7 Specification of regression variables: A preliminary skirmish

What are the consequences of including in the regression model a variable that should not be there? What are the consequences of leaving out a variable that should be included? What happens if you have difficulty finding data on a variable and use a proxy instead? This chapter is a preliminary skirmish with these issues in the sense that it focuses on the consequences of variable misspecification, rather than on procedures for model selection, a much more complex subject that is left to later in the text. The chapter concludes by showing how simple restrictions on the parameters can be tested.

7.1 Model specification

The construction of an economic model involves the specification of the relationships that constitute it, the specification of the variables that participate in each relationship, and the mathematical function representing each relationship. The last element was discussed in Chapter 5. In this chapter, we will consider the second element, and we will continue to assume that the model consists of just one equation. We will discuss the application of regression analysis to models consisting of systems of simultaneous relationships in Chapter 10.

If we know exactly which explanatory variables ought to be included in the equation when we undertake regression analysis, our task is limited to calculating estimates of their coefficients, confidence intervals for these estimates, and so on. In practice, however, we can never be sure that we have specified the equation correctly. Economic theory ought to provide a guide, but theory is never perfect. Without being aware of it, we might be including some variables that ought not to be in the model, and we might be leaving out others that ought to be included.

The properties of the regression estimates of the coefficients depend crucially on the validity of the specification of the model. The consequences of misspecification of the variables in a relationship are summarized in Table 7.1.

1. If you leave out a variable that ought to be included, the regression estimates are in general (but not always) biased. The standard errors of the coefficients and the corresponding t tests are in general invalid.

Table 7.1 Consequences of variable specification

Fitted model	True model	
	$Y = \beta_1 + \beta_2 X_2 + u$	$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$
$\hat{Y} = b_1 + b_2 X_2$	Correct specification, no problems	Coefficients are biased (in general). Standard errors are invalid
$\hat{Y} = b_1 + b_2 X_2 + b_3 X_3$	Coefficients are unbiased (in general) but inefficient. Standard errors are valid (in general).	Correct specification, no problems

2. If you include a variable that ought not to be in the equation, the regression coefficients are in general (but not always) inefficient but not biased. The standard errors are in general valid but, because the regression estimation is inefficient, they will be needlessly large.

We will begin by discussing these two cases and then come to some broader issues of model specification.

7.2 The effect of omitting a variable that ought to be included

The problem of bias

Suppose that the dependent variable Y depends on two variables X_2 and X_3 according to a relationship

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u, \tag{7.1}$$

but you are unaware of the importance of X_3. Thinking that the model should be

$$Y = \beta_1 + \beta_2 X_2 + u, \tag{7.2}$$

you use regression analysis to fit

$$\hat{Y} = b_1 + b_2 X_2, \tag{7.3}$$

and you calculate b_2 using the expression $\text{Cov}(X_2, Y)/\text{Var}(X_2)$, instead of the correct expression

$$b_2 = \frac{\text{Cov}(X_2, Y)\text{Var}(X_3) - \text{Cov}(X_3, Y)\text{Cov}(X_2, X_3)}{\text{Var}(X_2)\text{Var}(X_3) - [\text{Cov}(X_2, X_3)]^2}. \tag{7.4}$$

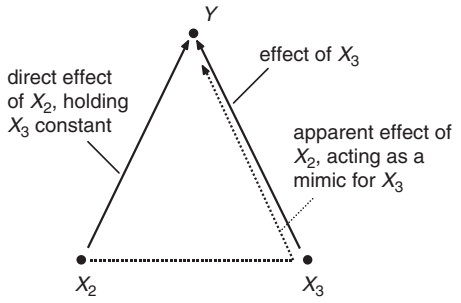

Figure 7.1

By definition, b_2 is an unbiased estimator of β_2 if and only if $E(b_2)$ is equal to β_2. In fact, if (7.1) is true,

$$E\left[\frac{\text{Cov}(X_2, Y)}{\text{Var}(X_2)}\right] = \beta_2 + \beta_3 \frac{\text{Cov}(X_2, X_3)}{\text{Var}(X_2)}. \tag{7.5}$$

We shall give first an intuitive explanation of this and then a formal proof.

If X_3 is omitted from the regression model, X_2 will appear to have a double effect, as illustrated in Figure 7.1. It will have a direct effect and also a proxy effect when it mimics the effect of X_3. The apparent indirect effect of X_2 on Y depends on two factors: the apparent ability of X_2 to mimic X_3, and the effect of X_3 on Y.

The apparent ability of X_2 to explain X_3 is determined by the slope coefficient h in the pseudo-regression

$$\hat{X}_3 = g + hX_2. \tag{7.6}$$

h of course is given by the usual simple regression formula, in this case $\text{Cov}(X_2, X_3)/\text{Var}(X_2)$. The effect of X_3 on Y is β_3, so the mimic effect via X_3 may be written $\beta_3 \text{Cov}(X_2, X_3)/\text{Var}(X_2)$. The direct effect of X_2 on Y is β_2, and hence when Y is regressed on X_2, omitting X_3, the coefficient of X_2 is given by

$$\beta_2 + \beta_3 \frac{\text{Cov}(X_2, X_3)}{\text{Var}(X_2)} + \text{sampling error}. \tag{7.7}$$

Provided that X_2 and X_3 are nonstochastic, the expected value of the coefficient will be the sum of the first two terms. The presence of the second term implies that in general the expected value of the coefficient will be different from the true value β_2 and therefore biased.

The formal proof of (7.5) is straightforward. We begin by making a theoretical expansion of the estimator b_2:

$$\begin{aligned} b_2 &= \frac{\text{Cov}(X_2, Y)}{\text{Var}(X_2)} = \frac{\text{Cov}(X_2, [\beta_1 + \beta_2 X_2 + \beta_3 X_3 + u])}{\text{Var}(X_2)} \\ &= \frac{1}{\text{Var}(X_2)}[\text{Cov}(X_2, \beta_1) + \text{Cov}(X_2, \beta_2 X_2) + \text{Cov}(X_2, \beta_3 X_3) + \text{Cov}(X_2, u)] \\ &= \frac{1}{\text{Var}(X_2)}[0 + \beta_2 \text{Var}(X_2) + \beta_3 \text{Cov}(X_2, X_3) + \text{Cov}(X_2, u)] \\ &= \beta_2 + \beta_3 \frac{\text{Cov}(X_2, X_3)}{\text{Var}(X_2)} + \frac{\text{Cov}(X_2, u)}{\text{Var}(X_2)}. \end{aligned} \tag{7.8}$$

Provided that X_2 and X_3 are nonstochastic, the first two terms are unaffected when we take expectations and the third is 0. Hence we obtain (7.5).

