Chapter 6
Specification of regression variables

6.1 Overview

This chapter treats a variety of topics relating to the specification of the variables in a regression model. First there are the consequences for the regression coefficients, their standard errors, and R^2 of failing to include a relevant variable, and of including an irrelevant one. This leads to a discussion of the use of proxy variables to alleviate a problem of omitted variable bias. Next come F and t tests of the validity of a restriction, the use of which was advocated in Chapter 3 as a means of improving efficiency and perhaps mitigating a problem of multicollinearity. The chapter concludes by outlining the potential benefit to be derived from examining observations with large residuals after fitting a regression model.

6.2 Learning outcomes

After working through the corresponding chapter in the text, studying the corresponding slideshows, and doing the starred exercises in the text and the additional exercises in this subject guide, you should be able to:

- derive the expression for the bias in an OLS estimator of a slope coefficient when the true model has two explanatory variables but the regression model has only one
- determine the likely direction of omitted variable bias, given data on the correlation between the explanatory variables
- explain the consequence of omitted variable bias for the standard errors of the coefficients and for t tests and F tests
- explain the consequences of including an irrelevant variable for the regression coefficients, their standard errors, and t and F tests
- explain how the regression results are affected by the substitution of a proxy variable for a missing explanatory variable
- perform an F test of a restriction, stating the null hypothesis for the test
- perform a t test of a restriction, stating the null hypothesis for the test.

6.3 Additional exercises

A6.1 *Does the omission of total household expenditure or household size give rise to omitted variable bias in your CES regressions?*

Regress *LGCATPC* (1) on both *LGEXPPC* and *LGSIZE*, (2) on *LGEXPPC* only, and (3) on *LGSIZE* only. Assuming that (1) is the correct specification, analyse the likely direction of the bias in the estimate of the coefficient of *LGEXPPC* in (2) and that of *LGSIZE* in (3). Check whether the regression results are consistent with your analysis.

A6.2 A school has introduced an extra course of reading lessons for children starting school and a researcher is evaluating the impact of the course on the scores on a literacy test taken at the age of seven. In the first year of its implementation, those children whose surnames begin A–M are assigned to the extra course, while the rest have the normal curriculum. The researcher hypothesises that:

$$Y = \beta_1 + \beta_2 D + \beta_3 A + u$$

where Y is the score on the literacy test, D is a dummy variable that is equal to 1 for those assigned to the extra course and 0 for the others, A is a measure of the cognitive ability of the child when starting school, and u is an iid (independently and identically distributed) disturbance term assumed to have a normal distribution. Unfortunately, the researcher has no data on A. Using OLS (ordinary least squares), she fits the regression:

$$\widehat{Y} = \widehat{\beta}_1 + \widehat{\beta}_2 D.$$

- Demonstrate that $\widehat{\beta}_2$ is an unbiased estimator of β_2.
- A commentator says that the standard error of $\widehat{\beta}_2$ will be invalid because an important variable, A, has been omitted from the specification. The researcher replies that the standard error will remain valid if A can be assumed to have a normal distribution. Explain whether the commentator or the researcher is correct.
- Another commentator says that whether the distribution of A is normal or not makes no difference to the validity of the standard error. Evaluate this assertion.

A6.3 A researcher obtains data on household annual expenditure on books, B, and annual household income, Y, for 100 households. He hypothesises that B is related to Y and the average cognitive ability of adults in the household, IQ, by the relationship:

$$\log B = \beta_1 + \beta_2 \log Y + \beta_3 \log IQ + u \qquad \text{(A)}$$

where u is a disturbance term that satisfies the regression model assumptions. He also considers the possibility that $\log B$ may be determined by $\log Y$ alone:

$$\log B = \beta_1 + \beta_2 \log Y + u. \qquad \text{(B)}$$

He does not have data on IQ and decides to use average years of schooling of the adults in the household, S, as a proxy in specification (A). It may be assumed that Y and S are both nonstochastic. In the sample the correlation between $\log Y$ and $\log S$ is 0.86. He performs the following regressions: (1) $\log B$ on both $\log Y$ and $\log S$, and (2) $\log B$ on $\log Y$ only, with the results shown in the table (standard errors in parentheses):

	(1)	(2)
$\log Y$	1.10	2.10
	(0.69)	(0.35)
$\log S$	0.59	—
	(0.35)	
constant	−6.89	−3.37
	(2.28)	(0.89)
R^2	0.29	0.27

- Assuming that (A) is the correct specification, explain, with a mathematical proof, whether you would expect the coefficient of $\log Y$ to be greater in regression (2).
- Assuming that (A) is the correct specification, describe the various benefits from using $\log S$ as a proxy for $\log IQ$, as in regression (1), if $\log S$ is a good proxy.
- Explain whether the low value of R^2 in regression (1) implies that $\log S$ is not a good proxy.
- Assuming that (A) is the correct specification, provide an explanation of why the coefficients of $\log Y$ and $\log S$ in regression (1) are not significantly different from zero, using two-sided t tests.
- Discuss whether the researcher would be justified in using one-sided t tests in regression (1).
- Assuming that (B) is the correct specification, explain whether you would expect the coefficient of $\log Y$ to be lower in regression (1).
- Assuming that (B) is the correct specification, explain whether the standard errors in regression (1) are valid estimates.

A6.4 A researcher has the following data for the year 2012: T, annual total sales of cinema tickets per household, and P, the average price of a cinema ticket in the city. She believes that the true relationship is:

$$\log T = \beta_1 + \beta_2 \log P + \beta_3 \log Y + u$$

where Y is average household income, but she lacks data on Y and fits the regression (standard errors in parentheses):

$$\widehat{\log T} = \underset{(0.52)}{13.74} + \underset{(0.23)}{0.17} \log P \qquad R^2 = 0.01$$

Explain analytically whether the slope coefficient is likely to be biased. You are told that if the researcher had been able to obtain data on Y, her regression would have been:

$$\widehat{\log T} = \underset{(2.93)}{-1.63} \underset{(0.21)}{- 0.48 \log P} + \underset{(0.35)}{1.83 \log Y} \qquad R^2 = 0.44$$

You are also told that Y and P are positively correlated.

The researcher is not able to obtain data on Y but, from local records, she is able to obtain data on H, the average value of a house in each city, and she decides to use it as a proxy for Y. She fits the following regression (standard errors in parentheses):

$$\widehat{\log T} = \underset{(3.22)}{-0.63} \underset{(0.22)}{- 0.37 \log P} + \underset{(0.38)}{1.69 \log H} \qquad R^2 = 0.36$$

Describe the theoretical benefits from using H as a proxy for Y, discussing whether they appear to have been obtained in this example.

A6.5 A researcher has data on years of schooling, S, weekly earnings in dollars, W, hours worked per week, H, and hourly earnings, E (computed as W/H) for a sample of 1,755 white males in the United States in the year 2000. She calculates LW, LE, and LH as the natural logarithms of W, E, and H, respectively, and fits the following regressions, with the results shown in the table below (standard errors in parentheses; RSS = residual sum of squares):

- Column 1: a regression of LE on S.
- Column 2: a regression of LW on S and LH.
- Column 3: a regression of LE on S and LH.

The correlation between S and LH is 0.06.

	(1)	(2)	(3)	(4)	(5)
Respondents	All	All	All	FT	PT
Dependent variable	*LE*	*LW*	*LE*	*LW*	*LW*
S	0.099	0.098	0.098	0.101	0.030
	(0.006)	(0.006)	(0.006)	(0.006)	(0.049)
LH	—	1.190	0.190	0.980	0.885
		(0.065)	(0.065)	(0.088)	(0.325)
constant	6.111	5.403	5.403	6.177	7.002
	(0.082)	(0.254)	(0.254)	(0.345)	(1.093)
RSS	741.5	737.9	737.9	626.1	100.1
Observations	1,755	1,755	1,755	1,669	86

- Explain why specification (1) is a restricted version of specification (2), stating and interpreting the restriction.
- Supposing the restriction to be valid, explain whether you expect the coefficient of S and its standard error to differ, or be similar, in specifications (1) and (2).

- Supposing the restriction to be invalid, how would you expect the coefficient of S and its standard error to differ, or be similar, in specifications (1) and (2)?
- Perform an F test of the restriction.
- Perform a t test of the restriction.
- Explain whether the F test and the t test could lead to different conclusions.
- At a seminar, a commentator says that part-time workers tend to be paid worse than full-time workers and that their earnings functions are different. Defining full-time workers as those working at least 35 hours per week, the researcher divides the sample and fits the earnings functions for full-time workers (column 4) and part-time workers (column 5). Test whether the commentator's assertion is correct.
- What are the implications of the commentator's assertion for the test of the restriction?

A6.6 A researcher investigating whether government expenditure tends to crowd out investment has data on government recurrent expenditure, G, investment, I, and gross domestic product, Y, all measured in US$ billion, for 30 countries in 2012. She fits two regressions (standard errors in parentheses; t statistics in square brackets; RSS = residual sum of squares).

