3. Multiple Regression Analysis

Chapters 1 and 2 were restricted to the simple regression model where it was assumed that the dependent variable in the model was related to only one explanatory variable. In general, there will be several, perhaps many, explanatory variables and we wish to quantify the impact of each, controlling for the effects of the others. In the natural sciences, one may perform controlled experiments, varying each explanatory variable, holding the others constant. In economics, this is usually not possible, but we may address the objective of discriminating between the effects of different explanatory variables using the technique known as multiple regression analysis that is treated in this chapter. Much of the discussion will be a straightforward extension of the simple regression model. Most of the issues can be explained within the context of a model with just two explanatory variables and we will start with such a model.

3.1 Illustration: a model with two explanatory variables

We will begin by considering an example, the determinants of earnings. We will extend the earlier model to allow for the possibility that earnings are influenced by years of work experience, as well as education, and we will assume that the true relationship can be expressed as

$$EARNINGS = \beta_1 + \beta_2 S + \beta_3 EXP + u, \tag{3.1}$$

where *EARNINGS* is hourly earnings, *S* is years of schooling (highest grade completed), *EXP* is years spent working after leaving full-time education, and u is a disturbance term. This model is still of course a great simplification, both in terms of the explanatory variables included in the relationship and in terms of its mathematical specification.

To illustrate the relationship geometrically, one needs a three-dimensional diagram with separate axes for *EARNINGS*, *S*, and *EXP* as in Figure 3.1. The base of Figure 3.1 shows the axes for *S* and *EXP*, and, if one neglects the effect of the disturbance term for the moment, the tilted plane above it shows the value of *EARNINGS* corresponding to any (*S*, *EXP*) combination, measured by the vertical height of the plane above the base at that point. Since earnings may be expected to increase with both schooling and work experience, the diagram

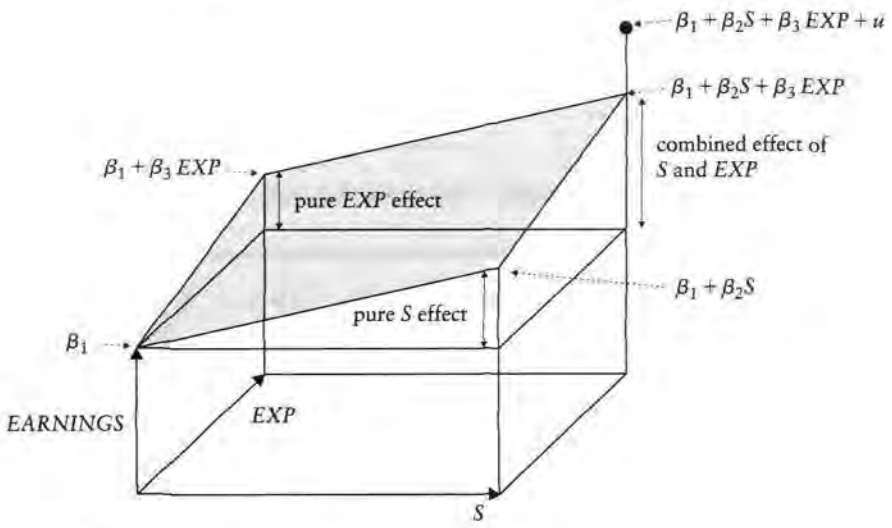

Figure 3.1 True model with two explanatory variables: earnings as a function of schooling and work experience

has been drawn on the assumption that β_2 and β_3 are both positive. Literally, the intercept β_1 gives the predicted earnings for zero schooling and zero work experience. However, such an interpretation would be dangerous because there was nobody with no schooling in the NLSY data set. Indeed very few individuals failed to complete eight years of schooling. Mathematically, (3.1) implies that, if *EXP* were zero, for any positive *S*, earnings would be equal to $\beta_1 + \beta_2 S$, the increase $\beta_2 S$ being marked 'pure *S* effect' in the figure. Keeping *S* at zero, the equation implies that for any positive value of *EXP*, earnings would be equal to $\beta_1 + \beta_3 EXP$, the increase $\beta_3 EXP$ being marked 'pure *EXP* effect'. The combined effect of schooling and work experience, $\beta_2 S + \beta_3 EXP$, is also indicated.

We have thus far neglected the disturbance term. If it were not for the presence of this in (3.1), the values of *EARNINGS* in a sample of observations on *EARNINGS*, *S*, and *EXP* would lie exactly on the tilted plane and it would be a trivial matter to deduce the exact values of β_1, β_2, and β_3 (not trivial geometrically, unless you are a genius at constructing three-dimensional models, but easy enough algebraically).

