigs} + \beta_2 \text{faminc} + u.
\]

(i) What is the most likely sign for \( \beta_2 \)?

(ii) Do you think \( \text{cigs} \) and \( \text{faminc} \) are likely to be correlated? Explain why the correlation might be positive or negative.

(iii) Now, estimate the equation with and without \( \text{faminc} \), using the data in BWGHT.RAW. Report the results in equation form, including the sample size and \( R \)-squared. Discuss your results, focusing on whether adding \( \text{faminc} \) substantially changes the estimated effect of \( \text{cigs} \) on \( \text{bwght} \).

**C2** Use the data in HPRICE1.RAW to estimate the model

\[
\text{price} = \beta_0 + \beta_1 \text{sqrmt} + \beta_2 \text{bdrms} + u,
\]

where \( \text{price} \) is the house price measured in thousands of euros.

(i) Write out the results in equation form.

(ii) What is the estimated increase in price for a house with one more bedroom, holding square metreage constant?

(iii) What is the estimated increase in price for a house with an additional bedroom that is 13.006 square metres in size? Compare this to your answer in part (ii).

(iv) What percentage of the variation in price is explained by square metreage and number of bedrooms?

(v) The first house in the sample has \( \text{sqrmt} = 2,438 \) and \( \text{bdrms} = 4 \). Find the predicted selling price for this house from the OLS regression line.

(vi) The actual selling price of the first house in the sample was €300,000 (so \( \text{price} = 300 \)). Find the residual for this house. Does it suggest that the buyer underpaid or overpaid for the house?
C3 The file CEOSAL2.RAW contains data on 177 chief executive officers and can be used to examine the effects of firm performance on CEO salary.

(i) Estimate a model relating annual salary to firm sales and market value. Make the model of the constant elasticity variety for both independent variables. Write the results out in equation form.

(ii) Add profits to the model from part (i). Why can this variable not be included in logarithmic form? Would you say that these firm performance variables explain most of the variation in CEO salaries?

(iii) Add the variable ceoten to the model in part (ii). What is the estimated percentage return for another year of CEO term of office, holding other factors fixed?

(iv) Find the sample correlation coefficient between the variables log(mktval) and profits. Are these variables highly correlated? What does this say about the OLS estimators?

C4 Use the data in ATTEND.RAW for this exercise.

(i) Obtain the minimum, maximum, and average values for the variables atndrte, priGPA, and ACT.

(ii) Estimate the model

\[ atndrte = \beta_0 + \beta_1 priGPA + \beta_2 ACT + u, \]

and write the results in equation form. Interpret the intercept. Does it have a useful meaning?

(iii) Discuss the estimated slope coefficients. Are there any surprises?

(iv) What is the predicted atndrte if priGPA = 3.65 and ACT = 20? What do you make of this result? Are there any students in the sample with these values of the explanatory variables?

(v) If Student A has priGPA = 3.1 and ACT = 21 and Student B has priGPA = 2.1 and ACT = 26, what is the predicted difference in their attendance rates?

C5 Confirm the partialling out interpretation of the OLS estimates by explicitly doing the partialing out for Example 3.2. This first requires regressing educ on exper and term and saving the residuals, \( \hat{r}_1 \). Then, regress \( \log(wage) \) on \( \hat{r}_1 \). Compare the coefficient on \( \hat{r}_1 \) with the coefficient on educ in the regression of \( \log(wage) \) on educ, exper, and term.

C6 Use the data set in WAGE2.RAW for this problem. As usual, be sure all of the following regressions contain an intercept.

(i) Run a simple regression of IQ on educ to obtain the slope coefficient, say, \( \hat{\delta}_1 \).

(ii) Run the simple regression of log(wage) on educ, and obtain the slope coefficient, \( \hat{\beta}_1 \).

(iii) Run the multiple regression of log(wage) on educ and IQ, and obtain the slope coefficients, \( \hat{\beta}_1 \) and \( \hat{\beta}_2 \), respectively.

(iv) Verify that \( \hat{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \hat{\delta}_1 \).

C7 Use BWGHT2.RAW to answer this question.

(i) Estimate the model

\[ \log(bwght) = \beta_0 + \beta_1 cigs + \beta_2 npvis + u, \]

and report the results in the usual form, including the sample size and R-squared. Are the signs of the slope coefficients what you expected? Explain.
(ii) If npvis increases by one sample standard deviation, what is the estimated effect on birth weight?

(iii) Now run the simple regression of log(bwght) on cigs, and compare the slope coefficient with the estimate obtained in part (i). Is the estimated effect of cigarette smoking much different than in part (i)?

(iv) Find the correlation between cigs and npvis and use it to explain the similarity of the simple and multiple regression estimates of \( \beta_1 \).

(v) Add the variables mage, meduc, fage, and feduc to the regression from part (i). Is birth weight [more precisely log(bwght)] an easy variable to explain?

**C8** Use the data in DISCRIM.RAW to answer this question. Assume these are postal code–level data on prices for various items at fast-food restaurants, along with characteristics of the post code population, in two regions of the UK. The idea is to see whether fast-food restaurants charge higher prices in areas with a larger concentration of ethnic minorities.

(i) Find the average values of prpeth and income in the sample, along with their standard deviations. What are the units of measurement of prpeth and income?

(ii) Consider a model to explain the price of fizzy drinks, pfizz, in terms of the proportion of the population that belongs to an ethnic minority and median income:

\[
pfizz = \beta_0 + \beta_1 \text{prpeth} + \beta_2 \text{income} + u.
\]

Estimate this model by OLS and report the results in equation form, including the sample size and \( R^2 \)-squared. (Do not use scientific notation when reporting the estimates.) Interpret the coefficient on prpeth. Do you think it is economically large?

(iii) Compare the estimate from part (ii) with the simple regression estimate from pfizz on prpeth. Is the discrimination effect larger or smaller when you control for income?

(iv) A model with a constant price elasticity with respect to income may be more appropriate. Report estimates of the model

\[
\log(pfizz) = \beta_0 + \beta_1 \text{prpeth} + \beta_2 \text{income} + u.
\]

If prpeth increases by .20 (20 percentage points), what is the estimated percentage change in pfizz? (Hint: The answer is 2.xx, where you fill in the “xx.”)

(v) Now add the variable prppov to the regression in part (iv). What happens to \( \hat{\beta}_{\text{prpeth}} \)?

(vi) Find the correlation between log(income) and prppov. Is it roughly what you expected?

(vii) Evaluate the following statement: “Because log(income) and prppov are so highly correlated, they have no business being in the same regression.”
Chapter 4

Problems

1 Which of the following can cause the usual OLS $t$ statistics to be invalid (that is, not to have $t$ distributions under $H_0$)?
   (i) Heteroskedasticity.
   (ii) A sample correlation coefficient of .95 between two independent variables that are in the model.
   (iii) Omitting an important explanatory variable.

2 Consider an equation to explain salaries of CEOs in terms of annual firm sales, return on equity ($roe$, in percentage form), and return on the firm’s stock ($ros$, in percentage form):

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + \beta_2 roe + \beta_3 ros + u.$$ 

   (i) In terms of the model parameters, state the null hypothesis that, after controlling for sales and roe, ros has no effect on CEO salary. State the alternative that better stock market performance increases a CEO’s salary.
   (ii) Using the data in CEOSAL1.RAW, the following equation was obtained by OLS:

$$\begin{align*}
\log(salary) &= 4.32 + .280 \log(sales) + .0174 roe + .00024 ros \\
&\quad (.32) \quad (.035) \quad (.0041) \quad (.00054)
\end{align*}$$

   $n = 209, R^2 = .283$.

   By what percentage is salary predicted to increase if ros increases by 50 points? Does ros have a practically large effect on salary?
   (iii) Test the null hypothesis that ros has no effect on salary against the alternative that ros has a positive effect. Carry out the test at the 10% significance level.
   (iv) Would you include ros in a final model explaining CEO compensation in terms of firm performance? Explain.

3 The variable rdintens is expenditures on research and development (R&D) as a percentage of sales. Sales are measured in millions of euros. The variable profmarg is profits as a percentage of sales.

Using the data in RDCHEM.RAW for 32 firms in the chemical industry, the following equation is estimated:

$$\begin{align*}
\text{rdintens} &= .472 + .321 \log(sales) + .050 \text{ profmarg} \\
&\quad (1.369) \quad (2.16) \quad (0.46)
\end{align*}$$

   $n = 32, R^2 = .999$.

   (i) Interpret the coefficient on log(sales). In particular, if sales increases by 10%, what is the estimated percentage point change in rdintens? Is this an economically large effect?
   (ii) Test the hypothesis that R&D intensity does not change with sales against the alternative that it does increase with sales. Do the test at the 5% and 10% levels.
   (iii) Interpret the coefficient on profmarg. Is it economically large?
   (iv) Does profmarg have a statistically significant effect on rdintens?
Are rent rates influenced by the student population in a university town? Let rent be the average monthly rent paid on rental units in a university town in the United Kingdom. Let pop denote the total city population, avginc the average city income, and pctstu the student population as a percentage of the total population. One model to test for a relationship is

\[
\log(\text{rent}) = \beta_0 + \beta_1 \log(\text{pop}) + \beta_2 \log(\text{avginc}) + \beta_3 \text{pctstu} + u.
\]

(i) State the null hypothesis that size of the student body relative to the population has no ceteris paribus effect on monthly rents. State the alternative that there is an effect.

(ii) What signs do you expect for \(\beta_1\) and \(\beta_2\)?

(iii) The equation estimated using data from RENTAL.RAW for 64 university towns is

\[
\log(\text{rent}) = .043 + .066 \log(\text{pop}) + .507 \log(\text{avginc}) + .0056 \text{pctstu}
\]

\[
(\text{.844}) (\text{.039}) (\text{.081}) (\text{.0017})
\]

\(n = 64, R^2 = .458\).

What is wrong with the statement: “A 10% increase in population is associated with about a 6.6% increase in rent”?

(iv) Test the hypothesis stated in part (i) at the 1% level.

Using the data in UNIECON.RAW, the following equation was estimated:

\[
\text{unigpa} = .028 + .659 \text{psgpa} + .0130 \text{actmth} + .0122 \text{acteng}
\]

\[
(\text{.168}) (\text{.053}) (\text{.0085}) (\text{.0050})
\]

\(n = 814, R^2 = .256\),

where \(\text{unigpa}\) is grade point average at a University, \(\text{psgpa}\) is post-sixteen GPA, \(\text{actmth}\) is the mathematics ACT score, and \(\text{acteng}\) is the English ACT score.

(i) Using the standard normal approximation, find the 99% confidence interval for \(\beta_{\text{psgpa}}\).

(ii) Can you reject \(H_0: \beta_{\text{psgpa}} = .55\) against the two-sided alternative at the 1% significance level? What about \(H_0: \beta_{\text{psgpa}} = 1\)?

(iii) Let \(\text{actsum}\) be the sum of the maths and English test scores, \(\text{actsum} = \text{actmth} + \text{acteng}\). Using \(\text{actsum}\) in place of \(\text{acteng}\) gives

\[
\text{unigpa} = .028 + .659 \text{psgpa} + .0008 \text{actmth} + .0122 \text{actsum}
\]

\[
(\text{.168}) (\text{.053}) (\text{.0085}) (\text{.0050})
\]

\(n = 814, R^2 = .256\).

Can you reject \(H_0: \beta_{\text{actmth}} = \beta_{\text{acteng}}\) at any reasonable significance level? Explain.

In Section 4.5, we used as an example testing the rationality of assessments of housing prices. There, we used a log-log model in price and assess [see equation (4.47)]. Here, we use a level-level formulation.

(i) In the simple regression model

\[
\text{price} = \beta_0 + \beta_\text{assess} + u,
\]

the assessment is rational if \(\beta_1 = 1\) and \(\beta_0 = 0\). The estimated equation is

\[
\text{price} = -14.47 + .976 \text{assess}
\]

\[
(16.27) (\text{.049})
\]

\(n = 88, SSR = 165,644.51, R^2 = .820\).
First, test the hypothesis that \( H_0: \beta_0 = 0 \) against the two-sided alternative. Then, test \( H_0: \beta_1 = 1 \) against the two-sided alternative. What do you conclude?

(ii) To test the joint hypothesis that \( \beta_0 = 0 \) and \( \beta_1 = 1 \), we need the SSR in the restricted model. This amounts to computing \( \sum_{i=1}^{n} (price_i - assess)^2 \), where \( n = 88 \), since the residuals in the restricted model are just \( price_i - assess \). (No estimation is needed for the restricted model because both parameters are specified under \( H_0 \).) This turns out to yield SSR = 209,448.99. Carry out the \( F \) test for the joint hypothesis.

(iii) Now, test \( H_0: \beta_2 = 0, \beta_3 = 0, \) and \( \beta_4 = 0 \) in the model

\[
price = \beta_0 + \beta_1 assess + \beta_2 lotsize + \beta_3 sqrmtr + \beta_4 bdrms + u.
\]

The \( R \)-squared from estimating this model using the same 88 houses is .829.

(iv) If the variance of \( price \) changes with \( assess \), \( lotsize \), \( sqrmtr \), or \( bdrms \), what can you say about the \( F \) test from part (iii)?

In Example 4.7, we used data on nonunionised manufacturing firms to estimate the relationship between the scrap rate and other firm characteristics. We now look at this example more closely and use all available firms.

(i) The population model estimated in Example 4.7 can be written as

\[
\log(scrap) = \beta_0 + \beta_1 hrsemp + \beta_2 \log(sales) + \beta_3 \log(employ) + u.
\]

Using the 43 observations available for 1987, the estimated equation is

\[
\begin{align*}
\log(scrap) &= 11.74 - .042 \ hrsemp - .951 \ \log(sales) + .992 \ \log(employ) \\
&= 11.74 (4.57) - .042 (.019) - .951 (.370) + .992 (.360) \\
&= .310 \\
n &= 43, R^2 = .310.
\end{align*}
\]

Compare this equation to that estimated using only the 29 nonunionized firms in the sample.

(ii) Show that the population model can also be written as

\[
\log(scrap) = \beta_0 + \beta_1 hrsemp + \beta_2 \log(sales/employ) + \theta_3 \log(employ) + u,
\]

where \( \theta_3 = \beta_2 + \beta_3 \). [Hint: Recall that \( \log(x_2/x_3) = \log(x_2) - \log(x_3) \).] Interpret the hypothesis \( H_0: \theta_3 = 0 \).

(iii) When the equation from part (ii) is estimated, we obtain

\[
\begin{align*}
\log(scrap) &= 11.74 - .042 \ hrsemp - .951 \ \log(sales/employ) + .041 \ \log(employ) \\
&= 11.74 (4.57) - .042 (.019) - .951 (.370) + .041 (.205) \\
&= .310 \\
n &= 43, R^2 = .310.
\end{align*}
\]

Controlling for worker training and for the sales-to-employee ratio, do bigger firms have larger statistically significant scrap rates?

(iv) Test the hypothesis that a 1% increase in \( sales/employ \) is associated with a 1% drop in the scrap rate.

Consider the multiple regression model with three independent variables, under the classical linear model assumptions MLR.1 through MLR.6:

\[
y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u.
\]

You would like to test the null hypothesis \( H_0: \beta_1 - 3\beta_3 = 1 \).
(i) Let $\hat{\beta}_1$ and $\hat{\beta}_2$ denote the OLS estimators of $\beta_1$ and $\beta_2$. Find $\text{Var}(\hat{\beta}_1 - 3\hat{\beta}_2)$ in terms of the variances of $\hat{\beta}_1$ and $\hat{\beta}_2$ and the covariance between them. What is the standard error of $\hat{\beta}_1 - 3\hat{\beta}_2$?

(ii) Write the $t$-statistic for testing $H_0: \beta_1 = 3\beta_2$.

(iii) Define $\theta_1 = \beta_1 - 3\beta_2$ and $\hat{\theta}_1 = \hat{\beta}_1 - 3\hat{\beta}_2$. Write a regression equation involving $\beta_0$, $\theta_1$, $\beta_2$, and $\beta_3$ that allows you to directly obtain $\hat{\theta}_1$ and its standard error.

9 In Problem 3 in Chapter 3, we estimated the equation

$$\text{sleep} = 3,638.25 - 0.148 \text{totwrk} - 11.13 \text{educ} + 2.20 \text{age}$$

$$n = 706, R^2 = .113,$$

where we now report standard errors along with the estimates.

(i) Is either $\text{educ}$ or $\text{age}$ individually significant at the 5% level against a two-sided alternative? Show your work.

(ii) Dropping $\text{educ}$ and $\text{age}$ from the equation gives

$$\text{sleep} = 3,586.38 - 0.151 \text{totwrk}$$

$$n = 706, R^2 = .103.$$

Are $\text{educ}$ and $\text{age}$ jointly significant in the original equation at the 5% level? Justify your answer.

(iii) Does including $\text{educ}$ and $\text{age}$ in the model greatly affect the estimated tradeoff between sleeping and working?

(iv) Suppose that the sleep equation contains heteroskedasticity. What does this mean about the tests computed in parts (i) and (ii)?

10 Regression analysis can be used to test whether the market efficiently uses information in valuing stocks. For concreteness, let $\text{return}$ be the total return from holding a firm’s stock over the four-year period from the end of 1990 to the end of 1994. The efficient markets hypothesis says that these returns should not be systematically related to information known in 1990. If firm characteristics known at the beginning of the period help predict stock returns, then we could use this information in choosing stocks.

For 1990, let $dkr$ be a firm’s debt to capital ratio, let $eps$ denote the earnings per share, let $netinc$ denote net income, and let $salary$ denote total compensation for the CEO.

(i) Using the data in RETURN.RAW, the following equation was estimated:

$$\text{return} = -14.37 + 0.321 dkr + 0.043 eps - 0.0051 netinc + 0.0035 salary$$

$$n = 142, R^2 = .0395.$$  

Test whether the explanatory variables are jointly significant at the 5% level. Is any explanatory variable individually significant?

(ii) Now, reestimate the model using the log form for $netinc$ and $salary$:

$$\text{return} = -36.30 + 0.327 dkr + 0.069 eps - 4.74 \log(\text{netinc}) + 7.24 \log(\text{salary})$$

$$n = 142, R^2 = .0330.$$  

Do any of your conclusions from part (i) change?
(iii) How come we do not also use logs of dkr and eps in part (ii)?
(iv) Overall, is the evidence for predictability of stock returns strong or weak?

The following table was created using the data in CEOSAL2.RAW:

<table>
<thead>
<tr>
<th>Dependent Variable: log(salary)</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>log(sales)</td>
<td>.224</td>
<td>.158</td>
<td>.188</td>
</tr>
<tr>
<td></td>
<td>(.027)</td>
<td>(.040)</td>
<td>(.040)</td>
</tr>
<tr>
<td>log(mktval)</td>
<td>——</td>
<td>.112</td>
<td>.100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.050)</td>
<td>(.049)</td>
</tr>
<tr>
<td>profmarg</td>
<td>——</td>
<td>-.0023</td>
<td>-.0022</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.0022)</td>
<td>(.0021)</td>
</tr>
<tr>
<td>ceoterm</td>
<td>——</td>
<td>——</td>
<td>.0171</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(.0055)</td>
</tr>
<tr>
<td>comterm</td>
<td>——</td>
<td>——</td>
<td>-.0092</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(.0033)</td>
</tr>
<tr>
<td>intercept</td>
<td>4.94</td>
<td>4.62</td>
<td>4.57</td>
</tr>
<tr>
<td></td>
<td>(0.20)</td>
<td>(0.25)</td>
<td>(0.25)</td>
</tr>
<tr>
<td>Observations</td>
<td>177</td>
<td>177</td>
<td>177</td>
</tr>
<tr>
<td>R-squared</td>
<td>.281</td>
<td>.304</td>
<td>.353</td>
</tr>
</tbody>
</table>

The variable mktval is market value of the firm, profmarg is profit as a percentage of sales, ceoterm is years as CEO with the current company, and comterm is total years with the company.

(i) Comment on the effect of profmarg on CEO salary.
(ii) Does market value have a significant effect? Explain.
(iii) Interpret the coefficients on ceoterm and comterm. Are these explanatory variables statistically significant?
(iv) What do you make of the fact that longer the term with the company, holding the other factors fixed, is associated with a lower salary?

**Computer Exercises**

C1 The following model can be used to study whether campaign expenditures affect election outcomes:

\[
vote_A = \beta_0 + \beta_1 \log(expend_A) + \beta_2 \log(expend_B) + \beta_3 prtystr_A + u,
\]

where vote_A is the percentage of the vote received by Candidate A, expend_A and expend_B are campaign expenditures by Candidates A and B, and prtystr_A is a measure of party strength for Candidate A (the percentage of the most recent vote that went to A’s party).

(i) What is the interpretation of \( \beta_1 \)?
(ii) In terms of the parameters, state the null hypothesis that a 1% increase in A’s expenditures is offset by a 1% increase in B’s expenditures.
(iii) Estimate the given model using the data in VOTE1.RAW and report the results in usual form. Do A’s expenditures affect the outcome? What about B’s expenditures? Can you use these results to test the hypothesis in part (ii)?

(iv) Estimate a model that directly gives the $t$ statistic for testing the hypothesis in part (ii). What do you conclude? (Use a two-sided alternative.)

C2 Use the data in LAWUNI85.RAW for this exercise.

(i) Using the same model as in Problem 4 in Chapter 3, state and test the null hypothesis that the rank of institutions offering law has no ceteris paribus effect on median starting salary.

(ii) Are features of the incoming cohort of students—namely, LSAT and GPA—individually or jointly significant for explaining salary? (Be sure to account for missing data on LSAT and GPA.)

(iii) Test whether the size of the entering cohort (cosize) or the size of the faculty (faculty) needs to be added to this equation; carry out a single test. (Be careful to account for missing data on cosize and faculty.)

(iv) What factors might influence the rank of the institution that are not included in the salary regression?

C3 Refer to Computer Exercise C2 in Chapter 3. Now, use the log of the housing price as the dependent variable:

\[
\log(price) = \beta_0 + \beta_1sqmtr + \beta_2bdrms + u.
\]

(i) You are interested in estimating and obtaining a confidence interval for the percentage change in price when a 150-square-metre bedroom is added to a house. In decimal form, this is $\theta_1 = 150\beta_1 + \beta_2$. Use the data in HPRICE1.RAW to estimate $\theta_1$.

(ii) Write $\beta_2$ in terms of $\theta_1$ and $\beta_1$ and plug this into the log(price) equation.

(iii) Use part (ii) to obtain a standard error for $\hat{\theta}_1$ and use this standard error to construct a 95% confidence interval.

C4 In Example 4.9, the restricted version of the model can be estimated using all 1,388 observations in the sample. Compute the $R$-squared from the regression of $bwght$ on $cigs$, $parity$, and $faminc$ using all observations. Compare this to the $R$-squared reported for the restricted model in Example 4.9.

C5 Use the data in HTV.RAW for this exercise.

(i) Consider a wage equation that includes parents’ education levels:

\[
\log(wage) = \beta_0 + \beta_1educ + \beta_2exper + \beta_3motheduc + \beta_4fatheduc + u.
\]

State the null hypothesis that mother’s and father’s education levels have the same effect on log(wage).

(ii) Estimate the model in part (i) and comment on the magnitudes of $\hat{\beta}_3$ and $\hat{\beta}_4$.

(iii) Test the hypothesis in part (i) against the two-sided alternative, at the 5% level. What do you conclude?

C6 The data set DCPSSUBS.RAW contains information on net financial wealth ($nettfa$), age of the survey respondent ($age$), annual family income ($inc$), family size ($fsize$), and information on participation in certain pension plans for people in a country. The wealth
and income variables are both recorded in thousands of euros. For this question, use only the data for married people without children living at home \((marr = 1, fsize = 2)\).

(i) How many married couples without children at home are in the data set?

(ii) Use OLS to estimate the model

\[ netfa = \beta_0 + \beta_1 \text{inc} + \beta_2 \text{age} + u, \]

and report the results using the usual format. Interpret the slope coefficients. Are there any surprises in the slope estimates?

(iii) Does the intercept from the regression in part (ii) have an interesting meaning? Explain.

(iv) Find the \(p\)-value for the test \(H_0: \beta_2 = 1\) against \(H_1: \beta_2 > 1\). Do you reject \(H_0\) at the 1% significance level?

(v) If you do a simple regression of \(netfa\) on \(inc\), is the estimated coefficient on \(inc\) much different from the estimate in part (ii)? Why or why not?

C7 Use the data in DISCRIM.RAW to answer this question. (See also Computer Exercise C8 in Chapter 3.)

(i) Use OLS to estimate the model

\[ \log(pfizz) = \beta_0 + \beta_1 \text{prpeth} + \beta_2 \log(income) + \beta_3 \text{prppov} + u, \]

and report the results in the usual form. Is \(\hat{\beta}_1\) statistically different from zero at the 5% level against a two-sided alternative? What about at the 1% level?

(ii) What is the correlation between \(\log(income)\) and \(\text{prppov}\)? Is each variable statistically significant in any case? Report the two-sided \(p\)-values.

(iii) To the regression in part (i), add the variable \(\log(hseval)\). Interpret its coefficient and report the two-sided \(p\)-value for \(H_0: \beta_{\log(hseval)} = 0\).

(iv) In the regression in part (iii), what happens to the individual statistical significance of \(\log(income)\) and \(\text{prppov}\)? Are these variables jointly significant? (Compute a \(p\)-value.) What do you make of your answers?

(v) Given the results of the previous regressions, which one would you report as most reliable in determining whether the racial makeup of a post code influences local fast-food prices?
Chapter 5

Problems

1 In the simple regression model under MLR.1 through MLR.4, we argued that the slope estimator, \( \hat{\beta}_1 \), is consistent for \( \beta_1 \). Using \( \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \), show that \( \text{plim} \hat{\beta}_0 = \beta_0 \). [You need to use the consistency of \( \hat{\beta}_1 \) and the law of large numbers, along with the fact that \( \beta_0 = E(y) = \beta_1 E(x) \).]

2 Suppose that the model

\[
\text{pctstck} = \beta_0 + \beta_1 \text{funds} + \beta_2 \text{risktol} + u
\]

satisfies the first four Gauss-Markov assumptions, where \( \text{pctstck} \) is the percentage of a worker’s pension invested in the stock market, \( \text{funds} \) is the number of mutual funds that the worker can choose from, and \( \text{risktol} \) is some measure of risk tolerance (larger \( \text{risktol} \) means the person has a higher tolerance for risk). If \( \text{funds} \) and \( \text{risktol} \) are positively correlated, what is the inconsistency in \( \hat{\beta}_1 \), the slope coefficient in the simple regression of \( \text{pctstck} \) on \( \text{funds} \)?

3 The data set SMOKE.RAW contains information on smoking behaviour and other variables for a random sample of single adults in a country. The variable \( \text{cigs} \) is the (average) number of cigarettes smoked per day. Do you think \( \text{cigs} \) has a normal distribution in the country’s adult population? Explain.

4 In the simple regression model (5.16), under the first four Gauss-Markov assumptions, we showed that estimators of the form (5.17) are consistent for the slope, \( \beta_1 \). Given such an estimator, define an estimator of \( \beta_0 \) by \( \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \). Show that \( \text{plim} \hat{\beta}_0 = \beta_0 \).

Computer Exercises

C1 Use the data in WAGE1.RAW for this exercise.
   (i) Estimate the equation
       \[
wage = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{exper} + \beta_3 \text{term} + u.
       \]
       Save the residuals and plot a histogram.
   (ii) Repeat part (i), but with \( \log(wage) \) as the dependent variable.
   (iii) Would you say that Assumption MLR.6 is closer to being satisfied for the level-level model or the log-level model?

C2 Use the data in GPA2.RAW for this exercise.
   (i) Using all 4,137 observations, estimate the equation
       \[
       \text{unigpa} = \beta_0 + \beta_1 \text{psperc} + \beta_2 \text{sat} + u
       \]
       and report the results in standard form.
   (ii) Reestimate the equation in part (i), using the first 2,070 observations.
   (iii) Find the ratio of the standard errors on \( \text{psperc} \) from parts (i) and (ii). Compare this with the result from (5.10).

C3 (i) For the equation in part (i) of Computer Exercise C6 in Chapter 4, obtain the \( LM \) statistic for \( H_0: \beta_3 = 0, \beta_4 = 0 \).
   (ii) Obtain the (asymptotic) \( p \)-value for the test in part (i).
Chapter 6

Problems

1. The following equation was estimated using the data in CEOSAL1.RAW:

\[
\log(salary) = 4.322 + .276 \log(sales) + .0215 \text{ roe} - .00008 \text{ roe}^2
\]

\[
(0.324) \quad (0.033) \quad (0.0129) \quad (0.00026)
\]

\[n = 209, R^2 = .282.\]

This equation allows roe to have a diminishing effect on log(salary). Is this generality necessary? Explain why or why not.