This confirms our earlier intuitive conclusion that b_2 is biased by an amount $\beta_3 \text{Cov}(X_2, X_3)/\text{Var}(X_2)$. The direction of the bias will depend on the signs of β_3 and $\text{Cov}(X_2, X_3)$. For example, if β_3 is positive and the covariance is positive, the bias will be positive and b_2 will tend to overestimate β_2. There is, however, one exceptional case where b_2 is unbiased after all. That is when the sample covariance between X_2 and X_3 happens to be exactly 0. If $\text{Cov}(X_2, X_3)$ is 0, the bias term disappears. Indeed, the regression coefficient obtained using simple regression will be exactly the same as if you had used a properly specified multiple regression. Of course, the bias term would also be 0 if β_3 were 0, but then the model is not misspecified.

Invalidation of the statistical tests

Another serious consequence of omitting a variable that ought to be included in the regression is that the standard errors of the coefficients and the test statistics are in general invalidated. This means of course that you are not in principle able to test any hypotheses with your regression results.

Example

The problem of omitted variable bias will first be illustrated with the educational attainment function using *EAEF* Data Set 21 (Table 7.2). For the present purposes, it will be assumed that the true model is

$$S = \beta_1 + \beta_2 ASVABC + \beta_3 SM + u, \tag{7.9}$$

although obviously this is a great oversimplification. The first part of the regression output shows the result of this regression. The second and third parts of the output then show the effects of omitting *SM* and *ASVABC*, respectively.

Table 7.2

```
. reg S ASVABC SM

  Source |       SS       df       MS          Number of obs =     570
---------+------------------------------       F(  2,   567) =  156.81
   Model |   1230.2039     2  615.101949        Prob > F      =  0.0000
Residual |  2224.04347   567  3.92247526        R-squared     =  0.3561
---------+------------------------------       Adj R-squared =  0.3539
   Total |  3454.24737   569  6.07073351        Root MSE      =  1.9805

------------------------------------------------------------------------------
       S |    Coef.   Std. Err.      t    P>|t|    [95% Conf. Interval]
---------+--------------------------------------------------------------------
  ASVABC | .1381062   .0097494   14.166   0.000    .1189567    .1572556
      SM |  .154783   .0350728    4.413   0.000    .0858946    .2236715
   _cons | 4.791277   .5102431    9.390   0.000     3.78908    5.793475
------------------------------------------------------------------------------

. reg S ASVABC

  Source |       SS       df       MS          Number of obs =     570
---------+------------------------------       F(  1,   568) =  284.89
   Model |  1153.80864     1  1153.80864        Prob > F      =  0.0000
Residual |  2300.43873   568  4.05006818        R-squared     =  0.3340
---------+------------------------------       Adj R-squared =  0.3329
   Total |  3454.24737   569  6.07073351        Root MSE      =  2.0125

------------------------------------------------------------------------------
       S |    Coef.   Std. Err.      t    P>|t|    [95% Conf. Interval]
---------+--------------------------------------------------------------------
  ASVABC | .1545378   .0091559   16.879   0.000    .1365543    .1725213
   _cons | 5.770845   .4668473   12.361   0.000    4.853888    6.687803
------------------------------------------------------------------------------

. reg S SM

  Source |       SS       df       MS          Number of obs =     570
---------+------------------------------       F(  1,   568) =   83.59
   Model |  443.110436     1  443.110436        Prob > F      =  0.0000
Residual |  3011.13693   568  5.30129742        R-squared     =  0.1283
---------+------------------------------       Adj R-squared =  0.1267
   Total |  3454.24737   569  6.07073351        Root MSE      =  2.3025

------------------------------------------------------------------------------
       S |    Coef.   Std. Err.      t    P>|t|    [95% Conf. Interval]
---------+--------------------------------------------------------------------
      SM | .3445198   .0376833    9.142   0.000    .2705041    .4185354
   _cons | 9.506491   .4495754   21.145   0.000    8.623458    10.38952
------------------------------------------------------------------------------
```

When *SM* is omitted,

$$E(b_2) = \beta_2 + \beta_3 \frac{\text{Cov}(ASVABC, SM)}{\text{Var}(ASVABC)}. \tag{7.10}$$

The correlation between *ASVABC* and *SM* is positive (0.38). Therefore the covariance term is positive. Since variances are always positive (unless equal to 0), the only other relevant factor for determining the sign of the bias is β_3. It is reasonable to assume that this is positive, and the fact that its estimate in the first regression is indeed positive and highly significant provides overwhelming corroborative evidence. One would therefore anticipate that the coefficient of *ASVABC* will be upwards biased when *SM* is omitted, and you can see that it is indeed higher. Not all of the difference should be attributed to bias. Part of it may be attributable to the effects of the disturbance term, which could go either way.

Similarly, when *ASVABC* is omitted,

$$E(b_3) = \beta_3 + \beta_2 \frac{\text{Cov}(ASVABC, SM)}{\text{Var}(SM)}. \tag{7.11}$$

Since β_2 is also likely to be positive, the coefficient of *SM* in the third regression should be upwards biased. The estimate in the third regression is indeed higher than that in the first.

In this example, the omission of one explanatory variable causes the coefficient of the other to be overestimated. However, the bias could just as easily be negative. The sign of the bias depends on the sign of the true coefficient of the omitted variable and on the sign of the sample covariance between the included and omitted variables, and these will depend on the nature of the model being investigated.

It should be emphasized that the analysis above applies only to the case where the true model is a multiple regression model with two explanatory variables. When there are more explanatory variables, it may be difficult to predict the impact of omitted variable bias mathematically. Nevertheless it may be possible to conclude that the estimates of the coefficients of some of the variables may have been inflated or deflated by the bias.

R^2 in the presence of omitted variable bias

In Section 4.5 it was asserted that in general it is impossible to determine the contribution to R^2 of each explanatory variable in multiple regression analysis, and we are now in a position to see why.

We will discuss the issue first with reference to the educational attainment model above. In the regression of *S* on *ASVABC* alone, R^2 was 0.33. In the regression on *SM* alone, it was 0.13. Does this mean that *ASVABC* explains 33 percent of the variance in *S*, and *SM* 13 percent? No, because this would

imply that together they would explain 46 percent of the variance, and this conflicts with the finding in the multiple regression that their joint explanatory power is 0.36.