The disturbance term causes the actual value of earnings to be sometimes above and sometimes below the value indicated by the tilted plane. Consequently, one now has a three-dimensional counterpart to the two-dimensional problem illustrated in Figure 1.2. Instead of locating a line to fit a two-dimensional scatter of points, we now have to locate a plane to fit a three-dimensional scatter. The equation of the fitted plane will be

$$\widehat{EARNINGS} = b_1 + b_2 S + b_3 EXP \tag{3.2}$$

Table 3.1

```
. reg EARNINGS S EXP

      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(2, 537)     =   67.54
       Model |  22513.6473     2  11256.8237            Prob > F      =  0.0000
    Residual |  89496.5838   537  166.660305            R-squared     =  0.2010
-------------+------------------------------           Adj R-squared =  0.1980
       Total |  112010.231   539  207.811189            Root MSE      =   12.91

------------------------------------------------------------------------------
    EARNINGS |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           S |   2.678125   .2336497    11.46   0.000     2.219146    3.137105
         EXP |   .5624326   .1285136     4.38   0.000     .3099816    .8148837
       _cons |  -26.48501   4.27251    -6.20   0.000    -34.87789   -18.09213
------------------------------------------------------------------------------
```

and its location will depend on the choice of b_1, b_2, and b_3, the estimates of β_1, β_2, and β_3, respectively. Using *EAEF* Data Set 21, we obtain the regression output shown in Table 3.1.

The equation should be interpreted as follows. For every additional year of schooling, holding work experience constant, hourly earnings increase by \$2.68. For every year of work experience, holding schooling constant, earnings increase by \$0.56. The constant has no meaningful interpretation. Literally, it suggests that a respondent with zero years of schooling (no respondent had fewer than six) and no work experience would earn *minus* \$26.49 per hour.

3.2 Derivation and interpretation of the multiple regression coefficients

We will initially examine the case where a dependent variable Y may be assumed to be determined by two explanatory variables, X_2 and X_3, the true relationship being

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i, \tag{3.3}$$

where u is a disturbance term. The X variables now have two subscripts. The first identifies the X variable (years of schooling, years of work experience, etc.) and the second the observation. The fitted model will be written

$$\hat{Y}_i = b_1 + b_2 X_{2i} + b_3 X_{3i}. \tag{3.4}$$

As in the simple regression case, we choose the values of the regression coefficients to make the fit as good as possible in the hope that we will obtain the most satisfactory estimates of the unknown true parameters. As before, our definition of goodness of fit is the minimization of $RSS = \sum_{i=1}^{n} e_i^2$, the sum of squares of the residuals, where e_i is the residual in observation i, the difference

BOX 3.1 Whatever happened to X_1?

You may have noticed that X_1 is missing from the general regression model

$$Y_i = \beta_1 + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i.$$

Why so? The reason is to make the notation consistent with that found in texts using linear algebra (matrix algebra), and your next course in econometrics will almost certainly use such a text. For analysis using linear algebra, it is essential that every term on the right side of the equation should consist of the product of a parameter and a variable. When there is an intercept in the model, as here, the anomaly is dealt with by writing the equation

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i,$$

where $X_{1i} = 1$ in every observation. In analysis using ordinary algebra, there is usually no point in introducing X_1 explicitly, and so it has been suppressed. One occasion in this text where it can help is in the discussion of the dummy variable trap in Section 5.2.

between the actual value Y_i in that observation and the value $\hat{Y}_i$ predicted by the regression equation:

$$e_i = Y_i - \hat{Y}_i = Y_i - b_1 - b_2 X_{2i} - b_3 X_{3i}. \tag{3.5}$$

Thus,

$$RSS = \sum_{i=1}^{n} (Y_i - b_1 - b_2 X_{2i} - b_3 X_{3i})^2 \tag{3.6}$$

The first-order conditions for a minimum, $\frac{\partial RSS}{\partial b_1} = 0$, $\frac{\partial RSS}{\partial b_2} = 0$, and $\frac{\partial RSS}{\partial b_3} = 0$, yield the following equations:

$$\frac{\partial RSS}{\partial b_1} = -2\sum_{i=1}^{n} (Y_i - b_1 - b_2 X_{2i} - b_3 X_{3i}) = 0 \tag{3.7}$$

$$\frac{\partial RSS}{\partial b_2} = -2\sum_{i=1}^{n} X_{2i}(Y_i - b_1 - b_2 X_{2i} - b_3 X_{3i}) = 0 \tag{3.8}$$

$$\frac{\partial RSS}{\partial b_3} = -2\sum_{i=1}^{n} X_{3i}(Y_i - b_1 - b_2 X_{2i} - b_3 X_{3i}) = 0. \tag{3.9}$$