(1) A regression of $\log I$ on $\log G$ and $\log Y$:

$$\begin{array}{rcl} \widehat{\log I} & = & -2.44 - 0.63 \log G + 1.60 \log Y \qquad R^2 = 0.98 \qquad (1) \\ & & \begin{array}{llll} (0.26) & (0.12) & (0.12) & RSS = 0.90 \\ [9.42] & [-5.23] & [12.42] & \end{array} \end{array}$$

(2) a regression of $\log(I/Y)$ on $\log(G/Y)$:

$$\begin{array}{rcl} \widehat{\log\left(\frac{I}{Y}\right)} & = & 2.65 - 0.63 \log\left(\frac{G}{Y}\right) \qquad R^2 = 0.48 \qquad (2) \\ & & \begin{array}{ll} (0.23) \ (0.12) & RSS = 0.99 \\ [11.58] \ [-5.07] & \end{array} \end{array}$$

The correlation between $\log G$ and $\log Y$ in the sample is 0.98. The table gives some further basic data on $\log G$, $\log Y$, and $\log(G/Y)$.

	Sample mean	Mean square deviation
$\log G$	3.75	2.00
$\log Y$	5.57	1.95
$\log (G/Y)$	−1.81	0.08

- Explain why the second specification is a restricted version of the first. State the restriction.
- Perform a test of the restriction.

- The researcher expected the standard error of the coefficient of $\log(G/Y)$ in (2) to be smaller than the standard error of the coefficient of $\log G$ in (1). Explain why she expected this.
- However, the standard error is the same, at least to two decimal places. Give an explanation.
- Show how the restriction could be tested using a t test in a reparameterised version of the specification for (1).

A6.7 *Is expenditure per capita on your CES category related to total household expenditure per capita?*

The model specified in Exercise A4.1 is a restricted version of that in Exercise 4.5 in the text. Perform an F test of the restriction. Also perform a t test of the restriction.

[Exercise 4.5: regress *LGCAT* on *LGEXP* and *LGSIZE*; Exercise A4.1: regress *LGCATPC* on *LGEXPPC*.]

A6.8 A researcher is considering two regression specifications:

$$\log Y = \beta_1 + \beta_2 \log X + u \qquad (1)$$

and:

$$\log \frac{Y}{X} = \alpha_1 + \alpha_2 \log X + u \qquad (2)$$

where u is a disturbance term. Determine whether (2) is a reparameterised or a restricted version of (1).

A6.9 Three researchers investigating the determinants of hourly earnings have the following data for a sample of 104 male workers in the United States in 2006: E, hourly earnings in dollars; S, years of schooling; *NUM*, score on a test of numeracy; and *VERB*, score on a test of literacy. The *NUM* and *VERB* tests are marked out of 100. The correlation between them is 0.81. Defining *LGE* to be the natural logarithm of E, Researcher 1 fits the following regression (standard errors in parentheses; *RSS* = residual sum of squares):

$$\begin{array}{rcll} \widehat{LGE} & = & 2.02 + 0.063S + 0.0044NUM + 0.0026\,VERB & RSS = 2{,}000 \\ & & (1.81)\ \ (0.007)\ \ (0.0011) \qquad (0.0010) & \end{array}$$

Researcher 2 defines a new variable *SCORE* as the average of *NUM* and *VERB*. She fits the regression:

$$\begin{array}{rcll} \widehat{LGE} & = & 1.72 + 0.050S + 0.0068SCORE & RSS = 2{,}045 \\ & & (1.78)\ \ (0.005)\ \ (0.0010) & \end{array}$$

Researcher 3 fits the regression:

$$\begin{array}{rcll} \widehat{LGE} & = & 2.02 + 0.063S + 0.0088SCORE - 0.0018\,VERB & RSS = 2{,}000 \\ & & (1.81)\ \ (0.007)\ \ (0.0022) \qquad (0.0012) & \end{array}$$

- Show that the specification of Researcher 2 is a restricted version of the specification of Researcher 1, stating the restriction.
- Perform an F test of the restriction.
- Show that the specification of Researcher 3 is a reparameterised version of the specification of Researcher 1 and hence perform a t test of the restriction in the specification of Researcher 2.
- Explain whether the F test and the t test could have led to different results.
- Perform a test of the hypothesis that the numeracy score has a greater effect on earnings than the literacy score.
- Compare the regression results of the three researchers.

A6.10 It is assumed that manufacturing output is subject to the production function:

$$Q = AK^{\alpha}L^{\beta} \qquad (1)$$

where Q is output and K and L are capital and labour inputs. The cost of production is:

$$C = \rho K + wL \qquad (2)$$

where ρ is the cost of capital and w is the wage rate. It can be shown that, if the cost is minimised, the wage bill wL will be given by the relationship:

$$\log wL = \frac{1}{\alpha+\beta}\log Q + \frac{\alpha}{\alpha+\beta}\log \rho + \frac{\beta}{\alpha+\beta}\log w + \text{constant}. \qquad (3)$$

(Note: You are not expected to prove this.)

A researcher has annual data for 2002 for Q, K, L, ρ, and w (all monetary measures being converted into US dollars) for the manufacturing sectors of 30 industrialised countries and regresses $\log wL$ on $\log Q$, $\log \rho$, and $\log w$.

- Demonstrate that relationship (3) embodies a testable restriction and show how the model may be reformulated to take advantage of it.
- Explain how the restriction could be tested using an F test.
- Explain how the restriction could be tested using a t test.
- Explain the theoretical benefits of making use of a valid restriction. How could the researcher assess whether there are any benefits in practice, in this case?
- At a seminar, someone suggests that it is reasonable to hypothesise that manufacturing output is subject to constant returns to scale, so that $\alpha + \beta = 1$. Explain how the researcher could test this hypothesis (1) using an F test, (2) using a t test.

A6.11 A researcher hypothesises that the net annual growth of private sector purchases of government bonds, B, is positively related to the nominal rate of interest on the bonds, I, and negatively related to the rate of price inflation, P:

$$B = \beta_1 + \beta_2 I + \beta_3 P + u$$

where u is a disturbance term. The researcher anticipates that $\beta_2 > 0$ and $\beta_3 < 0$. She also considers the possibility that B depends on the real rate of interest on the bonds, R, where $R = I - P$. Using a sample of observations for 40 countries, she regresses B:

- (1) on I and P
- (2) on R
- (3) on I
- (4) on P and R

with the results shown in the corresponding columns of the table below (standard errors in parentheses; *RSS* is the residual sum of squares). The correlation coefficient for I and P was 0.97.

	(1)	(2)	(3)	(4)
I	2.17	—	0.69	—
	(1.04)		(0.25)	
P	−3.19	—	—	−1.02
	(2.17)			(1.19)
R	—	1.37	—	2.17
		(0.44)		(1.04)
constant	−5.14	−3.15	−1.53	−5.14
	(2.62)	(1.21)	(0.92)	(2.62)
R^2	0.22	0.20	0.17	0.22
RSS	967.9	987.1	1,024.3	967.9

- Explain why the researcher was dissatisfied with the results of regression (1).
- Demonstrate that specification (2) may be considered to be a restricted version of specification (1).
- Perform an F test of the restriction, stating carefully your null hypothesis and conclusion.
- Perform a t test of the restriction.
- Demonstrate that specification (3) may also be considered to be a restricted version of specification (1).
- Perform both an F test and a t test of the restriction in specification (3), stating your conclusion in each case.
- At a seminar, someone suggests that specification (4) is also a restricted version of specification (1). Is this correct? If so, state the restriction.
- State, with an explanation, which would be your preferred specification.

A6.12 A researcher has a sample of 43 observations on a dependent variable, Y, and two potential explanatory variables, X and Z. He defines two further variables V and W as the sum of X and Z and the difference between them:

$$V_i = X_i + Z_i$$

$$W_i = X_i - Z_i.$$

He fits the following four regressions:

(1) A regression of Y on X and Z.

(2) A regression of Y on V and W.

(3) A regression of Y on V.

(4) A regression of Y on Z and V.

The table shows the regression results (standard errors in parentheses; RSS = residual sum of squares; there was an intercept, not shown, in each regression). Unfortunately, a goat ate part of the regression output and some of the numbers are missing. These are indicated by letters.

	(1)	(2)	(3)	(4)
X	0.60	—	—	—
	(0.04)			
Z	0.80	—	—	**H**
	(0.04)		**(I)**	
V	—	**A**	0.72	**J**
		(B)	(0.02)	**(K)**
W	—	**C**	—	—
		(D)		
R^2	0.60	**E**	**G**	**L**
RSS	200	**F**	220	**M**

Each regression included an intercept (not shown).

Reconstruct each missing number if this is possible, giving a brief explanation. If it is not possible to reconstruct a number, give a brief explanation.

A6.13 In Exercise A6.7, a researcher proposes to test the restriction using variations in R^2 instead of variations in RSS. For food consumed at home, the unrestricted regression of $LGFDHO$ on $LGEXP$ and $LGSIZE$ had $R^2 = 0.4831$. For the regression of $LGFDHOPC$ on $LGEXPPC$, $R^2 = 0.4290$. Hence the researcher's statistic is:

$$F = \frac{(0.4831 - 0.4290)/1}{(1 - 0.4290)/6331} = 599.8.$$

Explain why this is different from the F statistic reported for food consumed at home in the answer to Exercise A6.7.