The explanation is that in the simple regression of *S* on *ASVABC*, *ASVABC* is acting partly as a variable in its own right and partly as a proxy for the missing *SM*, as in Figure 7.1. R^2 for that regression therefore reflects the combined explanatory power of *ASVABC* in both of these roles, and not just its direct explanatory power. Hence 0.33 overestimates the latter.

Similarly, in the simple regression of *S* on *SM*, *SM* is acting partly as a proxy for the missing *ASVABC*, and the level of R^2 in that regression reflects the combined explanatory power of *SM* in both those roles, and not just its direct explanatory power.

In this example, the explanatory power of the two variables overlapped, with the consequence that R^2 in the multiple regression was less than the sum of R^2 in the individual simple regressions. However it is also possible for R^2 in the multiple regression to be greater than the sum of R^2 in the individual simple regressions, as is shown in the accompanying regression output for an earnings function model. It is assumed that the true model is

$$LGEARN = \beta_1 + \beta_2 S + \beta_3 MALE + u, \tag{7.12}$$

where *MALE* is a dummy variable equal to 1 for males and 0 for females. The first part of the regression output shows the result of fitting (7.12), and the second and third parts show the results of omitting, first *MALE*, and then *S* (see Table 7.3). R^2 in the multiple regression is 0.188, while it is 0.141 and 0.038 in the simple regressions, the sum being 0.179. As in the previous example, it can be assumed that both β_2 and β_3 are positive. However *S* and *MALE* are negatively correlated, so in this case the coefficients of *S* and *MALE* in the second and third regressions may be expected to be biased downwards. As a consequence, the apparent explanatory power of *S* and *MALE* in the simple regressions is underestimated.

Exercises

7.1 Using your *EAEF* data set, regress *LGEARN* (1) on *S* and *ASVABC*, (2) on *S* only, and (3) on *ASVABC* only. Calculate the correlation between *S* and *ASVABC*. Compare the coefficients of *S* in regressions (1) and (2). Give both mathematical and intuitive explanations of the direction of the change. Also compare the coefficients of *ASVABC* in regressions (1) and (3) and explain the direction of the change.

7.2* 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 868 households in the 1995 Consumer Expenditure Survey.

Table 7.3

```
. reg LGEARN S MALE

  Source |       SS       df       MS          Number of obs =     570
---------+------------------------------       F(  2,   567) =   65.74
   Model |   28.951332     2   14.475666        Prob > F      =  0.0000
Residual |  124.850561   567  .220194992        R-squared     =  0.1882
---------+------------------------------       Adj R-squared =  0.1854
   Total |  153.801893   569  .270302096        Root MSE      =  .46925

------------------------------------------------------------------------------
  LGEARN |    Coef.   Std. Err.      t     P>|t|    [95% Conf. Interval]
---------+--------------------------------------------------------------------
       S | .0818944   .0079976   10.240   0.000    .0661858     .097603
    MALE | .2285156   .0397695    5.746   0.000    .1504021    .3066291
   _cons |  1.19254   .1134845   10.508   0.000    .9696386    1.415441
------------------------------------------------------------------------------

. reg LGEARN S

  Source |       SS       df       MS          Number of obs =     570
---------+------------------------------       F(  1,   568) =   93.21
   Model |   21.681253     1   21.681253        Prob > F      =  0.0000
Residual |   132.12064   568   .23260676        R-squared     =  0.1410
---------+------------------------------       Adj R-squared =  0.1395
   Total |  153.801893   569  .270302096        Root MSE      =  .48229

------------------------------------------------------------------------------
  LGEARN |    Coef.   Std. Err.      t     P>|t|    [95% Conf. Interval]
---------+--------------------------------------------------------------------
       S | .0792256   .0082061    9.655   0.000    .0631077    .0953435
   _cons | 1.358919   .1127785   12.049   0.000    1.137406    1.580433
------------------------------------------------------------------------------

. reg LGEARN MALE

  Source |       SS       df       MS          Number of obs =     570
---------+------------------------------       F(  1,   568) =   22.51
   Model |  5.86288165     1  5.86288165        Prob > F      =  0.0000
Residual |  147.939011   568  .260456005        R-squared     =  0.0381
---------+------------------------------       Adj R-squared =  0.0364
   Total |  153.801893   569  .270302096        Root MSE      =  .51035

------------------------------------------------------------------------------
  LGEARN |    Coef.   Std. Err.      t     P>|t|    [95% Conf. Interval]
---------+--------------------------------------------------------------------
    MALE | .2048652   .0431797    4.744   0.000    .1200538    .2896767
   _cons | 2.313324    .032605   70.950   0.000    2.249282    2.377365
------------------------------------------------------------------------------
```

	(1)	(2)	(3)
LGEXP	0.29	0.48	—
	(0.02)	(0.02)	
LGSIZE	0.49	—	0.63
	(0.03)		(0.02)
constant	4.72	3.17	7.50
	(0.22)	(0.24)	(0.02)
R^2	0.52	0.31	0.42

The correlation coefficient for *LGEXP* and *LGSIZE* was 0.45. Explain the variations in the regression coefficients.

7.3 Suppose that Y is determined by X_2 and X_3 according to the relationship

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u,$$

and that $\text{Cov}(X_2, X_3)$ is 0. Use this to simplify the multiple regression coefficient b_2 given by

$$b_2 = \frac{\text{Cov}(X_2, Y)\text{Var}(X_3) - \text{Cov}(X_3, Y)\text{Cov}(X_2, X_3)}{\text{Var}(X_2)\text{Var}(X_3) - [\text{Cov}(X_2, X_3)]^2}$$

and show that it reduces to the simple regression expression. What are the implications for the specification of the regression equation?

7.4 In a Monte Carlo experiment, a variable Y was generated as a linear function of two variables X_2 and X_3:

$$Y = 10.0 + 10.0X_2 + 0.5X_3 + u,$$

where X_2 was the sequence of integers $1, 2,$ *ldots*$, 30$, X_3 was generated from X_2 by adding random numbers, and u was a normally distributed disturbance term with mean 0 and standard deviation 100. The correlation between X_2 and X_3 was 0.95. The sample variance of X_2 was 74.92 and that of X_3 was 82.67. The sample covariance between X_2 and X_3 was 74.94. The table shows the result of fitting the following regressions for ten samples:

$$\text{Model A} \quad \hat{Y} = b_1 + b_2 X_2 + b_3 X_3.$$

$$\text{Model B} \quad \hat{Y} = b_1 + b_2 X_2.$$

Comment on all aspects of the regression results, giving full explanations of what you observe.