Hence, we have three equations in the three unknowns, b_1, b_2, and b_3. The first can easily be rearranged to express b_1 in terms of b_2, b_3, and the data on Y, X_2, and X_3:

$$b_1 = \bar{Y} - b_2 \bar{X}_2 - b_3 \bar{X}_3. \tag{3.10}$$

Using this expression and the other two equations, with a little work one can obtain the following expression for b_2:

$$b_2 = \frac{\sum_{i=1}^{n}(X_{2i}-\bar{X}_2)(Y_i-\bar{Y})\sum_{i=1}^{n}(X_{3i}-\bar{X}_3)^2 - \sum_{i=1}^{n}(X_{3i}-\bar{X}_3)(Y_i-\bar{Y})\sum_{i=1}^{n}(X_{2i}-\bar{X}_2)(X_{3i}-\bar{X}_3)}{\sum_{i=1}^{n}(X_{2i}-\bar{X}_2)^2\sum_{i=1}^{n}(X_{3i}-\bar{X}_3)^2 - \left(\sum_{i=1}^{n}(X_{2i}-\bar{X}_2)(X_{3i}-\bar{X}_3)\right)^2}. \quad (3.11)$$

A parallel expression for b_3 can be obtained by interchanging X_2 and X_3 in (3.11).

The intention of this discussion is to make two basic points. First, the principles behind the derivation of the regression coefficients are the same for multiple regression as for simple regression. Second, the expressions, however, are different. The expression for the intercept, b_1, is an extension of that for simple regression analysis, but the expressions for the slope coefficients are more complex.

The general model

When there are more than two explanatory variables, it is not possible to give a geometrical representation of the model, but the extension of the algebra is in principle quite straightforward. We assume that a variable Y depends on $k - 1$ explanatory variables $X_2, \ldots, X_k$ according to a true, unknown relationship

$$Y_i = \beta_1 + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i. \quad (3.12)$$

Given a set of n observations on $Y, X_2, \ldots, X_k$, we use least squares regression analysis to fit the equation

$$\hat{Y}_i = b_1 + b_2 X_{2i} + \cdots + b_k X_{ki}. \quad (3.13)$$

This again means minimizing the sum of the squares of the residuals, which are given by

$$e_i = Y_i - \hat{Y}_i = Y_i - b_1 - b_2 X_{2i} - \cdots - b_k X_{ki}. \quad (3.14)$$

We now choose $b_1, \ldots, b_k$ so as to minimize RSS, the sum of the squares of the residuals, $\sum e_i^2$. We obtain k first-order conditions $\partial RSS/\partial b_1 = 01, \ldots, \partial RSS/\partial b_k = 0$, and these provide k equations for solving for the k unknowns.

It can readily be shown that the first of these equations yields a counterpart to (3.10) in the case with two explanatory variables:

$$b_1 = \bar{Y} - b_2\bar{X}_2 - \cdots - b_k\bar{X}_k. \quad (3.15)$$

The expressions for $b_2, \ldots, b_k$ become very complicated and the mathematics will not be presented explicitly here. The analysis should be done with linear (matrix) algebra.

Interpretation of the multiple regression coefficients

Multiple regression analysis allows one to discriminate between the effects of the explanatory variables, making allowance for the fact that they may be correlated. The regression coefficient of each X variable provides an estimate of its influence on Y, controlling for the effects of all the other X variables.

This can be demonstrated in two ways. One is to show that the estimators are unbiased, if the model is correctly specified and the assumptions relating to the regression model are valid. We shall do this in the next section for the case where there are only two explanatory variables. A second method is to run a simple regression of Y on one of the X variables, having first purged both Y and the X variable of the components that could be accounted for by the other explanatory variables. The estimate of the slope coefficient and its standard error thus obtained are exactly the same as in the multiple regression, a result that is proved by the Frisch–Waugh–Lovell theorem (Frisch and Waugh, 1933; Lovell, 1963). It follows that a scatter diagram plotting the purged Y against the purged X variable will provide a valid graphical representation of their relationship that can be obtained in no other way. This result will not be proved but it will be illustrated using the earnings function in Section 3.1:

$$EARNINGS = \beta_1 + \beta_2 S + \beta_3 EXP + u. \tag{3.16}$$

Suppose that we are particularly interested in the relationship between earnings and schooling and that we would like to illustrate it graphically. A straightforward plot of *EARNINGS* on *S*, as in Figure 1.7, would give a distorted view of the relationship because *EXP* is negatively correlated with *S*. Among those of similar age, individuals who have spent more time in school will tend to have spent less time working. As a consequence, as *S* increases, (1) *EARNINGS* will tend to increase, because β_2 is positive; (2) *EXP* will tend to decrease, because *S* and *EXP* are negatively correlated; and (3) *EARNINGS* will be reduced by the decrease in *EXP* and the fact that β_3 is positive. In other words, the variations in *EARNINGS* will not fully reflect the influence of the variations in *S* because in part they will be undermined by the associated variations in *EXP*. As a consequence, in a simple regression the estimator of β_2 will be biased downwards. We will investigate the bias analytically in Section 6.2.