6.4 Answers to the starred exercises in the textbook

6.4 The table gives the results of multiple and simple regressions of $LGFDHO$, the logarithm of annual household expenditure on food eaten at home, on $LGEXP$, the logarithm of total annual household expenditure, and $LGSIZE$, the logarithm of the number of persons in the household, using a sample of 6,334 households in the 2013 Consumer Expenditure Survey. The correlation coefficient for $LGEXP$ and $LGSIZE$ was 0.32. Explain the variations in the regression coefficients.

	(1)	(2)	(3)
LGEXP	0.58	0.67	—
	(0.01)	(0.01)	
LGSIZE	0.33	—	0.58
	(0.01)		(0.02)
constant	1.16	0.70	6.04
	(0.08)	(0.08)	(0.01)
R^2	0.48	0.43	0.19

Answer:

If the model is written as:

$$LGFDHO = \beta_1 + \beta_2 LGEXP + \beta_3 LGSIZE + u$$

the expected value of $\widehat{\beta}_2$ in the second regression is given by:

$$E(\widehat{\beta}_2) = \beta_2 + \beta_3 \frac{\sum \left(LGEXP_i - \overline{LGEXP}\right)\left(LGSIZE_i - \overline{LGSIZE}\right)}{\sum \left(LGEXP_i - \overline{LGEXP}\right)^2}.$$

We know that the covariance is positive because the correlation is positive, and it is reasonable to suppose that β_3 is also positive, especially given the highly significant positive estimate in the first regression, and so $\widehat{\beta}_2$ is biased upwards. This accounts for the large increase in its size in the second regression. In the third regression:

$$E(\widehat{\beta}_3) = \beta_3 + \beta_2 \frac{\sum \left(LGEXP_i - \overline{LGEXP}\right)\left(LGSIZE_i - \overline{LGSIZE}\right)}{\sum \left(LGSIZE_i - \overline{LGSIZE}\right)^2}.$$

β_2 is certainly positive, especially given the highly significant positive estimate in the first regression, and so $\widehat{\beta}_3$ is also biased upwards. As a consequence, the estimate in the third regression is greater than that in the first.

6.7 A researcher thinks that the level of activity in the shadow economy, Y, depends either positively on the level of the tax burden, X, or negatively on the level of government expenditure to discourage shadow economy activity, Z. Y might also depend on both X and Z. International cross-sectional data on Y, X, and Z, all measured in US$ million, are obtained for a sample of 30 industrialised countries and a second sample of 30 developing countries. The researcher regresses (1) $\log Y$ on both $\log X$ and $\log Z$, (2) $\log Y$ on $\log X$ alone, and (3) $\log Y$ on $\log Z$ alone, for each sample, with the following results (standard errors in parentheses):

	Industrialised countries			Developing countries		
	(1)	(2)	(3)	(1)	(2)	(3)
$\log X$	0.699	0.201	—	0.806	0.727	—
	(0.154)	(0.112)		(0.137)	(0.090)	
$\log Z$	−0.646	—	−0.053	−0.091	—	0.427
	(0.162)		(0.124)	(0.117)		(0.116)
constant	−1.137	−1.065	1.230	−1.122	−1.024	2.824
	(0.863)	(1.069)	(0.896)	(0.873)	(0.858)	(0.835)
R^2	0.44	0.10	0.01	0.71	0.70	0.33

X was positively correlated with Z in both samples. Having carried out the appropriate statistical tests, write a short report advising the researcher how to interpret these results.

Answer:

One way to organise an answer to this exercise is, for each sample, to consider the evidence for and against each of the three specifications in turn. The t statistics for the slope coefficients are given in the following table. * indicates significance at the 5 per cent level, ** at the 1 per cent level, and *** at the 0.1 per cent level, using one-sided tests. (Justification for one-sided tests: one may rule out a negative coefficient for X and a positive one for Y.)

	Industrialised countries			Developing countries		
	(1)	(2)	(3)	(1)	(2)	(3)
$\log X$	4.54***	1.79*	—	5.88****	8.08***	—
$\log Z$	−3.99***	—	−0.43	−0.78	—	3.68***

Industrialised countries:

The first specification is clearly the only satisfactory one for this sample, given the t statistics. Writing the model as:

$$\log Y = \beta_1 + \beta_2 \log X + \beta_3 \log Z + u$$

in the second specification:

$$E(\widehat{\beta}_2) = \beta_2 + \beta_3 \frac{\sum \left(\log X_i - \overline{\log X}\right)\left(\log Z_i - \overline{Z}\right)}{\sum \left(\log X_i - \overline{\log X}\right)^2}.$$

Anticipating that β_3 is negative, and knowing that X and Z are positively correlated, the bias term should be negative. The estimate of β_2 is indeed lower in the second specification. In the third specification:

$$E(\widehat{\beta}_3) = \beta_3 + \beta_2 \frac{\sum \left(\log X_i - \overline{\log X}\right)\left(\log Z_i - \overline{Z}\right)}{\sum \left(\log Z_i - \overline{\log Z}\right)^2}$$

and the bias should be positive, assuming β_2 is positive. $\widehat{\beta}_3$ is indeed less negative than in the first specification.

Note that the sum of the R^2 statistics for the second and third specifications is less than R^2 in the first. This is because the bias terms undermine the apparent explanatory power of X and Z in the second and third specifications. In the third specification, the bias term virtually neutralises the true effect and R^2 is very low indeed.

Developing countries:

In principle the first specification is acceptable. The failure of the coefficient of Z to be significant might be due to a combination of a weak effect of Z and a relatively small sample.

The second specification is also acceptable since the coefficient of Z and its t statistic in the first specification are very low. Because the t statistic of Z is low, R^2 is virtually unaffected when it is omitted.

The third specification is untenable because it cannot account for the highly significant coefficient of X in the first. The omitted variable bias is now so large that it overwhelms the negative effect of Z with the result that the estimated coefficient is positive.

6.11 A researcher has data on output per worker, Y, and capital per worker, K, both measured in thousands of dollars, for 50 firms in the textiles industry in 2012. She hypothesises that output per worker depends on capital per worker and perhaps also the technological sophistication of the firm, *TECH*:

$$Y = \beta_1 + \beta_2 K + \beta_3 TECH + u$$

where u is a disturbance term. She is unable to measure *TECH* and decides to use expenditure per worker on research and development in 2012, *R&D*, as a proxy for it. She fits the following regressions (standard errors in parentheses):

$$\begin{aligned} \widehat{Y} &= 1.02 + 0.32K \qquad R^2 = 0.749 \\ &\quad (0.45)\ \ (0.04) \\ \widehat{Y} &= 0.34 + 0.29K + 0.05R\&D \qquad R^2 = 0.750 \\ &\quad (0.61)\ \ (0.22)\ \ (0.15) \end{aligned}$$

The correlation coefficient for K and *R&D* is 0.92. Discuss these regression results:

1. assuming that Y does depend on both K and *TECH*
2. assuming that Y depends only on K.

Answer:

If Y depends on both K and *TECH*, the first specification is subject to omitted variable bias, with the expected value of $\widehat{\beta}_2$ being given by:

$$E(\widehat{\beta}_2) = \beta_2 + \beta_3 \frac{\sum \left(K_i - \overline{K}\right)\left(TECH_i - \overline{TECH}\right)}{\sum \left(K_i - \overline{K}\right)^2}.$$

Since K and *R&D* have a high positive correlation, it is reasonable to assume that K and *TECH* are positively correlated. It is also reasonable to assume that β_3 is positive. Hence one would expect $\widehat{\beta}_2$ to be biased upwards. It is indeed greater than in the second equation, but not by much. The second specification is clearly subject to multicollinearity, with the consequence that, although the estimated coefficients remain unbiased, they are erratic, this being reflected in large standard errors. The large variance of the estimate of the coefficient of K means that much of the difference between it and the estimate in the first specification is likely to be purely random, and this could account for the fact that the omitted variable bias appears to be so small.

If Y depends only on K, the inclusion of *R&D* in the second specification gives rise to inefficiency. Since the standard errors in both equations remain valid, they can

be compared and it is evident that the loss of efficiency is severe. As expected in this case, the coefficient of *R&D* is not significantly different from zero and the increase in R^2 in the second specification is minimal.