2. Let \( \hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k \) be the OLS estimates from the regression of \( y_i \) on \( x_{i1}, \ldots, x_{ik}, i = 1, 2, \ldots, n \). For nonzero constants \( c_1, \ldots, c_k \), argue that the OLS intercept and slopes from the regression of \( c_0 y_i \) on \( c_1 x_{i1}, \ldots, c_k x_{ik}, i = 1, 2, \ldots, n \), are given by \( \hat{\beta}_0 = c_0 \hat{\beta}_0, \hat{\beta}_1 = (c_0/c_1) \hat{\beta}_1, \ldots, \hat{\beta}_k = (c_0/c_k) \hat{\beta}_k \). [Hint: Use the fact that the \( \hat{\beta}_j \) solve the first order conditions in (3.13), and the \( \hat{\beta}_j \) must solve the first order conditions involving the rescaled dependent and independent variables.]

3. Using the data in RDCHEM.RAW, the following equation was obtained by OLS:

\[
\text{rdintens} = 2.613 + .00030 \text{ sales} - .0000000070 \text{ sales}^2
\]

\[
(0.429) \quad (0.0014) \quad (0.000000037)
\]

\[n = 32, R^2 = .1484.\]

(i) At what point does the marginal effect of sales on rdintens become negative? (ii) Would you keep the quadratic term in the model? Explain. (iii) Define salesbil as sales measured in billions of euros: salesbil = sales/1,000. Rewrite the estimated equation with salesbil and salesbil^2 as the independent variables. Be sure to report standard errors and the R-squared. [Hint: Note that salesbil^2 = sales^2/(1,000)^2.] (iv) For the purpose of reporting the results, which equation do you prefer?

4. The following model allows the return to education to depend upon the total amount of the parents’ education, called pareduc:

\[
\log(wage) = \beta_0 + \beta_1 \text{ educ} + \beta_2 \text{ pareduc} + \beta_3 \text{ exper} + u.
\]

(i) Show that the proportionate effect on wage of another year of education is \( \beta_1 + \beta_2 \text{ pareduc} \). (ii) Using the data in HTV.RAW, the estimated equation is

\[
\begin{align*}
\log(wage) &= .515 + .093 \text{ educ} + .00099 \text{ educ} \cdot \text{ pareduc} + .034 \text{ exper} \\
(0.179) &\quad (0.013) \quad (0.0026) \quad (0.007)
\end{align*}
\]

\[n = 1,230, R^2 = .186.\]
If someone’s parents have 32 years of total education, by what percentage is that person’s return to education estimated to exceed that of someone whose parents have 24 years of education? Is the difference statistically significant?

(iii) When $pareduc$ is added as a separate regressor, we get

$$\log(wage) = -0.772 + 0.186\ educ + 0.054\ pareduc - 0.0028\ educ \cdot pareduc$$

$$(0.401)\quad (0.029)\quad (0.015)\quad (0.0011)$$

$$+ 0.034\ exper$$

$$(0.007)$$

$n = 1,230, R^2 = .195$.

Now how does the return to education depend on parent education? Find the two-sided $p$-value for testing the null hypothesis that the return to education does not depend on parent education. What do you conclude?

5 In Example 4.2, where the percentage of students receiving a passing score on a maths exam at age 16 ($math16$) is the dependent variable, does it make sense to include $sci17$—the percentage of 17-year olds passing a science exam—as an additional explanatory variable?

6 When $atndrte^2$ and $ACT\cdot atndrte$ are added to the equation estimated in (6.19), the $R$-squared becomes .232. Are these additional terms jointly significant at the 10% level? Would you include them in the model?

7 The following three equations were estimated using the 1,534 observations in DCPP.RAW:

$$\hat{prate} = 80.29 + 5.44\ mrate + .269\ age - .00013\ totemp$$

$$(.78)\quad (.52)\quad (.045)\quad (.0004)$$

$R^2 = .100, R^2 = .098.$

$$\hat{prate} = 97.32 + 5.02\ mrate + .314\ age - 2.66\ \log(totemp)$$

$$(1.95)\quad (0.51)\quad (0.044)\quad (0.28)$$

$R^2 = .144, R^2 = .142.$

$$\hat{prate} = 80.62 + 5.34\ mrate + .290\ age - .00043\ totemp$$

$$(.78)\quad (.52)\quad (.045)\quad (.00009)$$

$$+ .0000000039\ totemp^2$$

$$(.0000000010)$$

$R^2 = .108, R^2 = .106.$

Which of these three models do you prefer? Why?

8 Suppose we want to estimate the effects of alcohol consumption ($alcohol$) on university grade point average ($uniGPA$). In addition to collecting information on grade point averages and alcohol usage, we also obtain attendance information (say, percentage of lectures attended, called $attend$). A standardized test score (say, SAT) and post-sixteen GPA ($psGPA$) are also available.

(i) Should we include $attend$ along with $alcohol$ as explanatory variables in a multiple regression model? (Think about how you would interpret $\beta_{alcohol}$.)

(ii) Should SAT and $psGPA$ be included as explanatory variables? Explain.
Computer Exercises

C1 Use the data in KIELMC.RAW, only for the year 1981, to answer the following questions. Assume the data are for houses that sold during a particular year in a town when that year was the year construction began on a local refuse incinerator.

(i) To study the effects of the incinerator location on housing price, consider the simple regression model

$$\log(price) = \beta_0 + \beta_1 \log(dist) + u,$$

where price is housing price in euros and dist is distance from the house to the incinerator measured in metres. Interpreting this equation causally, what sign do you expect for $\beta_1$ if the presence of the incinerator depresses housing prices? Estimate this equation and interpret the results.

(ii) To the simple regression model in part (i), add the variables $\log(intst)$, $\log(area)$, $\log(land)$, rooms, baths, and age, where intst is distance from the home to the interstate, area is square metreage of the house, land is the lot size in square metres, rooms is total number of rooms, baths is number of bathrooms, and age is age of the house in years. Now, what do you conclude about the effects of the incinerator? Explain why (i) and (ii) give conflicting results.

(iii) Add $[\log(intst)]^2$ to the model from part (ii). Now what happens? What do you conclude about the importance of functional form?

(iv) Is the square of $\log(dist)$ significant when you add it to the model from part (iii)?

C2 Use the data in WAGE1.RAW for this exercise.

(i) Use OLS to estimate the equation

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + u$$

and report the results using the usual format.

(ii) Is $exper^2$ statistically significant at the 1% level?

(iii) Using the approximation

$$\%\Delta wage \approx 100(\hat{\beta}_2 + 2\hat{\beta}_3 exper)\Delta exper,$$

find the approximate return to the fifth year of experience. What is the approximate return to the twentieth year of experience?

(iv) At what value of exper does additional experience actually lower predicted $\log(wage)$? How many people have more experience in this sample?

C3 Consider a model where the return to education depends upon the amount of work experience (and vice versa):

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 educ \cdot exper + u.$$ 

(i) Show that the return to another year of education (in decimal form), holding exper fixed, is $\beta_1 + \beta_3 exper$.

(ii) State the null hypothesis that the return to education does not depend on the level of exper. What do you think is the appropriate alternative?

(iii) Use the data in WAGE2.RAW to test the null hypothesis in (ii) against your stated alternative.
(iv) Let $\theta_i$ denote the return to education (in decimal form), when $\text{exper} = 10$; $\theta_i = \beta_1 + 10\beta_3$. Obtain $\hat{\theta}_i$ and a 95% confidence interval for $\theta_i$. (Hint: Write $\beta_1 = \theta_1 - 10\beta_3$ and plug this into the equation; then rearrange. This gives the regression for obtaining the confidence interval for $\theta_i$.)

C4 Use the data in GPA2.RAW for this exercise.
(i) Estimate the model

$$ sat = \beta_0 + \beta_1 hsize + \beta_2 hsize^2 + u, $$

where $hsize$ is the size of a final year post-sixteen school cohort (in hundreds), and write the results in the usual form. Is the quadratic term statistically significant?
(ii) Using the estimated equation from part (i), what is the “optimal” post-sixteen school size? Justify your answer.
(iii) Is this analysis representative of the academic performance of all post-sixteen school students? Explain.
(iv) Find the estimated optimal post-sixteen school size, using log$(sat)$ as the dependent variable. Is it much different from what you obtained in part (ii)?

C5 Use the housing price data in HPRICE1.RAW for this exercise.
(i) Estimate the model

$$ \log(price) = \beta_0 + \beta_1 \log(lotsize) + \beta_2 \log(sqrmtr) + \beta_3 bdrms + u $$

and report the results in the usual OLS format.
(ii) Find the predicted value of log$(price)$, when lotsize = 20,000, sqrmtr = 2,500, and bdrms = 4. Using the methods in Section 6.4, find the predicted value of price at the same values of the explanatory variables.
(iii) For explaining variation in price, decide whether you prefer the model from part (i) or the model

$$ price = \beta_0 + \beta_1 lotsize + \beta_2 sqrmtr + \beta_3 bdrms + u. $$

C6 Use the data in VOTE1.RAW for this exercise.
(i) Consider a model with an interaction between expenditures:

$$ voteA = \beta_0 + \beta_1 prtystrA + \beta_2 expendA + \beta_3 expendB + \beta_4 expendA \cdot expendB + u. $$

What is the partial effect of $\text{expendB}$ on $voteA$, holding $prtystrA$ and $\text{expendA}$ fixed? What is the partial effect of $\text{expendA}$ on $voteA$? Is the expected sign for $\beta_4$ obvious?
(ii) Estimate the equation in part (i) and report the results in the usual form. Is the interaction term statistically significant?
(iii) Find the average of $\text{expendA}$ in the sample. Fix $\text{expendA}$ at 300 (for €300,000). What is the estimated effect of another €100,000 spent by Candidate B on $voteA$? Is this a large effect?
(iv) Now fix $\text{expendB}$ at 100. What is the estimated effect of $\Delta \text{expendA} = 100$ on $voteA$? Does this make sense?
(v) Now, estimate a model that replaces the interaction with $shareA$, Candidate A’s percentage share of total campaign expenditures. Does it make sense to hold both $\text{expendA}$ and $\text{expendB}$ fixed, while changing $shareA$?
(vi) (Requires calculus) In the model from part (v), find the partial effect of \(\text{expendB}\) on \(\text{voteA}\), holding \(\text{prtystrA}\) and \(\text{expendA}\) fixed. Evaluate this at \(\text{expendA} = 300\) and \(\text{expendB} = 0\) and comment on the results.

C7 Use the data in ATTEND.RAW for this exercise.
(i) In the model of Example 6.3, argue that

\[ \Delta \text{stndfll}/\Delta \text{priGPA} = \beta_2 + 2\beta_4 \text{priGPA} + \beta_6 \text{atndrte}. \]

Use equation (6.19) to estimate the partial effect when \(\text{priGPA} = 2.59\) and \(\text{atndrte} = 82\). Interpret your estimate.
(ii) Show that the equation can be written as

\[
\text{stndfll} = \theta_0 + \beta_1 \text{atndrte} + \theta_2 \text{priGPA} + \beta_4 \text{ACT} + \beta_6 (\text{priGPA} - 2.59)^2 \\
+ \beta_5 \text{ACT}^2 + \beta_6 \text{priGPA}(\text{atndrte} - 82) + u,
\]

where \(\theta_2 = \beta_2 + 2\beta_4(2.59) + \beta_6(82)\). (Note that the intercept has changed, but this is unimportant.) Use this to obtain the standard error of \(\hat{\beta}_2\) from part (i).
(iii) Suppose that, in place of \(\text{priGPA}(\text{atndrte} - 82)\), you put \((\text{priGPA} - 2.59)\cdot(\text{atndrte} - 82)\). Now how do you interpret the coefficients on \(\text{atndrte}\) and \(\text{priGPA}\)?

C8 Use the data in HPRICE1.RAW for this exercise.
(i) Estimate the model

\[
\text{price} = \beta_0 + \beta_1 \text{lotsize} + \beta_2 \text{sqrmartr} + \beta_3 \text{bdrms} + u
\]

and report the results in the usual form, including the standard error of the regression. Obtain predicted price, when we plug in \(\text{lotsize} = 10,000\), \(\text{sqrmartr} = 2,300\), and \(\text{bdrms} = 4\); round this price to the nearest euro.
(ii) Run a regression that allows you to put a 95% confidence interval around the predicted value in part (i). Note that your prediction will differ somewhat due to rounding error.
(iii) Let \(\text{price}^0\) be the unknown future selling price of the house with the characteristics used in parts (i) and (ii). Find a 95% CI for \(\text{price}^0\) and comment on the width of this confidence interval.

C9 The data set NBASAL.RAW contains salary information and career statistics for 269 players in the National Basketball Association (NBA).
(i) Estimate a model relating points-per-game (points) to years in the league (exper), age, and years played in university (univ). Include a quadratic in exper; the other variables should appear in level form. Report the results in the usual way.
(ii) Holding university years and age fixed, at what value of experience does the next year of experience actually reduce points-per-game? Does this make sense?
(iii) Why do you think \(\text{univ}\) has a negative and statistically significant coefficient? (Hint: NBA players can be drafted before finishing their university careers and even directly out of school.)
(iv) Add a quadratic in age to the equation. Is it needed? What does this appear to imply about the effects of age, once experience and education are controlled for?
(v) Now regress \(\log(\text{wage})\) on \(\text{points}, \text{exper}, \text{exper}^2, \text{age}, \text{and univ}\). Report the results in the usual format.
(vi) Test whether age and univ are jointly significant in the regression from part (v). What does this imply about whether age and education have separate effects on wage, once productivity and seniority are accounted for?

C10 Use the data in BWGHT2.RAW for this exercise.
(i) Estimate the equation
\[ \log(bwght) = \beta_0 + \beta_1 npvis + \beta_2 npvis^2 + u \]
by OLS, and report the results in the usual way. Is the quadratic term significant?
(ii) Show that, based on the equation from part (i), the number of prenatal visits that maximises \( \log(bwght) \) is estimated to be about 22. How many women had at least 22 prenatal visits in the sample?
(iii) Does it make sense that birth weight is actually predicted to decline after 22 prenatal visits? Explain.
(iv) Add mother’s age to the equation, using a quadratic functional form. Holding npvis fixed, at what mother’s age is the birth weight of the child maximised? What fraction of women in the sample are older than the “optimal” age?
(v) Would you say that mother’s age and number of prenatal visits explain a lot of the variation in \( \log(bwght) \)?
(vi) Using quadratics for both npvis and age, decide whether using the natural log or the level of bwght is better for predicting bwght.

C11 Use APPLE.RAW to verify some of the claims made in Section 6.3.
(i) Run the regression ecokgs on ecoprc, regprc and report the results in the usual form, including the R-squared and adjusted R-squared. Interpret the coefficients on the price variables and comment on their signs and magnitudes.
(ii) Are the price variables statistically significant? Report the \( p \)-values for the individual \( t \) tests.
(iii) What is the range of fitted values for ecokgs? What fraction of the sample reports ecolbs = 0? Comment.
(iv) Do you think the price variables together do a good job of explaining variation in ecokgs? Explain.
(v) Add the variables faminc, hhsize (household size), educ, and age to the regression from part (i). Find the \( p \)-value for their joint significance. What do you conclude?

C12 Use the subset of DCPPSUBS.RAW with marr = 1 and fsize = 2; this restricts the analysis to married couples without children living at home. (See also Computer Exercise C8 in Chapter 4.)
(i) What is the youngest age of the household head in the sample? How many household heads are at that age?
(ii) In the model
\[ nettfa = \beta_0 + \beta_1 inc + \beta_2 age + \beta_3 age^2 + u, \]
what is the literal interpretation of \( \beta_2 \)? By itself, is it of much interest?
(iii) Estimate the model from part (ii) and report the results in standard form. Are you concerned that the coefficient on age is negative? Explain.
(iv) Since the youngest people in the sample are 25, it makes sense to think that, for a given level of income, the lowest average amount of net total financial assets is at
age 25. Recall that the partial effect of age on netfa is \( \beta_2 + 2\beta_3 \cdot \text{age} \), so the partial effect at age 25 is \( \beta_2 + 2\beta_3(25) = \beta_2 + 50\beta_3 \); call this \( \theta_2 \). Find \( \theta_2 \) and obtain the two-sided \( p \)-value for testing \( H_0: \theta_2 = 0 \). You should conclude that \( \hat{\theta}_2 \) is small and very statistically insignificant. [\textit{Hint}: One way to do this is to estimate the model \( \text{netfa} = \alpha_0 + \beta_1 \cdot \text{inc} + \theta_2 \cdot \text{age} + \beta_3(\text{age} - 25)^2 + u \), where the intercept, \( \alpha_0 \), is different from \( \beta_0 \). There are other ways, too.]

(v) Because the evidence against \( H_0: \theta_2 = 0 \) is very weak, set it to zero and estimate the model

\[
\text{netfa} = \alpha_0 + \beta_1 \cdot \text{inc} + \beta_3(\text{age} - 25)^2 + u.
\]

In terms of goodness-of-fit, does this model fit better than that in part (ii)?

(vi) For the estimated equation in part (v), set \( \text{inc} = 50 \) (roughly, the average value) and graph the relationship between \( \text{netfa} \) and \( \text{age} \), but only for \( \text{age} \geq 25 \). Describe what you see.

(vii) Verify that \( \text{inc}^2 \) is statistically significant when added to the equation. Explain how you can use an approach similar to parts (iv) and (v) to estimate a strictly increasing effect of \( \text{inc} \) on \( \text{netfa} \). (\textit{Hint}: You would need to find the minimum level of income in the sample.)
Chapter 7

Problems

1 Using the data in SLEEP75.RAW (see also Problem 3 in Chapter 3), we obtain the estimated equation

\[
\hat{\text{sleep}} = 3.840.83 - .163 \text{totwrk} - 11.71 \text{educ} - 8.70 \text{age}
\]

\[
\begin{align*}
(235.11) & \quad (.018) \quad (5.86) \quad (11.21) \\
+ .128 \text{age}^2 + 87.75 \text{male}
\end{align*}
\]

\[
(134) \quad (34.33)
\]

\[n = 706, R^2 = .123, \bar{R}^2 = .117.\]

The variable \(\text{sleep}\) is total minutes per week spent sleeping at night, \(\text{totwrk}\) is total weekly minutes spent working, \(\text{educ}\) and \(\text{age}\) are measured in years, and \(\text{male}\) is a gender dummy.

(i) All other factors being equal, is there evidence that men sleep more than women? How strong is the evidence?

(ii) Is there a statistically significant tradeoff between working and sleeping? What is the estimated tradeoff?

(iii) What other regression do you need to run to test the null hypothesis that, holding other factors fixed, age has no effect on sleeping?

2 The following equations were estimated using the data in BWGHT.RAW:

\[
\log(\text{bwght}) = 4.66 - .0044 \text{cigs} + .0093 \log(\text{faminc}) + .016 \text{parity}
\]

\[
(22) \quad (.0099) \quad (.0059) \quad (.006)
\]

\[
+ .027 \text{male} + .055 \text{white}
\]

\[
(.010) \quad (.013)
\]

\[n = 1,388, R^2 = .0472\]

and

\[
\log(\text{bwght}) = 4.65 - .0052 \text{cigs} + .0110 \log(\text{faminc}) + .017 \text{parity}
\]

\[
(38) \quad (.0010) \quad (.0085) \quad (.006)
\]

\[
+ .054 \text{male} + .045 \text{white} - .030 \text{motheduc} + .0032 \text{fatheduc}
\]

\[
(111) \quad (.015) \quad (.0030) \quad (.0026)
\]

\[n = 1,191, R^2 = .0493.\]

The variables are defined as in Example 4.9, but we have added a dummy variable for whether the child is male and a dummy variable indicating whether the child is classified as white.

(i) In the first equation, interpret the coefficient on the variable \(\text{cigs}\). In particular, what is the effect on birth weight from smoking 10 more cigarettes per day?

(ii) How much more is a white child predicted to weigh than a nonwhite child, holding the other factors in the first equation fixed? Is the difference statistically significant?

(iii) Comment on the estimated effect and statistical significance of \(\text{motheduc}\).

(iv) From the given information, why are you unable to compute the \(F\) statistic for joint significance of \(\text{motheduc}\) and \(\text{fatheduc}\)?: What would you have to do to compute the \(F\) statistic?
3 Using the data in GPA2.RAW, the following equation was estimated:

$$\bar{sat} = 1,028.10 + 19.30 hsize - 2.19 hsize^2 - 45.09 female$$

$$- 169.81 ethnic + 62.31 female \cdot ethnic$$

$$n = 4,137, R^2 = .0858.$$ 

The variable sat is the combined SAT score, psize is size of the student’s post-16 school cohort, in hundreds, female is a gender dummy variable, and ethnic is a race dummy variable equal to one for ethnic groups and zero otherwise.

(i) Is there strong evidence that psize^2 should be included in the model? From this equation, what is the optimal post-16 school size?

(ii) Holding psize fixed, what is the estimated difference in SAT score between non-ethnic females and nonethnic males? How statistically significant is this estimated difference?

(iii) What is the estimated difference in SAT score between nonethnic males and ethnic males? Test the null hypothesis that there is no difference between their scores, against the alternative that there is a difference.

(iv) What is the estimated difference in SAT score between ethnic females and nonethnic females? What would you need to do to test whether the difference is statistically significant?

4 Using the data in UNIECON.RAW, which contains information on students in a large principles of microeconomics course at a university, the following equation was estimated:

$$\bar{score} = 13.98 + 11.25 unigpa + 2.57 psgpa + .742 act + .157 work$$

$$+ 4.41 calculus - .728 mothcoll + .218 fathcoll$$

$$n = 814, R^2 = .4194,$$

where the dependent variable, score, is the course total, as a percentage of total points possible. The explanatory variables are, in the order that they appear in the equation, UNI grade point average (at the beginning of the term), post-16 grade point average, ACT score, hours of work per week, a binary variable for whether the student has taken a calculus course, and binary indicators for whether mother and father have bachelor’s degrees.

(i) Interpret the coefficient on calculus and decide whether its estimated effect seems reasonable.

(ii) After controlling for unigpa, does high school performance (grade point average or ACT score) help predict performance in microeconomics principles?

(iii) When mothcoll and fathcoll are dropped from the equation, the R^2 becomes .4188. Is there any evidence that having a parent with a college degree helps predict performance in microeconomics principles, having controlled for the other explanatory variables?

5 In Example 7.2, let noPC be a dummy variable equal to one if the student does not own a PC, and zero otherwise.
If noPC is used in place of PC in equation (7.6), what happens to the intercept in the estimated equation? What will be the coefficient on noPC? (Hint: Write PC = 1 - noPC and plug this into the equation colGPA = \( \hat{\beta}_0 + \hat{\delta}_0\text{PC} + \hat{\beta}_1\text{hsGPA} + \hat{\beta}_2\text{ACT} \).)

(ii) What will happen to the R-squared if noPC is used in place of PC?

(iii) Should PC and noPC both be included as independent variables in the model? Explain.

6 To test the effectiveness of a job training programme on the subsequent wages of workers, we specify the model

\[
\log(\text{wage}) = \beta_0 + \beta_1\text{train} + \beta_2\text{educ} + \beta_3\text{exper} + u,
\]

where train is a binary variable equal to unity if a worker participated in the programme. Think of the error term u as containing unobserved worker ability. If less able workers have a greater chance of being selected for the programme, and you use an OLS analysis, what can you say about the likely bias in the OLS estimator of \( \beta_1 \)? (Hint: Refer back to Chapter 3 of the textbook.)

7 In the example in equation (7.29), suppose that we define outlf to be one if the woman is out of the labour force, and zero otherwise.

(i) If we regress outlf on all of the independent variables in equation (7.29), what will happen to the intercept and slope estimates? (Hint: inlf = 1 - outlf. Plug this into the population equation inlf = \( \beta_0 + \beta_1\text{nwifeinc} + \beta_2\text{educ} + \ldots \) and rearrange.)

(ii) What will happen to the standard errors on the intercept and slope estimates?

(iii) What will happen to the R-squared?

8 Suppose you collect data from a survey on wages, education, experience, and gender. In addition, you ask for information about marijuana usage. The original question is: “On how many separate occasions last month did you smoke marijuana?”

(i) Write an equation that would allow you to estimate the effects of marijuana usage on wage, while controlling for other factors. You should be able to make statements such as, “Smoking marijuana five more times per month is estimated to change wage by \( x\% \).”

(ii) Write a model that would allow you to test whether drug usage has different effects on wages for men and women. How would you test that there are no differences in the effects of drug usage for men and women?

(iii) Suppose you think it is better to measure marijuana usage by putting people into one of four categories: nonuser, light user (1 to 5 times per month), moderate user (6 to 10 times per month), and heavy user (more than 10 times per month). Now, write a model that allows you to estimate the effects of marijuana usage on wage.

(iv) Using the model in part (iii), explain in detail how to test the null hypothesis that marijuana usage has no effect on wage. Be very specific and include a careful listing of degrees of freedom.

(v) What are some potential problems with drawing causal inference using the survey data that you collected?

9 Let \( d \) be a dummy (binary) variable and let \( z \) be a quantitative variable. Consider the model

\[
y = \beta_0 + \delta_0d + \beta_1z + \delta_1d \cdot z + u;
\]

this is a general version of a model with an interaction between a dummy variable and a quantitative variable. [An example is in equation (7.17).]
(i) Since it changes nothing important, set the error to zero, \( u = 0 \). Then, when \( d = 0 \) we can write the relationship between \( y \) and \( z \) as the function \( f_0(z) = \beta_0 + \beta_1 z \). Write the same relationship when \( d = 1 \), where you should use \( f_1(z) \) on the left-hand side to denote the linear function of \( z \).

(ii) Assuming that \( \delta_1 \neq 0 \) (which means the two lines are not parallel), show that the value of \( z^* \) such that \( f_0(z^*) = f_1(z^*) \) is \( z^* = -\delta_0/\delta_1 \). This is the point at which the two lines intersect [as in Figure 7.2(b)]. Argue that \( z^* \) is positive if and only if \( \delta_0 \) and \( \delta_1 \) have opposite signs.

**Computer Exercises**

C1 Use the data in GPA1.RAW for this exercise.

(i) Add the variables \( mothcoll \) and \( fathcoll \) to the equation estimated in (7.6) and report the results in the usual form. What happens to the estimated effect of PC ownership? Is PC still statistically significant?

(ii) Test for joint significance of \( mothcoll \) and \( fathcoll \) in the equation from part (i) and be sure to report the \( p \)-value.

(iii) Add \( psGPA^2 \) to the model from part (i) and decide whether this generalisation is needed.

C2 Use the data in WAGE2.RAW for this exercise.

(i) Estimate the model

\[
\log(\text{wage}) = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{exper} + \beta_3 \text{term} + \beta_4 \text{married} \\
+ \beta_5 \text{ethnic} + \beta_6 \text{south} + \beta_7 \text{urban} + u
\]

and report the results in the usual form. Holding other factors fixed, what is the approximate difference in monthly salary between ethnic minorities and nonethnic minorities? Is this difference statistically significant?

(ii) Add the variables \( \text{exper}^2 \) and \( \text{term}^2 \) to the equation and show that they are jointly insignificant at even the 20% level.

(iii) Extend the original model to allow the return to education to depend on ethnicity and test whether the return to education does depend on ethnicity.

(iv) Again, start with the original model, but now allow wages to differ across four groups of people: married and ethnic, married and nonethnic, single and ethnic, and single and nonethnic. What is the estimated wage differential between married ethnics and married nonethnics?

C3 Use the data in GPA2.RAW for this exercise.

(i) Consider the equation

\[
\text{unigpa} = \beta_0 + \beta_1 \text{pssize} + \beta_2 \text{pssize}^2 + \beta_3 \text{pssperc} + \beta_4 \text{sat} \\
+ \beta_5 \text{female} + \beta_6 \text{athlete} + u,
\]

where \( \text{unigpa} \) is cumulative university grade point average, \( \text{pssize} \) is size of post-16 cohort, in hundreds, \( \text{pssperc} \) is academic percentile in the cohort, \( \text{sat} \) is combined SAT score, \( \text{female} \) is a binary gender variable, and \( \text{athlete} \) is a binary variable, which is one for student-athletes. What are your expectations for the coefficients in this equation? Which ones are you unsure about?
(ii) Estimate the equation in part (i) and report the results in the usual form. What is the estimated GPA differential between athletes and nonathletes? Is it statistically significant?

(iii) Drop sat from the model and reestimate the equation. Now, what is the estimated effect of being an athlete? Discuss why the estimate is different than that obtained in part (ii).

(iv) In the model from part (i), allow the effect of being an athlete to differ by gender and test the null hypothesis that there is no ceteris paribus difference between women athletes and women nonathletes.

(v) Does the effect of sat on unigpa differ by gender? Justify your answer.