In this example, there is only one other explanatory variable, *EXP*. To purge *EARNINGS* and *S* of their *EXP* components, we first regress them on *EXP*:

$$\widehat{EARNINGS} = c_1 + c_2 EXP \tag{3.17}$$

$$\hat{S} = d_1 + d_2 EXP. \tag{3.18}$$

We then subtract the fitted values from the actual values:

$$EEARN = EARNINGS - \widehat{EARNINGS}. \tag{3.19}$$

Table 3.2

```
. reg EEARN ES

      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(1, 538)     =  131.63
       Model |  21895.9298     1  21895.9298            Prob > F      =  0.0000
    Residual |  89496.5838   537  166.350527            R-squared     =  0.1966
-------------+------------------------------           Adj R-squared =  0.1951
       Total |  111392.513   539  206.665145            Root MSE      =  12.898

------------------------------------------------------------------------------
       EEARN |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          ES |   2.678125   .2334325    11.47   0.000     2.219574    3.136676
       _cons |   8.10e-09   .5550284     0.00   1.000    -1.090288    1.090288
------------------------------------------------------------------------------
```

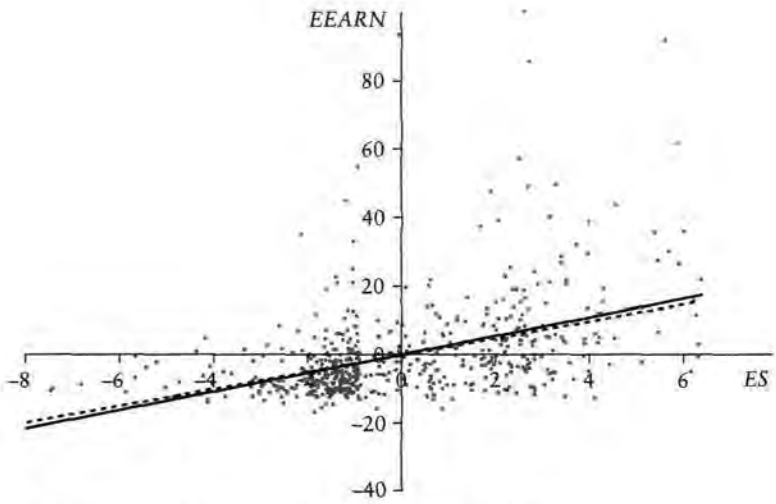

Figure 3.2 Regression of *EARNINGS* residuals on *S* residuals

$$ES = S - \hat{S}. \tag{3.20}$$

The purged variables *EEARN* and *ES* are of course just the residuals from the regressions (3.17) and (3.18). We now regress *EEARN* on *ES* and obtain the output in Table 3.2.

The estimate of the intercept in the regression uses a common convention for fitting very large numbers or very small ones into a field with a predefined number of digits. e+*n* indicates that the coefficient should be multiplied by 10^n. Similarly e−*n* indicates that it should be multiplied by 10^{-n}. Thus, in this regression the intercept is effectively zero.

You can verify that the coefficient of *ES* is identical to that of *S* in the multiple regression in Section 3.1. Figure 3.2 shows the regression line in a scatter diagram. The dotted line in the figure is the regression line from a simple regression

of *EARNINGS* on *S*, shown for comparison. The latter is a little flatter than the true relationship between *EARNINGS* and *S* because it does not control for the effect of *EXP*. In this case, the bias is small because the correlation between *S* and *EXP*, –0.22, is small. Even so, the diagram is valuable because it allows a direct inspection of the relationship between earnings and schooling, controlling for experience. The presence of outliers for large values of *S* suggests that the model is misspecified in some way.

EXERCISES

3.1 The output is the result of fitting an educational attainment function, regressing *S* on *ASVABC*, a measure of cognitive ability, *SM*, and *SF*, years of schooling (highest grade completed) of the respondent's mother and father, respectively, using *EAEF* Data Set 21. Give an interpretation of the regression coefficients.

```
. reg S ASVABC SM SF

      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(3, 536)     =  104.30
       Model |  1181.36981     3  393.789935            Prob > F      =  0.0000
    Residual |  2023.61353   536  3.77539837           R-squared     =  0.3686
-------------+------------------------------           Adj R-squared =  0.3651
       Total |  3204.98333   539  5.94616574           Root MSE      =  1.943

------------------------------------------------------------------------------
           S |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      ASVABC |   .1257087   .0098533    12.76   0.000     .1063528   .1450646
          SM |   .0492424   .0390901     1.26   0.208    -.027546    .1260309
          SF |   .1076825   .0309522     3.48   0.001     .04688     .1684851
       _cons |   5.370631   .4882155    11.00   0.000     4.41158    6.329681
------------------------------------------------------------------------------
```