6.14 The first regression shows the result of regressing *LGFDHO*, the logarithm of annual household expenditure on food eaten at home, on *LGEXP*, the logarithm of total annual household expenditure, and *LGSIZE*, the logarithm of the number of persons in the household, using a sample of 6,334 households in the 2013 Consumer Expenditure Survey. In the second regression, *LGFDHOPC*, the logarithm of food expenditure per capita (*FDHO*/*SIZE*), is regressed on *LGEXPPC*, the logarithm of total expenditure per capita (*EXP*/*SIZE*). In the third regression *LGFDHOPC* is regressed on *LGEXPPC* and *LGSIZE*.

```
. reg LGFDHO LGEXP LGSIZE
------------------------------------------------------------------------------
    Source |       SS       df       MS              Number of obs =    6334
-----------+------------------------------           F(  2,  6331) = 2958.94
     Model |  1858.61471     2  929.307357           Prob> F       =  0.0000
  Residual |  1988.36474  6331  .314068037           R-squared     =  0.4831
-----------+------------------------------           Adj R-squared =  0.4830
     Total |  3846.97946  6333   .60744978           Root MSE      =  .56042
------------------------------------------------------------------------------
    LGFDHO |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------+------------------------------------------------------------------
     LGEXP |   .5842097   .0097174    60.12   0.000     .5651604    .6032591
    LGSIZE |   .3343475   .0127587    26.21   0.000     .3093362    .3593589
     _cons |   1.158326   .0820119    14.12   0.000     .9975545    1.319097
------------------------------------------------------------------------------
```

```
. gen LGFDHOPC = ln(FDHO/SIZE)
. gen LGEXPPC = ln(EXP/SIZE)

. reg LGFDHOPC LGEXPPC
------------------------------------------------------------------------------
    Source |       SS       df       MS              Number of obs =    6334
-----------+------------------------------           F(  1,  6332) = 4757.00
     Model |  1502.58928     1  1502.58928           Prob> F       =  0.0000
  Residual |   2000.0827  6332   .31586903           R-squared     =  0.4290
-----------+------------------------------           Adj R-squared =  0.4289
     Total |  3502.67197  6333  .553082579           Root MSE      =  .56202
------------------------------------------------------------------------------
  LGFDHOPC |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------+------------------------------------------------------------------
   LGEXPPC |   .6092734   .0088338    68.97   0.000     .5919562    .6265905
     _cons |   .8988292   .0703516    12.78   0.000     .7609162    1.036742
------------------------------------------------------------------------------
```

```
. reg LGFDHOPC LGEXPPC LGSIZE
------------------------------------------------------------------------------
      Source |       SS       df       MS              Number of obs =    6334
-------------+------------------------------           F(  2,  6331) = 2410.79
       Model |  1514.30723     2  757.153617            Prob> F       =  0.0000
    Residual |  1988.36474  6331  .314068037            R-squared     =  0.4323
-------------+------------------------------           Adj R-squared =  0.4321
       Total |  3502.67197  6333  .553082579            Root MSE      =  .56042
------------------------------------------------------------------------------
    LGFDHOPC |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     LGEXPPC |   .5842097   .0097174    60.12   0.000     .5651604    .6032591
      LGSIZE |  -.0814427   .0133333    -6.11   0.000    -.1075805   -.0553049
       _cons |   1.158326   .0820119    14.12   0.000     .9975545    1.319097
------------------------------------------------------------------------------
```

1. Explain why the second model is a restricted version of the first, stating the restriction.
2. Perform an F test of the restriction.
3. Perform a t test of the restriction.
4. Summarise your conclusions from the analysis of the regression results.

Answer:

Write the first specification as:

$$LGFDHO = \beta_1 + \beta_2 LGEXP + \beta_3 LGSIZE + u.$$

Then the restriction implicit in the second specification is $\beta_3 = 1 - \beta_2$, for:

$$\begin{aligned} LGFDHO &= \beta_1 + \beta_2 LGEXP + (1 - \beta_2) LGSIZE + u \\ LGFDHO - LGSIZE &= \beta_1 + \beta_2 (LGEXP - LGSIZE) + u \\ \log \frac{FDHO}{SIZE} &= \beta_1 + \beta_2 \log \frac{EXP}{SIZE} + u \\ LGFDHOPC &= \beta_1 + \beta_2 LGEXPPC + u \end{aligned}$$

the last equation being the second specification. The F statistic for the null hypothesis $H_0 : \beta_3 = 1 - \beta_2$ is:

$$F(1,\, 6331) = \frac{(2000.1 - 1988.4)/1}{1988.4/6331} = 37.3.$$

The critical value of $F(1,\, 1000)$ at the 0.1 per cent level is 10.9, and hence the restriction is rejected at that significance level.

Alternatively, we could use the t test approach. Under the null hypothesis that the restriction is valid, $\theta = 1 - \beta_2 - \beta_3 = 0$. Substituting for β_3, the unrestricted version may be rewritten:

$$LGFDHO = \beta_1 + \beta_2 LGEXP + (1 - \beta_2 - \theta) LGSIZE + u.$$

This may be rewritten:

$$\log \frac{FDHO}{SIZE} = \beta_1 + \beta_2 \log \frac{EXP}{SIZE} - \theta \log SIZE + u$$

that is:

$$LGFDHOPC = \beta_1 + \beta_2 LGEXPPC - \theta LGSIZE + u.$$

The t statistic for the coefficient of $LGSIZE$ is -6.11, so we reject the restriction at a very high significance level. Note that the t statistic is the square root of the F statistic and the critical value of t at the 0.1 per cent level will be the square root of the critical value of F.

6.5 Answers to the additional exercises

A6.1 The output below gives the results of a simple regression of $LGFDHOPC$ on $LGSIZE$. See Exercise A4.1 for the simple regression of $LGFDHOPC$ on $LGEXPPC$ and Exercise A4.2 for the multiple regression of $LGFDHOPC$ on $LGEXPPC$ and $LGSIZE$.

```
. reg LGFDHOPC LGSIZE
------------------------------------------------------------------------------
      Source |       SS       df       MS              Number of obs =    6334
-------------+------------------------------           F(  1,  6332) =  768.56
       Model |  379.128845     1  379.128845           Prob> F       =  0.0000
    Residual |  3123.54316  6332  .493294877           R-squared     =  0.1082
-------------+------------------------------           Adj R-squared =  0.1081
       Total |  3502.67201  6333  .553082585           Root MSE      =  .70235
------------------------------------------------------------------------------
    LGFDHOPC |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      LGSIZE |  -.4199282   .0151473   -27.72   0.000     -.449622   -.3902344
       _cons |   6.040547   .0143586   420.69   0.000     6.012399    6.068695
------------------------------------------------------------------------------
```

If the true model is assumed to be:

$$LGFDHOPC = \beta_1 + \beta_2 LGEXPPC + \beta_3 LGSIZE + u$$

the expected value of $\widehat{\beta}_2$ in the simple regression of $LGFDHOPC$ on $LGEXPPC$ is given by:

$$E(\widehat{\beta}_2) = \beta_2 + \beta_3 \frac{\sum \left(LGEXPPC_i - \overline{LGEXPPC}\right)\left(LGSIZE_i - \overline{LGSIZE}\right)}{\sum \left(LGEXPPC_i - \overline{LGEXPPC}\right)^2}.$$

We know that the numerator of the second factor in the bias term is negative because the correlation is negative:

```
. cor LGEXPPC LGSIZE
(obs=6334)
           |  LGEXPPC   LGSIZE
-----------+------------------
   LGEXPPC |   1.0000
    LGSIZE |  -0.4223   1.0000
```

It is reasonable to suppose that economies of scale will cause β_3 to be negative, and the highly significant negative estimate in the multiple regression provides empirical support, so $\widehat{\beta}_2$ is biased upwards. This accounts for the increase in its size in the second regression. In the third regression:

$$E(\widehat{\beta}_3) = \beta_3 + \beta_2 \frac{\sum \left(LGEXPPC_i - \overline{LGEXPPC}\right)\left(LGSIZE_i - \overline{LGSIZE}\right)}{\sum \left(LGSIZE_i - \overline{LGSIZE}\right)^2}.$$

β_2 is certainly positive, especially given the highly significant positive estimate in the first regression, and so $\widehat{\beta}_3$ is biased downwards. As a consequence, the estimate in the third regression is lower than that in the first.

Similar results are obtained for the other categories of expenditure. The correlation between *LGEXPPC* and *LGSIZE* varies because the missing observations are different for different categories.

Omitted variable bias, dependent variable *LGCATPC*					
		Multiple regression		Simple regressions	
	n	*LGEXPPC*	*LGSIZE*	*LGEXPPC*	*LGSIZE*
ADM	2,815	1.080	−0.055	1.098	−0.678
CLOT	4,500	0.842	0.146	0.794	−0.375
DOM	1,661	0.941	0.415	0.812	−0.150
EDUC	561	1.229	−0.437	1.382	−1.243
ELEC	5,828	0.472	−0.362	0.586	−0.645
FDAW	5,102	0.879	−0.213	0.947	−0.735
FDHO	6,334	0.584	−0.081	0.609	−0.420
FOOT	1,827	0.396	−0.560	0.608	−0.842
FURN	487	0.807	−0.246	0.912	−0.848
GASO	5,710	0.676	−0.004	0.677	−0.410
HEAL	4,802	0.779	−0.306	0.868	−0.723
HOUS	6,223	0.989	−0.140	1.033	−0.716
LIFE	1,253	0.464	−0.461	0.607	−0.701
LOCT	692	0.389	−0.396	0.510	−0.639
MAPP	399	0.721	−0.264	0.817	−0.717
PERS	3,817	0.824	−0.217	0.891	−0.703
READ	2,287	0.764	−0.503	0.909	−0.923
SAPP	1,037	0.467	−0.592	0.665	−0.879
TELE	5,788	0.640	−0.222	0.710	−0.603
TEXT	992	0.388	−0.713	0.629	−0.959
TOB	1,155	0.563	−0.515	0.721	−0.822
TOYS	2,504	0.638	−0.304	0.733	−0.691
TRIP	516	0.681	−0.142	0.723	−0.492