C4 In Problem 2 in Chapter 4, we added the return on the firm’s stock, ros, to a model explaining CEO salary; ros turned out to be insignificant. Now, define a dummy variable, rosneg, which is equal to one if $ros \leq 0$ and equal to zero if $ros > 0$. Use CEOSAL1.RAW to estimate the model

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + \beta_2 \text{roe} + \beta_3 \text{rosneg} + u.$$ 

Discuss the interpretation and statistical significance of $\hat{\beta}_3$.

C5 Use the data in FERTIL2.RAW for this exercise. A simple model to explain number of living children is

$$children = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{age} + \beta_3 \text{age}^2 + \beta_4 \text{urban} + \beta_5 \text{electric} + \beta_6 \text{tv} + u,$$

where the explanatory variables are woman’s years of education, age (in years), and binary variables indicating whether the woman’s household has electricity and a television, respectively.

(i) Estimate this equation by OLS and report the results in the usual form. Discuss the coefficients and statistical significance of the variables electric and tv.

(ii) Are there differences in fertility between urban and nonurban residents? Explain.

(iii) Now estimate the equation separately for urban and nonurban residents (dropping urban, of course). Other than the intercepts, do there seem to be important differences in the coefficients?

(iv) Obtain the Chow statistic that allows different intercepts for urban and nonurban residents under the null. What do you conclude? [Hint: You are testing five restrictions, and the SSR form is easily obtained from parts (ii) and (iii).]

C6 Use the data in WAGE1.RAW for this exercise.

(i) Use equation (7.18) to estimate the gender differential when $educ = 12.5$. Compare this with the estimated differential when $educ = 0$.

(ii) Run the regression used to obtain (7.18), but with female-$(educ - 12.5)$ replacing female-$educ$. How do you interpret the coefficient on female now?

(iii) Is the coefficient on female in part (ii) statistically significant? Compare this with (7.18) and comment.

C7 Use the data in LOANAPP.RAW for this exercise. The binary variable to be explained is approve, which is equal to one if a mortgage loan to an individual was approved. The key explanatory variable is white, a dummy variable equal to one if the applicant was white. The other applicants in the data set are black and Asian.
To test for discrimination in the mortgage loan market, a linear probability model can be used:

\[ \text{approve} = \beta_0 + \beta_1 \text{white} + \text{other factors}. \]

(i) If there is discrimination against minorities, and the appropriate factors have been controlled for, what is the sign of \( \beta_1 \)?

(ii) Regress \( \text{approve} \) on \( \text{white} \) and report the results in the usual form. Interpret the coefficient on \( \text{white} \). Is it statistically significant? Is it practically large?

(iii) As controls, add the variables \( \text{hrat}, \text{obrat}, \text{loanprc}, \text{unem}, \text{male}, \text{married}, \text{dep}, \text{sch}, \text{cosign}, \text{chist}, \text{pubrec}, \text{mortlat1}, \text{mortlat2}, \text{and vr} \). What happens to the coefficient on \( \text{white} \)? Is there still evidence of discrimination against nonwhites?

(iv) Now, allow the effect of ethnicity to interact with the variable measuring other obligations as a percentage of income (\( \text{obrat} \)). Is the interaction term significant?

(v) Using the model from part (iv), what is the effect of being white on the probability of approval when \( \text{obrat} = 32 \), which is roughly the mean value in the sample? Obtain a 95% confidence interval for this effect.

C8 There has been much interest in whether the presence of defined contribution pension plans, available to workers, increases net savings. The data set DCPPSUBS.RAW contains information on net financial assets (\( \text{nettfa} \)), family income (\( \text{inc} \)), a binary variable for eligibility in a defined contribution plan (\( \text{eDCPP} \)), and several other variables.

(i) What fraction of the families in the sample are eligible for participation in a DCPP plan?

(ii) Estimate a linear probability model explaining DCPP eligibility in terms of income, age, and gender. Include income and age in quadratic form, and report the results in the usual form.

(iii) Would you say that DCPP eligibility is independent of income and age? What about gender? Explain.

(iv) Obtain the fitted values from the linear probability model estimated in part (ii). Are any fitted values negative or greater than one?

(v) Using the fitted values \( \hat{\text{DCPP}} \) from part (iv), define \( \hat{\text{DCPP}} = 1 \) if \( \hat{\text{DCPP}} \geq .5 \) and \( \hat{\text{DCPP}} = 0 \) if \( \hat{\text{DCPP}} < .5 \). Out of 9,275 families, how many are predicted to be eligible for a defined contribution pension plan?

(vi) For the 5,638 families not eligible for a DCPP, what percentage of these are predicted not to have a pension plan, using the predictor \( \hat{\text{DCPP}} \)? For the 3,637 families eligible for a defined contribution pension plan, what percentage are predicted to have one? (It is helpful if your econometrics package has a “tabulate” command.)

(vii) The overall percent correctly predicted is about 64.9%. Do you think this is a complete description of how well the model does, given your answers in part (vi)?

(viii) Add the variable \( \text{pira} \) as an explanatory variable to the linear probability model. Other things equal, if a family has someone with an individual retirement account, how much higher is the estimated probability that the family is eligible for a defined contribution pension plan? Is it statistically different from zero at the 10% level?

C9 Use the data for women only in AFFAIRS.RAW for this exercise.

(i) Estimate a linear probability model for \( \text{affair} \), a binary indicator equal to one if a woman reports having at least one extramarital affair during her marriage. Include as explanatory variables \( \text{yrsmarr}, \text{age}, \text{and educ} \). Interpret the coefficient on \( \text{yrsmarr} \).
(ii) Do age and educ seem to have an effect on affair once yrsmarr is controlled for? Carry out an appropriate test.

(iii) Add the variable kids to the model in part (i). Interpret its coefficient, and decide whether the estimate is statistically significant.

(iv) To the model from part (iii)—that is, keeping kids in the model—add the four religion dummy variables. The base group consists of women reporting that they are “anti-religious.” Does the probability of reporting an affair differ between women who are “very religious” and women who are “anti-religion”? How big is the effect?

(v) Does the probability of reporting an affair differ between women who report any degree of religious beliefs and women who are “not religious”? [Hint: It is easiest to change the base group from part (iv).]

C10 Use the data in DCPPSUBS.RAW for this exercise.

(i) Compute the average, standard deviation, minimum, and maximum values of nettfa in the sample.

(ii) Test the hypothesis that average nettfa does not differ by pension plan eligibility status; use a two-sided alternative. What is the euro amount of the estimated difference?

(iii) From part (ii) of Computer Exercise C8, it is clear that eDCPP is not exogenous in a simple regression model; at a minimum, it changes by income and age. Estimate a multiple linear regression model for nettfa that includes income, age, and eDCPP as explanatory variables. The income and age variables should appear as quadratics. Now, what is the estimated euro effect of pension plan eligibility?

(iv) To the model estimated in part (iii), add the interactions eDCPP · (age – 41) and eDCPP · (age – 41)^2. Note that the average age in the sample is about 41, so that in the new model, the coefficient on eDCPP is the estimated effect of pension plan eligibility at the average age. Which interaction term is significant?

(v) Comparing the estimates from parts (iii) and (iv), do the estimated effects of pension plan eligibility at age 41 differ much? Explain.

(vi) Now, drop the interaction terms from the model, but define five family size dummy variables: fsize1, fsize2, fsize3, fsize4, and fsize5. The variable fsize5 is unity for families with five or more members. Include the family size dummies in the model estimated from part (iii); be sure to choose a base group. Are the family dummies significant at the 1% level?

(vii) Now, do a Chow test for the model

\[ nettfa = \beta_0 + \beta_1 \text{inc} + \beta_2 \text{inc}^2 + \beta_3 \text{age} + \beta_4 \text{age}^2 + \beta_5 \text{eDCPP} + u \]

across the five family size categories, allowing for intercept differences. The restricted sum of squared residuals, SSR_r, is obtained from part (vi) because that regression assumes all slopes are the same. The unrestricted sum of squared residuals is SSR_u = SSR_1 + SSR_2 + ... + SSR_5, where SSR_f is the sum of squared residuals for the equation estimated using only family size f. You should convince yourself that there are 30 parameters in the unrestricted model (5 intercepts plus 25 slopes) and 10 parameters in the restricted model (5 intercepts plus 5 slopes). Therefore, the number of restrictions being tested is q = 20, and the df for the unrestricted model is 9,275 − 30 = 9,245.

C11 Use the data set in BEAUTY.RAW, which contains a subset of the variables (but more usable observations than in the regressions) reported by Hamermesh and Biddle (1994).
(i) Find the separate fractions of men and women that are classified as having above average looks. Are more people rated as having above average or below average looks?

(ii) Test the null hypothesis that the population fractions of above-average-looking women and men are the same. Report the one-sided p-value that the fraction is higher for women. (Hint: Estimating a simple linear probability model is easiest.)

(iii) Now estimate the model
\[ \log(\text{wage}) = \beta_0 + \beta_{1\text{belavg}} + \beta_{2\text{abvavg}} + u \]
separately for men and women, and report the results in the usual form. In both cases, interpret the coefficient on \(\text{belavg}\). Explain in words what the hypothesis \(H_0: \beta_1 = 0\) against \(H_1: \beta_1 \neq 0\) means, and find the p-values for men and women.

(iv) Is there convincing evidence that women with above average looks earn more than women with average looks? Explain.

(v) For both men and women, add the explanatory variables \(\text{educ}, \text{exper}, \text{exper}^2, \text{union}, \text{goodhlth}, \text{ethnic}, \text{married}, \text{south}, \text{bigcity}, \text{smllcity}, \text{and service}\). Do the effects of the “looks” variables change in important ways?

C12 Use the data in APPLE.RAW to answer this question.

(i) Define a binary variable as \(\text{ecobuy} = 1\) if \(\text{ecokg} > 0\) and \(\text{ecobuy} = 0\) if \(\text{ecokg} = 0\). In other words, \(\text{ecobuy}\) indicates whether, at the prices given, a family would buy any ecologically friendly apples. What fraction of families claim they would buy ecolabeled apples?

(ii) Estimate the linear probability model
\[ \text{ecobuy} = \beta_0 + \beta_{1\text{ecoprc}} + \beta_{2\text{regprc}} + \beta_{3\text{faminc}} + \beta_{4\text{hhsize}} + \beta_{5\text{educ}} + \beta_{6\text{age}} + u, \]
and report the results in the usual form. Carefully interpret the coefficients on the price variables.

(iii) Are the nonprice variables jointly significant in the LPM? (Use the usual F statistic, even though it is not valid when there is heteroskedasticity.) Which explanatory variable other than the price variables seems to have the most important effect on the decision to buy ecolabeled apples? Does this make sense to you?

(iv) In the model from part (ii), replace \(\text{faminc}\) with \(\log(\text{faminc})\). Which model fits the data better, using \(\text{faminc}\) or \(\log(\text{faminc})\)? Interpret the coefficient on \(\log(\text{faminc})\).

(v) In the estimation in part (iv), how many estimated probabilities are negative? How many are bigger than one? Should you be concerned?

(vi) For the estimation in part (iv), compute the percent correctly predicted for each outcome, \(\text{ecobuy} = 0\) and \(\text{ecobuy} = 1\). Which outcome is best predicted by the model?
Problems

1 Which of the following are consequences of heteroskedasticity?
   (i) The OLS estimators, \( \hat{\beta}_j \), are inconsistent.
   (ii) The usual \( F \) statistic no longer has an \( F \) distribution.
   (iii) The OLS estimators are no longer BLUE.

2 Consider a linear model to explain monthly beer consumption:

\[
beer = \beta_0 + \beta_1 inc + \beta_2 price + \beta_3 educ + \beta_4 female + u
\]

\[
E(u|inc,price,educ,female) = 0
\]

\[
Var(u|inc,price,educ,female) = \sigma^2 inc^2.
\]

Write the transformed equation that has a homoskedastic error term.

3 True or False: WLS is preferred to OLS when an important variable has been omitted from the model.

4 Using the data in GPA3.RAW, the following equation was estimated for the fall and second semester students:

\[
\text{trmgpa} = -2.12 + .900 \text{crsgpa} + .193 \text{cumgpa} + .0014 \text{tothrs}
\]

\[
(\text{.55}) (\text{.175}) (\text{.064}) (\text{.0012})
\]

\[
(\text{.55}) (\text{.166}) (\text{.074}) (\text{.0012})
\]

\[
+.0018 \text{ sat} - .0039 \text{ psperc} + .351 \text{ female} - .157 \text{ season}
\]

\[
(\text{.0002}) (\text{.0018}) (\text{.085}) (\text{.098})
\]

\[
(\text{.0002}) (\text{.0019}) (\text{.079}) (\text{.080})
\]

\[
n = 269, R^2 = .465.
\]

Here, \( \text{trmgpa} \) is term GPA, \( \text{crsgpa} \) is a weighted average of overall GPA in courses taken, \( \text{cumgpa} \) is GPA prior to the current semester, \( \text{tothrs} \) is total credit hours prior to the semester, \( \text{sat} \) is SAT score, \( \text{psperc} \) is percentile in post-16 education, \( \text{female} \) is a gender dummy, and \( \text{season} \) is a dummy variable equal to unity if the student’s sport is in season during autumn/winter. The usual and heteroskedasticity-robust standard errors are reported in parentheses and brackets, respectively.

(i) Do the variables \( \text{crsgpa}, \text{cumgpa}, \) and \( \text{tothrs} \) have the expected estimated effects? Which of these variables are statistically significant at the 5% level? Does it matter which standard errors are used?

(ii) Why does the hypothesis \( H_0: \beta_{\text{crsgpa}} = 1 \) make sense? Test this hypothesis against the two-sided alternative at the 5% level, using both standard errors. Describe your conclusions.

(iii) Test whether there is an in-season effect on term GPA, using both standard errors. Does the significance level at which the null can be rejected depend on the standard error used?
The following equation was estimated using the data in BWGHT2.RAW:

\[
\bar{lbwght} = 7.96 - .0023 \text{ cigs} + .0121 \text{ npvis} - .00024 \text{ npvis}^2
\]

\[
(0.05) \quad (.0012) \quad (.0037) \quad (.00012)
\]

\[
- .00098 \text{ mage} + .0022 \text{ fage} - .0014 \text{ meduc} + .0027 \text{ feduc}
\]

\[
(0.015) \quad (.0012) \quad (.0030) \quad (.0027)
\]

\[n = 1,624, R^2 = .0194,\]

where \(lbwght\) is the log of the birth weight, \(npvis\) is the number of prenatal visits, \(mage\) is mother’s age, \(fage\) is father’s age, \(meduc\) is mother’s education, and \(feduc\) is father’s education. The usual standard errors are in parentheses and the heteroskedasticity-robust standard errors are in brackets.

(i) Interpret the coefficient on \(cigs\). Does the 95% confidence interval for \(\beta_{cigs}\) depend on which standard error you use?

(ii) Comment on the statistical significance of \(npvis^2\), using both the usual and heteroskedasticity-robust standard errors.

(iii) If the four age and education terms are dropped from the regression (and the same set of observations is used), the \(R^2\) becomes .0162. Is this enough information to compute the heteroskedasticity-robust test of \(H_0: \beta_{mage} = 0, \beta_{fage} = 0, \beta_{meduc} = 0, \beta_{feduc} = 0\)? Explain.

There are different ways to combine features of the Breusch-Pagan and White tests for heteroskedasticity. One possibility not covered in the text is to run the regression

\[
\hat{u}_i^2 \text{ on } x_{i1}, x_{i2}, \ldots, x_{ik}, \hat{y}^2_i, i = 1, \ldots, n,
\]

where the \(\hat{u}_i\) are the OLS residuals and the \(\hat{y}^2_i\) are the OLS fitted values. Then, we would test joint significance of \(x_{i1}, x_{i2}, \ldots, x_{ik}\) and \(\hat{y}^2_i\). (Of course, we always include an intercept in this regression.)

(i) What are the \(df\) associated with the proposed \(F\) test for heteroskedasticity?

(ii) Explain why the \(R\)-squared from the regression above will always be at least as large as the \(R\)-squareds for the BP regression and the special case of the White test.

(iii) Does part (ii) imply that the new test always delivers a smaller \(p\)-value than either the BP or special case of the White statistic? Explain.

(iv) Suppose someone suggests also adding \(\hat{y}\) to the newly proposed test. What do you think of this idea?

Consider a model at the employee level,

\[
y_{i,e} = \beta_0 + \beta_1 x_{i,e1} + \beta_2 x_{i,e2} + \ldots + \beta_k x_{i,ek} + f_i + \nu_{i,e},
\]

where the unobserved variable \(f_i\) is a “firm effect” to each employee at a given firm \(i\). The error term \(\nu_{i,e}\) is specific to employee \(e\) at firm \(i\). The composite error is \(u_{i,e} = f_i + \nu_{i,e}\), such as in equation (8.28).

(i) Assume that \(\text{Var}(f_i) = \sigma_f^2\), \(\text{Var}(\nu_{i,e}) = \sigma_v^2\), and \(f_i\) and \(\nu_{i,e}\) are uncorrelated. Show that \(\text{Var}(u_{i,e}) = \sigma_f^2 + \sigma_v^2\); call this \(\sigma^2\).

(ii) Now suppose that for \(e \neq g\), \(\nu_{i,e}\) and \(\nu_{i,g}\) are uncorrelated. Show that \(\text{Cov}(u_{i,e}, u_{i,g}) = \sigma_f^2\).

(iii) Let \(\bar{u}_i = m_i^{-1} \sum_{e \in m_i} u_{i,e}\) be the average of the composite errors within a firm. Show that \(\text{Var}(u_i) = \sigma_f^2 + \sigma_v^2 / m_i\).
Discuss the relevance of part (iii) for WLS estimation using data averaged at the firm level, where the weight used for observation \( i \) is the usual firm size.

**Computer Exercises**

**C1** Consider the following model to explain sleeping behaviour:

\[
\text{sleep} = \beta_0 + \beta_1 \text{totwrk} + \beta_2 \text{educ} + \beta_3 \text{age} + \beta_4 \text{age}^2 + \beta_5 \text{yngkid} + \beta_6 \text{male} + u.
\]

(i) Write down a model that allows the variance of \( u \) to differ between men and women. The variance should not depend on other factors.

(ii) Use the data in SLEEP75.RAW to estimate the parameters of the model for heteroskedasticity. (You have to estimate the \text{sleep} equation by OLS, first, to obtain the OLS residuals.) Is the estimated variance of \( u \) higher for men or for women?

(iii) Is the variance of \( u \) statistically different for men and for women?

**C2**

(i) Use the data in HPRICE1.RAW to obtain the heteroskedasticity-robust standard errors for equation (8.17). Discuss any important differences with the usual standard errors.

(ii) Repeat part (i) for equation (8.18).

(iii) What does this example suggest about heteroskedasticity and the transformation used for the dependent variable?

**C3** Apply the full White test for heteroskedasticity [see equation (8.19)] to equation (8.18). Using the chi-square form of the statistic, obtain the \( p \)-value. What do you conclude?

**C4** Use VOTE1.RAW for this exercise.

(i) Estimate a model with \( \text{voteA} \) as the dependent variable and \( \text{prtystrA} \), \( \text{democA} \), \( \log(\text{expendA}) \), and \( \log(\text{expendB}) \) as independent variables. Obtain the OLS residuals, \( \hat{u} \), and regress these on all of the independent variables. Explain why you obtain \( R^2 = 0 \).

(ii) Now, compute the Breusch-Pagan test for heteroskedasticity. Use the \( F \) statistic version and report the \( p \)-value.

(iii) Compute the special case of the White test for heteroskedasticity, again using the \( F \) statistic form. How strong is the evidence for heteroskedasticity now?

**C5** In Example 7.12, we estimated a linear probability model for whether a young man was arrested during 2012:

\[
\text{arr12} = \beta_0 + \beta_1 \text{pcnv} + \beta_2 \text{avgson} + \beta_3 \text{tottime} + \beta_4 \text{ptime12} + \beta_5 \text{qemp12} + u.
\]

(i) Estimate this model by OLS and verify that all fitted values are strictly between zero and one. What are the smallest and largest fitted values?

(ii) Estimate the equation by weighted least squares, as discussed in Section 8.5.

(iii) Use the WLS estimates to determine whether \( \text{avgson} \) and \( \text{tottime} \) are jointly significant at the 5% level.

**C6** Use the data in LOANAPP.RAW for this exercise.

(i) Estimate the equation in part (iii) of Computer Exercise C7 in Chapter 7, computing the heteroskedasticity-robust standard errors. Compare the 95% confidence interval on \( \beta_{\text{white}} \) with the nonrobust confidence interval.
Obtain the fitted values from the regression in part (i). Are any of them less than zero? Are any of them greater than one? What does this mean about applying weighted least squares?

C7 Use the data set GPA1.RAW for this exercise.

(i) Use OLS to estimate a model relating uniGPA to psGPA, ACT, skipped, and PC. Obtain the OLS residuals.

(ii) Compute the special case of the White test for heteroskedasticity. In the regression of \( \hat{u}_i^2 \) on uniGPA\(_i\), uniGPA\(_i^2\), obtain the fitted values, say \( \hat{h}_i \).

(iii) Verify that the fitted values from part (ii) are all strictly positive. Then, obtain the weighted least squares estimates using weights \( \frac{1}{\hat{h}_i} \). Compare the weighted least squares estimates for the effect of skipping lectures and the effect of PC ownership with the corresponding OLS estimates. What about their statistical significance?

(iv) In the WLS estimation from part (iii), obtain heteroskedasticity-robust standard errors. In other words, allow for the fact that the variance function estimated in part (ii) might be misspecified. (See Question 8.4.) Do the standard errors change much from part (iii)?

C8 In Example 8.7, we computed the OLS and a set of WLS estimates in a cigarette demand equation.

(i) Obtain the OLS estimates in equation (8.35).

(ii) Obtain the \( \hat{h}_i \) used in the WLS estimation of equation (8.36) and reproduce equation (8.36). From this equation, obtain the unweighted residuals and fitted values; call these \( \hat{u}_i \) and \( \hat{y}_i \), respectively. (For example, in Stata, the unweighted residuals and fitted values are given by default.)

(iii) Let \( \hat{u}_i = \hat{u}_i / \sqrt{\hat{h}_i} \) and \( \hat{y}_i = \hat{y}_i / \sqrt{\hat{h}_i} \) be the weighted quantities. Carry out the special case of the White test for heteroskedasticity by regressing \( \hat{u}_i^2 \) on \( \hat{y}_i, \hat{y}_i^2 \), being sure to include an intercept, as always. Do you find heteroskedasticity in the weighted residuals?

(iv) What does the finding from part (iii) imply about the proposed form of heteroskedasticity used in obtaining (8.36)?

(v) Obtain valid standard errors for the WLS estimates that allow the variance function to be misspecified.

C9 Use the data set DCPPSUBS.RAW for this exercise.

(i) Using OLS, estimate a linear probability model for eDCPP, using as explanatory variables inc, inc\(^2\), age, age\(^2\), and male. Obtain both the usual OLS standard errors and the heteroskedasticity-robust versions. Are there any important differences?

(ii) In the special case of the White test for heteroskedasticity, where we regress the squared OLS residuals on a quadratic in the OLS fitted values, \( \hat{u}_i^2 \) on \( \hat{y}_i, \hat{y}_i^2 \), \( i = 1, \ldots, n \), argue that the probability limit of the coefficient on \( \hat{y}_i \) should be one, the probability limit of the coefficient on \( \hat{y}_i^2 \) should be \(-1\), and the probability limit of the intercept should be zero. [Hint: Remember that \( \text{Var}(\gamma|x_1, \ldots, x_k) = p(x) [1 - p(x)] \), where \( p(x) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k \).]

(iii) For the model estimated from part (i), obtain the White test and see if the coefficient estimates roughly correspond to the theoretical values described in part (ii).

(iv) After verifying that the fitted values from part (i) are all between zero and one, obtain the weighted least squares estimates of the linear probability model. Do they differ in important ways from the OLS estimates?
C10 Use the data in DCPPSUBS.RAW to answer this question, restricting attention to married couples with no children living at home \((marr = 1, fsize = 2)\).

(i) Estimate the equation

\[
\text{nettfa} = \beta_0 + \beta_1 \text{inc} + \beta_2 (\text{inc} - 10)^2 + \beta_3 \text{age} + \beta_4 (\text{age} - 25)^2 + \beta_5 \text{eDCPP} + u
\]

by OLS. The quadratic terms have been adjusted by subtracting off the minimum values of \text{inc} and \text{age}, so that \(\beta_1\) is the partial effect of \text{inc} on \text{nettfa} at \text{inc} = 10 and \(\beta_3\) is the partial effect of \text{age} on \text{nettfa} at \text{age} = 25. Report the usual standard errors as well as the heteroskedasticity-robust standard errors.

(ii) Test \text{inc} and \text{age} for joint significance, using a heteroskedasticity-robust test.

(iii) Estimate the model from part (i) using WLS, where the \(\hat{h}_i\) are obtained from equation (8.33). Obtain the usual WLS standard errors and the standard errors that are robust to misspecification of the heteroskedasticity function. Compare the usual WLS test of joint significance of \text{inc} and \text{age} with the robust WLS test.

(iv) Compare the OLS estimate of \(\beta_{e401k}\) with the WLS estimate. What do you think is happening?
Problems

1 In Problem 11 in Chapter 4 of the textbook, the $R^2$-squared from estimating the model

$$
\log(salary) = \beta_0 + \beta_1 \log(sales) + \beta_2 \log(mktval) + \beta_3 \text{profmarg} \\
+ \beta_4 \text{ceoterm} + \beta_5 \text{comterm} + u,
$$

using the data in CEOSAL2.RAW, was $R^2 = .353$ ($n = 177$). When $\text{ceoterm}^2$ and $\text{comterm}^2$ are added, $R^2 = .375$. Is there evidence of functional form misspecification in this model?

2 Let us modify Computer Exercise C4 in Chapter 8 by using voting outcomes in one year for incumbents who were elected two-years earlier. Assume Candidate A was elected in 2010 and was seeking reelection in 2012; $\text{voteA12}$ is Candidate A’s share of the two-party vote in 2012. The 2010 voting share of Candidate A is used as a proxy variable for quality of the candidate. Assume all other variables are for the 2012 election. The following equations were estimated, using the data in VOTE2.RAW:

$$
\widetilde{\text{voteA12}} = 75.71 + .312 prtystrA + 4.93 democA \\
\quad + -.929 \log(\text{expendA}) - 1.950 \log(\text{expendB}) \\
\quad + (9.25) \quad (1.01) \quad (1.01) \\
\quad + (0.64) \quad (0.281) \\
\quad n = 186, \widetilde{R^2} = .495, \widetilde{R^2} = .483,
$$

and

$$
\widetilde{\text{voteA12}} = 70.81 + .282 prtystrA + 4.52 democA \\
\quad + -.839 \log(\text{expendA}) - 1.846 \log(\text{expendB}) + .067 \text{voteA88} \\
\quad + (10.01) \quad (0.52) \quad (1.06) \\
\quad + (0.687) \quad (0.292) \quad (0.053) \\
\quad n = 186, \widetilde{R^2} = .499, \widetilde{R^2} = .485.
$$

(i) Interpret the coefficient on $\text{voteA10}$ and discuss its statistical significance.
(ii) Does adding $\text{voteA10}$ have much effect on the other coefficients?

3 Let $\text{math16}$ denote the percentage of students at a school receiving a passing score on a standardised maths test. We are interested in estimating the effect of per student spending on maths performance. A simple model is

$$
\text{math16} = \beta_0 + \beta_1 \log(\text{expend}) + \beta_2 \log(\text{enroll}) + \beta_3 \text{poverty} + u,
$$

where $\text{poverty}$ is the percentage of students living in poverty.

(i) The variable $\text{lnchprg}$ is the percentage of students eligible for a government funded school lunch programme. Why is this a sensible proxy variable for $\text{poverty}$?
(ii) The table that follows contains OLS estimates, with and without $\text{lnchprg}$ as an explanatory variable.
Explain why the effect of expenditures on math16 is lower in column (2) than in column (1). Is the effect in column (2) still statistically greater than zero?

(iii) Does it appear that pass rates are lower at larger schools, other factors being equal? Explain.

(iv) Interpret the coefficient on lnchprg in column (2).

(v) What do you make of the substantial increase in $R^2$ from column (1) to column (2)?

4 The following equation explains weekly hours a child plays video games, gamehours, in terms of child’s age, mother’s education, father’s education, and number of siblings:

$$gamehours = \beta_0 + \beta_1 \text{age} + \beta_2 \text{age}^2 + \beta_3 \text{motheduc} + \beta_4 \text{fatheduc} + \beta_5 \text{sibs} + u.$$ 

We are worried that gamehours is measured with error in our survey. The gamehours denote the reported hours of video game playing per week.

(i) Explain what the classical errors-in-variables (CEV) setup entails in this application.

(ii) Do you think the CEV assumptions are likely to hold? Explain.

5 In Example 4.4, we estimated a model relating number of campus crimes to student enrollment for a sample of universities. Assume the sample we used was not a random sample of universities because many institutions in 1992 did not report campus crimes. Do you think that university failure to report crimes can be viewed as exogenous sample selection? Explain.