3.2 Fit an educational attainment function parallel to that in Exercise 3.1, using your *EAEF* data set. First regress *S* on *ASVABC* and *SM* and interpret the regression results. Repeat the regression using *SF* instead of *SM*, and then again, including both *SM* and *SF* as regressors. There is a saying that if you educate a male, you educate an individual, while if you educate a female, you educate a nation. The premise is that the education of a future mother has a beneficial knock-on effect on the educational attainment of her children. Do your regression results support this view?

3.3 Fit an earnings function parallel to that in Section 3.1, using your *EAEF* data set. Regress *EARNINGS* on *S* and *EXP* and interpret the regression results.

3.4 Using your *EAEF* data set, make a graphical representation of the relationship between *S* and *SM* using the Frisch–Waugh–Lovell technique, assuming that the true model is as in Exercise 3.2. To do this, regress *S* on *ASVABC* and *SF* and save the residuals. Do the same with *SM*. Plot the *S* and *SM* residuals. Also regress the former on the latter, and verify that the slope coefficient is the same as that obtained in Exercise 3.2.

3.5* Explain why the intercept in the regression of *EEARN* on *ES* is equal to zero.

3.6 Show that in the general case, with true model (3.12) and fitted regression (3.13), the fitted regression will pass through the point represented by $\bar{Y}$ and the means of the X variables, provided that the equation includes an intercept. (This is a generalization of Exercise 1.1.)

3.7 Two researchers are investigating the effects of time spent studying on the examination marks earned by students on a certain course. For a sample of 100 students, they have the examination mark, M, total hours spent studying, H, hours on primary study, P, and hours spent on revision, R. By definition, $H = P + R$. Researcher A decides to regress M on P and R and fits the following regression:

$$\hat{M} = 45.6 + 0.15\,P + 0.21\,R.$$

Researcher B decides to regress M on H and P, with regression output

$$\hat{M} = 45.6 + 0.21\,H - 0.06\,P.$$

Give an interpretation of the coefficients of both regressions.

3.3 Properties of the multiple regression coefficients

As in the case of simple regression analysis, the regression coefficients should be thought of as special kinds of random variables whose random components are attributable to the presence of the disturbance term in the model. Each regression coefficient is calculated as a function of the values of Y and the explanatory variables in the sample, and Y in turn is determined by the explanatory variables and the disturbance term. It follows that the regression coefficients are really determined by the values of the explanatory variables and the disturbance term and that their properties depend critically upon the properties of the latter.

We are continuing to work within the framework of Model A, where the explanatory variables are nonstochastic. We shall make the following six assumptions, which are a restatement of those in Chapter 2 in terms appropriate for the multiple regression model.

A.1 The model is linear in parameters and correctly specified.

$$Y = \beta_1 + \beta_2 X_2 + \cdots + \beta_k X_k + u. \tag{3.21}$$

This is the same as before, except that we have multiple explanatory variables.

A.2 There does not exist an exact linear relationship among the regressors in the sample.

This is the only assumption that requires a new explanation. It will be deferred to Section 3.4 on multicollinearity.

Assumptions A.3–A.6 are the same as before.

A.3 The disturbance term has zero expectation.

$$E(u_i) = 0 \quad \text{for all } i. \tag{3.22}$$

A.4 The disturbance term is homoscedastic.

$$\sigma_{u_i}^2 = \sigma_u^2 \quad \text{for all } i. \tag{3.23}$$

A.5 The values of the disturbance term have independent distributions.

$$u_i \text{ is distributed independently of } u_j \text{ for all } j \neq i. \tag{3.24}$$

A.6 The disturbance term has a normal distribution.

Unbiasedness

We saw that in the case of the simple regression model

$$b_2 = \beta_2 + \sum_{i=1}^{n} a_i u_i, \tag{3.25}$$

where

$$a_i = \frac{X_i - \bar{X}}{\sum_{j=1}^{n} (X_j - \bar{X})^2}. \tag{3.26}$$

Similar relationships obtain in the multiple regression case. The coefficient of X_j can be decomposed as

$$b_j = \beta_j + \sum_{i=1}^{n} a_{ij}^* u_i, \tag{3.27}$$

where the a_{ij}^* terms are functions of the data on the explanatory variables in the model. The difference is that the a_{ij}^* terms are more complex than the a_i terms in the simple regression model and the proof of the decomposition likewise more complex. When one makes the transition to matrix algebra, the results are very easily obtained. We shall take them on trust. Taking (3.27) as given, unbiasedness follows as a formality:

$$E(b_j) = \beta_j + E\left\{\sum_{i=1}^{n} a_{ij}^* u_i\right\} = \beta_j + \sum_{i=1}^{n} a_{ij}^* E(u_i) = \beta_j, \tag{3.28}$$

applying Assumption A.3.