Sample	Model A					Model B		
	b_2	s.e.(b_2)	b_3	s.e.(b_3)	R^2	b_2	s.e.(b_2)	R^2
1	10.68	6.05	0.60	5.76	0.5800	11.28	1.82	0.5799
2	7.52	7.11	3.74	6.77	0.5018	11.26	2.14	0.4961
3	7.26	6.58	2.93	6.26	0.4907	10.20	1.98	0.4865
4	11.47	8.60	0.23	8.18	0.4239	11.70	2.58	0.4239
5	13.07	6.07	−3.04	5.78	0.5232	10.03	1.83	0.5183
6	16.74	6.63	−4.01	6.32	0.5966	12.73	2.00	0.5906
7	15.70	7.50	−4.80	7.14	0.4614	10.90	2.27	0.4523
8	8.01	8.10	1.50	7.71	0.3542	9.51	2.43	0.3533
9	1.08	6.78	9.52	6.45	0.5133	10.61	2.11	0.4740
10	13.09	7.58	−0.87	7.21	0.5084	12.22	2.27	0.5081

7.3 The effect of including a variable that ought not to be included

Suppose that the true model is

$$Y = \beta_1 + \beta_2 X_2 + u \tag{7.13}$$

and you think it is

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u, \tag{7.14}$$

and you estimate b_2 using (7.4) instead of $\text{Cov}(X_2,Y)/\text{Var}(X_2)$.

In general there is no problem of bias, even though b_2 has been calculated incorrectly. $E(b_2)$ will still be equal to β_2, but in general b_2 will be an inefficient estimator. It will be more erratic, in the sense of having a larger variance about β_2, than if it had been calculated correctly. This is illustrated in Figure 7.2

This is easy to explain intuitively. The true model may be rewritten

$$Y = \beta_1 + \beta_2 X_2 + 0X_3 + u. \tag{7.15}$$

So if you regress Y on X_2 and X_3, b_2 will be an unbiased estimator of β_2 and b_3 will be an unbiased estimator of 0, provided that the Gauss–Markov conditions are satisfied. Effectively, you are discovering for yourself that β_3 is 0. If you realized beforehand that β_3 is 0, you would be able to exploit this information to exclude X_3 and use simple regression, which in this context is more efficient.

The loss of efficiency caused by including X_3 when it ought not to be included depends on the correlation between X_2 and X_3. Compare the expressions for the variances of b_2 using simple and multiple regression in Table 7.4. The variance

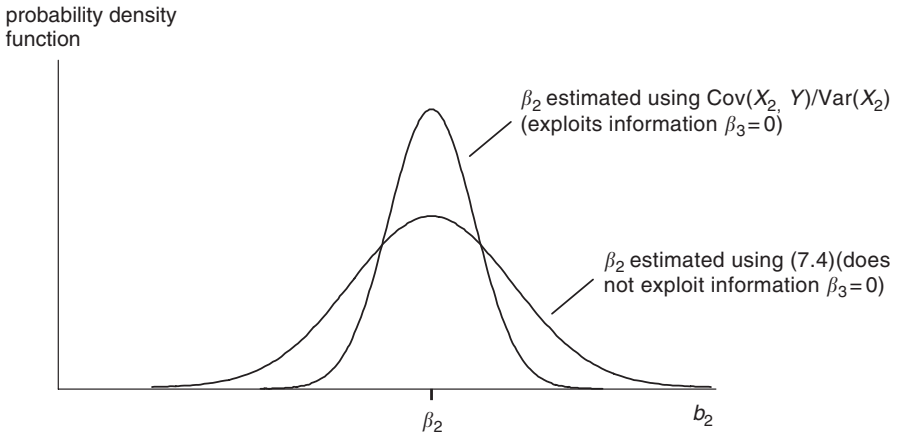

Figure 7.2

Table 7.4

Simple regression	*Multiple regression*
$\sigma_{b_2}^2 = \dfrac{\sigma_u^2}{n\text{Var}(X_2)}$	$\sigma_{b_2}^2 = \dfrac{\sigma_u^2}{n\text{Var}(X_2)} \dfrac{1}{1 - r_{X_2X_3}^2}$

will in general be larger in the case of multiple regression, and the difference will be the greater the closer the correlation coefficient is to plus or minus 1. The one exception to the loss of efficiency occurs when the correlation coefficient happens to be exactly equal to 0. In that case the estimator b_2 for multiple regression will be identical to that for simple regression. The proof of this will be left as an easy exercise.

There is one exception to the unbiasedness conclusion that ought to be kept in mind. If X_3 is correlated with u, the regression coefficients will be biased after all. Writing the model as (7.15), this amounts to the fourth Gauss–Markov condition not being satisfied with respect to X_3.

Example

The regression output shows the results of 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 868 households in the 1995 Consumer Expenditure Survey (Table 7.5). In the second regression, *LGHOUS*, the logarithm of annual expenditure on housing services, has been added. It is safe to assume that *LGHOUS* is an irrelevant variable and, not surprisingly, its coefficient is not significantly different from 0. It is however highly correlated with *LGEXP* (correlation coefficient 0.81), and also, to a

Table 7.5

```
. reg LGFDHO LGEXP LGSIZE

  Source |       SS       df       MS          Number of obs =     868
---------+------------------------------       F(  2,   865) =  460.92
   Model |  138.776549     2  69.3882747        Prob > F      =  0.0000
Residual |  130.219231   865  .150542464        R-squared     =  0.5159
---------+------------------------------       Adj R-squared =  0.5148
   Total |  268.995781   867  .310260416        Root MSE      =    .388

------------------------------------------------------------------------------
  LGFDHO |    Coef.   Std. Err.      t    P>|t|    [95% Conf. Interval]
---------+--------------------------------------------------------------------
   LGEXP | .2866813   .0226824   12.639   0.000    .2421622    .3312003
  LGSIZE | .4854698   .0255476   19.003   0.000    .4353272    .5356124
   _cons | 4.720269   .2209996   21.359   0.000    4.286511    5.154027
------------------------------------------------------------------------------

. reg LGFDHO LGEXP LGSIZE LGHOUS

  Source |       SS       df       MS          Number of obs =     868
---------+------------------------------       F(  3,   864) =  307.22
   Model |  138.841976     3  46.2806586        Prob > F      =  0.0000
Residual |  130.153805   864  .150640978        R-squared     =  0.5161
---------+------------------------------       Adj R-squared =  0.5145
   Total |  268.995781   867  .310260416        Root MSE      =  .38812

------------------------------------------------------------------------------
  LGFDHO |    Coef.   Std. Err.      t    P>|t|    [95% Conf. Interval]
---------+--------------------------------------------------------------------
   LGEXP | .2673552   .0370782    7.211   0.000    .1945813     .340129
  LGSIZE | .4868228   .0256383   18.988   0.000    .4365021    .5371434
  LGHOUS | .0229611   .0348408    0.659   0.510   -.0454214    .0913436
   _cons | 4.708772   .2217592   21.234   0.000    4.273522    5.144022
------------------------------------------------------------------------------
```

lesser extent, with *LGSIZE* (correlation coefficient 0.33). Its inclusion does not cause the coefficients of those variables to be biased but it does increase their standard errors, particularly that of *LGEXP*, as you would expect, given the loss of efficiency.