Computer Exercises

C1 (i) Apply RESET from equation (9.3) to the model estimated in Computer Exercise C4 in Chapter 7. Is there evidence of functional form misspecification in the equation?

(ii) Compute a heteroskedasticity-robust form of RESET. Does your conclusion from part (i) change?

C2 Use the data set WAGE2.RAW for this exercise.

(i) Use the variable KWW (the “knowledge of the world of work” test score) as a proxy for ability in place of IQ in Example 9.3. What is the estimated return to education in this case?
(ii) Now, use IQ and KWW together as proxy variables. What happens to the estimated return to education?

(iii) In part (ii), are IQ and KWW individually significant? Are they jointly significant?

C3 Use the data from JTRAIN.RAW for this exercise.

(i) Consider the simple regression model

\[ \log(\text{scrap}) = \beta_0 + \beta_1 \text{grant} + u, \]

where \(\text{scrap}\) is the firm scrap rate and \(\text{grant}\) is a dummy variable indicating whether a firm received a job training grant. Can you think of some reasons why the unobserved factors in \(u\) might be correlated with \(\text{grant}\)?

(ii) Estimate the simple regression model using the data for 1988. (You should have 54 observations.) Does receiving a job training grant significantly lower a firm’s scrap rate?

(iii) Now, add as an explanatory variable \(\log(\text{scrap}_{87})\). How does this change the estimated effect of \(\text{grant}\)? Interpret the coefficient on \(\text{grant}\). Is it statistically significant at the 5% level against the one-sided alternative \(H_1: \beta_{\text{grant}} < 0\)?

(iv) Test the null hypothesis that the parameter on \(\log(\text{scrap}_{87})\) is one against the two-sided alternative. Report the \(p\)-value for the test.

(v) Repeat parts (iii) and (iv), using heteroskedasticity-robust standard errors, and briefly discuss any notable differences.

C4 Use the data for the year 1990 in INFMRTRAW.RAW for this exercise.

(i) Reestimate equation (9.37), but now include a dummy variable for the observation on a region (called \(NR\)). Interpret the coefficient on \(NR\) and comment on its size and significance.

(ii) Compare the estimates and standard errors from part (i) with those from equation (9.38). What do you conclude about including a dummy variable for a single observation?

C5 Use the data in RDCHEMRAW.RAW to further examine the effects of outliers on OLS estimates and to see how LAD is less sensitive to outliers. The model is

\[ \text{rdintens} = \beta_0 + \beta_1 \text{sales} + \beta_2 \text{sales}^2 + \beta_3 \text{profmarg} + u, \]

where you should first change \(\text{sales}\) to be in billions of euros to make the estimates easier to interpret.

(i) Estimate the above equation by OLS, both with and without the firm having annual sales of almost €40 billion. Discuss any notable differences in the estimated coefficients.

(ii) Estimate the same equation by LAD, again with and without the largest firm. Discuss any important differences in estimated coefficients.

(iii) Based on your findings in (i) and (ii), would you say OLS or LAD is more resilient to outliers?

C6 Redo Example 4.10 by dropping schools where teacher benefits are less than 1% of salary.

(i) How many observations are lost?

(ii) Does dropping these observations have any important effects on the estimated tradeoff?
C7 Use the data in LOANAPP.RAW for this exercise.
   (i) How many observations have obrat > 40, that is, other debt obligations more than 40% of total income?
   (ii) Reestimate the model in part (iii) of Computer Exercise C7 in Chapter 7, excluding observations with obrat > 40. What happens to the estimate and t statistic on white?
   (iii) Does it appear that the estimate of \( \beta_{\text{white}} \) is overly sensitive to the sample used?

C8 Use the data in DCPPSUBS.RAW to answer this question, using only single people (\( f_{\text{size}} = 1 \)). The equation we are interested in is

\[
\text{nettfa} = \beta_0 + \beta_1 \text{inc} + \beta_2 (\text{age} - 25)^2 + \beta_4 \text{male} + \beta_5 \text{eDCPP} + u.
\]

(i) Estimate the equation by OLS, report the results in the usual form, and interpret the coefficient on eDCPP.
(ii) Use the OLS residuals to test for heteroskedasticity using the Breusch-Pagan test. Does it appear \( u \) is independent of the explanatory variables?
(iii) Estimate the equation by LAD, and report the results in the same form as for OLS. Interpret the coefficient on eDCPP.
(iv) Reconcile your findings from parts (ii) and (iii).

C9 You need to use two data sets for this exercise, JTRAIN2.RAW and JTRAIN3.RAW. The former is the outcome of a job training experiment. The file JTRAIN3.RAW contains observational data, where individuals themselves largely determine whether they participate in job training. The data sets cover the same time period.

   (i) In the data set JTRAIN2.RAW, what fraction of the men received job training? What is the fraction in JTRAIN3.RAW? Why do you think there is such a big difference?
   (ii) Using JTRAIN2.RAW, run a simple regression of \( \text{re78} \) on \( \text{train} \). What is the estimated effect of participating in job training on real earnings?
   (iii) Now add as controls to the regression in part (ii) the variables \( \text{re74}, \text{re75}, \text{educ}, \text{age}, \text{black}, \text{and hisp} \). Does the estimated effect of job training on \( \text{re78} \) change much? How come? (Hint: Remember that these are experimental data.)
   (iv) Do the regressions in parts (ii) and (iii) using the data in JTRAIN3.RAW, reporting only the estimated coefficients on \( \text{train} \), along with their t statistics. What is the effect now of controlling for the extra factors, and why?
   (v) Define \( \text{avgre} = (\text{re74} + \text{re75})/2 \). Find the sample averages, standard deviations, and minimum and maximum values in the two data sets. Are these data sets representative of the same populations in 1978?
   (vi) Almost 96% of men in the data set JTRAIN2.RAW have \( \text{avgre} \) less than €10,000. Using only these men, run the regression

\[
\text{re78} \text{ on } \text{train}, \text{re74}, \text{re75}, \text{educ}, \text{age}, \text{black, asian}
\]

   and report the training estimate and its t statistic. Run the same regression for JTRAIN3.RAW, using only men with \( \text{avgre} \leq 10 \). For the subsample of low-income men, how do the estimated training effects compare across the experimental and nonexperimental data sets?
   (vii) Now use each data set to run the simple regression \( \text{re78} \) on \( \text{train} \), but only for men who were unemployed in 1974 and 1975. How do the training estimates compare now?
   (viii) Using your findings from the previous regressions, discuss the potential importance of having comparable populations underlying comparisons of experimental and nonexperimental estimates.
Chapter 10

Problems

1 Decide if you agree or disagree with each of the following statements and give a brief explanation of your decision:
(i) Like cross-sectional observations, we can assume that most time series observations are independently distributed.
(ii) The OLS estimator in a time series regression is unbiased under the first three Gauss-Markov assumptions.
(iii) A trending variable cannot be used as the dependent variable in multiple regression analysis.
(iv) Seasonality is not an issue when using annual time series observations.

2 Suppose that, for a province or district, the crime rate, crime, is a two-year distributed lag of the clear-up rate (percentage of crimes resulting in a conviction):
\[\text{crime}_t = \alpha_0 + \delta_0\text{clearup}_t + \delta_1\text{clearup}_{t-1} + \delta_2\text{clearup}_{t-2} + u_t,\]
where \(u_t\) is uncorrelated with \(\text{clearup}_t, \text{clearup}_{t-1}, \text{clearup}_{t-2}\), and all other past values of the arrest rate. Suppose that, through expenditures on law enforcement, the clear-up rate can be made to react to last year’s crime rate:
\[\text{clearup}_t = \gamma_0 + \gamma_\text{crime}_{t-1} + v_t.\]
(i) Explain, in behavioural terms, what \(\gamma_1 \neq 0\) means.
(ii) If \(v_t\) is uncorrelated with all past values of clearup, and \(u_t\), argue that clearup, and \(u_{t-1}\) must be correlated. (Hint: Lag the first equation one period and substitute for \(\text{crime}_{t-1}\) in the second equation.)
(iii) Which Gauss-Markov assumption does \(\text{Corr}\left(\text{clearup}_t, u_{t-1}\right) \neq 0\) violate?

3 Suppose \(y_t\) follows a second order FDL model:
\[y_t = \alpha_0 + \delta_0z_t + \delta_1z_{t-1} + \delta_2z_{t-2} + u_t.\]
Let \(z^*\) denote the equilibrium value of \(z\), and let \(y^*\) be the equilibrium value of \(y\), such that
\[y^* = \alpha_0 + \delta_0z^* + \delta_1z^* + \delta_2z^*.\]
Show that the change in \(y^*\), due to a change in \(z^*\), equals the long-run propensity times the change in \(z^*\):
\[\Delta y^* = LRP\cdot \Delta z^*.\]
This gives an alternative way of interpreting the LRP.

4 When the three event indicators \(\text{befile6}, \text{affile6},\) and \(\text{afdec6}\) are dropped from equation (10.22), we obtain \(R^2 = .281\) and \(\bar{R}^2 = .264\). Are the event indicators jointly significant at the 10% level?

5 Suppose you have quarterly data on new housing starts, interest rates, and real per capita income. Specify a model for housing starts that accounts for possible trends and seasonality in the variables.

6 In Example 10.4, we saw that our estimates of the individual lag coefficients in a distributed lag model were very imprecise. One way to alleviate the multicollinearity problem is
to assume that the $\delta_j$ follow a relatively simple pattern. For concreteness, consider a model with four lags:

$$y_t = \alpha_0 + \delta_0 z_{t-1} + \delta_1 z_{t-2} + \delta_2 z_{t-3} + \delta_3 z_{t-4} + u_t.$$  

Now, let us assume that the $\delta_j$ follow a quadratic in the lag, $j$:

$$\delta_j = \gamma_0 + \gamma_1 j + \gamma_2 j^2,$$

for parameters $\gamma_0$, $\gamma_1$, and $\gamma_2$. This is an example of a polynomial distributed lag (PDL) model.

(i) Plug the formula for each $\delta_j$ into the distributed lag model and write the model in terms of the parameters $\gamma_h$, for $h = 0,1,2$.

(ii) Explain the regression you would run to estimate the $\gamma_h$.

(iii) The polynomial distributed lag model is a restricted version of the general model. How many restrictions are imposed? How would you test these? (Hint: Think $F$ test.)

7 In Example 10.4, we wrote the model that explicitly contains the long-run propensity, $\theta_0$, as

$$gr_t = \alpha_0 + \theta_0 p_{e_t} + \delta_1 (p_{e_{t-1}} - p_{e_t}) + \delta_2 (p_{e_{t-2}} - p_{e_t}) + u_t,$$

where we omit the other explanatory variables for simplicity. As always with multiple regression analysis, $\theta_0$ should have a ceteris paribus interpretation. Namely, if $p_{e_t}$ increases by one (dollar) holding $(p_{e_{t-1}} - p_{e_t})$ and $(p_{e_{t-2}} - p_{e_t})$ fixed, $gr_t$ should change by $\theta_0$.

(i) If $(p_{e_{t-1}} - p_{e_t})$ and $(p_{e_{t-2}} - p_{e_t})$ are held fixed but $p_{e_t}$ is increasing, what must be true about changes in $p_{e_{t-1}}$ and $p_{e_{t-2}}$?

(ii) How does your answer in part (i) help you interpret $\theta_0$ in the above equation as the LRP?

Computer Exercises

C1 Assume that in October 1979, a country’s central bank changed its policy of targeting the money supply and instead began to focus directly on short-term interest rates. Using the data in INTDEF.RAW, define a dummy variable equal to 1 for years after 1979. Include this dummy in equation (10.15) to see if there is a shift in the interest rate equation after 1979. What do you conclude?

C2 Use the data in BARIUM.RAW for this exercise.

(i) Add a linear time trend to equation (10.22). Are any variables, other than the trend, statistically significant?

(ii) In the equation estimated in part (i), test for joint significance of all variables except the time trend. What do you conclude?

(iii) Add monthly dummy variables to this equation and test for seasonality. Does including the monthly dummies change any other estimates or their standard errors in important ways?

C3 Add the variable log(prgnp) to the minimum wage equation in (10.38). Is this variable significant? Interpret the coefficient. How does adding log(prgnp) affect the estimated minimum wage effect?

C4 Use the data in FERTIL3.RAW to verify that the standard error for the LRP in equation (10.19) is about .030.
C5 Use the data in EZANDERS.RAW for this exercise. The data are on monthly unemploy-
ment claims in a city, from January 1980 through November 1988. Assume that in
1984, an enterprise zone (EZ) was located in the city.
(i) Regress log(uclms) on a linear time trend and 11 monthly dummy variables. What
was the overall trend in unemployment claims over this period? (Interpret the
coefficient on the time trend.) Is there evidence of seasonality in unemployment
claims?
(ii) Add ez, a dummy variable equal to 1 in the months the city had an EZ, to the regres-
sion in part (i). Does having the enterprise zone seem to decrease unemployment
claims? By how much? [You should use formula (7.10) from Chapter 7.]
(iii) What assumptions do you need to make to attribute the effect in part (ii) to the cre-
ation of an EZ?

C6 Use the data in FERTIL3.RAW for this exercise.
(i) Regress gfr$_t$ on $t$ and $t^2$ and save the residuals. This gives a detrended gfr$_{t}$, say, gfr$_{t}$. 
(ii) Regress $gfr_{t}$ on all of the variables in equation (10.35), including $t$ and $t^2$. Compare
the R-squared with that from (10.35). What do you conclude?
(iii) Reestimate equation (10.35) but add $t^3$ to the equation. Is this additional term sta-
tistically significant?

C7 Use the data set CONSUMP.RAW for this exercise.
(i) Estimate a simple regression model relating the growth in real per capita con-
sumption (of nondurables and services) to the growth in real per capita disposable
income. Use the change in the logarithms in both cases. Report the results in the
usual form. Interpret the equation and discuss statistical significance.
(ii) Add a lag of the growth in real per capita disposable income to the equation
from part (i). What do you conclude about adjustment lags in consumption
growth?
(iii) Add the real interest rate to the equation in part (i). Does it affect consumption
growth?

C8 Use the data in FERTIL3.RAW for this exercise.
(i) Add $pe_{t-3}$ and $pe_{t-4}$ to equation (10.19). Test for joint significance of these lags.
(ii) Find the estimated long-run propensity and its standard error in the model from
part (i). Compare these with those obtained from equation (10.19).
(iii) Estimate the polynomial distributed lag model from Problem 6. Find the estimated
LRP and compare this with what is obtained from the unrestricted model.

C9 Use the data in VOLAT.RAW for this exercise. The variable $rsp500$ is the monthly
return on the Standard & Poor’s 500 stock market index, at an annual rate. (This in-
cludes price changes as well as dividends.) The variable $i3$ is the return on three-month
treasury bills (T-bills), and $pcip$ is the percentage change in industrial production; these
are also at an annual rate.
(i) Consider the equation

$$rsp500_t = \beta_0 + \beta_1 pcip_t + \beta_2 i3_t + u_t.$$

What signs do you think $\beta_1$ and $\beta_2$ should have?
(ii) Estimate the previous equation by OLS, reporting the results in standard form. Interpret the signs and magnitudes of the coefficients.

(iii) Which of the variables is statistically significant?

(iv) Does your finding from part (iii) imply that the return on the S&P 500 is predictable? Explain.

C10 Consider the model estimated in (10.15); use the data in INTDEF.RAW.

(i) Find the correlation between inf and def over this sample period and comment.

(ii) Add a single lag of inf and def to the equation and report the results in the usual form.

(iii) Compare the estimated LRP for the effect of inflation with that in equation (10.15). Are they vastly different?

(iv) Are the two lags in the model jointly significant at the 5% level?

C11 The file TRAFFIC2.RAW contains 108 monthly observations on car accidents, traffic laws, and some other variables in a region of a country from January 1981 through December 1989. Use this data set to answer the following questions.

(i) During what month and year did the region’s seat belt law take effect? When did the highway speed limit increase to 65 kilometres per hour?

(ii) Regress the variable log(totacc) on a linear time trend and 11 monthly dummy variables, using January as the base month. Interpret the coefficient estimate on the time trend. Would you say there is seasonality in total accidents?

(iii) Add to the regression from part (ii) the variables wkends, unem, spdlaw, and beltlaw. Discuss the coefficient on the unemployment variable. Does its sign and magnitude make sense to you?

(iv) In the regression from part (iii), interpret the coefficients on spdlaw and beltlaw. Are the estimated effects what you expected? Explain.

(v) The variable prcfat is the percentage of accidents resulting in at least one fatality. Note that this variable is a percentage, not a proportion. What is the average of prcfat over this period? Does the magnitude seem about right?

(vi) Run the regression in part (iii) but use prcfat as the dependent variable in place of log(totacc). Discuss the estimated effects and significance of the speed and seat belt law variables.

C12 Use the data in MINWAGE.RAW for this exercise. In particular, use the employment and wage series for sector 232 (Men’s and Boy’s Furnishings). The variable gwage232 is the monthly growth (change in logs) in the average wage in sector 232, gemp232 is the growth in employment in sector 232, gmwage is the growth in a government minimum wage, and gcpi is the growth in the (urban) Consumer Prices Index.

(i) Run the regression gwage232 on gmwage, gcpi. Do the sign and magnitude of $\hat{\beta}_{gmwage}$ make sense to you? Explain. Is gmwage statistically significant?

(ii) Add lags one through 12 of gmwage to the equation in part (i). Do you think it is necessary to include these lags to estimate the long-run effect of minimum wage growth on wage growth in sector 232? Explain.

(iii) Run the regression gemp232 on gmwage, gcpi. Does minimum wage growth appear to have a contemporaneous effect on gemp232?

(iv) Add lags one through 12 to the employment growth equation. Does growth in the minimum wage have a statistically significant effect on employment growth, either in the short run or the long run? Explain.
Chapter 11

Problems

1 Let \( \{x_t; \ t = 1, 2, \ldots\} \) be a covariance stationary process and define \( \gamma_h = \text{Cov}(x_t, x_{t+h}) \) for \( h \geq 0 \). [Therefore, \( \gamma_0 = \text{Var}(x_t) \).] Show that \( \text{Corr}(x_t, x_{t+h}) = \gamma_h/\gamma_0 \).

2 Let \( \{e_t; \ t = -1, 0, 1, \ldots\} \) be a sequence of independent, identically distributed random variables with mean zero and variance one. Define a stochastic process by \( x_t = e_t - (1/2)e_{t-1} + (1/2)e_{t-2}, \ t = 1, 2, \ldots \).

(i) Find \( \text{E}(x_t) \) and \( \text{Var}(x_t) \). Do either of these depend on \( t \)?

(ii) Show that \( \text{Corr}(x_t, x_{t+1}) = -1/2 \) and \( \text{Corr}(x_t, x_{t+2}) = 1/3 \). (Hint: It is easiest to use the formula in Problem 1.)

(iii) What is \( \text{Corr}(x_t, x_{t+h}) \) for \( h > 2 \)?

(iv) Is \( \{x_t\} \) an asymptotically uncorrelated process?

3 Suppose that a time series process \( \{y_t\} \) is generated by \( y_t = z \cdot e_t \), for all \( t = 1, 2, \ldots \), where \( \{e_t\} \) is an i.i.d. sequence with \( \text{E}(e_t) = 1 \) and \( \text{Var}(e_t) = 2 \). The random variable \( z \) does not change over time; it also has \( \text{E}(z) = 1 \), \( \text{Var}(z) = 2 \), and is independent of \( \{e_t\} \).

(i) Find the expected value and variance of \( y_t \). Does it depend on \( t \)?

(ii) Find \( \text{Var}(y_t) \). Does it depend on \( t \)?

(iii) Find \( \text{Cov}(y_t, y_s) \) for \( t \neq s \).

(iv) Is the sequence \( \{y_t\} \) asymptotically uncorrelated?

4 Let \( \{y_t; \ t = 1, 2, \ldots\} \) follow a random walk, as in (11.20), with \( y_0 = 0 \). Show that \( \text{Corr}(y_t, y_{t+h}) = \rho(t + h) \) for \( t \geq 1, h > 0 \).

5 For an economy, let \( g\text{price} \) denote the monthly growth in the overall price level and let \( gw\text{age} \) be the monthly growth in hourly wages. [These are both obtained as differences of logarithms: \( g\text{price} = \Delta \log(\text{price}) \) and \( gw\text{age} = \Delta \log(\text{wage}) \).] Using the monthly data in WAGEPRC.RAW, we estimate the following distributed lag model:

\[
g\text{price} = -0.00093 + 0.119 gw\text{age} + 0.097 gw\text{age}_{-1} + 0.040 gw\text{age}_{-2}
\]

\[
\begin{array}{cccc}
\text{gwage} & \text{gwage}_{-1} & \text{gwage}_{-2} \\
\text{gwage} & 0.052 & 0.039 & 0.039 \\
0.039 & 0.095 & 0.039 & 0.039 \\
0.107 & 0.039 & 0.039 & 0.095 \\
0.103 & 0.159 & 0.110 & 0.103 \\
0.039 & 0.039 & 0.039 & 0.039 \\
0.103 & 0.016 & 0.103 & 0.103 \\
0.039 & 0.039 & 0.039 & 0.039 \\
\end{array}
\]

\( n = 273, R^2 = .317, R^2 = .283. \)

(i) Sketch the estimated lag distribution. At what lag is the effect of \( gw\text{age} \) on \( g\text{price} \) largest? Which lag has the smallest coefficient?

(ii) For which lags are the \( t \) statistics less than two?

(iii) What is the estimated long-run propensity? Is it much different than one? Explain what the LRP tells us in this example.
(iv) What regression would you run to obtain the standard error of the LRP directly?

(v) How would you test the joint significance of six more lags of gwage? What would be the dfs in the F distribution? (Be careful here; you lose six more observations.)

6 Let $hy_6$ denote the three-month holding yield (in percent) from buying a six-month T-bill at time $(t-1)$ and selling it at time $t$ (three months hence) as a three-month T-bill. Let $hy_3_{t-1}$ be the three-month holding yield from buying a three-month T-bill at time $(t-1)$. At time $(t-1)$, $hy_3_{t-1}$ is known, whereas $hy_6$ is unknown because $p_3$, the price of three-month T-bills is unknown at time $(t-1)$. The expectations hypothesis (EH) says that these two different three-month investments should be the same, on average. Mathematically, we can write this as a conditional expectation:

$$E(hy_6 | I_{t-1}) = hy_3_{t-1},$$

where $I_{t-1}$ denotes all observable information up through time $t-1$. This suggests estimating the model

$$hy_6 = \beta_0 + \beta_1 hy_3_{t-1} + u_t,$$

and testing $H_0: \beta_1 = 1$. (We can also test $H_0: \beta_0 = 0$, but we often allow for a term premium for buying assets with different maturities, so that $\beta_0 \neq 0$.)

(i) Estimating the previous equation by OLS using the data in INTQRT.RAW (spaced every three months) gives

$$\hat{hy_6} = -0.058 + 1.104 hy_3_{t-1}
\begin{pmatrix}
(.070) & (.039)
\end{pmatrix}
\begin{pmatrix}
n = 123, R^2 = .866.
\end{pmatrix}

Do you reject $H_0: \beta_1 = 1$ against $H_0: \beta_1 \neq 1$ at the 1% significance level? Does the estimate seem practically different from one?

(ii) Another implication of the EH is that no other variables dated as $t-1$ or earlier should help explain $hy_6$, once $hy_3_{t-1}$ has been controlled for. Including one lag of the spread between six-month and three-month T-bill rates gives

$$\hat{hy_6} = -0.123 + 1.053 hy_3_{t-1} + .480 (r6_{t-1} - r3_{t-1})
\begin{pmatrix}
(.067) & (.039) & (.109)
\end{pmatrix}
\begin{pmatrix}
n = 123, R^2 = .885.
\end{pmatrix}

Now, is the coefficient on $hy_3_{t-1}$ statistically different from one? Is the lagged spread term significant? According to this equation, if, at time $t-1$, $r6$ is above $r3$, should you invest in six-month or three-month T-bills?

(iii) The sample correlation between $hy_3$ and $hy_3_{t-1}$ is .914. Why might this raise some concerns with the previous analysis?

(iv) How would you test for seasonality in the equation estimated in part (ii)?

7 A partial adjustment model is

$$y_t = y_0 + \gamma_1 x_t + \epsilon_t,$$

$$y_t - y_{t-1} = \lambda (y_t^* - y_{t-1}) + a_t,$$
where $y_t^*$ is the desired or optimal level of $y$, and $y_t$ is the actual (observed) level. For example, $y_t^*$ is the desired growth in firm inventories, and $x_t$ is growth in firm sales. The parameter $\gamma_t$ measures the effect of $x_t$ on $y_t^*$. The second equation describes how the actual $y_t$ adjusts depending on the relationship between the desired $y_t$ in time $t$ and the actual $y_t$ in time $t-1$. The parameter $\lambda$ measures the speed of adjustment and satisfies $0 < \lambda < 1$.

(i) Plug the first equation for $y_t^*$ into the second equation and show that we can write

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 x_t + u_t.$$ 

In particular, find the $\beta_j$ in terms of the $g_j$ and $l$ and find $u_t$ in terms of $e_t$ and $a_t$.

Therefore, the partial adjustment model leads to a model with a lagged dependent variable and a contemporaneous $x$.

(ii) If $E(e_t|x_t,y_{t-1},x_{t-1},\ldots) = E(a_t|x_t,y_{t-1},x_{t-1},\ldots) = 0$ and all series are weakly dependent, how would you estimate the $\beta_j$?

(iii) If $\hat{\beta}_1 = .7$ and $\hat{\beta}_2 = .2$, what are the estimates of $g_1$ and $\lambda$?

### Computer Exercises

**C1** Use the data in HSEINV.RAW for this exercise.

(i) Find the first order autocorrelation in log($invpc$). Now, find the autocorrelation after linearly detrending log($invpc$). Do the same for log($price$). Which of the two series may have a unit root?

(ii) Based on your findings in part (i), estimate the equation

$$\log(invpc_t) = \beta_0 + \beta_1 \Delta \log(price_t) + \beta_2 t + u_t$$

and report the results in standard form. Interpret the coefficient $\hat{\beta}_1$ and determine whether it is statistically significant.

(iii) Linearly detrend log($invpc$) and use the detrended version as the dependent variable in the regression from part (ii) (see Section 10.5). What happens to $R^2$?

(iv) Now use $\Delta \log(invpc)$ as the dependent variable. How do your results change from part (ii)? Is the time trend still significant? Why or why not?

**C2** In Example 11.7, define the growth in hourly wage and output per hour as the change in the natural log:

$$ghrwage = \Delta \log(hrwage) \quad \text{and} \quad goutphr = \Delta \log(outphr).$$

Consider a simple extension of the model estimated in (11.29):

$$ghrwage_t = \beta_0 + \beta_1 goutphr_t + \beta_2 goutphr_{t-1} + u_t.$$ 

This allows an increase in productivity growth to have both a current and lagged effect on wage growth.

(i) Estimate the equation using the data in EARNS.RAW and report the results in standard form. Is the lagged value of $goutphr$ statistically significant?

(ii) If $\beta_1 + \beta_2 = 1$, a permanent increase in productivity growth is fully passed on in higher wage growth after one year. Test $H_0: \beta_1 + \beta_2 = 1$ against the two-sided alternative. Remember, one way to do this is to write the equation so that $\theta = \beta_1 + \beta_2$ appears directly in the model, as in Example 10.4 from Chapter 10 of the textbook.

(iii) Does $goutphr_{t-2}$ need to be in the model? Explain.
C3

(i) In Example 11.4 of the textbook, it may be that the expected value of the return at time $t$, given past returns, is a quadratic function of $\text{return}_{t-1}$. To check this possibility, use the data in NYSE.RAW to estimate

$$\text{return}_t = \beta_0 + \beta_1 \text{return}_{t-1} + \beta_2 \text{return}^2_{t-1} + u_t,$$

report the results in standard form.

(ii) State and test the null hypothesis that $E(\text{return}_t | \text{return}_{t-1})$ does not depend on $\text{return}_{t-1}$. (Hint: There are two restrictions to test here.) What do you conclude?

(iii) Drop $\text{return}^2_{t-1}$ from the model, but add the interaction term $\text{return}_{t-1} \cdot \text{return}_{t-2}$. Now test the efficient markets hypothesis.

(iv) What do you conclude about predicting weekly stock returns based on past stock returns?

C4

Use the data in PHILLIPS.RAW for this exercise, but only through 1996.

(i) In Example 11.5, we assumed that the natural rate of unemployment is constant. An alternative form of the expectations augmented Phillips curve allows the natural rate of unemployment to depend on past levels of unemployment. In the simplest case, the natural rate at time $t$ equals $\text{unem}_{t-1}$. If we assume adaptive expectations, we obtain a Phillips curve where inflation and unemployment are in first differences:

$$\Delta \text{inf} = \beta_0 + \beta_1 \Delta \text{unem} + u.$$

Estimate this model, report the results in the usual form, and discuss the sign, size, and statistical significance of $\hat{\beta}_1$.