A6.2 *Demonstrate that $\widehat{\beta}_2$ is an unbiased estimator of β_2.*

$$\begin{aligned}\widehat{\beta}_2 &= \frac{\sum \left(D_i - \overline{D}\right)\left(Y_i - \overline{Y}\right)}{\sum \left(D_i - \overline{D}\right)^2} \\ &= \frac{\sum \left(D_i - \overline{D}\right)\left((\beta_1 + \beta_2 D_i + \beta_3 A_i + u_i) - (\beta_1 + \beta_2 \overline{D} + \beta_3 \overline{A} + \overline{u})\right)}{\sum \left(D_i - \overline{D}\right)^2} \\ &= \beta_2 + \beta_3 \frac{\sum \left(D_i - \overline{D}\right)\left(A_i - \overline{A}\right)}{\sum \left(D_i - \overline{D}\right)^2} + \frac{\sum \left(D_i - \overline{D}\right)(u_i - \overline{u})}{\sum \left(D_i - \overline{D}\right)^2}.\end{aligned}$$

Hence:

$$\widehat{\beta}_2 = \beta_2 + \beta_3 \sum d_i(A_i - \overline{A}) + \sum d_i(u_i - \overline{u})$$

where:

$$d_i = \frac{D_i - \overline{D}}{\sum (D_j - \overline{D})^2}.$$

Hence:

$$E(\widehat{\beta}_2) = \beta_2 + \beta_3 \sum E(d_i(A_i - \overline{A})) + \sum E(d_i(u_i - \overline{u})).$$

Now, since the assignment to the course was random, D is distributed independently of both A and u, and hence:

$$E(d_i(A_i - \overline{A})) = E(d_i)E(A_i - \overline{A}) = 0$$

and:

$$E(d_i(u_i - \overline{u})) = E(d_i)E(u_i - \overline{u}) = 0.$$

Hence $\widehat{\beta}_2$ is an unbiased estimator of β_2.

A commentator says that the standard error of $\widehat{\beta}_2$ will be invalid because an important variable, A, has been omitted from the specification. The researcher replies that the standard error will remain valid if A can be assumed to have a normal distribution. Explain whether the commentator or the researcher is correct.

The researcher is nearly correct. Given the random selection of the sample, A will be distributed independently of D and so it can be treated as part of the disturbance term and the standard error will remain valid. The requirement that A have a normal distribution is too strong, since the expression for the standard error does not depend on it. However, if the standard error is to be used for t tests, then it is important that the enlarged standard error should have a normal distribution, and this will be the case if an only if A has a normal distribution (assuming that u has one). If both A and u have normal distributions, a linear combination will also have one.

Another commentator says that whether the distribution of A is normal or not makes no difference to the validity of the standard error. Evaluate this assertion.

The commentator is correct for the reasons just explained.

A6.3 *Assuming that (A) is the correct specification, explain, with a mathematical proof, whether you would expect the coefficient of log Y to be greater in regression (2).*

To simplify the algebra, throughout this answer $\log B$, $\log Y$, $\log S$ and $\log IQ$ will be written as B, Y, S and IQ, it being understood that these are logarithms.

$$\begin{aligned}
\widehat{\beta}_2 &= \frac{\sum \left(B_i - \overline{B}\right)\left(Y_i - \overline{Y}\right)}{\sum \left(Y_i - \overline{Y}\right)^2} \\
&= \frac{\sum \left(\beta_1 + \beta_2 Y_i + \beta_3 IQ_i + u_i - \beta_1 - \beta_2 \overline{Y} - \beta_3 \overline{IQ} - \bar{u}\right)\left(Y_i - \overline{Y}\right)}{\sum \left(Y_i - \overline{Y}\right)^2} \\
&= \frac{\sum \left(\beta_2 Y_i - \beta_2 \overline{Y}\right)\left(Y_i - \overline{Y}\right) + \sum \left(\beta_3 IQ_i - \beta_3 \overline{IQ}\right)\left(Y_i - \overline{Y}\right) + \sum \left(u_i - \bar{u}\right)\left(Y_i - \overline{Y}\right)}{\sum \left(Y_i - \overline{Y}\right)^2} \\
&= \beta_2 + \beta_3 \frac{\sum \left(IQ_i - \overline{IQ}\right)\left(Y_i - \overline{Y}\right)}{\sum \left(Y_i - \overline{Y}\right)^2} + \frac{\sum \left(u_i - \bar{u}\right)\left(Y_i - \overline{Y}\right)}{\sum \left(Y_i - \overline{Y}\right)^2}.
\end{aligned}$$

Hence:

$$\begin{aligned}
E(\widehat{\beta}_2) &= \beta_2 + \beta_3 \frac{\sum \left(IQ_i - \overline{IQ}\right)\left(Y_i - \overline{Y}\right)}{\sum \left(Y_i - \overline{Y}\right)^2} + \frac{1}{\sum \left(Y_i - \overline{Y}\right)^2} E\left(\sum (u_i - \bar{u})(Y_i - \overline{Y})\right) \\
&= \beta_2 + \beta_3 \frac{\sum \left(IQ_i - \overline{IQ}\right)\left(Y_i - \overline{Y}\right)}{\sum \left(Y_i - \overline{Y}\right)^2} + \frac{1}{\sum \left(Y_i - \overline{Y}\right)^2} \sum E\left((u_i - \bar{u})(Y_i - \bar{Y})\right) \\
&= \beta_2 + \beta_3 \frac{\sum \left(IQ_i - \overline{IQ}\right)\left(Y_i - \overline{Y}\right)}{\sum \left(Y_i - \overline{Y}\right)^2} + \frac{1}{\sum \left(Y_i - \overline{Y}\right)^2} \sum (Y_i - \overline{Y}) E(u_i - \bar{u}) \\
&= \beta_2 + \beta_3 \frac{\sum \left(IQ_i - \overline{IQ}\right)\left(Y_i - \overline{Y}\right)}{\sum \left(Y_i - \overline{Y}\right)^2}
\end{aligned}$$

assuming that Y and IQ are nonstochastic. Thus $\widehat{\beta}_2$ is biased, the direction of the bias depending on the signs of β_3 and $\sum \left(IQ_i - \overline{IQ}\right)\left(Y_i - \overline{Y}\right)$. We would expect the former to be positive and we expect the latter to be positive since we are told that the correlation between S and Y is positive and S is a proxy for IQ. So we would expect an upward bias in regression (2).

Assuming that (A) is the correct specification, describe the various benefits from using log S as a proxy for log IQ, as in regression (1), if log S is a good proxy.

The use of S as a proxy for IQ will alleviate the problem of omitted variable bias. In particular, comparing the results of regression (1) with those that would have been obtained if B had been regressed on Y and IQ:

- the coefficient of Y will be approximately the same
- its standard error will be approximately the same
- the t statistic for S will be approximately equal to that of IQ
- R^2 will be approximately the same.

Explain whether the low value of R^2 in regression (1) implies that log S is not a good proxy.

Not necessarily. It could be that S is a poor proxy for IQ, but it could also be that the original model had low explanatory power.

Assuming that (A) is the correct specification, provide an explanation of why the coefficients of log Y and log S in regression (1) are not significantly different from zero, using two-sided t tests.

The high correlation between Y and S has given rise to multicollinearity, the standard errors being so large that the coefficients are not significantly different from zero.

Discuss whether the researcher would be justified in using one-sided t tests in regression (1).

Yes. It is reasonable to suppose that expenditure on books should not be negatively influenced by either income or cognitive ability. (Note that one should *not* say that it is reasonable to suppose that expenditure on books is positively influenced by them. This rules out the null hypothesis.)

Assuming that (B) is the correct specification, explain whether you would expect the coefficient of log Y to be lower in regression (1).

No. It would be randomly higher or lower, if S is an irrelevant variable.

Assuming that (B) is the correct specification, explain whether the standard errors in regression (1) are valid estimates.

Yes. The inclusion of an irrelevant variable in general does not invalidate the standard errors. It causes them to be larger than those in the correct specification.

A6.4 *Explain analytically whether the slope coefficient is likely to be biased.*

If the fitted model is:

$$\widehat{\log T} = \widehat{\beta}_1 + \widehat{\beta}_2 \log P$$

then:

$$\begin{aligned}
\widehat{\beta}_2 &= \frac{\sum \left(\log P_i - \overline{\log P}\right)\left(\log T_i - \overline{\log T}\right)}{\sum \left(\log P_i - \overline{\log P}\right)^2} \\
&= \frac{\sum \left(\log P_i - \overline{\log P}\right)\left(\beta_1 + \beta_2 \log P_i + \beta_3 \log Y_i + u_i - \beta_1 - \beta_2\overline{\log P} - \beta_3\overline{\log Y} - \bar{u}\right)}{\sum \left(\log P_i - \overline{\log P}\right)^2} \\
&= \beta_2 + \beta_3 \frac{\sum \left(\log P_i - \overline{\log P}\right)\left(\log Y_i - \overline{\log Y}\right)}{\sum \left(\log P_i - \overline{\log P}\right)^2} + \frac{\sum \left(\log P_i - \overline{\log P}\right)(u_i - \bar{u})}{\sum \left(\log P_i - \overline{\log P}\right)^2}.
\end{aligned}$$

Hence:

$$E(\widehat{\beta}_2) = \beta_2 + \beta_3 \frac{\sum \left(\log P_i - \overline{\log P}\right)\left(\log Y_i - \overline{\log Y}\right)}{\sum \left(\log P_i - \overline{\log P}\right)^2}$$

provided that any random component of $\log P$ is distributed independently of u. Since it is reasonable to assume $\beta_3 > 0$, and since we are told that Y and P are positively correlated, the bias will be upwards. This accounts for the nonsensical positive price elasticity in the fitted equation.