It is important to note that the proof of unbiasedness does not require the explanatory variables to be uncorrelated. It is natural to suppose, for example, that if two variables are positively correlated, this will somehow undermine the estimation of their coefficients and give rise to bias. It does not. Even if the variables are highly correlated, the OLS estimators of their coefficients will remain unbiased. As we will see in a moment, correlations will affect the precision of the estimates, but that is another matter. They will not be responsible for a tendency to underestimation or overestimation. This is the reason that multiple regression

is such a powerful and popular tool. It enables us to estimate the effect of one variable, controlling for the effects of others, in the non-experimental conditions that characterize the work of most applied economists.

Efficiency

The Gauss–Markov theorem proves that, for multiple regression analysis, as for simple regression analysis, the ordinary least squares (OLS) technique yields the most efficient linear estimators of the parameters, in the sense that it is impossible to find other unbiased estimators with lower variances, using the same sample information, provided that the regression model assumptions are satisfied. We will not attempt to prove this theorem since matrix algebra is required.

Precision of the multiple regression coefficients

We will investigate the factors governing the likely precision of the regression coefficients for the case where there are two explanatory variables. Similar considerations apply in the more general case, but with more than two variables the analysis becomes complex and one needs to switch to matrix algebra.

If the true relationship is

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i, \tag{3.29}$$

and the fitted regression line is

$$\hat{Y}_i = b_1 + b_2 X_{2i} + b_3 X_{3i}, \tag{3.30}$$

the population variance of the probability distribution of b_2, $\sigma_{b_2}^2$, is given by

$$\sigma_{b_2}^2 = \frac{\sigma_u^2}{\sum_{i=1}^{n} (X_{2i} - \bar{X}_2)^2} \times \frac{1}{1 - r_{X_2X_3}^2}, \tag{3.31}$$

where σ_u^2 is the population variance of u and $r_{X_2X_3}$ is the correlation between X_2 and X_3. A parallel expression may be obtained for the population variance of b_3, replacing $\sum (X_{2i} - \bar{X}_2)^2$ with $\sum (X_{3i} - \bar{X}_3)^2$. Rewriting this as

$$\sigma_{b_2}^2 = \frac{\sigma_u^2}{n \, \mathrm{MSD}(X_2)} \times \frac{1}{1 - r_{X_2X_3}^2}, \tag{3.32}$$

where $\mathrm{MSD}(X_2)$, the mean square deviation of X_2, is given by

$$\mathrm{MSD}(X_2) = \frac{1}{n} \sum (X_{2i} - \bar{X}_2)^2, \tag{3.33}$$

we can see that, as in the case of simple regression analysis, it is desirable for n and $\mathrm{MSD}(X_2)$ to be large and for σ_u^2 to be small. However, we now have the

further term $\left(1-r_{X_2X_3}^2\right)$ and clearly it is desirable that the correlation between X_2 and X_3 should be low.

It is easy to give an intuitive explanation of this. The greater the correlation, the harder it is to discriminate between the effects of the explanatory variables on Y, and the less accurate will be the regression estimates. This can be a serious problem and it is discussed in the next section.

The standard deviation of the distribution of b_2 is the square root of the variance. As in the simple regression case, the standard error of b_2 is the estimate of the standard deviation. For this we need to estimate σ_u^2. The sample average of the squared residuals provides a biased estimator:

$$E\left\{\frac{1}{n}\sum_{i=1}^{n}e_i^2\right\}=\frac{n-k}{n}\sigma_u^2, \tag{3.34}$$

where k is the number of parameters in the regression equation. We will not attempt to prove this. In view of (3.34), we can obtain an unbiased estimator, s_u^2, by dividing *RSS* by $n-k$, instead of n, thus neutralizing the bias:

$$s_u^2=\frac{1}{n-k}\sum_{i=1}^{n}e_i^2. \tag{3.35}$$

The standard error is then given by

$$\text{s.e.}(b_2)=\sqrt{\frac{s_u^2}{\sum_{i=1}^{n}\left(X_{2i}-\bar{X}_{2i}\right)^2}\times\frac{1}{1-r_{X_2,X_3}^2}}. \tag{3.36}$$

The determinants of the standard error will be illustrated by comparing them in earnings functions fitted to two subsamples of the respondents in *EAEF* Data Set 21: those who reported that their wages were set by collective bargaining and the remainder. Regression output for the two subsamples is shown in Tables 3.3