(ii) Which model fits the data better, (11.19) or the model from part (i)? Explain.

C5

(i) Add a linear time trend to equation (11.27). Is a time trend necessary in the first-difference equation?

(ii) Drop the time trend and add the variables $\text{ww2}$ and $\text{pill}$ to (11.27) (do not difference these dummy variables). Are these variables jointly significant at the 5% level?

(iii) Using the model from part (ii), estimate the LRP and obtain its standard error. Compare this to (10.19), where $\text{gfr}$ and $\text{pe}$ appeared in levels rather than in first differences.

C6

Let $\text{inven}$ be the real value inventories in Germany during year $t$, let $\text{GDP}$ denote real gross domestic product, and let $r3_t$ denote the (ex post) real interest rate on three-month T-bills. The ex post real interest rate is (approximately) $r3_t = i3_t - \text{inf}_t$, where $i3_t$ is the rate on three-month T-bills and $\text{inf}_t$ is the annual inflation rate. The change in inventories, $\Delta \text{inven}$, is the inventory investment for the year. The accelerator model of inventory investment relates $\text{cinven}$ to the $c\text{GDP}$, the change in $\text{GDP}$:

$$\Delta \text{inven} = \beta_0 + \beta_1 \Delta \text{GDP}_t + u_t,$$

where $\beta_1 > 0$.

(i) Use the data in INVEN.RAW to estimate the accelerator model. Report the results in the usual form and interpret the equation. Is $\hat{\beta}_1$ statistically greater than zero?

(ii) If the real interest rate rises, then the opportunity cost of holding inventories rises, and so an increase in the real interest rate should decrease inventories. Add the real interest rate to the accelerator model and discuss the results.
(iii) Does the level of the real interest rate work better than the first difference, $\Delta r_3$?

C7 Use CONSUMP.RAW for this exercise. One version of the permanent income hypothesis (PIH) of consumption is that the growth in consumption is unpredictable. (Another version is that the change in consumption itself is unpredictable; see Mankiw [1994, Chapter 15] for discussion of the PIH.) Let $gc_t = \log(c_t) - \log(c_{t-1})$ be the growth in real per capita consumption (of nondurables and services). Then the PIH implies that $E(gc_t | I_{t-1}) = E(gc_t)$, where $I_{t-1}$ denotes information known at time $(t - 1)$; in this case, $t$ denotes a year.

(i) Test the PIH by estimating $gc_t = \beta_0 + \beta_1 gc_{t-1} + u_t$. Clearly state the null and alternative hypotheses. What do you conclude?

(ii) To the regression in part (i), add $g_y_{t-1}$ and $i_3_{t-1}$, where $g_y$ is the growth in real per capita disposable income and $i_3$ is the interest rate on three-month T-bills; note that each must be lagged in the regression. Are these two additional variables jointly significant?

C8 Use the data in PHILLIPS.RAW for this exercise.

(i) Estimate an AR(1) model for the unemployment rate. Use this equation to predict the unemployment rate for 2004.

(ii) Add a lag of inflation to the AR(1) model from part (i). Is $inf_{t-1}$ statistically significant?

(iii) Use the equation from part (ii) to predict the unemployment rate for 2004. Is the result better or worse than in the model from part (i)?

(iv) Use the method from Section 6.4 to construct a 95% prediction interval for the 2004 unemployment rate.

C9 Use the data in TRAFFIC2.RAW for this exercise. Computer Exercise C11 in Chapter 10 previously asked for an analysis of these data.

(i) Compute the first order autocorrelation coefficient for the variable $prcfat$. Are you concerned that $prcfat$ contains a unit root? Do the same for the unemployment rate.

(ii) Estimate a multiple regression model relating the first difference of $prcfat$, $\Delta prcfat$, to the same variables in part (vi) of Computer Exercise C11 in Chapter 10, except you should first difference the unemployment rate, too. Then, include a linear time trend, monthly dummy variables, the weekend variable, and the two policy variables; do not difference these. Do you find any interesting results?

(iii) Comment on the following statement: “We should always first difference any time series we suspect of having a unit root before doing multiple regression because it is the safe strategy and should give results similar to using the levels.” [In answering this, you may want to do the regression from part (vi) of Computer Exercise C11 in Chapter 10, if you have not already.]

C10 Use the data in MINWAGE.RAW for this exercise, focusing on the wage and employment series for sector 232 (Men’s and Boy’s Furnishings). The variable $gwage232$ is the monthly growth (change in logs) in the average wage in sector 232, $gemp232$ is the growth in employment in sector 232, $gmwage$ is the growth in the federal minimum wage, and $gcpi$ is the growth in the (urban) Consumer Prices Index.

(i) Find the first order autocorrelation in $gwage232$. Does this series appear to be weakly dependent?

(ii) Estimate the dynamic model

$$gwage232_t = \beta_0 + \beta_1 gwage232_{t-1} + \beta_2 gmwage_t + \beta_3 gcpi_t + u_t$$
by OLS. Holding fixed last month’s growth in wage and the growth in the CPI, does an increase in a government minimum wage result in a contemporaneous increase in \( gwage_{232} \)? Explain.

(iii) Now add the lagged growth in employment, \( gemp_{232,-1} \), to the equation in part (ii). Is it statistically significant?

(iv) Compared with the model without \( gwage_{232,-1} \) and \( gemp_{232,-1} \), does adding the two lagged variables notably change the estimated effect of the minimum wage variable?

(v) Run the regression of \( gmwaged \) on \( gwage_{232,-1} \) and \( gemp_{232,-1} \), and report the \( R \)-squared. Comment on how the value of \( R \)-squared helps explain your answer to part (iv).
Chapter 12

Problems

1 When the errors in a regression model have AR(1) serial correlation, why do the OLS standard errors tend to underestimate the sampling variation in the $\hat{\beta}$? Is it always true that the OLS standard errors are too small?

2 Evaluate the following statement: “Practically important differences between the OLS and Prais-Winsten estimates indicate there is something wrong with OLS.”

3 In Example 10.6, we estimated a variant on Fair’s model for predicting presidential election outcomes in the United States.
   (i) What argument can be made for the error term in this equation being serially uncorrelated? (*Hint:* How often do presidential elections take place?)
   (ii) When the OLS residuals from (10.23) are regressed on the lagged residuals, we obtain $\hat{\rho} = -.068$ and $se(\hat{\rho}) = .240$. What do you conclude about serial correlation in the $u_t$?
   (iii) Does the small sample size in this application worry you in testing for serial correlation?

4 True or false: “If the errors in a regression model contain ARCH, they must be serially correlated.”

5 (i) In the enterprise zone event study in Computer Exercise C5 in Chapter 10, a regression of the OLS residuals on the lagged residuals produces $\hat{\rho} = .841$ and $se(\hat{\rho}) = .053$. What implications does this have for OLS?
   (ii) If you want not only to use OLS but also want to obtain a valid standard error for the EZ coefficient, what would you do?

6 In Example 12.8, we found evidence of heteroskedasticity in $u_t$ in equation (12.47). Thus, we compute the heteroskedasticity-robust standard errors (in $[\cdot]$) along with the usual standard errors:

\[
\begin{align*}
\text{return}_t &= .180 + .059 \text{return}_{t-1} \\
&= (.081) \ (0.038) \\
&= [.085] \ [.069]
\end{align*}
\]

$n = 689, R^2 = .0035, \tilde{R}^2 = .0020.$

What does using the heteroskedasticity-robust $t$ statistic do to the significance of $\text{return}_{t-1}$?

Computer Exercises

C1 In Example 11.6, we estimated a finite DL model in first differences:

\[
\Delta gfr_t = \gamma_0 + \delta_0 \Delta pe_t + \delta_1 \Delta pe_{t-1} + \delta_2 \Delta pe_{t-2} + u_t.
\]

Use the data in FERTIL3.RAW to test whether there is AR(1) serial correlation in the errors.
C2  (i) Using the data in WAGEPRC.RAW, estimate the distributed lag model from Problem 5 in Chapter 11. Use regression (12.14) to test for AR(1) serial correlation. (ii) Reestimate the model using iterated Cochrane-Orcutt estimation. What is your new estimate of the long-run propensity? (iii) Using iterated CO, find the standard error for the LRP. (This requires you to estimate a modified equation.) Determine whether the estimated LRP is statistically different from one at the 5% level.

C3  (i) In part (i) of Computer Exercise C6 in Chapter 11, you were asked to estimate the accelerator model for inventory investment. Test this equation for AR(1) serial correlation. (ii) If you find evidence of serial correlation, reestimate the equation by Cochrane-Orcutt and compare the results.

C4  (i) Use NYSE.RAW to estimate equation (12.48). Let $\hat{h}_t$ be the fitted values from this equation (the estimates of the conditional variance). How many $\hat{h}_t$ are negative? (ii) Add $\text{return}_{t-1}^2$ to (12.48) and again compute the fitted values, $\hat{h}_t$. Are any $\hat{h}_t$ negative? (iii) Use the $\hat{h}_t$ from part (ii) to estimate (12.47) by weighted least squares (as in Section 8.4). Compare your estimate of $\beta_1$ with that in equation (11.16). Test $H_0: \beta_1 = 0$ and compare the outcome when OLS is used. (iv) Now, estimate (12.47) by WLS, using the estimated ARCH model in (12.51) to obtain the $\hat{h}_t$. Does this change your findings from part (iii)?

C5  Consider the version of Fair’s model in Example 10.6. Now, rather than predicting the proportion of the two-party vote received by the Democrat, estimate a linear probability model for whether or not the Democrat wins. (i) Use the binary variable $\text{demwins}$ in place of $\text{demvote}$ in (10.23) and report the results in standard form. Which factors affect the probability of winning? Use the data only through 1992. (ii) How many fitted values are less than zero? How many are greater than one? (iii) Use the following prediction rule: if $\text{demwins} > .5$, you predict the Democrat wins; otherwise, the Republican wins. Using this rule, determine how many of the 20 elections are correctly predicted by the model. (iv) Plug in the values of the explanatory variables for 1996. What is the predicted probability that Clinton would win the election? Clinton did win; did you get the correct prediction? (v) Use a heteroskedasticity-robust $t$ test for AR(1) serial correlation in the errors. What do you find? (vi) Obtain the heteroskedasticity-robust standard errors for the estimates in part (i). Are there notable changes in any $t$ statistics?

C6  (i) In Computer Exercise C7 in Chapter 10, you estimated a simple relationship between consumption growth and growth in disposable income. Test the equation for AR(1) serial correlation (using CONSUMP.RAW). (ii) In Computer Exercise C7 in Chapter 11, you tested the permanent income hypothesis by regressing the growth in consumption on one lag. After running this regression, test for heteroskedasticity by regressing the squared residuals on $gc_{t-1}^2$ and $gc_{t-1}^2$. What do you conclude?

C7  (i) For Example 12.4, using the data in BARIUM.RAW, obtain the iterative Cochrane-Orcutt estimates.
(ii) Are the Prais-Winsten and Cochrane-Orcutt estimates similar? Did you expect them to be?

C8 Use the data in TRAFFIC2.RAW for this exercise.
(i) Run an OLS regression of prcfat on a linear time trend, monthly dummy variables, and the variables wkends, unem, spdlaw, and beltlaw. Test the errors for AR(1) serial correlation using the regression in equation (12.14). Does it make sense to use the test that assumes strict exogeneity of the regressors?
(ii) Obtain serial correlation- and heteroskedasticity-robust standard errors for the coefficients on spdlaw and beltlaw, using four lags in the Newey-West estimator. How does this affect the statistical significance of the two policy variables?
(iii) Now, estimate the model using iterative Prais-Winsten and compare the estimates with the OLS estimates. Are there important changes in the policy variable coefficients or their statistical significance?

C9 Assume the file FISH.RAW contains 97 daily price and quantity observations on fish prices at the Billingsgate fish market in London. Use the variable log(avgprc) as the dependent variable.
(i) Regress log(avgprc) on four daily dummy variables, with Friday as the base. Include a linear time trend. Is there evidence that price varies systematically within a week?
(ii) Now, add the variables wave2 and wave3, which are measures of wave heights over the past several days. Are these variables individually significant? Describe a mechanism by which stormier seas would increase the price of fish.
(iii) What happened to the time trend when wave2 and wave3 were added to the regression? What must be going on?
(iv) Explain why all explanatory variables in the regression are safely assumed to be strictly exogenous.
(v) Test the errors for AR(1) serial correlation.
(vi) Obtain the Newey-West standard errors using four lags. What happens to the \( t \) statistics on wave2 and wave3? Did you expect a bigger or smaller change compared with the usual OLS \( t \) statistics?
(vii) Now, obtain the Prais-Winsten estimates for the model estimated in part (ii). Are wave2 and wave3 jointly statistically significant?

C10 Use the data in PHILLIPS.RAW to answer these questions.
(i) Using the entire data set, estimate the static Phillips curve equation \( \inf_t = \beta_0 + \beta_1 \text{unem}_t + u_t \) by OLS and report the results in the usual form.
(ii) Obtain the OLS residuals from part (i), \( \hat{u}_t \), and obtain \( \rho \) from the regression \( \hat{u}_t \) on \( \hat{u}_{t-1} \). (It is fine to include an intercept in this regression.) Is there strong evidence of serial correlation?
(iii) Now estimate the static Phillips curve model by iterative Prais-Winsten. Compare the estimate of \( \beta_1 \) with that obtained in Table 12.2. Is there much difference in the estimate when the later years are added?
(iv) Rather than using Prais-Winsten, use iterative Cochrane-Orcutt. How similar are the final estimates of \( \rho \)? How similar are the PW and CO estimates of \( \beta_1 \)?

C11 Use the data in NYSE.RAW to answer these questions.
(i) Estimate the model in equation (12.47) and obtain the squared OLS residuals. Find the average, minimum, and maximum values of \( \hat{u}_t^2 \) over the sample.
(ii) Use the squared OLS residuals to estimate the following model of heteroskedasticity:
\[ \text{Var}(u_t | return_{t-1}, return_{t-2}, \ldots) = \delta_0 + \delta_1 return_{t-1} + \delta_2 return_{t-1}^2. \]
Report the estimated coefficients, the reported standard errors, the R-squared, and the adjusted R-squared.

(iii) Sketch the conditional variance as a function of the lagged return\(_{-1}\). For what value of return\(_{-1}\) is the variance the smallest, and what is the variance?

(iv) For predicting the dynamic variance, does the model in part (ii) produce any negative variance estimates?

(v) Does the model in part (ii) seem to fit better or worse than the ARCH(1) model in Example 12.9? Explain.

(vi) To the ARCH(1) regression in equation (12.51), add the second lag, \(u_{t-2}^2\). Does this lag seem important? Does the ARCH(2) model fit better than the model in part (ii)?

C12 Use the data in MINWAGE.RAW for this exercise. As in Computer Exercise C12 in Chapter 10, use the series for sector 232 (Men’s and Boy’s Furnishings).

(i) Estimate the model
\[ gwage_{232t} = \beta_0 + \beta_1 gwage_t + \beta_2 gcpi_t + u_t \]
by OLS and test the errors for AR(1) serial correlation. Assume that the regressors are strictly exogenous. Are the errors positively or negatively serially correlated?

(ii) Estimate the model in part (i) by iterative Prais-Winsten. How does the coefficient on gwage compare with the OLS estimate?

(iii) Add lags one through 12 of gwage to the equation in part (ii) and estimate the model by iterative Prais-Winsten. Test whether the 12 lags are jointly significant.

(iv) Go back to the static model estimated in part (i), but compute the Newey-West standard errors using a lag equal to six. How does the Newey-West standard error compare with the usual OLS standard error?

(v) Add lags one through 12 of gwage to the static model and check for joint significance of the 12 lags using the Newey-West test with six lags. Does your conclusion differ from part (iii)?
Chapter 13

Problems

1 In Example 13.1, assume that the averages of all factors other than educ have remained constant over time and that the average level of education is 12.2 for the 1972 sample and 13.3 in the 1984 sample. Using the estimates in Table 13.1, find the estimated change in average fertility between 1972 and 1984. (Be sure to account for the intercept change and the change in average education.)

2 Using the data in KIELMC.RAW, the following equations were estimated using the years 1978 and 1981:

\[
\log(\text{price}) = 11.49 - 0.547 \text{nearinc} + 0.394 y81 \cdot \text{nearinc} \\
(0.26) (0.058) (0.080)
\]

\(n = 321, R^2 = 0.220\)

and

\[
\log(\text{price}) = 11.18 + 0.563 y81 - 0.403 y81 \cdot \text{nearinc} \\
(0.27) (0.044) (0.067)
\]

\(n = 321, R^2 = 0.337\).

Compare the estimates on the interaction term y81 \cdot \text{nearinc} with those from equation (13.9). Why are the estimates so different?

3 Why can we not use first differences when we have independent cross sections in two years (as opposed to panel data)?

4 If we think that \(\beta_1\) is positive in (13.14) and that \(\Delta u_i\) and \(\Delta u\) are negatively correlated, what is the bias in the OLS estimator of \(\beta_1\) in the first-differenced equation? [Hint: Review equation (5.4).]

5 Suppose that we want to estimate the effect of several variables on annual saving and that we have a panel data set on individuals collected on January 31, 1990, and January 31, 1992. If we include a year dummy for 1992 and use first differencing, can we also include age in the original model? Explain.

6 Suppose that in 2003 a region in north east England with two sixth form colleges was given a sharp increase in funding. Call the colleges Bladen and Gallowgate. The local authority decided that the funding would go to Gallowgate, mostly to hire more teachers to reduce class sizes. To evaluate the effectiveness of the increased spending, the local authority collected random samples of final year students in 2002 and 2004 at both colleges. The measure of student performance is a compulsory standardised test score given to all final year students in the region.

(i) Let score be the test score. Without controlling for other factors, write down a model that allows you to determine whether the additional funding led to improved test scores. Be sure to state which coefficient measures the effect of the new funding.

(ii) How might estimating the model from (i) be biased if Gallowgate gets an influx of better students because of the increased funding?
(iii) Suppose you also can collect information on a test given when all students were 16 years old. How might this help isolate the spending effects given the potential problem in (ii)? What other factors might you control for?

7 (i) Using the data in INJURY.RAW, we find the estimated equation when afchnge is dropped from (13.12) is

\[
\log(\text{durat}) = 1.129 + .253 \text{ highearn} + .198 \text{ afchnge-highearn}
\]

(0.022) (0.042) (0.052)

\[ n = 5,626, R^2 = .021. \]

Is it surprising that the estimate on the interaction is fairly close to that in (13.12)? Explain.

(ii) When afchnge is included but highearn is dropped, the result is

\[
\log(\text{durat}) = 1.233 - .100 \text{ afchnge} + .447 \text{ afchnge-highearn}
\]

(0.023) (0.040) (0.050)

\[ n = 5,626, R^2 = .016. \]

Why is the coefficient on the interaction term now so much larger than in (13.12)? [Hint: In equation (13.10), what is the assumption being made about the treatment and control groups if \( \beta_1 = 0 \)?]

**Computer Exercises**

**C1** Use the data in FERTIL1.RAW for this exercise.

(i) In the equation estimated in Example 13.1, test whether living environment at age 16 has an effect on fertility. (The base group is a large city.) Report the value of the \( F \) statistic and the \( p \)-value.

(ii) Test whether the region of the country at age 16 (South is the base group) has an effect on fertility.

(iii) Let \( u \) be the error term in the population equation. Suppose you think that the variance of \( u \) changes over time (but not with educ, age, and so on). A model that captures this is

\[ u^2 = \gamma_0 + \gamma_1 y74 + \gamma_2 y76 + \ldots + \gamma_6 y84 + v. \]

Using this model, test for heteroskedasticity in \( u \). (Hint: Your \( F \) test should have 6 and 1,122 degrees of freedom.)

(iv) Add the interaction terms \( y74 \cdot \text{educ}, y76 \cdot \text{educ}, \ldots, y84 \cdot \text{educ} \) to the model estimated in Table 13.1. Explain what these terms represent. Are they jointly significant?

**C2** Use the data in CPS78_85.RAW for this exercise.

(i) How do you interpret the coefficient on \( y85 \) in equation (13.2)? Does it have an interesting interpretation? (Be careful here; you must account for the interaction terms \( y85 \cdot \text{educ} \) and \( y85 \cdot \text{female} \).)

(ii) Holding other factors fixed, what is the estimated percent increase in nominal wage for a male with 12 years of education? Propose a regression to obtain a confidence interval for this estimate. [Hint: To get the confidence interval, replace \( y85 \cdot \text{educ} \) with \( y85 \cdot (\text{educ} - 12) \); refer to Example 6.3.]

(iii) Reestimate equation (13.2) but let all wages be measured in 1978 pounds. In particular, define the real wage as \( \text{rwage} = \text{wage} \) for 1978 and as \( \text{rwage} = \text{wage}/1.65 \)
for 1985. Now, use log(rwage) in place of log(wage) in estimating (13.2). Which coefficients differ from those in equation (13.2)?

(iv) Explain why the R-squared from your regression in part (iii) is not the same as in equation (13.2). (Hint: The residuals, and therefore the sum of squared residuals, from the two regressions are identical.)


(vi) Starting with equation (13.2), test whether the union wage differential changed over time. (This should be a simple t test.)

(vii) Do your findings in parts (v) and (vi) conflict? Explain.

C3 Use the data in KIELMC.RAW for this exercise.

(i) The variable dist is the distance from each home to the incinerator site, in feet. Consider the model

\[
\log(\text{price}) = \beta_0 + \delta_0 \gamma 81 + \beta_1 \log(\text{dist}) + \delta_1 \gamma 81 \cdot \log(\text{dist}) + u.
\]

If building the incinerator reduces the value of homes closer to the site, what is the sign of \(\delta_1\)? What does it mean if \(\beta_1 > 0\)?

(ii) Estimate the model from part (i) and report the results in the usual form. Interpret the coefficient on \(\gamma 81 \cdot \log(\text{dist})\). What do you conclude?

(iii) Add age, age^2, rooms, baths, log(inst), log(land), and log(area) to the equation. Now, what do you conclude about the effect of the incinerator on housing values?

C4 Use the data in INJURY.RAW for this exercise.

(i) Using the data for region 1, reestimate equation (13.12), adding as explanatory variables male, married, and a full set of industry and injury type dummy variables. How does the estimate on \(\text{afchnge} \cdot \text{highearn}\) change when these other factors are controlled for? Is the estimate still statistically significant?

(ii) What do you make of the small R-squared from part (i)? Does this mean the equation is useless?

(iii) Estimate equation (13.12) using the data for region 2. Compare the estimates on the interaction term for region 2 and region 1. Is the region 2 estimate statistically significant? What do you make of this?

C5 Use the data in RENTAL.RAW for this exercise. The data for the years 1980 and 1990 include rental prices and other variables for university towns. The idea is to see whether a stronger presence of students affects rental rates. The unobserved effects model is

\[
\log(\text{rent}) = \beta_0 + \delta_0 \gamma 90 + \beta_1 \log(\text{pop}) + \beta_2 \log(\text{avginc}) + \beta_3 \text{pctstu} + a_i + u_i,
\]

where \(\text{pop}\) is city population, \(\text{avginc}\) is average income, and \(\text{pctstu}\) is student population as a percentage of city population (during the university year).

(i) Estimate the equation by pooled OLS and report the results in standard form. What do you make of the estimate on the 1990 dummy variable? What do you get for \(\beta_{\text{pctstu}}\)?

(ii) Are the standard errors you report in part (i) valid? Explain.

(iii) Now, difference the equation and estimate by OLS. Compare your estimate of \(\beta_{\text{pctstu}}\) with that from part (ii). Does the relative size of the student population appear to affect rental prices?
(iv) Obtain the heteroskedasticity-robust standard errors for the first-differenced equation in part (iii). Does this change your conclusions?

C6 Use CRIME3.RAW for this exercise.
(i) In the model of Example 13.6, test the hypothesis $H_0: \beta_1 = \beta_2$. (Hint: Define $\theta_1 = \beta_1 - \beta_2$ and write $\beta_1$ in terms of $\theta_1$ and $\beta_2$. Substitute this into the equation and then rearrange. Do a t test on $\theta_1$.)
(ii) If $\beta_1 = \beta_2$, show that the differenced equation can be written as

$$\Delta \log(\text{crime}_i) = \delta_0 + \delta_1 \Delta \text{avgclr}_i + \Delta u_i,$$

where $\delta_1 = 2\beta_1$ and $\text{avgclr}_i = (\text{clrprc}_{i-1} + \text{clrprc}_{i-2})/2$ is the average clear-up percentage over the previous two years.
(iii) Estimate the equation from part (ii). Compare the adjusted $R$-squared with that in (13.22). Which model would you finally use?

C7 Use GPA3.RAW for this exercise. The data set is for 366 student-athletes from a large university for fall and spring semesters. Because you have two terms of data for each student, an unobserved effects model is appropriate. The primary question of interest is this: Do athletes perform more poorly in university during the semester their sport is in season?
(i) Use pooled OLS to estimate a model with term GPA ($\text{trmgpa}$) as the dependent variable. The explanatory variables are $\text{spring}$, $\text{sat}$, $\text{psperc}$, $\text{female}$, $\text{ethnic}$, $\text{white}$, $\text{frstsem}$, $\text{toths}$, $\text{crsgpa}$, and $\text{season}$. Interpret the coefficient on $\text{season}$. Is it statistically significant?
(ii) Most of the athletes who play their sport only in the autumn are rugby players. Suppose the ability levels of rugby players differ systematically from those of other athletes. If ability is not adequately captured by SAT score and post-16 percentile, explain why the pooled OLS estimators will be biased.
(iii) Now, use the data differenced across the two terms. Which variables drop out? Now, test for an in-season effect.
(iv) Can you think of one or more potentially important, time-varying variables that have been omitted from the analysis?

C8 VOTE2.RAW includes panel data on local government elections in 1988 and 1990. Only winners from 1988 who are also running in 1990 appear in the sample; these are the incumbents. An unobserved effects model explaining the share of the incumbent’s vote in terms of expenditures by both candidates is

$$\text{vote}_i = \beta_0 + \delta_i \text{d90}_i + \beta_1 \log(\text{inexp}_i) + \beta_2 \log(\text{chexp}_i) + \beta_3 \text{incshr}_i + a_i + u_i,$$

where $\text{incshr}_i$ is the incumbent’s share of total campaign spending (in percentage form). The unobserved effect $a_i$ contains characteristics of the incumbent—such as “quality”—as well as things about the local area that are constant. The incumbent’s gender and party are constant over time, so these are subsumed in $a_i$. We are interested in the effect of campaign expenditures on election outcomes.
(i) Difference the given equation across the two years and estimate the differenced equation by OLS. Which variables are individually significant at the 5% level against a two-sided alternative?
(ii) In the equation from part (i), test for joint significance of $\Delta \log(\text{inexp})$ and $\Delta \log(\text{chexp})$. Report the $p$-value.
(iii) Reestimate the equation from part (i) using $\Delta \text{incshr}$ as the only independent variable. Interpret the coefficient on $\Delta \text{incshr}$. For example, if the incumbent’s share of spending increases by 10 percentage points, how is this predicted to affect the incumbent’s share of the vote?

(iv) Redo part (iii), but now use only the pairs that have repeat challengers. [This allows us to control for characteristics of the challengers as well, which would be in $a$. Levitt (1994) conducts a much more extensive analysis.]

C9 Use CRIME4.RAW for this exercise.

(i) Add the logs of each wage variable in the data set and estimate the model by first differencing. How does including these variables affect the coefficients on the criminal justice variables in Example 13.9?

(ii) Do the wage variables in (i) all have the expected sign? Are they jointly significant? Explain.

C10 Use WAGEPAN to answer this question.

(i) Use pooled OLS to estimate a log(wage) equation using as explanatory variables $\text{educ}$, $\text{black}$, $\text{asian}$, $\text{exper}$, $\text{married}$, $\text{union}$, and a full set of year dummies (using 1980 as the base year). Interpret and discuss the coefficients on the $\text{married}$ and $\text{union}$ variables.

(ii) Explain why the usual standard errors reported in part (i) are almost certainly too small. Obtain standard errors that are robust to serial correlation and heteroskedasticity, and for $\text{married}$ and $\text{union}$, compare these with the usual OLS standard errors.

(iii) Now first difference the variables lwage, exper, married, and union. (The time-constant variables $\text{educ}$, $\text{black}$, and $\text{asian}$ drop out of the estimation, and so does exper because it always increases by one year.) Be sure to exclude differences for the first year, 1980, as there is no earlier year.