Exercises

7.5* A social scientist 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-section data on Y, X, and Z, all measured in US$ million, are obtained for

a sample of thirty industrialized countries and a second sample of thirty developing countries. The social scientist 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 results as shown in the table (standard errors in parentheses).

	Industrialized 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 social scientist how to interpret these results.

7.6 Regress *LGEARN* on *S*, *ASVABC*, *MALE*, *ETHHISP*, and *ETHBLACK* using your *EAEF* data set. Repeat the regression, adding *SIBLINGS*. Calculate the correlations between *SIBLINGS* and the other explanatory variables. Compare the results of the two regressions.

7.4 Proxy variables

It frequently happens that you are unable to obtain data on a variable that you would like to include in a regression equation. Some variables, such as socioeconomic status and quality of education, are so vaguely defined that it may be impossible even in principle to measure them. Others might be measurable, but require so much time and energy that in practice they have to be abandoned. Sometimes you are frustrated because you are using survey data collected by someone else, and an important variable (from your point of view) has been omitted.

Whatever the reason, it is usually a good idea to use a proxy for the missing variable, rather than leave it out entirely. For socioeconomic status, you might use income as a substitute if data on it are available. For quality of education,

you might use the staff–student ratio or expenditure per student. For a variable omitted in a survey, you will have to look at the data actually collected to see if there is a suitable substitute.

There are two good reasons for trying to find a proxy. First, if you simply leave the variable out, your regression is likely to suffer from omitted variable bias of the type described in Section 7.2, and the statistical tests will be invalidated. Second, the results from your proxy regression may indirectly shed light on the influence of the missing variable.

Suppose that the true model is

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_k X_k + u. \tag{7.16}$$

Suppose that we have no data for X_2, but another variable Z is an ideal proxy for it in the sense that there exists an exact linear relationship between X_2 and Z:

$$X_2 = \lambda + \mu Z. \tag{7.17}$$

λ and μ being fixed, but unknown, constants. (Note that if λ and μ were known, we could calculate X_2 from Z, and so there would be no need to use Z as a proxy. Note further that we cannot estimate λ and μ by regression analysis, because to do that we need data on X_2.)

Substituting for X_2 from (7.17) into (7.16), the model may be rewritten

$$\begin{aligned} Y &= \beta_1 + \beta_2(\lambda + \mu Z) + \beta_3 X_3 + \cdots + \beta_k X_k + u \\ &= \beta_1 + \beta_2\lambda + \beta_2\mu Z + \beta_3 X_3 + \cdots + \beta_k X_k + u. \end{aligned} \tag{7.18}$$

The model is now formally correctly specified in terms of observable variables, and if we fit it, the following results will obtain:

1. The coefficients of $X_3, \ldots, X_k$, their standard errors, and their t statistics will be the same as if X_2 had been used instead of Z.
2. R^2 will be the same as if X_2 had been used instead of Z.
3. The coefficient of Z will be an estimate of $\beta_2\mu$ and so it will not be possible to obtain an estimate of β_2, unless you are able to guess the value of μ.
4. However, the t statistic for Z will be the same as that which would have been obtained for X_2, and so you are able to assess the significance of X_2, even though you are not able to estimate its coefficient.
5. It will not be possible to obtain an estimate of β_1, since the intercept is now $(\beta_1 + \beta_2\lambda)$, but usually the intercept is of secondary interest, anyway.

With regard to the third point, suppose that you are investigating migration from country A to country B and you are using the (very naïve) model

$$M = \beta_1 + \beta_2 W + u, \tag{7.19}$$

where M is the rate of migration of a certain type of worker from A to B, and W is the ratio of the wage rate in B to the wage rate in A. The higher the relative

wage rate, you think the higher is migration. But suppose that you only have data on GDP per capita, not wages. You might define a proxy variable G that is the ratio of GDP in B to GDP in A.

In this case it might be reasonable to assume, as a first approximation, that relative wages are proportional to relative GDP. If that were true, one could write (7.17) with λ equal to 0 and μ equal to 1. In this case the coefficient of relative GDP would yield a direct estimate of the coefficient of relative wages. Since variables in regression analysis are frequently defined in relative terms, this special case actually has quite a wide application.

In this discussion we have assumed that Z is an ideal proxy for X_2, and the validity of all the foregoing results depends on this condition. In practice it is unusual to find a proxy that is exactly linearly related to the missing variable, but if the relationship is close the results will hold approximately. A major problem is posed by the fact that there is never any means of testing whether the condition is or is not approximated satisfactorily. One has to justify the use of the proxy subjectively.

Example

The main determinants of educational attainment appear to be the cognitive ability of an individual and the support and motivation provided by the family background. The NLSY data set is exceptional in that cognitive ability measures are available for virtually all the respondents, the data being obtained when the Department of Defense, needing to re-norm the Armed Services Vocational Aptitude Battery scores, sponsored the administration of the tests. However, there are no data that bear directly on support and motivation provided by the family background. This factor is difficult to define and probably has several dimensions. Accordingly, it is unlikely that a single proxy could do justice to it. The NLSY data set includes data on parental educational attainment and the number of siblings of the respondent, both of which could be used as proxies, the rationale for the latter being that parents who are ambitious for their children tend to limit the family size in order to concentrate resources. The data set also contains three dummy variables specifically intended to capture family background effects: whether anyone in the family possessed a library card, whether anyone in the family bought magazines, and whether anyone in the family bought newspapers, when the respondent was aged 14. However the explanatory power of these variables appears to be very limited.