Describe the theoretical benefits from using H as a proxy for Y, discussing whether they appear to have been obtained in this example.

Suppose that H is a perfect proxy for Y:

$$\log Y = \lambda + \mu \log H.$$

Then the relationship may be rewritten:

$$\log T = \beta_1 + \beta_3\lambda + \beta_2 \log P + \beta_3\mu \log H + u.$$

The coefficient of $\log P$ ought to be the same as in the true relationship. However in this example it is not the same. However it is of the right order of magnitude and much more plausible than the estimate in the first regression. The standard error of the coefficient ought to be the same as in the true relationship, and this is the case.

The coefficient of $\log H$ will be an estimate of $\beta_3\mu$, and since μ is unknown, β_3 is not identified. However, if it can be assumed that the average household income in a city is proportional to average house values, it could be asserted that μ is equal to 1, in which case the coefficient of $\log H$ will be a direct estimate of β_3 after all. The coefficient of $\log H$ is indeed quite close to that of $\log Y$. The t statistic for the coefficient of $\log H$ ought to be the same as that for $\log Y$, and this is approximately true, being a little lower. R^2 ought to be the same, but it is somewhat lower, suggesting that H appears to have been a good proxy, but not a perfect one.

A6.5 *Explain why specification (1) is a restricted version of specification (2), stating and interpreting the restriction.*

First note that, since $E = W/H$, $LE = \log(W/H) = LW - LH$.

Write specification (2) as:

$$LW = \beta_1 + \beta_2 S + \beta_3 LH + u.$$

If one imposes the restriction $\beta_3 = 1$, the model becomes specification (1):

$$LW - LH = \beta_1 + \beta_2 S + u.$$

The restriction implies that weekly earnings are proportional to hours worked, controlling for schooling.

Supposing the restriction to be valid, explain whether you expect the coefficient of S and its standard error to differ, or be similar, in specifications (1) and (2).

If the restriction is valid, the coefficient of S should be similar in the restricted specification (1) and the unrestricted specification (2). Both estimates will be

unbiased, but that in specification (1) will be more efficient. The gain in efficiency in specification (1) should be reflected in a smaller standard error. However, the gain will be small, given the low correlation.

Supposing the restriction to be invalid, how would you expect the coefficient of S and its standard error to differ, or be similar, in specifications (1) and (2)?

The estimate of the coefficient of S would be biased. The standard error in specification (1) would be invalid and so a comparison with the standard error in specification (2) would be illegitimate.

Perform an F test of the restriction.

The null and alternative hypotheses are $H_0 : \beta_3 = 1$ and $H_1 : \beta_3 \neq 1$.

$$F(1, 1752) = \frac{(741.5 - 737.9)/1}{737.9/1752} = 8.5.$$

The critical value of $F(1, 1000)$ at the 1 per cent level is 6.66. The critical value of $F(1, 1752)$ must be lower. Thus we reject the restriction at the 1 per cent level. (The critical value at the 0.1 per cent level is about 10.8.)

Perform a t test of the restriction.

The restriction is so simple that it can be tested with no reparameterisation: a simple t test on the coefficient of LH in specification (2), $H_0 : \beta_3 = 1$.

Alternatively, mechanically following the standard procedure, we rewrite the restriction as $\beta_3 - 1 = 0$. The reparameterisation will be:

$$\theta = \beta_3 - 1$$

and so:

$$\beta_3 = \theta + 1.$$

Substituting this into the unrestricted specification, the latter may be rewritten:

$$LW = \beta_1 + \beta_2 S + (\theta + 1)LH + u.$$

Hence:

$$LW - LH = \beta_1 + \beta_2 S + \theta LH + u.$$

This is regression specification (3) and the restriction may be tested with a t test on the coefficient of LH, the null hypothesis being $H_0 : \theta = \beta_3 - 1 = 0$. The t statistic is 2.92, which is significant at the 1 per cent level, implying that the restriction should be rejected.

Explain whether the F test and the t test could lead to different conclusions.

The tests must lead to the same conclusion since the F statistic is the square of the t statistic and the critical value of F is the square of the critical value of t.

At a seminar, a commentator says that part-time workers tend to be paid worse than full-time workers and that their earnings functions are different. Defining full-time workers as those working at least 35 hours per week, the researcher divides the sample and fits the earnings functions for full-time workers (column 4) and part-time workers (column 5). Test whether the commentators assertion is correct.

The appropriate test is a Chow test. The test statistic under the null hypothesis of no difference in the earnings functions is:

$$F(3, 1749) = \frac{(737.9 - 626.1 - 100.1)/3}{(626.1 + 100.1)/1749} = 9.39.$$

The critical value of $F(3, 1000)$ at the 0.1 per cent level is 5.46. Hence we reject the null hypothesis and conclude that the commentator is correct.

What are the implications of the commentators assertion for the test of the restriction?

The elasticity of LH is now not significantly different from 1 for either full-time or part-time workers, so the restriction is no longer rejected.

A6.6 *Explain why the second specification is a restricted version of the first. State the restriction.*

Write the second equation as:

$$\log \frac{I}{Y} = \beta_1 + \beta_2 \log \left(\frac{G}{Y} \right) + u.$$

It may be re-written as:

$$\log I = \beta_1 + \beta_2 \log G + (1 - \beta_2) \log Y + u.$$

This is a special case of the specification of the first equation:

$$\log I = \beta_1 + \beta_2 \log G + \beta_3 \log Y + u$$

with the restriction $\beta_3 = 1 - \beta_2$.

Perform a test of the restriction.

The null hypothesis is $H_0 : \beta_2 + \beta_3 = 1$. The test statistic is:

$$F(1, 27) = \frac{(0.99 - 0.90)/1}{0.90/27} = 2.7.$$

The critical value of $F(1, 27)$ is 4.21 at the 5 per cent level. Hence we do not reject the null hypothesis that the restriction is valid.

The researcher expected the standard error of the coefficient of log (G/Y) in (2) to be smaller than the standard error of the coefficient of log G in (1). Explain why she expected this.

The imposition of the restriction, if valid, should lead to a gain in efficiency and this should be reflected in lower standard errors.

However the standard error is the same, at least to two decimal places. Give an explanation.

The standard errors of the coefficients of G in (1) and G/Y in (2) are given by:

$$\sqrt{\frac{\widehat{\sigma}_u^2}{n\text{MSD}(G)} \times \frac{1}{1 - r_{G,Y}^2}} \quad \text{and} \quad \sqrt{\frac{\widehat{\sigma}_u^2}{n\text{MSD}(G/Y)}}$$

respectively, where $\widehat{\sigma}_u^2$ is an estimate of the variance of the disturbance term, n is the number of observations, MSD is the mean square deviation in the sample, and $r_{G,Y}$ is the sample correlation coefficient of G and Y. n is the same for both standard errors and $\widehat{\sigma}_u^2$ will be very similar. We are told that $r_{G,Y} = 0.98$, so its square is 0.96 and the second factor in the expression for the standard error of G is $(1/0.04) = 25$. Hence, other things being equal, the standard error of G/Y should be much lower than that of G. However the table shows that the MSD of G/Y is only 1/25 as great as that of G. This just about exactly negates the gain in efficiency attributable to the elimination of the correlation between G and Y.

Show how the restriction could be tested using a t test in a reparameterised version of the specification for (1).

Define $\theta = \beta_2 + \beta_3 - 1$, so that the restriction may be written $\theta = 0$. Then $\beta_3 = \theta - \beta_2 + 1$. Use this to substitute for β_3 in the unrestricted model:

$$\begin{aligned} \log I &= \beta_1 + \beta_2 \log G + \beta_3 \log Y + u \\ &= \beta_1 + \beta_2 \log G + (\theta - \beta_2 + 1) \log Y + u. \end{aligned}$$

Then:

$$\log I - \log Y = \beta_1 + \beta_2(\log G - \log Y) + \theta \log Y + u$$

and:

$$\log\left(\frac{I}{Y}\right) = \beta_1 + \beta_2\left(\frac{G}{Y}\right) + \theta \log Y + u.$$

Hence the restriction may be tested by a t test of the coefficient of $\log Y$ in a regression using this specification.