(iv) Run the regression $\Delta \text{lwage}_i$ on $\Delta \text{marriedit}_{i}$, $\Delta \text{unionit}_{i}$, $d_{82}t$, $d_{83}t$, ..., $d_{87}t$, $t = 2, \ldots, T$; $i = 1, \ldots, N$, being sure to include a constant. Including a constant and year dummies for 1982 through 1987 is, as far as estimating the other parameters, the same as differencing the year dummies. Report the coefficients and standard errors for $\Delta \text{marriedit}_{i}$ and $\Delta \text{unionit}_{i}$.

(v) Compare the estimated marriage and union premiums from the levels and first-difference estimations, and comment.

C11 Assume the file MATHPNL.RAW contains panel data on school districts in England for the years 1992 through 1998. The response variable of interest in this question is $\text{math10}$, the percentage of 9–10 year-olds in a district receiving a passing score on a standardised maths test. The key explanatory variable is $\text{rexpp}$, which is real expenditures per pupil in the district. The amounts are in 1997 pounds. The spending variable will appear in logarithmic form.

(i) Consider the static unobserved effects model

$$\text{math10}_i = \delta_1y93t + \ldots + \delta_6y98t + \beta_1\log(\text{rexpp}_i) + \beta_2\log(\text{enrol}_i) + \beta_3\text{lunch}_i + a_i + u_i,$$

where $\text{enrol}_i$ is total district enrollment and $\text{lunch}_i$ is the percentage of students in the district eligible for the school lunch programme. (So $\text{lunch}_i$ is a pretty
good measure of the district-wide poverty rate.) Argue that $\beta_i/10$ is the percentage point change in $\text{math10}_i$ when real per-student spending increases by roughly 10%.

(ii) Use first differencing to estimate the model in part (i). The simplest approach is to allow an intercept in the first-differenced equation and to include dummy variables for the years 1994 through 1998. Interpret the coefficient on the spending variable.

(iii) Now, add one lag of the spending variable to the model and reestimate using first differencing. Note that you lose another year of data, so you are only using changes starting in 1994. Discuss the coefficients and significance on the current and lagged spending variables.

(iv) Obtain heteroskedasticity-robust standard errors for the first-differenced regression in part (iii). How do these standard errors compare with those from part (iii) for the spending variables?

(v) Now, obtain standard errors robust to both heteroskedasticity and serial correlation. What does this do to the significance of the lagged spending variable?

(vi) Verify that the differenced errors $\Delta u_i$ have negative serial correlation by carrying out a test of AR(1) serial correlation.

(vii) Based on a fully robust joint test, does it appear necessary to include the enrollment and lunch variables in the model?

Use the data in MURDER.RAW for this exercise.

(i) Using the years 1990 and 1993, estimate the equation

$$mrdrte_{it} = \beta_0 + \beta_1 d93_t + \beta_2 \text{exec}_{it} + \beta_3 \text{unem}_{it} + a_i + u_{it}, \ t = 1, 2$$

by pooled OLS and report the results in the usual form. Do not worry that the usual OLS standard errors are inappropriate because of the presence of $a_i$. Do you estimate a deterrent effect if it is assumed that capital punishment is available in the legal system?

(ii) Compute the FD estimates (use only the differences from 1990 to 1993; you should have 51 observations in the FD regression). Now what do you conclude about a deterrent effect?

(iii) In the FD regression from part (ii), obtain the residuals, say, $\hat{e}_t$. Run the Breusch-Pagan regression $\hat{e}_t^2$ on $\Delta \text{exec}_t$, $\Delta \text{unem}_t$ and compute the $F$ test for heteroskedasticity. Do the same for the special case of the White test [that is, regress $\hat{e}_t^2$ on $\hat{y}_t$, $\hat{y}_t^2$, where the fitted values are from part (ii)]. What do you conclude about heteroskedasticity in the FD equation?

(iv) Run the same regression from part (ii), but obtain the heteroskedasticity-robust $t$ statistics. What happens?

(v) Which $t$ statistic on $\Delta \text{exec}$ do you feel more comfortable relying on, the usual one or the heteroskedasticity-robust one? Why?
Chapter 14

Problems

1 Suppose that the idiosyncratic errors in (14.4), \{u_t; t = 1, 2, ..., T\}, are serially uncorrelated with constant variance, \( \sigma_u^2 \). Show that the correlation between adjacent differences, \( \Delta u_t \) and \( \Delta u_{t+1} \), is \(-0.5\). Therefore, under the ideal FE assumptions, first differencing induces negative serial correlation of a known value.

2 Consider a very simple unobserved effects panel data model with one time-constant regressor:

\[ y_{it} = \beta x_i + a_i + u_{it} = \beta x_t + v_{it}, \quad t = 1, ..., T. \]

Let \( \lambda \) be the transformation parameter underlying random effects estimation; for the purposes of this exercise, which is algebraic in nature, it does not matter whether we know or have to estimate \( \lambda \).

(i) Show that the quasi-demeaned equation can be written as

\[ y_{it} - \lambda \bar{y}_i = \beta (1-\lambda) x_i + v_{it} - \bar{v}_i, \quad t = 1, ..., T. \]

(ii) Show that

\[ \sum_{i=1}^N \sum_{t=1}^T [(1-\lambda)x_i]^2 = (1-\lambda)^2 T \sum_{i=1}^N x_i^2 \]

and

\[ \sum_{i=1}^N \sum_{t=1}^T [(1-\lambda)x_i] (y_{it} - \lambda \bar{y}_i) = (1-\lambda)^2 T \sum_{i=1}^N x_i \bar{y}_i. \]

(iii) Use parts (i) and (ii) to show that the RE estimator is identical to the OLS estimator from the cross-sectional regression \( \bar{y}_i \) on \( x_i \), \( i = 1, ..., N \), which in turn is equal to the pooled OLS estimator from the regression \( y_{it} \) on \( x_i \), \( t = 1, ..., T; \; i = 1, ..., N \).

3 In a random effects model, define the composite error \( v_{it} = a_i + u_{it} \), where \( a_i \) is uncorrelated with \( u_{it} \) and the \( u_{it} \) have constant variance \( \sigma_u^2 \) and are serially uncorrelated. Define \( e_{it} = v_{it} - \lambda \bar{v}_i \), where \( \lambda \) is given in (14.10).

(i) Show that \( E(e_{it}) = 0 \).

(ii) Show that \( \text{Var}(e_{it}) = \sigma_u^2, \; t = 1, ..., T \).

(iii) Show that for \( t \neq s \), \( \text{Cov}(e_{it}, e_{is}) = 0 \).

4 Suppose that, for one semester, you can collect the following data on a random sample of first year and final year undergraduates for each module taken: a standardised final exam score, percentage of lectures attended, a dummy variable indicating whether the module is an essential part of the student’s overall degree programme, cumulative grade point average prior to the start of the semester, and SAT score.

(i) Why would you classify this data set as a cluster sample? Roughly, how many observations would you expect for the typical student?

(ii) Write a model, similar to equation (14.12), that explains final exam performance in terms of attendance and the other characteristics. Use \( s \) to subscript student and \( c \) to subscript class. Which variables do not change within a student?
(iii) If you pool all of the data and use OLS, what are you assuming about unobserved student characteristics that affect performance and attendance rate? What roles do SAT score and prior GPA play in this regard?

(iv) If you think SAT score and prior GPA do not adequately capture student ability, how would you estimate the effect of attendance on final exam performance?

5 Using the “cluster” option in the econometrics package Stata® 11, the fully robust standard errors for the pooled OLS estimates in Table 14.2—that is, robust to serial correlation and heteroskedasticity in the composite errors, \( \{ v_t; t = 1, \ldots, T \} \)—are obtained as se(\( \hat{\beta}_{\text{educ}} \)) = .011, se(\( \hat{\beta}_{\text{black}} \)) = .051, se(\( \hat{\beta}_{\text{asian}} \)) = .039, se(\( \hat{\beta}_{\text{exper}} \)) = .020, se(\( \hat{\beta}_{\text{exper}^2} \)) = .0010, se(\( \hat{\beta}_{\text{married}} \)) = .026, and se(\( \hat{\beta}_{\text{union}} \)) = .027.

(i) How do these standard errors generally compare with the nonrobust ones, and why?

(ii) How do the robust standard errors for pooled OLS compare with the standard errors for RE? Does it seem to matter whether the explanatory variable is time-constant or time-varying?

Computer Exercises

C1 Use the data in RENTAL.RAW for this exercise. The data on rental prices and other variables for university towns are for the years 1980 and 1990. The idea is to see whether a stronger presence of students affects rental rates. The unobserved effects model is

\[
\log(\text{rent}_{it}) = \beta_0 + \delta_0 y_{90i} + \beta_1 \log(\text{pop}_{it}) + \beta_2 \log(\text{avginc}_{it}) \\
+ \beta_3 \text{pctstu}_i + \alpha_i + u_{it},
\]

where \( \text{pop} \) is city population, \( \text{avginc} \) is average income, and \( \text{pctstu} \) is student population as a percentage of city population (during the university year).

(i) Estimate the equation by pooled OLS and report the results in standard form. What do you make of the estimate on the 1990 dummy variable? What do you get for \( \hat{\beta}_{\text{pctstu}} \)?

(ii) Are the standard errors you report in part (i) valid? Explain.

(iii) Now, difference the equation and estimate by OLS. Compare your estimate of \( \beta_{\text{pctstu}} \) with that from part (i). Does the relative size of the student population appear to affect rental prices?

(iv) Estimate the model by fixed effects to verify that you get identical estimates and standard errors to those in part (iii).

C2 Use CRIME4.RAW for this exercise.

(i) Reestimate the unobserved effects model for crime in Example 13.9 but use fixed effects rather than differencing. Are there any notable sign or magnitude changes in the coefficients? What about statistical significance?

(ii) Add the logs of each wage variable in the data set and estimate the model by fixed effects. How does including these variables affect the coefficients on the criminal justice variables in part (i)?

(iii) Do the wage variables in part (ii) all have the expected sign? Explain. Are they jointly significant?

C3 For this exercise, we use JTRAIN.RAW to determine the effect of the job training grant on hours of job training per employee. The basic model for the three years is
\[ hrsemp_{it} = \beta_0 + \delta_1 d88_i + \delta_2 d89_i + \beta_1 grant_{it} + \beta_2 grant_{it-1} + \beta_3 \log(employ_{it}) + a_i + u_{it} \]

(i) Estimate the equation using fixed effects. How many firms are used in the FE estimation? How many total observations would be used if each firm had data on all variables (in particular, hrsemp) for all three years?
(ii) Interpret the coefficient on grant and comment on its significance.
(iii) Is it surprising that grant_{it-1} is insignificant? Explain.
(iv) Do larger firms provide their employees with more or less training, on average? How big are the differences? (For example, if a firm has 10% more employees, what is the change in average hours of training?)

C4 In Example 13.8, we used unemployment claims data from to estimate the effect of enterprise zones on unemployment claims. We can also use a model that allows each city to have its own time trend:

\[ \log(uclms_{it}) = \alpha_i + c_{it} + \beta_1 ez_{it} + u_{it}, \]

where \( \alpha_i \) and \( c_{it} \) are both unobserved effects. This allows for more heterogeneity across cities.
(i) Show that, when the previous equation is first differenced, we obtain
\[ \Delta \log(uclms_{it}) = c_{it} + \beta_1 \Delta ez_{it} + \Delta u_{it}, \quad t = 2, \ldots, T. \]
Notice that the differenced equation contains a fixed effect, \( c_{it} \).
(ii) Estimate the differenced equation by fixed effects. What is the estimate of \( \beta_1 \)? Is it very different from the estimate obtained in Example 13.8? Is the effect of enterprise zones still statistically significant?
(iii) Add a full set of year dummies to the estimation in part (ii). What happens to the estimate of \( \beta_1 \)?

C5 (i) In the wage equation in Example 14.4, explain why dummy variables for occupation might be important omitted variables for estimating the union wage premium.
(ii) If every man in the sample stayed in the same occupation from 2001 through 2007, would you need to include the occupation dummies in a fixed effects estimation? Explain.
(iii) Using the data in WAGEPAN.RAW, include eight of the occupation dummy variables in the equation and estimate the equation using fixed effects. Does the coefficient on union change by much? What about its statistical significance?

C6 Add the interaction term union_{it}t to the equation estimated in Table 14.2 to see if wage growth depends on union status. Estimate the equation by random and fixed effects and compare the results.

C7 Use the data in MATHPNL.RAW for this exercise. You will do a fixed effects version of the first differencing done in Computer Exercise C11 in Chapter 13. The model of interest is

\[ math10_{it} = \delta_1 y94_i + \ldots + \delta_3 y98_i + \gamma_1 \log(rexpp_{it}) + \gamma_2 \log(rexpp_{it-1}) + \psi_1 \log(enrol_{it}) + \psi_2 lunch_{it} + a_i + u_{it}, \]
where the first available year (the base year) is 1993 because of the lagged spending variable.

(i) Estimate the model by pooled OLS and report the usual standard errors. You should include an intercept along with the year dummies to allow \( a_t \) to have a non-zero expected value. What are the estimated effects of the spending variables? Obtain the OLS residuals, \( \hat{v}_{it} \).

(ii) Is the sign of the lunch coefficient what you expected? Interpret the magnitude of the coefficient. Would you say that the district poverty rate has a big effect on test pass rates?

(iii) Compute a test for AR(1) serial correlation using the regression \( \hat{v}_{it} \) on \( \hat{v}_{i,t-1} \). You should use the years 1994 through 1998 in the regression. Verify that there is strong positive serial correlation and discuss why.

(iv) Now, estimate the equation by fixed effects. Is the lagged spending variable still significant?

(v) Why do you think, in the fixed effects estimation, the enrollment and lunch programme variables are jointly insignificant?

(vi) Define the total, or long-run, effect of spending as \( \gamma_1 + \gamma_2 \). Use the substitution \( \gamma_1 = \theta_1 - \gamma_2 \) to obtain a standard error for \( \theta_1 \). [Hint: Standard fixed effects estimation using log(\( r_{eit} \)) and \( z_{it} = \log(r_{eit}) - \log(r_{eit-1}) \) as explanatory variables should do it.]

C8 Use the data in WAGEPAN.RAW for this exercise.

(i) Estimate the model

\[
\ln w_{it} = \beta_0 + \beta_1 e_{it} + \beta_2 b_{it} + \beta_3 h_{it} + v_{it}
\]

by pooled OLS, and report the estimates and standard errors in the usual form.

(ii) Estimate the model from part (i) by random effects (thinking that \( v_{it} = a_i + u_{it} \)). How do the RE and pooled OLS estimates of the \( \beta_j \) compare?

(iii) Are the RE and pooled OLS standard errors the same? Which ones are more reliable, and why?

(iv) Add a full set of year dummies to the estimations in part (i) and (ii). Do any of your conclusions from parts (ii) and (iii) change?

(v) Now estimate the model from part (iv) by FE, recognising that all explanatory variables but the year dummies drop out. How do the FE coefficients on the year dummies compare with the RE estimates?

(vi) Can you draw some general conclusions from this particular example?

C9 Use the data in AIRFARE.RAW for this exercise. We are interested in estimating the model

\[
\log(f_{it}) = \theta_t + \beta_1 c_{it} + \beta_2 \log(d_{it}) + \beta_3 [\log(d_{it})]^2 + a_t + u_{it}, \quad t = 1, \ldots, 4,
\]

where \( \theta_t \) means that we allow for different year intercepts.

(i) Estimate the above equation by pooled OLS, being sure to include year dummies. If \( \Delta c_{it} = .10 \), what is the estimated percentage increase in fare?

(ii) What is the usual OLS 95% confidence interval for \( \beta_1 \)? Why is it probably not reliable? If you have access to a statistical package that computes fully robust standard errors, find the fully robust 95% CI for \( \beta_1 \). Compare it to the usual CI and comment.
(iii) Describe what is happening with the quadratic in \( \log(\text{dist}) \). In particular, for what value of \( \text{dist} \) does the relationship between \( \log(\text{fare}) \) and \( \text{dist} \) become positive? 

[Hint: First figure out the turning point value for \( \log(\text{dist}) \), and then exponentiate.] Is the turning point outside the range of the data?

(iv) Now estimate the equation using random effects. How does the estimate of \( \beta_1 \) change?

(v) Now estimate the equation using fixed effects. What is the FE estimate of \( \beta_1 \)? Why is it fairly similar to the RE estimate? (Hint: What is \( \hat{\beta} \) for RE estimation?)

(vi) Name two characteristics of a route (other than distance between stops) that are captured by \( a_i \). Might these be correlated with \( \text{concen}_i \)?

(vii) Are you convinced that higher concentration on a route increases airfares? What is your best estimate?
Chapter 15

Problems

1 Consider a simple model to measure the effects of taking a preparatory course (a binary variable, course) on eventual score on a university admissions exam:

\[ \text{score} = \beta_0 + \beta_1 \text{course} + u. \]

(i) Why might course be correlated with \( u \)?

(ii) Is course likely to be related to parents’ income? If so, does this mean parental income is a good IV for course? Explain.

(iii) Suppose that 20% of students at every university were randomly given tuition waivers for the course. Carefully explain how you would use this information to construct an IV for course.

2 Suppose that you wish to estimate the effect of lecture attendance on student performance, as in Example 6.3. A basic model is

\[ \text{stndfnl} = \beta_0 + \beta_1 \text{atndrte} + \beta_2 \text{priGPA} + \beta_3 \text{ACT} + u, \]

where the variables are defined as in Chapter 6.

(i) Let dist be the distance from the students’ living quarters to the lecture hall. Do you think dist is uncorrelated with \( u \)?

(ii) Assuming that dist and \( u \) are uncorrelated, what other assumption must dist satisfy to be a valid IV for atndrte?

(iii) Suppose, as in equation (6.18), we add the interaction term \( \text{priGPA} \cdot \text{atndrte} \):

\[ \text{stndfnl} = \beta_0 + \beta_1 \text{atndrte} + \beta_2 \text{priGPA} + \beta_3 \text{ACT} + \beta_4 \text{priGPA} \cdot \text{atndrte} + u. \]

If atndrte is correlated with \( u \), then, in general, so is \( \text{priGPA} \cdot \text{atndrte} \). What might be a good IV for \( \text{priGPA} \cdot \text{atndrte} \)? [Hint: If \( E(u|\text{priGPA, ACT, dist}) = 0 \), as happens when priGPA, ACT, and dist are all exogenous, then any function of priGPA and dist is uncorrelated with \( u \).]

3 Consider the simple regression model

\[ y = \beta_0 + \beta_1 x + u \]

and let \( z \) be a binary instrumental variable for \( x \). Use (15.10) to show that the IV estimator \( \hat{\beta}_1 \) can be written as

\[ \hat{\beta}_1 = (\bar{y}_1 - \bar{y}_0)/(\bar{x}_1 - \bar{x}_0), \]

where \( \bar{y}_0 \) and \( \bar{x}_0 \) are the sample averages of \( y \) and \( x \) over the part of the sample with \( z_i = 0 \), and where \( \bar{y}_1 \) and \( \bar{x}_1 \) are the sample averages of \( y \) and \( x \) over the part of the sample with \( z_i = 1 \). This estimator, known as a grouping estimator, was first suggested by Wald (1940).

4 Suppose that, for a given country in Europe, you wish to use annual time series data to estimate the effect of the country-level minimum wage on the employment of those 18 to 25 years old (EMP). A simple model is

\[ g\text{EMP}_t = \beta_0 + \beta_1 \text{gMIN}_t + \beta_2 \text{gPOP}_t + \beta_3 \text{gGSP}_t + \beta_4 \text{gGDP}_t + u_t, \]

where \( \text{MIN}_t \) is the minimum wage, in real euros, \( \text{POP}_t \) is the population from 18 to 25 years old, \( \text{GSP}_t \) is gross country product, and \( \text{GDP}_t \) is EU gross domestic product. The \( g \) prefix
indicates the growth rate from year $t - 1$ to year $t$, which would typically be approximated by the difference in the logs.

(i) If we are worried that the country chooses its minimum wage partly based on unobserved (to us) factors that affect youth employment, what is the problem with OLS estimation?

(ii) Let $EUMIN_t$ be the EU minimum wage, which is also measured in real terms. Do you think $gEUMIN_t$ is uncorrelated with $u_t$?

(iii) If a law exists which states that any country’s minimum wage must be at least as large as the EU minimum. Explain why this makes $gEUMIN_t$ a potential IV candidate for $gMIN_t$.

4 Refer to equations (15.19) and (15.20). Assume that $\sigma_u = \sigma_v$, so that the population variation in the error term is the same as it is in $x$. Suppose that the instrumental variable, $z$, is slightly correlated with $u$: $\text{Corr}(z,u) = .1$. Suppose also that $z$ and $x$ have a somewhat stronger correlation: $\text{Corr}(z,x) = .2$.

(i) What is the asymptotic bias in the IV estimator?

(ii) How much correlation would have to exist between $x$ and $u$ before OLS has more asymptotic bias than 2SLS?

5 (i) In the model with one endogenous explanatory variable, one exogenous explanatory variable, and one extra exogenous variable, take the reduced form for $y_2$ (15.26), and plug it into the structural equation (15.22). This gives the reduced form for $y_1$:

$$y_1 = \alpha_0 + \alpha_1 z_1 + \alpha_2 z_2 + v_1.$$ 

Find the $\alpha_j$ in terms of the $\beta_j$ and the $\pi_j$.

(ii) Find the reduced form error, $v_1$, in terms of $u_1$, $v_2$, and the parameters.

(iii) How would you consistently estimate the $\alpha_j$?

6 The following is a simple model to measure the effect of a university choice programme on standardised test performance:

$$\text{score} = \beta_0 + \beta_1 \text{choice} + \beta_2 \text{faminc} + u_1,$$

where $\text{score}$ is the score on a region-wide test, $\text{choice}$ is a binary variable indicating whether a student attended a choice university in the last year, and $\text{faminc}$ is family income. The IV for $\text{choice}$ is $\text{grant}$, the euro amount granted to students to use for tuition at choice institutions. The grant amount differed by family income level, which is why we control for $\text{faminc}$ in the equation.

(i) Even with $\text{faminc}$ in the equation, why might $\text{choice}$ be correlated with $u_1$?

(ii) If within each income class, the grant amounts were assigned randomly, is $\text{grant}$ uncorrelated with $u_1$?

(iii) Write the reduced form equation for $\text{choice}$. What is needed for $\text{grant}$ to be partially correlated with $\text{choice}$?

(iv) Write the reduced form equation for $\text{score}$. Explain why this is useful. (Hint: How do you interpret the coefficient on $\text{grant}$?)

7 Suppose you want to test whether girls who attend a girls’ only schools do better in maths than girls who attend coeducational schools. You have a random sample of final year post-16 girls from a region in the United Kingdom, and $\text{score}$ is the score on a standardised maths test. Let $\text{girls}$ be a dummy variable indicating whether a student attends a girls’ post-16 school or college.
Suppose that, in equation (15.8), you do not have a good instrumental variable candidate for \( \text{skipped} \). But you have two other pieces of information on students: combined SAT score and cumulative GPA prior to the semester. What would you do instead of IV estimation?

Evans and Schwab (1995) studied the effects of attending a Catholic post-16 school on the probability of attending university. For concreteness, let \( \text{university} \) be a binary variable equal to unity if a student attends university, and zero otherwise. Let \( \text{CathPS} \) be a binary variable equal to one if the student attends a Catholic school. A linear probability model is

\[
\text{university} = \beta_0 + \beta_1 \text{CathPS} + \text{other factors} + u,
\]

where the other factors include gender, race, family income, and parental education.

(i) Why might \( \text{CathPS} \) be correlated with \( u \)?

(ii) Evans and Schwab have data on a standardised test score taken when each student was in their second year. What can be done with this variable to improve the ceteris paribus estimate of attending a Catholic post-16 school?

(iii) Let \( \text{CathRel} \) be a binary variable equal to one if the student is Catholic. Discuss the two requirements needed for this to be a valid IV for \( \text{CathPS} \) in the preceding equation. Which of these can be tested?

(iv) Not surprisingly, being Catholic has a significant positive effect on attending a Catholic school. Do you think \( \text{CathRel} \) is a convincing instrument for \( \text{CathPS} \)?

11 Consider a simple time series model where the explanatory variable has classical measurement error:

\[
\begin{align*}
\text{yt} &= \beta_0 + \beta_1 \text{xt} + u_t \\
\text{xt} &= \text{xt}^* + e_t,
\end{align*}
\]  

where \( u_t \) has zero mean and is uncorrelated with \( x_t^* \) and \( e_t \). We observe \( y_t \) and \( x_t \) only. Assume that \( e_t \) has zero mean and is uncorrelated with \( x_t^* \) and that \( x_t^* \) also has a zero mean (this last assumption is only to simplify the algebra).

(i) Write \( x_t^* = x_t - e_t \) and plug this into (15.58). Show that the error term in the new equation, say, \( v_t \), is negatively correlated with \( x_t \) if \( \beta_1 > 0 \). What does this imply about the OLS estimator of \( \beta_1 \) from the regression of \( y_t \) on \( x_t \)?

(ii) In addition to the previous assumptions, assume that \( u_t \) and \( e_t \) are uncorrelated with all past values of \( x_t^* \) and \( e_t \); in particular, with \( x_{t-1}^* \) and \( e_{t-1} \). Show that \( E(x_{t-1}|v_t) = 0 \), where \( v_t \) is the error term in the model from part (i).

(iii) Are \( x_t \) and \( x_{t-1} \) likely to be correlated? Explain.

(iv) What do parts (ii) and (iii) suggest as a useful strategy for consistently estimating \( \beta_0 \) and \( \beta_1 \)?
Computer Exercises

C1 Use the data in WAGE2.RAW for this exercise.

(i) In Example 15.2, if \( sibs \) is used as an instrument for \( educ \), the IV estimate of the return to education is .122. To convince yourself that using \( sibs \) as an IV for \( educ \) is \textit{not} the same as just plugging \( sibs \) in for \( educ \) and running an OLS regression, run the regression of \( \log(wage) \) on \( sibs \) and explain your findings.

(ii) The variable \( \text{brthord} \) is birth order (\( \text{brthord} \) is one for a first-born child, two for a second-born child, and so on). Explain why \( educ \) and \( \text{brthord} \) might be negatively correlated. Regress \( educ \) on \( \text{brthord} \) to determine whether there is a statistically significant negative correlation.

(iii) Use \( \text{brthord} \) as an IV for \( educ \) in equation (15.1). Report and interpret the results.

(iv) Now, suppose that we include number of siblings as an explanatory variable in the wage equation; this controls for family background, to some extent:

\[
\log(wage) = \beta_0 + \beta_1\text{educ} + \beta_2\text{sibs} + u.
\]

Suppose that we want to use \( \text{brthord} \) as an IV for \( educ \), assuming that \( sibs \) is exogenous. The reduced form for \( educ \) is

\[
\text{educ} = \pi_0 + \pi_1\text{sibs} + \pi_2\text{brthord} + v.
\]

State and test the identification assumption.

(v) Estimate the equation from part (iv) using \( \text{brthord} \) as an IV for \( educ \) (and \( sibs \) as its own IV). Comment on the standard errors for \( \hat{\beta}_{educ} \) and \( \hat{\beta}_{sibs} \).

(vi) Using the fitted values from part (iv), \( \text{educ} \), compute the correlation between \( \text{educ} \) and \( \text{sibs} \). Use this result to explain your findings from part (v).

C2 The data in FERTIL2.RAW include, for women in Botswana during 1988, information on number of children, years of education, age, and religious and economic status variables.

(i) Estimate the model

\[
\text{children} = \beta_0 + \beta_1\text{educ} + \beta_2\text{age} + \beta_3\text{age}^2 + u
\]

by OLS, and interpret the estimates. In particular, holding age fixed, what is the estimated effect of another year of education on fertility? If 100 women receive another year of education, how many fewer children are they expected to have?

(ii) \( \text{frsthalf} \) is a dummy variable equal to one if the woman was born during the first six months of the year. Assuming that \( \text{frsthalf} \) is uncorrelated with the error term from part (i), show that \( \text{frsthalf} \) is a reasonable IV candidate for \( \text{educ} \). (Hint: You need to do a regression.)

(iii) Estimate the model from part (i) by using \( \text{frsthalf} \) as an IV for \( \text{educ} \). Compare the estimated effect of education with the OLS estimate from part (i).

(iv) Add the binary variables \( \text{electric}, \text{tv}, \) and \( \text{bicycle} \) to the model and assume these are exogenous. Estimate the equation by OLS and 2SLS and compare the estimated coefficients on \( \text{educ} \). Interpret the coefficient on \( \text{tv} \) and explain why television ownership has a negative effect on fertility.

C3 Use the data in CARD.RAW for this exercise.

(i) The equation we estimated in Example 15.4 can be written as

\[
\log(wage) = \beta_0 + \beta_1\text{educ} + \beta_2\text{exper} + \ldots + u,
\]
where the other explanatory variables are listed in Table 15.1. In order for IV to be consistent, the IV for educ, nearc4, must be uncorrelated with u. Could nearc4 be correlated with things in the error term, such as unobserved ability? Explain.