The regression output (Table 7.7) shows the results of regressing *S* on *ASVABC* only and on *ASVABC*, parental education, number of siblings, and the library card dummy variable. *ASVABC* is positively correlated with *SM*, *SF*, and *LIBRARY* (correlation coefficients 0.38, 0.42, and 0.22, respectively), and negatively correlated with *SIBLINGS* (correlation coefficient –0.19). Its coefficient is therefore unambiguously biased upwards in the first regression. However, there

Table 7.7

```
. reg S ASVABC

  Source |       SS       df       MS          Number of obs =     570
---------+------------------------------       F(  1,   568) =  284.89
   Model |  1153.80864     1  1153.80864        Prob > F      =  0.0000
Residual |  2300.43873   568  4.05006818        R-squared     =  0.3340
---------+------------------------------       Adj R-squared =  0.3329
   Total |  3454.24737   569  6.07073351        Root MSE      =  2.0125

------------------------------------------------------------------------------
       S |    Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
  ASVABC | .1545378   .0091559    16.879   0.000     .1365543    .1725213
   _cons | 5.770845   .4668473    12.361   0.000     4.853888    6.687803
------------------------------------------------------------------------------

. reg S ASVABC SM SF LIBRARY SIBLINGS

  Source |       SS       df       MS          Number of obs =     570
---------+------------------------------       F(  5,   564) =   66.87
   Model |  1285.58208     5  257.116416        Prob > F      =  0.0000
Residual |  2168.66529   564  3.84515122        R-squared     =  0.3722
---------+------------------------------       Adj R-squared =  0.3666
   Total |  3454.24737   569  6.07073351        Root MSE      =  1.9609

------------------------------------------------------------------------------
       S |     Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
  ASVABC |  .1277852    .010054   12.710   0.000     .1080373     .147533
      SM |  .0619975   .0427558    1.450   0.148    -.0219826    .1459775
      SF |  .1045035   .0314928    3.318   0.001      .042646     .166361
 LIBRARY |  .1151269   .1969844    0.584   0.559    -.2717856    .5020394
SIBLINGS | -.0509486    .039956   -1.275   0.203    -.1294293     .027532
   _cons |  5.236995   .5665539    9.244   0.000     4.124181    6.349808
------------------------------------------------------------------------------
```

may still be an element of bias in the second, given the weakness of the proxy variables.

Unintentional Proxies

It sometimes happens that you use a proxy without realizing it. You think that Y depends upon Z, but in reality it depends upon X.

If the correlation between Z and X is low, the results will be poor, so you may realize that something is wrong, but, if the correlation is good, the results may appear to be satisfactory (R^2 up to the anticipated level, etc) and you may remain blissfully unaware that the relationship is false.

Does this matter? Well, it depends on why you are running the regression in the first place. If the purpose of fitting the regression line is to predict future values of Y, the use of a proxy will not matter much, provided of course that the correlation remains high and was not a statistical fluke in the sample period. However, if your intention is to use the explanatory variable as a policy instrument for influencing the dependent variable, the consequences could be serious. Unless there happens to be a functional connection between the proxy and the true explanatory variable, manipulating the proxy will have no effect at all on the dependent variable. If the motive for your regression is scientific curiosity, the outcome is equally unsatisfactory.

Unintentional proxies are especially common in time-series analysis, particularly in macroeconomic models. If the true explanatory variable is subject to a time trend, you will probably get a good fit if you substitute (intentionally or otherwise) any other variable with a time trend. Even if you relate changes in your dependent variable to changes in your explanatory variable, you are likely to get similar results whether you are using the correct explanatory variable or a proxy, since macroeconomic variables tend to change in concert over the trade cycle.

Exercises

7.7 Length of work experience is generally found to be an important determinant of earnings. The data set does not contain this variable, but *TENURE*, tenure with the current employer, could be taken as a proxy. An alternative is to calculate years of potential work experience, *PWE*, as a proxy. This is defined to be current age, *AGE*, less age of completion of full-time education. The latter can be estimated as years of schooling plus 5, assuming that schooling begins at the age of 6. Hence

$$PWE = AGE - S - 5.$$

Using your *EAEF* data set, regress *LGEARN* on *S*, *ASVABC*, *MALE*, *ETHBLACK*, *ETHHISP*, and *PWE*. Compare the results with the corresponding regression without *PWE*. You are likely to find that the coefficient of *S* is greater than before. Can you explain why?

The data set includes *TENURE*, tenure with current employer. This allows one to divide *PWE* into two components: potential work experience with previous employers, *PWEBEF*, and *TENURE*. Define *PWEBEF* as

$$PWEBEF = PWE - TENURE$$

and regress *LGEARN* on the variables as before, replacing *PWE* by *PWEBEF* and *TENURE*. Compare the result with that of the previous regression.

Variation: *PWE* is not likely to be a satisfactory proxy for work experience for females because it does not take into account time spent out of the labor force rearing children. Investigate this by running the regressions

with *PWE* for the male and female subsamples separately. You must drop the *MALE* dummy from the specification (explain why). Do the same for the regressions with *PWEBEF* and *TENURE*.

7.8* A researcher has data on output per worker, Y, and capital per worker, K, both measured in thousands of dollars, for fifty firms in the textiles industry in 2001. She hypothesizes 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 2001, *R&D*, as a proxy for it. She fits the following regressions (standard errors in parentheses):

$$\hat{Y} = 1.02 + 0.32K. \quad R^2 = 0.79$$
$$\quad\quad (0.45) \quad (0.04)$$
$$\hat{Y} = 0.34 + 0.29K + 0.05R\&D. \quad R^2 = 0.750$$
$$\quad\quad (0.61) \quad (0.22) \quad (0.15)$$

The correlation coefficient for K and *R& D* was 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.

7.5 Testing a linear restriction

In Section 4.4 it was demonstrated that you can reduce the number of explanatory variables in a regression equation by one if you believe that there exists a linear relationship between the parameters in it. By exploiting the information about the relationship, you will make the regression estimates more efficient. If there was previously a problem of multicollinearity, it may be alleviated. Even if the original model was not subject to this problem, the gain in efficiency may yield a welcome improvement in the precision of the estimates, as reflected by their standard errors.

The example discussed in Section 4.4 was an educational attainment model with *S* related to *ASVABC*, *SM*, and *SF* (Table 7.7).

Somewhat surprisingly, the coefficient of *SM* is not significant, even at the 5 percent level, using a one-tailed test. However assortive mating leads to a high correlation between *SM* and *SF* and the regression appeared to be suffering from multicollinearity (Table 7.8).