Table 3.3

```
. reg EARNINGS S EXP if COLLBARG==1

      Source |       SS       df       MS              Number of obs =     101
-------------+------------------------------           F(2, 98)      =    9.72
       Model |  3076.31726     2  1538.15863            Prob > F      =  0.0001
    Residual |  15501.9762    98   158.18343            R-squared     =  0.1656
-------------+------------------------------           Adj R-squared =  0.1486
       Total |  18578.2934   100  185.782934            Root MSE      =  12.577

------------------------------------------------------------------------------
    EARNINGS |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           S |   2.333846   .5492604     4.25   0.000     1.243857    3.423836
         EXP |   .2235095   .3389455     0.66   0.511    -.4491169    .8961358
       _cons |  -15.12427   11.38141    -1.33   0.187    -37.71031    7.461779
------------------------------------------------------------------------------
```

Table 3.4

```
.reg EARNINGS S EXP if COLLBARG==0

      Source |       SS       df       MS              Number of obs =     439
-------------+------------------------------           F(2, 436)     =   57.77
       Model |  19540.1761     2  9770.08805            Prob > F      =  0.0000
    Residual |  73741.593    436  169.132094           R-squared     =  0.2095
-------------+------------------------------           Adj R-squared =  0.2058
       Total |  93281.7691    438  212.972076           Root MSE      =  13.005

------------------------------------------------------------------------------
    EARNINGS |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           S |   2.721698   .2604411    10.45   0.000     2.209822    3.233574
         EXP |   .6077342   .1400846     4.34   0.000     .3324091    .8830592
       _cons |  -28.00805   4.643211    -6.03   0.000    -37.13391   -18.88219
------------------------------------------------------------------------------
```

and 3.4. In Stata, subsamples may be selected by adding an 'if' expression to a command. *COLLBARG* is a variable in the data set defined to be 1 for the collective bargaining subsample and 0 for the others. Note that in tests for equality, Stata requires the = sign to be duplicated.

The standard error of the coefficient of *S* in the first regression is 0.5493, twice as large as that in the second, 0.2604. We will investigate the reasons for the difference. It will be convenient to rewrite (3.36) in such a way as to isolate the contributions of the various factors:

$$\text{s.e.}(b_2) = s_u \times \frac{1}{\sqrt{n}} \times \frac{1}{\sqrt{\text{MSD}(X_2)}} \times \frac{1}{\sqrt{1 - r_{X_2,X_3}^2}}. \tag{3.37}$$

The first element we need, s_u, can be obtained directly from the regression output in any serious regression application. In the Stata output, *RSS* is given in the top left quarter of the regression output, as part of the decomposition of the total sum of squares into the explained ('model') sum of squares and the residual sum of squares. The value of $n-k$ is given to the right of *RSS*, and the ratio $RSS/(n-k)$, the estimator of s_u^2, to the right of that. The square root, s_u, is listed as the Root MSE (root mean square error) in the top right quarter of the regression output, 12.577 for the collective bargaining subsample and 13.005 for the regression with the other respondents.

The number of observations, 101 in the first regression and 439 in the second, is also listed in the top right quarter of the regression output. The mean squared deviations of *S*, 6.2325 and 5.8666, were calculated as the squares of the standard deviations reported using the Stata 'sum' command, multiplied by $(n-1)/n$. The correlations between *S* and *EXP*, –0.4087 and –0.1784 respectively, were calculated using the Stata 'cor' command. The factors of the components of the standard error in equation (3.37) were then derived and are shown in the lower half of Table 3.5.

Table 3.5 Decomposition of the standard error of S

Component	s_u	n	MSD(S)	$r_{S,EXP}$	s.e.
Collective bargaining	12.577	101	6.2325	−0.4087	0.5493
Not collective bargaining	13.005	439	5.8666	−0.1784	0.2604
Factor					
Collective bargaining	12.577	0.0995	0.4006	1.0957	0.5493
Not collective bargaining	13.005	0.0477	0.4129	1.0163	0.2603

It can be seen that, in this example, the reason that the standard error of S in the collective bargaining subsample is relatively large is that the number of observations in that subsample is relatively small. The larger correlation coefficient for S and EXP reinforces the difference while the smaller s_u and the larger MSD(S) reduce it, but these are relatively minor factors.

t tests and confidence intervals

t tests on the regression coefficients are performed in the same way as for simple regression analysis. Note that when you are looking up the critical level of t at any given significance level, it will depend on the number of degrees of freedom, $n - k$: the number of observations minus the number of parameters estimated. The confidence intervals are also constructed in exactly the same way as in simple regression analysis, subject to the above comment about the number of degrees of freedom. As can be seen from the regression output, Stata automatically calculates confidence intervals for the coefficients (95 percent by default, other levels if desired), but this is not a standard feature of regression applications.