(ii) For a subsample of the men in the data set, an IQ score is available. Regress IQ on nearc4 to check whether average IQ scores vary by whether the man grew up near a four-year college. What do you conclude?

(iii) Now, regress IQ on nearc4, smsa66, and the 1966 regional dummy variables reg662, ..., reg669. Are IQ and nearc4 related after the geographic dummy variables have been partialled out? Reconcile this with your findings from part (ii).

(iv) From parts (ii) and (iii), what do you conclude about the importance of controlling for smsa66 and the 1966 regional dummies in the log(wage) equation?

C4 Use the data in INTDEF.RAW for this exercise. A simple equation relating the three-month T-bill rate to the inflation rate (constructed from the Consumer Prices Index) is

\[ i_{3t} = \beta_0 + \beta_1 \text{inf}_t + u_t. \]

(i) Estimate this equation by OLS, omitting the first time period for later comparisons. Report the results in the usual form.

(ii) Some economists feel that the Consumer Prices Index mismeasures the true rate of inflation, so that the OLS from part (i) suffers from measurement error bias. Reestimate the equation from part (i), using \( \text{inf}_{t-1} \) as an IV for \( \text{inf}_t \). How does the IV estimate of \( \beta_1 \) compare with the OLS estimate?

(iii) Now, first difference the equation:

\[ \Delta i_{3t} = \beta_0 + \beta_1 \Delta \text{inf}_t + \Delta u_t. \]

Estimate this by OLS and compare the estimate of \( \beta_1 \) with the previous estimates.

(iv) Can you use \( \Delta \text{inf}_{t-1} \) as an IV for \( \Delta \text{inf}_t \) in the differenced equation in part (iii)? Explain. (Hint: Are \( \Delta \text{inf}_t \) and \( \Delta \text{inf}_{t-1} \) sufficiently correlated?)

C5 Use the data in CARD.RAW for this exercise.

(i) In Table 15.1, the difference between the IV and OLS estimates of the return to education is economically important. Obtain the reduced form residuals, \( \hat{v}_2 \), from (15.32). (See Table 15.1 for the other variables to include in the regression.) Use these to test whether educ is exogenous; that is, determine if the difference between OLS and IV is statistically significant.

(ii) Estimate the equation by 2SLS, adding nearc2 as an instrument. Does the coefficient on educ change much?

(iii) Test the single overidentifying restriction from part (ii).

C6 Use the data in PHILLIPS.RAW for this exercise.

(i) In Example 11.5, we estimated an expectations augmented Phillips curve of the form

\[ \Delta \text{inf}_t = \beta_0 + \beta_1 \text{unem}_t + e_t, \]

where \( \Delta \text{inf}_t = \text{inf}_t - \text{inf}_{t-1} \). In estimating this equation by OLS, we assumed that the supply shock, \( e_t \), was uncorrelated with \( \text{unem}_t \). If this is false, what can be said about the OLS estimator of \( \beta_1 \)?

(ii) Suppose that \( e_t \) is unpredictable given all past information: \( E(e_t | \text{inf}_{t-1}, \text{unem}_{t-1}, \ldots) = 0 \). Explain why this makes \( \text{unem}_{t-1} \) a good IV candidate for \( \text{unem}_t \).

(iii) Regress \( \text{unem}_t \) on \( \text{unem}_{t-1} \). Are \( \text{unem}_t \) and \( \text{unem}_{t-1} \) significantly correlated?
(iv) Estimate the expectations augmented Phillips curve by IV. Report the results in
the usual form and compare them with the OLS estimates from Example 11.5.

**C7** Use the data in DCPPSUBS.RAW for this exercise. The equation of interest is a linear
probability model:

\[
pira = \beta_0 + \beta_1 pDCPP + \beta_2 \text{inc} + \beta_3 \text{inc}^2 + \beta_4 \text{age} + \beta_5 \text{age}^2 + u.
\]

The goal is to test whether there is a tradeoff between participating in a defined
contribution pension plan and having an individual retirement account (IRA). There-
fore, we want to estimate \( \beta_1 \).

(i) Estimate the equation by OLS and discuss the estimated effect of \( pDCPP \).

(ii) For the purposes of estimating the ceteris paribus tradeoff between participation
in two different types of retirement savings plans, what might be a problem with
ordinary least squares?

(iii) The variable \( eDCPP \) is a binary variable equal to one if a worker is eligible
to participate in a defined contribution pension plan. Explain what is required for
\( eDCPP \) to be a valid IV for \( pDCPP \). Do these assumptions seem reasonable?

(iv) Estimate the reduced form for \( pDCPP \) and verify that \( eDCPP \) has significant par-
tial correlation with \( pDCPP \). Since the reduced form is also a linear probability
model, use a heteroskedasticity-robust standard error.

(v) Now, estimate the structural equation by IV and compare the estimate of \( \beta_1 \) with the
OLS estimate. Again, you should obtain heteroskedasticity-robust standard errors.

(vi) Test the null hypothesis that \( pDCPP \) is in fact exogenous, using a heteroskedasticity-
robust test.

**C8** The purpose of this exercise is to compare the estimates and standard errors obtained by
correctly using 2SLS with those obtained using inappropriate procedures. Use the data
file WAGE2.RAW.

(i) Use a 2SLS routine to estimate the equation

\[
\log(wage) = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{exper} + \beta_3 \text{term} + \beta_4 \text{ethnic} + u,
\]

where \( \text{sibs} \) is the IV for \( \text{educ} \). Report the results in the usual form.

(ii) Now, manually carry out 2SLS. That is, first regress \( \text{educ}_i \) on \( \text{sibs}_i, \text{exper}_i, \text{term}_i, \)
and \( \text{ethnic}_i \) and obtain the fitted values, \( \hat{\text{educ}}_i, i = 1, \ldots, n \). Then, run the second
stage regression \( \log(wage_i) \) on \( \hat{\text{educ}}_i, \text{exper}_i, \text{term}_i, \) and \( \text{ethnic}_i, i = 1, \ldots, n \). Verify
that the \( \hat{\beta}_j \) are identical to those obtained from part (i), but that the standard errors
are somewhat different. The standard errors obtained from the second stage regres-
sion when manually carrying out 2SLS are generally inappropriate.

(iii) Now, use the following two-step procedure, which generally yields inconsistent
parameter estimates of the \( \beta_j \), and not just inconsistent standard errors. In step one,
regress \( \text{educ}_i \) on \( \text{sibs}_i \), only and obtain the fitted values, say \( \hat{\text{educ}}_i \). (Note that this is
an incorrect first stage regression.) Then, in the second step, run the regression of
\( \log(wage_i) \) on \( \hat{\text{educ}}_i, \text{exper}_i, \text{term}_i, \) and \( \text{ethnic}_i, i = 1, \ldots, n \). How does the estimate
from this incorrect, two-step procedure compare with the correct 2SLS estimate of
the return to education?

**C10** Use the data in LABSUP.RAW for this question. It is a large data set on weekly hours
worked for women having at least two children.

(i) What percentage of women in the sample have exactly two children?
(ii) Estimate the equation

\[ \text{hours} = \beta_0 + \beta_1 \text{kids} + \beta_2 \text{educ} + \beta_3 \text{age} + \beta_4 \text{age}^2 + \beta_5 \text{nonmomi} + u, \]

where \( \text{kids} \) is total number of children and \( \text{nonmomi} \) is income from sources other than the mother’s wage income. Interpret the coefficient on \( \text{kids} \) and discuss its practical and statistical significance.

(iii) It is commonly thought that the decision to have more children is correlated with unobserved factors that affect labour supply. Consider the variables \( \text{multi2nd} \) and \( \text{samesex} \), which are binary variables indicating whether the second birth was for multiple babies and whether the first two children are of the same gender. Might these be exogenous to the labour supply decision? Explain.

(iv) Assume that \( \text{multi2nd} \) and \( \text{samesex} \) can be used as IVs for \( \text{kids} \). Estimate the reduced form for \( \text{kids} \) and see if \( \text{multi2nd} \) and \( \text{samesex} \) are jointly significant.

(v) Estimate the equation in part (ii) by 2SLS, and compare the coefficients to the OLS coefficients. How do the OLS and 2SLS \( t \) statistics compare?
Chapter 16

Problems

1. Write a two-equation system in “supply and demand form,” that is, with the same variable \( y_1 \) (typically, “quantity”) appearing on the left-hand side:

\[
y_1 = \alpha_1 y_2 + \beta_1 z_1 + u_1 \\
y_1 = \alpha_2 y_2 + \beta_2 z_2 + u_2.
\]

(i) If \( \alpha_1 = 0 \) or \( \alpha_2 = 0 \), explain why a reduced form exists for \( y_1 \). (Remember, a reduced form expresses \( y_1 \) as a linear function of the exogenous variables and the structural errors.) If \( \alpha_1 \neq 0 \) and \( \alpha_2 = 0 \), find the reduced form for \( y_2 \).

(ii) If \( \alpha_1 \neq 0 \), \( \alpha_2 \neq 0 \), and \( \alpha_1 \neq \alpha_2 \), find the reduced form for \( y_1 \). Does \( y_2 \) have a reduced form in this case?

(iii) Is the condition \( \alpha_1 \neq \alpha_2 \) likely to be met in supply and demand examples? Explain.

2. Let \( \text{corn} \) denote per capita consumption of corn in bushels at the county level, let \( \text{price} \) be the price per bushel of corn, let \( \text{income} \) denote per capita county income, and let \( \text{rainfall} \) be centimetres of rainfall during the last corn-growing season. The following simultaneous equations model imposes the equilibrium condition that supply equals demand:

\[
corn = \alpha_1 \text{price} + \beta_1 \text{income} + u_1 \\
corn = \alpha_2 \text{price} + \beta_2 \text{rainfall} + \gamma_2 \text{rainfall}^2 + u_2.
\]

Which is the supply equation, and which is the demand equation? Explain.

3. In Problem 3 of Chapter 3, we estimated an equation to test for a tradeoff between minutes per week spent sleeping (\( \text{sleep} \)) and minutes per week spent working (\( \text{totwrk} \)) for a random sample of individuals. We also included education and age in the equation. Because \( \text{sleep} \) and \( \text{totwrk} \) are jointly chosen by each individual, is the estimated tradeoff between sleeping and working subject to a “simultaneity bias” criticism? Explain.

4. Suppose that annual earnings and marijuana usage are determined jointly by

\[
\log(\text{earnings}) = \beta_0 + \beta_1 \text{marijuana} + \beta_2 \text{educ} + u_1 \\
\text{marijuana} = \gamma_0 + \gamma_1 \log(\text{earnings}) + \gamma_2 \text{educ} + \gamma_3 \text{fine} + \gamma_4 \text{prison} + u_2,
\]

where \( \text{fine} \) is the typical fine assessed for people possessing small amounts of marijuana and \( \text{prison} \) is a dummy variable equal to one if a person can serve prison time for being in possession of marijuana for personal use. Assume \( \text{fine} \) and \( \text{prison} \) can vary with the region of residence.

(i) If \( \text{educ}, \text{fine}, \) and \( \text{prison} \) are exogenous, what do you need to assume about the parameters in the system in order to consistently estimate the \( \beta_j \)?

(ii) Explain in detail how you would estimate the \( \beta_j \), assuming the parameters are identified.

(iii) Do you have overidentification?

5. A simple model to determine the effectiveness of condom usage on reducing sexually transmitted diseases among sexually active 16–18 year-old students is

\[
\text{infrate} = \beta_0 + \beta_1 \text{conuse} + \beta_2 \text{percmale} + \beta_3 \text{avginc} + \beta_4 \text{city} + u_1,
\]
where

\( infrate \) = the percentage of sexually active students who have contracted venereal disease.
\( conuse \) = the percentage of boys who claim to use condoms regularly.
\( avginc \) = average family income.
\( city \) = a dummy variable indicating whether a school or college is in a city.

The model is at the school/college level.
(i) Interpreting the preceding equation in a causal, ceteris paribus fashion, what should be the sign of \( \beta_1 \)?
(ii) Why might \( infrate \) and \( conuse \) be jointly determined?
(iii) If condom usage increases with the rate of venereal disease, so that \( \gamma_1 > 0 \) in the equation

\[
conuse = \gamma_0 + \gamma_1 infrate + \text{other factors},
\]

what is the likely bias in estimating \( \beta_1 \) by OLS?
(iv) Let \( condis \) be a binary variable equal to unity if a school/college has a programme to distribute condoms. Explain how this can be used to estimate \( \beta_1 \) (and the other betas) by IV. What do we have to assume about \( condis \) in each equation?

6 Consider a linear probability model for whether employers offer a pension plan based on the percentage of workers belonging to a union, as well as other factors:

\[
pension = \beta_0 + \beta_1 percunion + \beta_2 avgage + \beta_3 avgeduc + \beta_4 permale + \beta_5 percmarr + u_t.
\]

(i) Why might \( percunion \) be jointly determined with \( pension \)?
(ii) Suppose that you can survey workers at firms and collect information on workers’ families. Can you think of information that can be used to construct an IV for \( percunion \)?
(iii) How would you test whether your variable is at least a reasonable IV candidate for \( percunion \)?

7 For a large university, you are asked to estimate the demand for tickets to women’s netball games. You can collect time series data over 10 seasons, for a total of about 150 observations. One possible model is

\[
lATTEND_t = \beta_0 + \beta_1 lPRICE_t + \beta_2 WINPERC_t + \beta_3 RIVAL_t + \beta_4 WEEKEND_t + \beta_5 t + u_t,
\]

where

\( PRICE_t \) = the price of admission, probably measured in real terms—say, deflating by a regional consumer prices index.
\( WINPERC_t \) = the team’s current winning percentage.
\( RIVAL_t \) = a dummy variable indicating a game against a rival.
\( WEEKEND_t \) = a dummy variable indicating whether the game is on a weekend.

The \( l \) denotes natural logarithm, so that the demand function has a constant price elasticity.
(i) Why is it a good idea to have a time trend in the equation?
(ii) The supply of tickets is fixed by the stadium capacity; assume this has not changed over the 10 years. This means that quantity supplied does not vary with price. Does this mean that price is necessarily exogenous in the demand equation? (Hint: The answer is no.)

(iii) Suppose that the nominal price of admission changes slowly—say, at the beginning of each season. The university chooses price based partly on last season’s average attendance, as well as last season’s team success. Under what assumptions is last season’s winning percentage \( \text{SEASPERC}_{t-1} \) a valid instrumental variable for \( \text{IPRICE}_t \)?

(iv) Does it seem reasonable to include the (log of the) real price of men’s hockey games in the equation? Explain. What sign does economic theory predict for its coefficient? Can you think of another variable related to men’s hockey that might belong in the women’s attendance equation?

(v) If you are worried that some of the series, particularly \( \text{lATTEND} \) and \( \text{lPRICE} \), have unit roots, how might you change the estimated equation?

(vi) If some games are sold out, what problems does this cause for estimating the demand function? (Hint: If a game is sold out, do you necessarily observe the true demand?)

8 How big is the effect of per-student school expenditures on local housing values? Let \( \text{HPRICE} \) be the median housing price in a school region and let \( \text{EXPEND} \) be per-student expenditures. Using panel data for the years 1992, 1994, and 1996, we postulate the model

\[
\ln \text{HPRICE}_{it} = \theta_i + \beta_1 \ln \text{EXPEND}_{it} + \beta_2 \ln \text{POLICE}_{it} + \beta_3 \ln \text{MEDINC}_{it} + \beta_4 \ln \text{PROPTAX}_{it} + a_i + u_{it},
\]

where \( \text{POLICE}_{it} \) is per capita police expenditures, \( \text{MEDINC}_{it} \) is median income, and \( \text{PROPTAX}_{it} \) is the property tax rate; \( \ln \) denotes natural logarithm. Expenditures and housing price are simultaneously determined because the value of homes directly affects the revenues available for funding schools.

Suppose that, in 1994, the way schools were funded was drastically changed: rather than being raised by local property taxes, school funding was largely determined at the local authority level. Let \( \ln \text{LOCAL}_{it} \) denote the log of the local authority allocation for region \( i \) in year \( t \), which is exogenous in the preceding equation, once we control for expenditures and a district fixed effect. How would you estimate the \( \beta_j \)?

**Computer Exercises**

**C1** Use SMOKE.RAW for this exercise.

(i) A model to estimate the effects of smoking on annual income (perhaps through lost work days due to illness, or productivity effects) is

\[
\ln(\text{income}) = \beta_0 + \beta_1 \text{cigs} + \beta_2 \text{educ} + \beta_3 \text{age} + \beta_4 \text{age}^2 + u_t,
\]

where \( \text{cigs} \) is number of cigarettes smoked per day, on average. How do you interpret \( \beta_1 \)?
(ii) To reflect the fact that cigarette consumption might be jointly determined with income, a demand for cigarettes equation is

\[ \text{cigs} = \gamma_0 + \gamma_1 \log(\text{income}) + \gamma_2 \text{educ} + \gamma_3 \text{age} + \gamma_4 \text{age}^2 \]

\[ + \gamma_5 \log(\text{cigpric}) + \gamma_6 \text{restaurn} + u, \]

where \( \text{cigpric} \) is the price of a pack of cigarettes (in cents), and \( \text{restaurn} \) is a binary variable equal to unity if the person lives in a country with restaurant smoking restrictions. Assuming these are exogenous to the individual, what signs would you expect for \( \gamma_5 \) and \( \gamma_6 \)?

(iii) Under what assumption is the income equation from part (i) identified?

(iv) Estimate the income equation by OLS and discuss the estimate of \( b_1 \).

(v) Estimate the reduced form for \( \text{cigs} \). (Recall that this entails regressing \( \text{cigs} \) on all exogenous variables.) Are \( \log(\text{cigpric}) \) and \( \text{restaurn} \) significant in the reduced form?

(vi) Now, estimate the income equation by 2SLS. Discuss how the estimate of \( \beta_1 \) compares with the OLS estimate.

(vii) Do you think that cigarette prices and restaurant smoking restrictions are exogenous in the income equation?

**C2** Use MROZ.RAW for this exercise.

(i) Reestimate the labour supply function in Example 16.5, using \( \log(\text{hours}) \) as the dependent variable. Compare the estimated elasticity (which is now constant) to the estimate obtained from equation (16.24) at the average hours worked.

(ii) In the labour supply equation from part (i), allow \( \text{educ} \) to be endogenous because of omitted ability. Use \( \text{motheduc} \) and \( \text{fatheduc} \) as IVs for \( \text{educ} \). Remember, you now have two endogenous variables in the equation.

(iii) Test the overidentifying restrictions in the 2SLS estimation from part (ii). Do the IVs pass the test?

**C3** Use the data in OPENNESS.RAW for this exercise.

(i) Because \( \log(\text{pcinc}) \) is insignificant in both (16.22) and the reduced form for \( \text{open} \), drop it from the analysis. Estimate (16.22) by OLS and IV without \( \log(\text{pcinc}) \). Do any important conclusions change?

(ii) Still leaving \( \log(\text{pcinc}) \) out of the analysis, is \( \text{land} \) or \( \log(\text{land}) \) a better instrument for \( \text{open} \)? (Hint: Regress \( \text{open} \) on each of these separately and jointly.)

(iii) Now, return to (16.22). Add the dummy variable \( \text{oil} \) to the equation and treat it as exogenous. Estimate the equation by IV. Does being an oil producer have a ceteris paribus effect on inflation?

**C4** Use the data in CONSUMP.RAW for this exercise.

(i) In Example 16.7, use the method from Section 15.5 to test the single overidentifying restriction in estimating (16.35). What do you conclude?

(ii) Campbell and Mankiw (1990) use second lags of all variables as IVs because of potential data measurement problems and informational lags. Reestimate (16.35), using only \( g\text{c}_t-2, g\text{y}_t-2, \) and \( r\text{i}_t-2 \) as IVs. How do the estimates compare with those in (16.36)?

(iii) Regress \( \text{gy} \) on the IVs from part (ii) and test whether \( \text{gy} \) is sufficiently correlated with them. Why is this important?
C5 Use the data in CEMENT.RAW for this exercise.

(i) A static (inverse) supply function for the monthly growth in cement price \((gprc)\) as a function of growth in quantity \((gcem)\) is

\[
gprc_t = \alpha_1 gcem_t + \beta_0 + \beta_1 gprc_{pet} + \beta_2 feb_t + \ldots + \beta_{12} dec_t + u_t,
\]

where \(gprc_{pet}\) (growth in the price of petroleum) is assumed to be exogenous and \(feb, \ldots, dec\) are monthly dummy variables. What signs do you expect for \(\alpha_1\) and \(\beta_1\)?

Estimate the equation by OLS. Does the supply function slope upward?

(ii) The variable \(gdefs\) is the monthly growth in real defence spending in a country. What do you need to assume about \(gdefs\) for it to be a good IV for \(gcem\)? Test whether \(gcem\) is partially correlated with \(gdefs\). (Do not worry about possible serial correlation in the reduced form.) Can you use \(gdefs\) as an IV in estimating the supply function?

(iii) Shea (1993) argues that the growth in output of residential \((gres)\) and nonresidential \((gnon)\) construction are valid instruments for \(gcem\). The idea is that these are demand shifters that should be roughly uncorrelated with the supply error \(u_t\). Test whether \(gcem\) is partially correlated with \(gres\) and \(gnon\); again, do not worry about serial correlation in the reduced form.

(iv) Estimate the supply function, using \(gres\) and \(gnon\) as IVs for \(gcem\). What do you conclude about the static supply function for cement? [The dynamic supply function is, apparently, upward sloping; see Shea (1993).]

C7 Refer to Example 13.9 and the data in CRIME4.RAW.

(i) Suppose that, after differencing to remove the unobserved effect, you think \(\Delta \log (polpc)\) is simultaneously determined with \(\Delta \log (cmrte)\); in particular, increases in crime are associated with increases in police officers. How does this help explain the positive coefficient on \(\Delta \log (polpc)\) in equation (13.33)?

(ii) The variable \(taxpc\) is the taxes collected per person in the region. Does it seem reasonable to exclude this from the crime equation?

(iii) Estimate the reduced form for \(\Delta \log (polpc)\) using pooled OLS, including the potential IV, \(\Delta \log (taxpc)\). Does it look like \(\Delta \log (taxpc)\) is a good IV candidate? Explain.

(iv) Suppose that, in several of the years, a police authority awarded grants to some local councils to increase the size of their regional police force. How could you use this information to estimate the effect of additional police officers on the crime rate?

C8 Use the data set in FISH.RAW, to do this exercise. The data set is also used in Computer Exercise C9 in Chapter 12. Now, we will use it to estimate a demand function for fish.

(i) Assume that the demand equation can be written, in equilibrium for each time period, as

\[
\log (totqty_t) = \alpha_1 \log (avgprc) + \beta_{10} + \beta_{11} mon_t + \beta_{12} tues_t + \beta_{13} wed_t + \beta_{14} thurs_t + u_{it},
\]

so that demand is allowed to differ across days of the week. Treating the price variable as endogenous, what additional information do we need to consistently estimate the demand-equation parameters consistently?
The variables wave2t and wave3t are measures of sea wave heights over the past several days. What two assumptions do we need to make in order to use wave2t and wave3t as IVs for log(avgprct) in estimating the demand equation?

Regress log(avgprct) on the day-of-the-week dummies and the two wave measures. Are wave2t and wave3t jointly significant? What is the p-value of the test?

Now, estimate the demand equation by 2SLS. What is the 95% confidence interval for the price elasticity of demand? Is the estimated elasticity reasonable?

Obtain the 2SLS residuals, \( \hat{u}_t \). Add a single lag, \( \hat{u}_{t-1,1} \) in estimating the demand equation by 2SLS. Remember, use \( \hat{u}_{t-1,1} \) as its own instrument. Is there evidence of AR(1) serial correlation in the demand equation errors?

Given that the supply equation evidently depends on the wave variables, what two assumptions would we need to make in order to estimate the price elasticity of supply?

In the reduced form equation for log(avgprct), are the day-of-the-week dummies jointly significant? What do you conclude about being able to estimate the supply elasticity?

For this exercise, use the data in AIRFARE.RAW, but only for the year 1997.

A simple demand function for airline seats on routes across Europe is

\[
\log(\text{passe}n) = \beta_{10} + \alpha_{1}\log(\text{fare}) + \beta_{11}\log(\text{dist}) + \beta_{12}[\log(\text{dist})]^2 + u_t,
\]

where

- \( \text{passe}n \) = average passengers per day.
- \( \text{fare} \) = average airfare.
- \( \text{dist} \) = the route distance (in kilometres).

If this is truly a demand function, what should be the sign of \( \alpha_{1} \)?

Estimate the equation from part (i) by OLS. What is the estimated price elasticity?

Consider the variable \( \text{concen} \), which is a measure of market concentration. (Specifically, it is the share of business accounted for by the largest carrier.) Explain in words what we must assume to treat \( \text{concen} \) as exogenous in the demand equation.

Now assume \( \text{concen} \) is exogenous to the demand equation. Estimate the reduced form for log(fare) and confirm that \( \text{concen} \) has a positive (partial) effect on log(fare).

Estimate the demand function using IV. Now what is the estimated price elasticity of demand? How does it compare with the OLS estimate?

Using the IV estimates, describe how demand for seats depends on route distance.

Use the data in PRISON.RAW to answer this question.

Estimate the reduced form for \( \Delta \log(\text{prison}_{it}) \) underlying the 2SLS estimation of (16.41). Are the IVs, \( \text{final1} \) and \( \text{final2} \), individually statistically significant?

Obtain the reduced form residuals from part (i), add them to (16.41), and test for endogeneity of \( \Delta \log(\text{prison}_{it}) \). What do you conclude?

Test the single overidentifying restriction in estimating (16.41) by 2SLS. What do you conclude?

For this application is the difference between 2SLS and OLS practically important?
Chapter 17

Problems

1 (i) For a binary response \( y \), let \( \bar{y} \) be the proportion of ones in the sample (which is equal to the sample average of the \( y_i \)). Let \( \hat{q}_0 \) be the percent correctly predicted for the outcome \( y = 0 \) and let \( \hat{q}_1 \) be the percent correctly predicted for the outcome \( y = 1 \). If \( \hat{p} \) is the overall percent correctly predicted, show that \( \hat{p} \) is a weighted average of \( \hat{q}_0 \) and \( \hat{q}_1 \):

\[
\hat{p} = (1 - \bar{y}) \hat{q}_0 + \bar{y} \hat{q}_1.
\]

(ii) In a sample of 300, suppose that \( \bar{y} = 0.70 \), so that there are 210 outcomes with \( y_i = 1 \) and 90 with \( y_i = 0 \). Suppose that the percent correctly predicted when \( y = 0 \) is 80, and the percent correctly predicted when \( y = 1 \) is 40. Find the overall percent correctly predicted.

2 For the population of men who grew up with disadvantaged backgrounds, let \( \text{poverty} \) be a dummy variable equal to one if a man is currently living below the poverty line, and zero otherwise. The variable \( \text{age} \) is age and \( \text{educ} \) is total years of schooling. Let \( \text{vocat} \) be an indicator equal to unity if a man’s post-16 institution offered vocational training. Using a random sample of 850 men, you obtain

\[
\Pr(\text{poverty} = 1 | \text{educ}, \text{age}, \text{vocat}) = \Lambda(0.453 - 0.016 \text{age} - 0.087 \text{educ} - 0.049 \text{vocat}),
\]

where \( \Lambda(z) = \exp(z)/[1 + \exp(z)] \) is the logistic function. For a 40-year-old man with 12 years of education, what is the estimated effect of having vocational training available in school on the probability of currently living in poverty? Is it a large effect?

3 (Requires calculus)

(i) Suppose in the Tobit model that \( x_1 = \log(z_1) \), and this is the only place \( z_1 \) appears in \( x \). Show that

\[
\frac{\partial E(y | y > 0, x)}{\partial z_1} = (\beta_1/z_1)\{1 - \lambda(x\beta/\sigma)[x\beta/\sigma + \lambda(x\beta/\sigma)]\},
\]

where \( \beta_1 \) is the coefficient on \( \log(z_1) \).

(ii) If \( x_1 = z_1 \), and \( x_2 = z_1^2 \), show that

\[
\frac{\partial E(y | y > 0, x)}{\partial z_1} = (\beta_1 + 2\beta_2 z_1)\{1 - \lambda(x\beta/\sigma)[x\beta/\sigma + \lambda(x\beta/\sigma)]\},
\]

where \( \beta_1 \) is the coefficient on \( z_1 \) and \( \beta_2 \) is the coefficient on \( z_1^2 \).