We then hypothesized that mother's and father's education are equally important for educational attainment, allowing us to impose the restriction $\beta_3 = \beta_4$

Table 7.7

```
. reg S ASVABC SM SF

  Source |       SS       df       MS          Number of obs =     570
---------+------------------------------       F(  3,   566) =  110.83
   Model |  1278.24153     3  426.080508        Prob > F      =  0.0000
Residual |  2176.00584   566  3.84453329        R-squared     =  0.3700
---------+------------------------------       Adj R-squared =  0.3667
   Total |  3454.24737   569  6.07073351        Root MSE      =  1.9607

------------------------------------------------------------------------------
       S |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
  ASVABC |   .1295006   .0099544    13.009   0.000     .1099486    .1490527
      SM |    .069403   .0422974     1.641   0.101    -.013676     .152482
      SF |   .1102684   .0311948     3.535   0.000     .0489967    .1715401
   _cons |   4.914654   .5063527     9.706   0.000     3.920094    5.909214
------------------------------------------------------------------------------
```

Table 7.8

```
. g SP=SM+SF
. reg S ASVABC SP

  Source |       SS       df       MS          Number of obs =     570
---------+------------------------------       F(  2,   567) =  166.22
   Model |  1276.73764     2  638.368819        Prob > F      =  0.0000
Residual |  2177.50973   567  3.84040517        R-squared     =  0.3696
---------+------------------------------       Adj R-squared =  0.3674
   Total |  3454.24737   569  6.07073351        Root MSE      =  1.9597

------------------------------------------------------------------------------
       S |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
  ASVABC |   .1295653   .0099485    13.024   0.000     .1100249    .1491057
      SP |    .093741   .0165688     5.658   0.000     .0611973    .1262847
   _cons |   4.823123   .4844829     9.955   0.000     3.871523    5.774724
------------------------------------------------------------------------------
```

and rewrite the equation as

$$\begin{aligned} S &= \beta_1 + \beta_2 ASVSABC + \beta_3 (SM + SF) + u \\ &= \beta_1 + \beta_2 ASVSABC + \beta_3 SP + u, \end{aligned} \tag{7.20}$$

where SP is the sum of SM and SF.

The standard error of SP is much smaller than those of SM and SF, indicating that the use of the restriction has led to a gain in efficiency, and as a consequence

the t statistic is very high. Thus the problem of multicollinearity has been eliminated. However, we are obliged to test the validity of the restriction, and there are two equivalent procedures.

F test of a restriction

Run the regression in both the restricted and the unrestricted forms and denote the sum of the squares of the residuals RSS_R in the restricted case and RSS_U in the unrestricted case. Since the imposition of the restriction makes it more difficult to fit the regression equation to the data, RSS_R cannot be less than RSS_U and will in general be greater. We would like to test whether the improvement in the fit on going from the restricted to the unrestricted version is significant. If it is, the restriction should be rejected.

For this purpose we can use an F test whose structure is the same as that described in Section 4.5:

$$F = \frac{\text{Improvement in fit/Extra degrees of freedom used up}}{\text{Residual sum of squares remaining/Degrees of freedom remaining}}. \tag{7.21}$$

In this case the improvement on going from the restricted to the unrestricted version is $(RSS_R - RSS_U)$, one extra degree of freedom is used up in the unrestricted version (because there is one more parameter to estimate), and the residual sum of squares remaining after the shift from the restricted to the unrestricted version is RSS_U. Hence the F statistic is in this case

$$F(1, n-k) = \frac{RSS_R - RSS_U}{RSS_U/(n-k)}, \tag{7.22}$$

where k is the number of parameters in the unrestricted version. It is distributed with 1 and $n-k$ degrees of freedom under the null hypothesis that the restriction is valid.

In the case of the educational attainment function, the null hypothesis was $H_0: \beta_3 = \beta_4$, where β_3 is the coefficient of SM and β_4 is the coefficient of SF. The residual sum of squares was 2177.51 in the restricted version and 2176.01 in the unrestricted version. Hence the F statistic is

$$F(1, n-k) = \frac{2177.51 - 2176.01}{2176.01/566} = 0.39. \tag{7.23}$$

Since the F statistic is less than 1, it is not significant at any significance level and we do not reject the null hypothesis that the coefficients of SM and SF are equal.

t test of a restriction

Linear restrictions can also be tested using a t test. This involves writing down the model for the restricted version and adding the term that would convert it

back to the unrestricted version. The test evaluates whether this additional term is needed. To find the conversion term, we write the restricted version of the model under the unrestricted version and subtract:

$$S = \beta_1 + \beta_2 ASVABC + \beta_3 SM + \beta_4 SF + u. \tag{7.24}$$

$$S = \beta_1 + \beta_2 ASVABC + \beta_3 SP + u. \tag{7.25}$$

$$\begin{aligned} 0 &= \beta_3 SM + \beta_4 SF - \beta_3 SP \\ &= \beta_3 SM + \beta_4 SF - \beta_3 (SM + SF) \\ &= (\beta_4 - \beta_3) SF. \end{aligned} \tag{7.26}$$

We add this term to the restricted model and investigate whether it is needed.

$$S = \beta_1 + \beta_2 ASVABC + \beta_3 SP + (\beta_4 - \beta_3) SF + u. \tag{7.27}$$

The null hypothesis, H_0: $\beta_4 - \beta_3 = 0$, is that the coefficient of the conversion term is 0, and the alternative hypothesis is that it is different from 0. Of course the null hypothesis is that the restriction is valid. If it is valid, the conversion term is not needed, and the restricted version is an adequate representation of the data.

Here is the corresponding regression for the educational attainment example (Table 7.9). We see that the coefficient of *SF* is not significantly different from 0, indicating that the term is not needed and that the restricted version is an adequate representation of the data.

Why is the *t* test approach equivalent to that of the *F* test? Well the *F* test tests the improvement in fit when you go from the restricted version to the unrestricted version. This is accomplished by adding the conversion term, but, as we know, an *F* test on the improvement in fit when you add an extra term is equivalent to the *t* test on the coefficient of that term (see Section 4.5).

Exercises

7.9 You will have found in Exercise 7.7 that the estimates of the coefficients of *PWEBEF* and *TENURE* are different. This raises the issue of whether the difference is due to random factors or whether the coefficients are significantly different. Set up the null hypothesis H_0: $\delta_1 = \delta_2$, where δ_1 is the coefficient of *PWEBEF* and δ_2 is the coefficient of *TENURE*. Explain why the regression with *PWE* is the correct specification if H_0 is true, while the regression with *PWEBEF* and *TENURE* should be used if H_0 is false. Perform an *F* test of the restriction using *RSS* for the two regressions. Do this for the combined sample and also for males and females separately.