A6.7 This is a generalisation of the example with *FDHO* in Exercise 6.14 in the text. The reason for the discrepancy in the number of observations is not known. Possibly it used an earlier version of the data set.

```
. reg LGFDHO LGEXP LGSIZE
------------------------------------------------------------------------------
      Source |       SS       df       MS              Number of obs =    6334
-------------+------------------------------           F(  2,  6331) = 2958.94
       Model |  1858.61471     2  929.307357           Prob> F       =  0.0000
    Residual |  1988.36474  6331  .314068037           R-squared     =  0.4831
-------------+------------------------------           Adj R-squared =  0.4830
       Total |  3846.97946  6333   .60744978           Root MSE      =  .56042
------------------------------------------------------------------------------
      LGFDHO |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       LGEXP |   .5842097   .0097174    60.12   0.000     .5651604    .6032591
      LGSIZE |   .3343475   .0127587    26.21   0.000     .3093362    .3593589
       _cons |   1.158326   .0820119    14.12   0.000     .9975545    1.319097
------------------------------------------------------------------------------
```

```
. reg LGFDHOPC LGEXPPC
------------------------------------------------------------------------------
    Source |       SS       df       MS              Number of obs =    6334
-----------+------------------------------           F(  1,  6332) = 4757.00
     Model |  1502.58932     1  1502.58932           Prob> F       =  0.0000
  Residual |  2000.08269  6332  .315869029           R-squared     =  0.4290
-----------+------------------------------           Adj R-squared =  0.4289
     Total |  3502.67201  6333  .553082585           Root MSE      =  .56202
------------------------------------------------------------------------------
  LGFDHOPC |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------+------------------------------------------------------------------
   LGEXPPC |   .6092734   .0088338    68.97   0.000     .5919562    .6265905
     _cons |   .8988291   .0703516    12.78   0.000     .7609161    1.036742
------------------------------------------------------------------------------
```

Write the first specification as:

$$LGFDHO = \beta_1 + \beta_2 LGEXP + \beta_3 LGSIZE + u.$$

Then the restriction implicit in the second specification is $\beta_3 = 1 - \beta_2$, for then:

$$\begin{aligned} LGFDHO &= \beta_1 + \beta_2 LGEXP + (1 - \beta_2) LGSIZE + u \\ LGFDHO - LGSIZE &= \beta_1 + \beta_2 (LGEXP - LGSIZE) + u \\ \log \frac{FDHO}{SIZE} &= \beta_1 + \beta_2 \log \frac{EXP}{SIZE} + u \\ LGFDHOPC &= \beta_1 + \beta_2 LGEXPPC + u \end{aligned}$$

the last equation being the second specification. The F statistic for the null hypothesis $H_0 : \beta_3 = 1 - \beta_2$ is:

$$F(1, 6331) = \frac{(2000.1 - 1988.4)/1}{1998.4/6331} = 37.25.$$

The critical value of $F(1, 1000)$ at the 0.1 per cent level is 10.9, and hence the restriction is rejected at that significance level. This is not a surprising result, given that the estimates of β_2 and β_3 in the unrestricted specification were 0.58 and 0.33, respectively, their sum being well short of 1, as implied by the restriction.

Summarising the results of the test for all the categories, we have:

- Restriction rejected at the 1 per cent level: *FDHO*, *FDAW*, *HOUS*, *TELE*, *FURN*, *MAPP*, *SAPP*, *CLOT*, *HEAL*, *ENT*, *FEES*, *READ*, *TOB*.
- Restriction rejected at the 5 per cent level: *TRIP*, *LOCT*.
- Restriction not rejected at the 5 per cent level: *DOM*, *TEXT*, *FOOT*, *GASO*, *TOYS*, *EDUC*.

	n	RSS restricted	RSS unrestricted	F	t
ADM	2,815	3,947.5	3,945.2	1.6	−1.26
CLOT	4,500	5,792.0	5,766.1	20.2	4.50
DOM	1,661	4,138.0	4,062.5	30.8	5.55
EDUC	561	1,404.6	1,380.1	9.9	−3.15
ELEC	5,828	2,842.9	2,636.3	456.4	−21.36
FDAW	5,102	3,430.9	3,369.1	93.6	−9.68
FDHO	6,334	2,000.1	1,988.4	37.2	−6.11
FOOT	1,827	1,506.4	1,373.5	176.4	−13.28
FURN	487	920.0	913.9	3.2	−1.80
GASO	5,710	2,879.4	2,879.3	0.0	−0.20
HEAL	4,802	6,183.4	6,062.5	95.7	−9.79
HOUS	6,223	4,859.4	4,825.6	43.6	−6.60
LIFE	1,253	1,622.7	1,559.2	50.9	−7.13
LOCT	692	1,108.1	1,075.1	21.1	−4.60
MAPP	399	583.5	576.8	4.6	−2.14
PERS	3,817	3,049.1	3,002.2	59.6	−7.72
READ	2,287	3,038.1	2,892.1	115.3	−10.74
SAPP	1,037	1,239.6	1,148.9	81.6	−9.03
TELE	5,788	3,133.1	3,055.1	147.6	−12.15
TEXT	992	1,150.5	1,032.9	112.6	−10.61
TOB	1,155	956.3	873.4	109.4	−10.46
TOYS	2,504	2,885.4	2,828.3	50.5	−7.11
TRIP	516	795.4	792.8	1.7	−1.30

For the t test, we first rewrite the restriction as $\beta_2 + \beta_3 - 1 = 0$. The test statistic is therefore $\theta = \beta_2 + \beta_3 - 1$. This allows us to write $\beta_3 = \theta - \beta_2 + 1$. Substituting for β_3, the unrestricted version becomes:

$$LGFDHO = \beta_1 + \beta_2 LGEXP + (\theta - \beta_2 + 1)LGSIZE + u.$$

Hence the unrestricted version may be rewritten:

$$LGFDHO - LGSIZE = \beta_1 + \beta_2(LGEXP - LGSIZE) + \theta LGSIZE + u$$

that is:

$$LGFDHOPC = \beta_1 + \beta_2 LGEXPPC + \theta LGSIZE + u.$$

We use a t test to see if the coefficient of $LGSIZE$ is significantly different from 0. If it is not, we can drop the $LGSIZE$ term and we conclude that the restricted specification is an adequate representation of the data. If it is, we have to stay with the unrestricted specification.

From the output for the third regression, we see that t is −6.11 and hence the null hypothesis $H_0 : \beta_2 + \beta_3 - 1 = 0$ is rejected (critical value of t at the 0.1 per cent level is 3.29). Note that the t statistic is the square root of the F statistic and the critical value of t at the 0.1 per cent level is the square root of the critical value of F. The results for the other categories are likewise identical to those for the F test.

```
. reg LGFDHOPC LGEXPPC LGSIZE
------------------------------------------------------------------------------
    Source |       SS       df       MS              Number of obs =    6334
-----------+------------------------------           F(  2,  6331) = 2410.79
     Model |  1514.30728     2   757.15364           Prob> F       =  0.0000
  Residual |  1988.36473  6331  .314068035           R-squared     =  0.4323
-----------+------------------------------           Adj R-squared =  0.4321
     Total |  3502.67201  6333  .553082585           Root MSE      =  .56042
------------------------------------------------------------------------------
  LGFDHOPC |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------+------------------------------------------------------------------
   LGEXPPC |   .5842097   .0097174    60.12   0.000     .5651604    .6032591
    LGSIZE |  -.0814427   .0133333    -6.11   0.000    -.1075806   -.0553049
     _cons |   1.158326   .0820119    14.12   0.000     .9975545    1.319097
------------------------------------------------------------------------------
```

A6.8 (2) may be rewritten:

$$\log Y = \alpha_1 + (\alpha_2 + 1) \log X + u$$

so it is a reparameterised version of (1) with $\beta_1 = \alpha_1$ and $\beta_2 = \alpha_2 + 1$.

A6.9 *Show that the specification of Researcher 2 is a restricted version of the specification of Researcher 1, stating the restriction.*

Let the model be written:

$$LGE = \beta_1 + \beta_2 S + \beta_3 NUM + \beta_4 VERB + u.$$

The restriction is $\beta_4 = \beta_3$ since NUM and $VERB$ are given equal weights in the construction of $SCORE$. Using the restriction, the model can be rewritten

$$\begin{aligned} LGE &= \beta_1 + \beta_2 S + \beta_3 (NUM + VERB) + u \\ &= \beta_1 + \beta_2 S + 2\beta_3 SCORE + u. \end{aligned}$$

Perform an F test of the restriction.

The null and alternative hypotheses are $H_0 : \beta_4 = \beta_3$ and $H_1 : \beta_4 \neq \beta_3$. The F statistic is:

$$F(1, 100) = \frac{(2045 - 2000)/1}{2000/100} = 2.25.$$

The critical value of $F(1, 100)$ is 3.94 at the 5 per cent level. Hence we do not reject the restriction at the 5 per cent level.

Show that the specification of Researcher 3 is a reparameterised version of the specification of Researcher 1 and hence perform a t test of the restriction in the specification of Researcher 2.

The restriction may be rewritten $\beta_4 - \beta_3 = 0$. The test statistic is therefore $\theta = \beta_4 - \beta_3$. Hence $\beta_4 = \theta + \beta_3$. Substituting for β_4 in the unrestricted model, one has:

$$\begin{aligned} LGE &= \beta_1 + \beta_2 S + \beta_3 NUM + (\theta + \beta_3) VERB + u \\ &= \beta_1 + \beta_2 S + \beta_3 (NUM + VERB) + \theta VERB + u \\ &= \beta_1 + \beta_2 S + 2\beta_3 SCORE + \theta VERB + u. \end{aligned}$$

This is the specification of Researcher 3. To test the hypothesis that the restriction is valid, we perform a t test on the coefficient of *VERB*. The t statistic is -1.5, so we do not reject the restriction at the 5 per cent level.

Explain whether the F test in (b) and the t test in (c) could have led to different results.