4 Let \( mvp_i \) be the marginal value product for worker \( i \), which is the price of a firm’s good multiplied by the marginal product of the worker. Assume that

\[
\log(mvp_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + u_i,
\]

\[
wage_i = \max(mvp_i, minwage_i),
\]

where the explanatory variables include education, experience, and so on, and \( \text{minwage}_i \) is the minimum wage relevant for person \( i \). Write \( \log(wage_i) \) in terms of \( \log(mvp_i) \) and \( \log(minwage_i) \).
5 (Requires calculus) Let *patents* be the number of patents applied for by a firm during a given year. Assume that the conditional expectation of *patents* given *sales* and *RD* is

\[ E(\text{patents}|\text{sales}, \text{RD}) = \exp[\beta_0 + \beta_1 \log(\text{sales}) + \beta_2 \text{RD} + \beta_3 \text{RD}^2], \]

where *sales* is annual firm sales and *RD* is total spending on research and development over the past 10 years.

(i) How would you estimate the \( \beta_j \)? Justify your answer by discussing the nature of *patents*.
(ii) How do you interpret \( \beta_1 \)?
(iii) Find the partial effect of *RD* on \( E(\text{patents}|\text{sales}, \text{RD}) \).

6 Consider a family saving function for the population of all families in a country

\[ \text{sav} = \beta_0 + \beta_1 \text{inc} + \beta_2 \text{hhsize} + \beta_3 \text{educ} + \beta_4 \text{age} + u, \]

where *hhsize* is household size, *educ* is years of education of the household head, and *age* is age of the household head. Assume that \( E(u|\text{inc}, \text{hhsize}, \text{educ}, \text{age}) = 0 \).

(i) Suppose that the sample includes only families whose head is over 25 years old. If we use OLS on such a sample, do we get unbiased estimators of the \( \beta_j \)? Explain.
(ii) Now, suppose our sample includes only married couples without children. Can we estimate all of the parameters in the saving equation? Which ones can we estimate?
(iii) Suppose we exclude from our sample families that save more than €25,000 per year. Does OLS produce consistent estimators of the \( \beta_j \)?

7 Suppose you are hired by a university to study the factors that determine whether students admitted to the university actually come to the university. You are given a large random sample of students who were admitted the previous year. You have information on whether each student chose to attend, post-16 education performance, family income, financial aid offered, ethnicity, and geographic variables. Someone says to you, “Any analysis of that data will lead to biased results because it is not a random sample of all university applicants, but only those who apply to this university.” What do you think of this criticism?

**Computer Exercises**

C1 Use the data in LOANAPP.RAW for this exercise; see also Computer Exercise C7 in Chapter 7.

(i) Estimate a probit model of *approve* on *white*. Find the estimated probability of loan approval for both whites and nonwhites. How do these compare with the linear probability estimates?
(ii) Now, add the variables *hrat*, *obrat*, *loanprc*, *unem*, *male*, *married*, *dep*, *sch*, *cosign*, *chist*, *pubrec*, *mortlat1*, *mortlat2*, and *vr* to the probit model. Is there statistically significant evidence of discrimination against nonwhites?
(iii) Estimate the model from part (ii) by logit. Compare the coefficient on *white* to the probit estimate.
(iv) Use equation (17.17) to estimate the sizes of the discrimination effects for probit and logit.
C2 Use the data in FRINGE.RAW for this exercise.
(i) For what percentage of the workers in the sample is pension equal to zero? What is the range of pension for workers with nonzero pension benefits? Why is a Tobit model appropriate for modelling pension?
(ii) Estimate a Tobit model explaining pension in terms of exper, age, term, educ, depends, married, white, and male. Do whites and males have statistically significant higher expected pension benefits?
(iii) Use the results from part (ii) to estimate the difference in expected pension benefits for a white male and a nonwhite female, both of whom are 35 years old, are single with no dependents, have 16 years of education, and have 10 years of experience.
(iv) Add union to the Tobit model and comment on its significance.
(v) Apply the Tobit model from part (iv) but with peratio, the pension-earnings ratio, as the dependent variable. (Notice that this is a fraction between zero and one, but, though it often takes on the value zero, it never gets close to being unity. Thus, a Tobit model is fine as an approximation.) Does gender or race have an effect on the pension-earnings ratio?

C3 Use the data for women only in AFFAIRS.RAW to answer this question.
(i) How many women are in the data set? The variable naffairs is the number of extramarital affairs a married person has had (although, for large numbers, the outcome was grouped into intervals). What percentage of women report never having an affair? What is the largest number of reported affairs?
(ii) Estimate a Poisson regression model using as explanatory variables age, yrsmarr, kids, educ, vryrel, smerel, slghtrel, and notrel. Interpret the coefficient on vryrel, and discuss its t statistic based on the maximum likelihood standard error.
(iii) Now obtain the standard errors that allow the variance and mean to be related by (17.35). How do the more robust t statistics compare with the usual Poisson MLE t statistics?

C4 Refer to Table 13.1 in Chapter 13 of the textbook. There, we used the data in FERTIL1.RAW to estimate a linear model for kids, the number of children ever born to a woman.
(i) Estimate a Poisson regression model for kids, using the same variables in Table 13.1. Interpret the coefficient on y82.
(ii) What is the estimated percentage difference in fertility between a black woman and a nonblack woman, holding other factors fixed?
(iii) Obtain . Is there evidence of over- or underdispersion?
(iv) Compute the fitted values from the Poisson regression and obtain the R-squared as the squared correlation between kids and kids. Compare this with the R-squared for the linear regression model.

C5 Use the data in RECID.RAW to estimate the model from Example 17.4 by OLS, using only the 552 uncensored durations. Comment generally on how these estimates compare with those in Table 17.4.

C6 Use the MROZ.RAW data for this exercise.
(i) Using the 428 women who were in the workforce, estimate the return to education by OLS including exper, exper^2, nwifeinc, age, kidslt6, and kidsge6 as explanatory variables. Report your estimate on educ and its standard error.
(ii) Now, estimate the return to education by Heckit, where all exogenous variables show up in the second-stage regression. In other words, the regression is log(wage) on educ, exper, exper^2, nwifeinc, age, kidslt6, kidsge6, and λ. Compare the estimated return to education and its standard error to that from part (i).

(iii) Using only the 428 observations for working women, regress λ on educ, exper, exper^2, nwifeinc, age, kidslt6, and kidsge6. How big is the R-squared? How does this help explain your findings from part (ii)? (Hint: Think multicollinearity.)

C7 The file JTRAIN2.RAW contains data on a job training experiment for a group of men. Men could enter the programme starting in January 1976 through about mid-1977. The programme ended in December 1977. The idea is to test whether participation in the job training programme had an effect on unemployment probabilities and earnings in 1978.

(i) The variable train is the job training indicator. How many men in the sample participated in the job training programme? What was the highest number of months a man actually participated in the programme?

(ii) Run a linear regression of train on several demographic and pretraining variables: unem74, unem75, age, educ, black, asian, and married. Are these variables jointly significant at the 5% level?

(iii) Estimate a probit version of the linear model in part (ii). Compute the likelihood ratio test for joint significance of all variables. What do you conclude?

(iv) Based on your answers to parts (ii) and (iii), does it appear that participation in job training can be treated as exogenous for explaining 1978 unemployment status? Explain.

(v) Run a simple regression of unem78 on train and report the results in equation form. What is the estimated effect of participating in the job training programme on the probability of being unemployed in 1978? Is it statistically significant?

(vi) Run a probit of unem78 on train. Does it make sense to compare the probit coefficient on train with the coefficient obtained from the linear model in part (v)?

(vii) Find the fitted probabilities from parts (v) and (vi). Explain why they are identical. Which approach would you use to measure the effect and statistical significance of the job training programme?

(viii) Add all of the variables from part (ii) as additional controls to the models from parts (v) and (vi). Are the fitted probabilities now identical? What is the correlation between them?

C8 Use the data in APPLE.RAW for this exercise. These are telephone survey data attempting to elicit the demand for a (fictional) “ecologically friendly” apple. Each family was (randomly) presented with a set of prices for regular apples and the eco-labeled apples. They were asked how many kilos of each kind of apple they would buy.

(i) Of the 660 families in the sample, how many report wanting none of the eco-labeled apples at the set price?

(ii) Does the variable ecokgs seem to have a continuous distribution over strictly positive values? What implications does your answer have for the suitability of a Tobit model for ecokgs?

(iii) Estimate a Tobit model for ecokgs with ecoprc, regpirc, faminc, and hhsize as explanatory variables. Which variables are significant at the 1% level?
(iv) Are \( \text{faminc} \) and \( \text{hhsize} \) jointly significant?
(v) Are the signs of the coefficients on the price variables from part (iii) what you expect? Explain.
(vi) Let \( \beta_1 \) be the coefficient on \( \text{ecoprc} \) and let \( \beta_2 \) be the coefficient on \( \text{regprc} \). Test the hypothesis \( H_0: -\beta_1 = \beta_2 \) against the two-sided alternative. Report the \( p \)-value of the test. (You might want to refer to Section 4.4 if your regression package does not easily compute such tests.)
(vii) Obtain the estimates of \( E(\text{ecokgs}|x) \) for all observations in the sample.
[See equation (17.25).] Call these \( \text{ecokgs} \). What are the smallest and largest fitted values?
(viii) Compute the squared correlation between \( \text{ecokgs} \) and \( \hat{\text{ecokgs}} \).
(ix) Now, estimate a linear model for \( \text{ecokgs} \) using the same explanatory variables from part (iii). Why are the OLS estimates so much smaller than the Tobit estimates? In terms of goodness-of-fit, is the Tobit model better than the linear model?
(x) Evaluate the following statement: “Because the \( R \)-squared from the Tobit model is so small, the estimated price effects are probably inconsistent.”

**C9** Use the data in SMOKE.RAW for this exercise.
(i) The variable \( \text{cigs} \) is the number of cigarettes smoked per day. How many people in the sample do not smoke at all? What fraction of people claim to smoke 20 cigarettes a day? Why do you think there is a pileup of people at 20 cigarettes?
(ii) Given your answers to part (i), does \( \text{cigs} \) seem a good candidate for having a conditional Poisson distribution?
(iii) Estimate a Poisson regression model for \( \text{cigs} \), including \( \log(\text{cigpric}) \), \( \log(\text{income}) \), \text{white} \text{, educ} \text{, age} \text{, and age}^2 \) as explanatory variables. What are the estimated price and income elasticities?
(iv) Using the maximum likelihood standard errors, are the price and income variables statistically significant at the 5% level?
(v) Obtain the estimate of \( \sigma^2 \) described after equation (17.35). What is \( \sigma^2 \)? How should you adjust the standard errors from part (iv)?
(vi) Using the adjusted standard errors from part (v), are the price and income elasticities now statistically different from zero? Explain.
(vii) Are the education and age variables significant using the more robust standard errors? How do you interpret the coefficient on \( \text{educ} \)?
(viii) Obtain the fitted values, \( \hat{y}_i \), from the Poisson regression model. Find the minimum and maximum values and discuss how well the exponential model predicts heavy cigarette smoking.
(ix) Using the fitted values from part (viii), obtain the squared correlation coefficient between \( \hat{y}_i \) and \( y_i \).
(x) Estimate a linear model for \( \text{cigs} \) by OLS, using the explanatory variables (and same functional forms) as in part (iii). Does the linear model or exponential model provide a better fit? Is either \( R \)-squared very large?

**C10** Use the data in CPS91.RAW for this exercise. These data are for married women, where we also have information on each husband’s income and demographics.
(i) What fraction of the women report being in the labour force?
(ii) Using only the data for working women—you have no choice—estimate the wage equation

\[ \log(wage) = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{exper} + \beta_3 \text{exper}^2 + \beta_4 \text{black} + \beta_5 \text{asian} + u \]

by ordinary least squares. Report the results in the usual form. Do there appear to be significant wage differences by race and ethnicity?

(iii) Estimate a probit model for \text{inlf} that includes the explanatory variables in the wage equation from part (ii) as well as \text{nwifeinc} and \text{kidlt6}. Do these last two variables have coefficients of the expected sign? Are they statistically significant?

(iv) Explain why, for the purposes of testing and, possibly, correcting the wage equation for selection into the workforce, it is important for \text{nwifeinc} and \text{kidlt6} to help explain \text{inlf}. What must you assume about \text{nwifeinc} and \text{kidlt6} in the wage equation?

(v) Compute the inverse Mills ratio (for each observation) and add it as an additional regressor to the wage equation from part (ii). What is its two-sided \( p \)-value? Do you think this is particularly small with 3,286 observations?

(vi) Does adding the inverse Mills ratio change the coefficients in the wage regression in important ways? Explain.
Chapter 18

Problems

1 Consider equation (18.15) with \( k = 2 \). Using the IV approach to estimating the \( \gamma_k \) and \( \rho \), what would you use as instruments for \( y_{t-1} \)?

2 An interesting economic model that leads to an econometric model with a lagged dependent variable relates \( y_t \) to the expected value of \( x_t \), say, \( x_t^* \), where the expectation is based on all observed information at time \( t - 1 \):

\[
y_t = \alpha_0 + \alpha_1 x_t^* + u_t.
\]  

[18.68]

A natural assumption on \( \{u_t\} \) is that \( E(u_t | I_{t-1}) = 0 \), where \( I_{t-1} \) denotes all information on \( y \) and \( x \) observed at time \( t - 1 \); this means that \( E(y_t | I_{t-1}) = \alpha_0 + \alpha_1 x_t^* \). To complete this model, we need an assumption about how the expectation \( x_t^* \) is formed. We saw a simple example of adaptive expectations in Section 11.2, where \( x_t^* = x_{t-1} \). A more complicated adaptive expectations scheme is

\[
x_t^* - x_{t-1}^* = \lambda (x_{t-1} - x_{t-1}^*),
\]  

[18.69]

where \( 0 < \lambda < 1 \). This equation implies that the change in expectations reacts to whether last period’s realised value was above or below its expectation. The assumption \( 0 < \lambda < 1 \) implies that the change in expectations is a fraction of last period’s error.

(i) Show that the two equations imply that

\[
y_t = \lambda \alpha_0 + (1 - \lambda) y_{t-1} + \lambda \alpha_1 x_{t-1} + u_t - (1 - \lambda) u_{t-1}.
\]  

[Hint: Lag equation (18.68) one period, multiply it by \( 1 - \lambda \), and subtract this from (18.68). Then, use (18.69).]

(ii) Under \( E(u_t | I_{t-1}) = 0 \), \( \{u_t\} \) is serially uncorrelated. What does this imply about the new errors, \( v_t = u_t - (1 - \lambda) u_{t-1} \)?

(iii) If we write the equation from part (i) as

\[
y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 x_{t-1} + v_t,
\]

how would you consistently estimate the \( \beta_j \)?

(iv) Given consistent estimators of the \( \beta_j \), how would you consistently estimate \( \lambda \) and \( \alpha_1 \)?

3 Suppose that \( \{y_t\} \) and \( \{z_t\} \) are I(1) series, but \( y_t - \beta z_t \) is I(0) for some \( \beta \neq 0 \). Show that for any \( \delta \neq \beta \), \( y_t - \delta z_t \) must be I(1).

4 Consider the error correction model in equation (18.37). Show that if you add another lag of the error correction term, \( y_{t-2} - \beta x_{t-2} \), the equation suffers from perfect collinearity. (Hint: Show that \( y_{t-2} - \beta x_{t-2} \) is a perfect linear function of \( y_{t-1} - \beta x_{t-1}, \Delta x_{t-1}, \text{and } \Delta y_{t-1} \).)

5 Suppose the process \( \{(x_t, y_t): t = 0, 1, 2, \ldots\} \) satisfies the equations

\[
y_t = \beta x_t + u_t
\]
and

$$\Delta x_t = \gamma \Delta x_{t-1} + v_t,$$

where $E(u_t|I_{t-1}) = E(v_t|I_{t-1}) = 0$, $I_{t-1}$ contains information on $x$ and $y$ dated at time $t - 1$ and earlier, $\beta \neq 0$, and $|\gamma| < 1$ [so that $x_t$, and therefore $y_t$, is $I(1)$]. Show that these two equations imply an error correction model of the form

$$\Delta y_t = \gamma_1 \Delta x_{t-1} + \delta (y_{t-1} - \beta x_{t-1}) + e_t,$$

where $\gamma_1 = \beta \gamma$, $\delta = -1$, and $e_t = u_t + \beta v_t$. (Hint: First subtract $y_{t-1}$ from both sides of the first equation. Then, add and subtract $\beta x_{t-1}$ from the right-hand side and rearrange. Finally, use the second equation to get the error correction model that contains $\Delta x_{t-1}$.)

6 The following equations were estimated using the data in TRAFFIC2.RAW, where $prcfat$ is the percentage of accidents resulting in a fatality, $spdlaw$ is a dummy variable taking on the value one when the speed limit was increased to 65 kilometres per hour, and $beltlaw$ is a dummy variable equal to one when a mandatory seat belt law took effect. Because the data are monthly, the regressions include a full set of monthly dummies (not shown), as well as the unemployment rate and the number of weekends in a month (not shown):

\[
\begin{align*}
\hat{prcfat} &= .891 + .026 spdlaw - .100 beltlaw + ... \\
&= .891 + .026 + .022 - .100 - .022 + ...
\end{align*}
\]

\[
\hat{prcfat} = 108, R^2 = .630
\]

and

\[
\begin{align*}
\hat{prcfat} &= .417 + .476 \hat{prcfat}_{t-1} + .011 spdlaw - .050 beltlaw + ... \\
&= .417 + .476 + .094 + .011 - .050 - .022 + ...
\end{align*}
\]

\[
\hat{prcfat} = 107, R^2 = .715
\]

where the standard errors are the usual OLS standard errors in both cases.

(i) From the static model, what is the estimated long-run effect of the seat belt law on $prcfat$? Is it statistically significant? What might you do to get a more reliable standard error?

(ii) Using the dynamic model, what is the estimated long-run effect of the seat belt law on $prcfat$? How does it compare with the estimate from the static model?

(iii) When a second lag $prcfat_{t-2}$ is added along with the first lag, the coefficient estimate on $prcfat_{t-2}$ is .098 with a standard error of .110. Does the second lag need to be included?

7 Let $gM_t$ be the annual growth in the money supply and let $unem_t$ be the unemployment rate. Assuming that $unem_t$ follows a stable AR(1) process, explain in detail how you would test whether $gM$ Granger causes $unem_t$.

8 Suppose that $y_t$ follows the model

$$y_t = \alpha + \delta_1 z_{t-1} + u_t$$

$$u_t = \rho u_{t-1} + e_t$$

$$E(e_t|I_{t-1}) = 0,$$

where $I_{t-1}$ contains $y$ and $z$ dated at $t - 1$ and earlier.
(i) Show that $E(y_{t+1}|I_t) = (1 - \rho)\alpha + \rho y_t + \delta_1 z_t - \rho \delta z_{t-1}$. (Hint: Write $u_{t-1} = y_{t-1} - \alpha - \delta z_{t-2}$ and plug this into the second equation; then, plug the result into the first equation and take the conditional expectation.)

(ii) Suppose that you use $n$ observations to estimate $\alpha$, $\delta_1$, and $\rho$. Write the equation for forecasting $y_{n+1}$.

(iii) Explain why the model with one lag of $z$ and AR(1) serial correlation is a special case of the model

$$y_t = \alpha_0 + \rho y_{t-1} + \gamma_1 z_{t-1} + \gamma_2 z_{t-2} + \epsilon_t.$$

(iv) What does part (iii) suggest about using models with AR(1) serial correlation for forecasting?

9 Let $\{y_t\}$ be an I(1) sequence. Suppose that $\hat{g}_n$ is the one-step-ahead forecast of $\Delta y_{n+1}$ and let $\hat{f}_n = \hat{g}_n + y_n$ be the one-step-ahead forecast of $y_{n+1}$. Explain why the forecast errors for forecasting $\Delta y_{n+1}$ and $y_{n+1}$ are identical.

**Computer Exercises**

**C1** Use the data in WAGEPRC.RAW for this exercise. Problem 5 in Chapter 11 gave estimates of a finite distributed lag model of $gprice$ on $gwage$, where 12 lags of $gwage$ are used.

(i) Estimate a simple geometric DL model of $gprice$ on $gwage$. In particular, estimate equation (18.11) by OLS. What are the estimated impact propensity and LRP? Sketch the estimated lag distribution.

(ii) Compare the estimated IP and LRP to those obtained in Problem 5 in Chapter 11. How do the estimated lag distributions compare?

(iii) Now, estimate the rational distributed lag model from (18.16). Sketch the lag distribution and compare the estimated IP and LRP to those obtained in part (ii).

**C2** Use the data in HSEINV.RAW for this exercise.

(i) Test for a unit root in log($invpc$), including a linear time trend and two lags of $\Delta log(invpc)$. Use a 5% significance level.

(ii) Use the approach from part (i) to test for a unit root in log($price$).

(iii) Given the outcomes in parts (i) and (ii), does it make sense to test for cointegration between log($invpc$) and log($price$)?

**C3** Use the data in VOLAT.RAW for this exercise.

(i) Estimate an AR(3) model for $pcip$. Now, add a fourth lag and verify that it is very insignificant.

(ii) To the AR(3) model from part (i), add three lags of $pcsp$ to test whether $pcsp$ Granger causes $pcip$. Carefully state your conclusion.

(iii) To the model in part (ii), add three lags of the change in $i^3$, the three-month T-bill rate. Does $pcsp$ Granger cause $pcip$ conditional on past $\Delta i^3$?

**C4** In testing for cointegration between $gfr$ and $pe$ in Example 18.5, add $i^2$ to equation (18.32) to obtain the OLS residuals. Include one lag in the augmented DF test. The 5% critical value for the test is $-4.15$. 
C5 Use INTQRT.RAW for this exercise.

(i) In Example 18.7, we estimated an error correction model for the holding yield on six-month T-bills, where one lag of the holding yield on three-month government bonds is the explanatory variable. We assumed that the cointegration parameter was one in the equation $hy_6 = \alpha + \beta hy_3 - 1 + u_t$. Now, add the lead change, $\Delta hy_3$, the contemporaneous change, $\Delta hy_3 - 1$, and the lagged change, $\Delta hy_3 - 2$, of $hy_3 - 1$. That is, estimate the equation

$$hy_6 = \alpha + \beta hy_3 - 1 + \phi_0 \Delta hy_3 + \phi_1 \Delta hy_3 - 1 + \rho_1 \Delta hy_3 - 2 + e_t$$

and report the results in equation form. Test $H_0: \beta = 1$ against a two-sided alternative. Assume that the lead and lag are sufficient so that $\{hy_3 - 1\}$ is strictly exogenous in this equation and do not worry about serial correlation.

(ii) To the error correction model in (18.39), add $\Delta hy_3 - 2$ and $(hy_6 - 2 - hy_3 - 3)$. Are these terms jointly significant? What do you conclude about the appropriate error correction model?

C6 Use the data in PHILLIPS.RAW to answer these questions.

(i) Estimate the models in (18.48) and (18.49) using the data through 1997. Do the parameter estimates change much compared with (18.48) and (18.49)?

(ii) Use the new equations to forecast $unem_{1998}$; round to two places after the decimal. Which equation produces a better forecast?

(iii) As we discussed in the text book, the forecast for $unem_{1998}$ using (18.49) is 4.90. Compare this with the forecast obtained using the data through 1997. Does using the extra year of data to obtain the parameter estimates produce a better forecast?

(iv) Use the model estimated in (18.48) to obtain a two-step-ahead forecast of $unem$. That is, forecast $unem_{1998}$ using equation (18.55) with $\tilde{\alpha} = 1.572, \tilde{\rho} = .732$, and $h = 2$. Is this better or worse than the one-step-ahead forecast obtained by plugging $unem_{1997} = 4.9$ into (18.48)?

C7 Use the data in BARIUM.RAW for this exercise.

(i) Estimate the linear trend model $chlnimp = \alpha + \beta t + u_t$, using the first 119 observations (this excludes the last 12 months of observations for 1988). What is the standard error of the regression?

(ii) Now, estimate an AR(1) model for $chlnimp$, again using all data but the last 12 months. Compare the standard error of the regression with that from part (i). Which model provides a better in-sample fit?

(iii) Use the models from parts (i) and (ii) to compute the one-step-ahead forecast errors for the 12 months in 1988. (You should obtain 12 forecast errors for each method.) Compute and compare the RMSEs and the MAEs for the two methods. Which forecasting method works better out-of-sample for one-step-ahead forecasts?

(iv) Add monthly dummy variables to the regression from part (i). Are these jointly significant? (Do not worry about the slight serial correlation in the errors from this regression when doing the joint test.)

C8 Use the data in FERTIL3.RAW for this exercise.

(i) Graph $gfr$ against time. Does it contain a clear upward or downward trend over the entire sample period?
(ii) Using the data through 1979, estimate a cubic time trend model for \( gfr \) (that is, regress \( gfr \) on \( t, t^2, \) and \( t^3 \), along with an intercept). Comment on the \( R \)-squared of the regression.

(iii) Using the model in part (ii), compute the mean absolute error of the one-step-ahead forecast errors for the years 1980 through 1984.

(iv) Using the data through 1979, regress \( \Delta gfr \) on a constant only. Is the constant statistically different from zero? Does it make sense to assume that any drift term is zero, if we assume that \( gfr \) follows a random walk?

(v) Now, forecast \( gfr \) for 1980 through 1984, using a random walk model: the forecast of \( gfr_{n+1} \) is simply \( gfr_n \). Find the MAE. How does it compare with the MAE from part (iii)? Which method of forecasting do you prefer?

(vi) Now, estimate an AR(2) model for \( gfr \), again using the data only through 1979. Is the second lag significant?

(vii) Obtain the MAE for 1980 through 1984, using the AR(2) model. Does this more general model work better out-of-sample than the random walk model?

C9 Use CONSUMP.RAW for this exercise.

(i) Let \( y_t \) be real per capita disposable income. Use the data through 1989 to estimate the model

\[ y_t = \alpha + \beta t + \rho y_{t-1} + u_t \]

and report the results in the usual form.

(ii) Use the estimated equation from part (i) to forecast \( y \) in 1990. What is the forecast error?

(iii) Compute the mean absolute error of the one-step-ahead forecasts for the 1990s, using the parameters estimated in part (i).

(iv) Now, compute the MAE over the same period, but drop \( y_{t-1} \) from the equation. Is it better to include \( y_{t-1} \) in the model or not?

C10 Use the data in INTQRT.RAW for this exercise.

(i) Using the data from all but the last four years (16 quarters), estimate an AR(1) model for \( \Delta r6_t \). (We use the difference because it appears that \( r6_t \) has a unit root.) Find the RMSE of the one-step-ahead forecasts for \( \Delta r6 \), using the last 16 quarters.

(ii) Now, add the error correction term \( spr_{t-1} = r6_{t-1} - r3_{t-1} \) to the equation from part (i). (This assumes that the cointegrating parameter is one.) Compute the RMSE for the last 16 quarters. Does the error correction term help with out-of-sample forecasting in this case?

(iii) Now, estimate the cointegrating parameter, rather than setting it to one. Use the last 16 quarters again to produce the out-of-sample RMSE. How does this compare with the forecasts from parts (i) and (ii)?

(iv) Would your conclusions change if you wanted to predict \( r6 \) rather than \( \Delta r6 \)? Explain.

C11 Use the data in VOLAT.RAW for this exercise.

(i) Confirm that \( lsn500 = \log(sp500) \) and \( lip = \log(ip) \) appear to contain unit roots. Use Dickey-Fuller tests with four lagged changes and do the tests with and without a linear time trend.
(ii) Run a simple regression of \( \text{lsp}500 \) on \( \text{lip} \). Comment on the sizes of the \( t \) statistic and \( R \)-squared.

(iii) Use the residuals from part (ii) to test whether \( \text{lsp}500 \) and \( \text{lip} \) are cointegrated. Use the standard Dickey-Fuller test and the ADF test with two lags. What do you conclude?

(iv) Add a linear time trend to the regression from part (ii) and now test for cointegration using the same tests from part (iii).

(v) Does it appear that stock prices and real economic activity have a long-run equilibrium relationship?

**C12** Use the data in MINWAGE.RAW for this question, using sector 314 (Footwear, Except Rubber).

(i) Estimate an AR(1) model for employment growth, \( g\text{emp}314 \). Is the lag practically and statistically significant?

(ii) Test whether wage growth Granger causes employment growth by adding a lag of wage growth to the AR(1) model. What do you conclude?

(iii) Add a second lag of employment growth in the Granger causality test. Do your conclusions change?

**C13** Use the data in FISH.RAW to answer this question.

(i) Obtain the standard Dickey-Fuller unit root test for \( \text{lavgprc} \), without allowing for a trend. What do you conclude?

(ii) Augment the DF test by including one lag of \( g\text{avgprc} = \text{lavgprc} - \text{lavgprc}_{-1} \). How does this affect your views on whether \( \text{lavgprc} \) contains a unit root?

(iii) Now test \( \text{ltotqty} \) for a unit root, using the DF and augmented DF tests allowing for one lag. What do you conclude?

(iv) Based on your findings in parts (i) through (iii), does differencing the price and quantity variables seem necessary before using them in regression (or instrumental variable) estimation?