7.10* 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

Table 7.9

```
. reg S ASVABC SP SF

  Source |       SS       df       MS          Number of obs =     570
---------+------------------------------       F(  3,   566) =  110.83
   Model |  1278.24153     3  426.080508        Prob > F      =  0.0000
Residual |  2176.00584   566  3.84453329        R-squared     =  0.3700
---------+------------------------------       Adj R-squared =  0.3667
   Total |  3454.24737   569  6.07073351        Root MSE      =  1.9607

------------------------------------------------------------------------------
       S |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
  ASVABC |   .1295006   .0099544    13.009   0.000    .1099486    .1490527
      SP |    .069403   .0422974     1.641   0.101    -.013676     .152482
      SF |   .0408654   .0653386     0.625   0.532   -.0874704    .1692012
   _cons |   4.914654   .5063527     9.706   0.000    3.920094    5.909214
------------------------------------------------------------------------------

. reg LGFDHO LGEXP LGSIZE

  Source |       SS       df       MS          Number of obs =     868
---------+------------------------------       F(  2,   865) =  460.92
   Model |  138.776549     2  69.3882747        Prob > F      =  0.0000
Residual |  130.219231   865  .150542464        R-squared     =  0.5159
---------+------------------------------       Adj R-squared =  0.5148
   Total |  268.995781   867  .310260416        Root MSE      =    .388

------------------------------------------------------------------------------
  LGFDHO |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
   LGEXP |  .2866813   .0226824    12.639   0.000    .2421622    .3312003
  LGSIZE |  .4854698   .0255476    19.003   0.000    .4353272    .5356124
   _cons |  4.720269   .2209996    21.359   0.000    4.286511    5.154027
------------------------------------------------------------------------------

. reg LGFDHOPC LGEXPPC

  Source |       SS       df       MS          Number of obs =     868
---------+------------------------------       F(  1,   866) =  313.04
   Model |  51.4364364     1  51.4364364        Prob > F      =  0.0000
Residual |  142.293973   866  .164311747        R-squared     =  0.2655
---------+------------------------------       Adj R-squared =  0.2647
   Total |   193.73041   867  .223449146        Root MSE      =  .40535

------------------------------------------------------------------------------
LGFDHOPC |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
 LGEXPPC |   .376283   .0212674    17.693   0.000    .3345414    .4180246
   _cons |  3.700667   .1978925    18.700   0.000    3.312262    4.089072
------------------------------------------------------------------------------
```

Table 7.9 *Continued*

```
. reg LGFDHOPC LGEXPPC LGSIZE

  Source |       SS       df       MS          Number of obs =     868
---------+------------------------------       F(  2,   865) =  210.94
   Model |  63.5111811     2  31.7555905        Prob > F      =  0.0000
Residual |  130.219229   865  .150542461        R-squared     =  0.3278
---------+------------------------------       Adj R-squared =  0.3263
   Total |   193.73041   867  .223449146        Root MSE      =    .388

------------------------------------------------------------------------------
LGFDHOPC |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
 LGEXPPC |   .2866813   .0226824    12.639   0.000     .2421622    .3312004
  LGSIZE |  -.2278489   .0254412    -8.956   0.000    -.2777826   -.1779152
   _cons |   4.720269   .2209996    21.359   0.000     4.286511    5.154027
```

LGSIZE, the logarithm of the number of persons in the household, using a sample of 868 households in the 1995 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*.

- Explain why the second model is a restricted version of the first, stating the restriction.
- Perform an F test of the restriction.
- Perform a t test of the restriction.
- Summarize your conclusions from the analysis of the regression results.

7.11 In his classic article, Nerlove (1963) derives the following cost function for electricity generation:

$$C = \beta_1 Y^{\beta_2} P_1^{\gamma_1} P_2^{\gamma_1} P_3^{\gamma_3} \upsilon,$$

where C is total production cost, Y is output (measured in kilowatt hours), P_1 is the price of labor input, P_2 is the price of capital input, P_3 is the price of fuel (all measured in appropriate units), and υ is a disturbance term. Theoretically, the sum of the price elasticities should be 1:

$$\gamma_1 + \gamma_2 + \gamma_3 = 1,$$

and hence the cost function may be rewritten

$$\frac{C}{P_3} = \beta_1 Y^{\beta_2} \left(\frac{P_1}{P_3}\right)^{\gamma_1} \left(\frac{P_2}{P_3}\right)^{\gamma_2} \upsilon.$$

The two versions of the cost function are fitted to the twenty-nine medium-sized firms in Nerlove's sample, with the following results (standard errors in parentheses):

$$\widehat{\log C} = -4.93 + 0.94 \log Y + 0.31 \log P_1 - 0.26 \log P_2 + 0.44 \log P_3.$$
$$(1.62) \quad (0.11) \quad (0.23) \quad (0.29) \quad (0.07)$$
$$RSS = 0.336$$
$$\log \frac{\widehat{C}}{P_3} = -6.55 + 0.91 \log Y + 0.51 \log \frac{P_1}{P_3} + 0.09 \log \frac{P_2}{P_3}.$$
$$(0.16) \quad (0.11) \quad (0.23) \quad (0.19)$$
$$RSS = 0.364$$

Compare the regression results for the two equations and perform a test of the validity of the restriction.

7.6 Getting the most out of your residuals

There are two ways of looking at the residuals obtained after fitting a regression equation to a set of data. If you are pessimistic and passive, you will simply see them as evidence of failure. The bigger the residuals, the worse is your fit, and the smaller is R_2. The whole object of the exercise is to fit the regression equation in such a way as to minimize the sum of the squares of the residuals. However, if you are enterprising, you will also see the residuals as a potentially fertile source of new ideas, perhaps even new hypotheses. They offer both a challenge and constructive criticism. The challenge is that providing the stimulus for most scientific research: evidence of the need to find a better explanation of the facts. The constructive criticism comes in because the residuals, taken individually, indicate when and where and by how much the existing model is failing to fit the facts.

Taking advantage of this constructive criticism requires patience on the part of the researcher. If the sample is small enough, you should look carefully at every observation with a large positive or negative residual, and try to hypothesize explanations for them. Some of these explanations may involve special factors specific to the observations in question. These are not of much use to the theorist. Other factors, however, may appear to be associated with the residuals in several observations. As soon as you detect a regularity of this kind, you have the makings of progress. The next step is to find a sensible way of quantifying the factor and of including it in the model.