No, the F test and the t test must give the same result because the F statistic must be the square of the t statistic and the critical value of F must be the square of the critical value of t for any given significance level. Note that this assumes a two-sided t test. If one is in a position to perform a one-sided test, the t test would be more powerful.

Perform a test of the hypothesis that the numeracy score has a greater effect on earnings than the literacy score.

One should perform a one-sided t test on the coefficient of *VERB* in regression 3 with the null hypothesis $H_0 : \theta = 0$ and the alternative hypothesis $H_1 : \theta < 0$. The null hypothesis is not rejected and hence one concludes that there is no significant difference.

Compare the regression results of the three researchers.

The regression results of Researchers 1 and 3 are equivalent, the only difference being that the coefficient of *VERB* provides a direct estimate of β_4 in the specification of Researcher 1 and $(\beta_4 - \beta_3)$ in the specification of Researcher 3. Assuming the restriction is valid, there is a large gain in efficiency in the estimation of β_3 in specification (2) because its standard error is effectively 0.0005, as opposed to 0.0011 in specifications (1) and (3).

A6.10 *Demonstrate that relationship (3) embodies a testable restriction and show how the model may be reformulated to take advantage of it.*

The coefficients of $\log \rho$ and $\log w$ sum to 1. Hence the model should be reformulated as:

$$\log L = \frac{1}{\alpha + \beta} \log Q + \frac{\alpha}{\alpha + \beta} \log \frac{\rho}{w} \qquad (4)$$

(plus a disturbance term).

Explain how the restriction could be tested using an F test.

Let RSS_U and RSS_R be the residual sums of squares from the unrestricted and restricted regressions. To test the null hypothesis that the coefficients of $\log \rho$ and $\log w$ sum to 1, one should calculate the F statistic:

$$F(1, 27) = \frac{(RSS_R - RSS_U)/1}{RSS_U/27}$$

and compare it with the critical values of $F(1, 27)$.

Explain how the restriction could be tested using a t test.

Alternatively, writing (3) as an unrestricted model:

$$\log wL = \gamma_1 \log Q + \gamma_2 \log \rho + \gamma_3 \log w + u \qquad (5)$$

the restriction is $\gamma_2 + \gamma_3 - 1 = 0$. Define $\theta = \gamma_2 + \gamma_3 - 1$. Then $\gamma_3 = \theta - \gamma_2 + 1$ and the unrestricted model may be rewritten as:

$$\log wL = \gamma_1 \log Q + \gamma_2 \log \rho + (\theta - \gamma_2 + 1) \log w + u.$$

Hence:

$$\log wL - \log w = \gamma_1 \log Q + \gamma_2(\log \rho - \log w) + \theta \log w + u.$$

Hence:

$$\log L = \gamma_1 \log Q + \gamma_2 \log \frac{\rho}{w} + \theta \log w + u.$$

Thus one should regress $\log L$ on $\log Q$, $\log(\rho/w)$, and $\log w$ and perform a t test on the coefficient of $\log w$.

Explain the theoretical benefits of making use of a valid restriction. How could the researcher assess whether there are any benefits in practice, in this case?

The main theoretical benefit of making use of a valid restriction is that one obtains more efficient estimates of the coefficients. The use of a restriction would eliminate the problem of duplicate estimates of the same parameter. Reduced standard errors should provide evidence of the gain in efficiency.

At a seminar, someone suggests that it is reasonable to hypothesise that manufacturing output is subject to constant returns to scale, so that $\alpha + \beta = 1$. Explain how the researcher could test this hypothesis (1) using an F test, (2) using a t test.

Under the assumption of constant returns to scale, the model becomes:

$$\log \frac{L}{Q} = \alpha \log \frac{\rho}{w}. \qquad (6)$$

One could test the hypothesis by computing the F statistic:

$$F(1, 28) = \frac{(RSS_R - RSS_U)/1}{RSS_U/28}$$

where RSS_U and RSS_R are for the specifications in (4) and (6) respectively.

Alternatively, one could perform a simple t test of the hypothesis that the coefficient of $\log Q$ in (4) is equal to 1.

A6.11 *Explain why the researcher was dissatisfied with the results of regression (1).*

The high correlation between I and P has given rise to a problem of multicollinearity. The standard errors are relatively large and the t statistics low.

Demonstrate that specification (2) may be considered to be a restricted version of specification (1).

The restriction is $\beta_3 = -\beta_2$. Imposing it, we have:

$$\begin{aligned} B &= \beta_1 + \beta_2 I + \beta_3 P + u \\ &= \beta_1 + \beta_2 I - \beta_2 P + u \\ &= \beta_1 + \beta_2 R + u. \end{aligned}$$

Perform an F test of the restriction, stating carefully your null hypothesis and conclusion.

The null hypothesis is $H_0 : \beta_3 = -\beta_2$. The test statistic is:

$$F(1, 37) = \frac{(987.1 - 967.9)/1}{967.9/37} = 0.73.$$

The null hypothesis is not rejected at any significance level since $F < 1$.

Perform a t test of the restriction

The unrestricted specification may be rewritten:

$$\begin{aligned} B &= \beta_1 + \beta_2 I + \beta_3 P + u \\ &= \beta_1 + \beta_2 (P + R) + \beta_3 P + u \\ &= \beta_1 + (\beta_2 + \beta_3) P + \beta_2 R + u. \end{aligned}$$

Thus a t test on the coefficient of P in this specification is a test of the restriction. The null hypothesis is not rejected, given that the t statistic is 0.86. Of course, the F statistic is the square of the t statistic and the tests are equivalent.

Demonstrate that specification (3) may also be considered to be a restricted version of specification (1)

The restriction is $\beta_3 = 0$.

Perform both an F test and a t test of the restriction in specification (3), stating your conclusion in each case.

$$F(1, 37) = \frac{(1024.3 - 967.9)/1}{967.9/37} = 2.16.$$

The critical value of $F(1, 37)$ at 5 per cent is approximately 4.08, so the null hypothesis that P does not influence B is not rejected. Of course, with $t = -1.47$, the t test, which is equivalent, leads to the same conclusion.

At a seminar, someone suggests that specification (4) is also a restricted version of specification (1). Is this correct? If so, state the restriction.

No, it is not correct. As shown above, it is an alternative form of the unrestricted specification.

State, with an explanation, which would be your preferred specification.

None of the specifications has been rejected. The second should be preferred because it should be more efficient than the unrestricted specification. The much lower standard error of the slope coefficient provides supportive evidence. The third specification should be eliminated on the grounds that price inflation ought to be a determinant.

A6.12 Write the original model:

$$Y = \beta_1 + \beta_2 X + \beta_3 Z + u. \qquad (1)$$

Then, with:

$$X = 0.5(V + W), \qquad Z = 0.5(V - W)$$

the other specifications are:

$$Y = \beta_1 + 0.5(\beta_2 + \beta_3)V + 0.5(\beta_2 - \beta_3)W + u \quad (2)$$

$$Y = \beta_1 + \beta_2 V + u \quad (3)$$

with the implicit restriction $\beta_3 = \beta_2$, and, using $X = V - Z$:

$$Y = \beta_1 + \beta_2 V + (\beta_3 - \beta_2)Z + u. \quad (4)$$

(2) and (4) are reparameterisations of (1), so the measures of fit are unchanged: **E** = **L** = 0.60, **F** = **M** = 200.

Given the relationships among the parameters, **A** = 0.70, **C** = −0.10, **J** = 0.60, **H** = 0.20.

The standard errors **B** and **D** cannot be reconstructed because the standard errors of $\widehat{\beta}_2$ and $\widehat{\beta}_3$ cannot be used (on their own) to construct standard errors of linear combinations (a loose explanation is acceptable because we have hardly touched on covariances between estimators).

K = 0.04 since **J** = coefficient of X in specification (1).

The F statistic for the restriction $\beta_3 = \beta_2$ implicit in specification (3) is:

$$F(1, 40) = \frac{(220 - 200)/1}{200/40} = 4.0.$$

In terms of R^2 it would be:

$$F(1, 40) = \frac{(0.60 - G)/1}{0.40/40}.$$

Hence **G** = 0.56.

A two-sided t test on the coefficient of Z in specification (4) provides an equivalent test of the restriction. The t statistic must therefore be $\sqrt{4.0} = 2.0$ and so **I** = 0.10.

[Note: One may also compute **G** using the t statistic for the coefficient of V in specification (3):

$$\frac{G}{(1 - G)/41} = t^2.$$

Yet another was of computing **G** is as follows. Since R^2 in specification (1) is 0.60, TSS must be 500, using:

$$R^2 = 1 - \frac{RSS}{TSS}.$$

TSS is the same in specification (3). Hence one obtains **G** = 0.56.]

A6.13 F statistics should always be computed using RSS, not R^2. Often the R^2 version is equivalent, but often it is not, and this is a case in point. The reason is very simple: the dependent variables in the two specifications are different, and so the R^2 for the specifications are not comparable. The RSS are comparable because:

$$\begin{aligned} LGFDHOPC - LGFDHOPC &= (LGFDHO - LGSIZE) - (\widehat{LGFDHO} - LGSIZE) \\ &= LGFDHO - \widehat{LGFDHO}. \end{aligned}$$