EXERCISES

3.8 Perform t tests on the coefficients of the variables in the educational attainment function reported in Exercise 3.1.

3.9 Perform t tests on the coefficients of the variables in the educational attainment and earnings functions fitted by you in Exercises 3.2 and 3.3.

3.10 The following earnings functions were fitted separately for males and females, using *EAEF* Data Set 21 (standard errors in parentheses):

males

$$\widehat{EARNINGS} = -31.5168 + 3.1408\ S + 0.6453\ EXP$$
$$(7.8708)\ (0.3693)\quad (0.2382)$$

females

$$\widehat{EARNINGS} = -17.2028 + 2.0772\,S + 0.3179\,EXP.$$
$$(4.5797)\ (0.2805)\quad (0.1388)$$

Using equation (3.37), explain why the standard errors of the coefficients of S and EXP are greater for the male subsample than for the female subsample, and why the difference in the standard errors is relatively large for EXP.

Further data:

	males	*females*
s_u	14.278	10.548
n	270	270
$r_{S,EXP}$	−0.4029	−0.0632
MSD(S)	6.6080	5.2573
MSD(EXP)	15.8858	21.4628

3.11* Demonstrate that $\bar{e} = 0$ in multiple regression analysis. (*Note*: The proof is a generalization of the proof for the simple regression model, given in Section 1.5.)

3.12 Investigate whether you can extend the determinants of weight model using your *EAEF* data set, taking *WEIGHT02* as the dependent variable, and *HEIGHT* and other continuous variables in the data set as explanatory variables. Provide an interpretation of the coefficients and perform t tests on them.

3.13 In Exercise 3.7, the sample means of H, P, and R are 100 hours, 95 hours, and 5 hours, respectively and the standard deviations of the distributions of H, P, and R are 10.1, 10.1, and 2.1, respectively. The standard errors of the coefficients of the regression of Researcher A are shown in parentheses under the coefficients.

$$\hat{M} = 45.6 + 0.15\,P + 0.21\,R.$$
$$(2.8)\ (0.03)\quad (0.14)$$

Perform t tests of the significance of the coefficients of P and R. The researcher says that the insignificant coefficient of R is to be expected because the students, on average, spent much less time on revision than on primary study. Explain whether this assertion is correct.

3.4 Multicollinearity

In the previous section, in the context of a model with two explanatory variables, it was seen that the higher is the correlation between the explanatory variables, the larger are the population variances of the distributions of their coefficients, and the greater is the risk of obtaining erratic estimates of the coefficients. If the correlation causes the regression model to become unsatisfactory in this respect, it is said to be suffering from multicollinearity.

A high correlation does not necessarily lead to poor estimates. If all the other factors determining the variances of the regression coefficients are helpful, that is, if the number of observations and the mean square deviations of the explanatory variables are large, and the variance of the disturbance term small, you may well obtain good estimates after all. Multicollinearity therefore must be caused by a *combination* of a high correlation and one or more of the other factors being unhelpful. And it is a matter of *degree*, not kind. Any regression will suffer from it to some extent, unless all the explanatory variables are uncorrelated. You only start to talk about it when you think that it is affecting the regression results seriously.

It is an especially common problem in time series regressions, that is, where the data consist of a series of observations on the variables over a number of time periods. If two or more of the explanatory variables have strong time trends, they will be highly correlated and this condition may give rise to multicollinearity.

It should be noted that the presence of multicollinearity does not mean that the model is misspecified. Accordingly, the regression coefficients remain unbiased and the standard errors remain valid. The standard errors will be larger than they would have been in the absence of multicollinearity, warning you that the regression estimates are unreliable.

We will consider first the case of exact multicollinearity where the explanatory variables are perfectly correlated. Suppose that the true relationship is

$$Y = 2 + 3X_2 + X_3 + u. \tag{3.38}$$

Suppose that there is a linear relationship between X_2 and X_3:

$$X_3 = 2X_2 - 1, \tag{3.39}$$

and suppose that X_2 increases by one unit in each observation. X_3 will increase by 2 units, and Y by approximately 5 units, for example as shown in Table 3.6 and illustrated in Figure 3.3 (where the effect of the disturbance term has been neglected).

Table 3.6

Observation	X_2	X_3	Y	*Change in* X_2	*Change in* X_3	*Approximate change in* Y
1	10	19	$51+u_1$	1	2	5
2	11	21	$56+u_2$	1	2	5
3	12	23	$61+u_3$	1	2	5
4	13	25	$66+u_4$	1	2	5
5	14	27	$71+u_5$	1	2	5
6	15	29	$76+u_6$	1	2	5

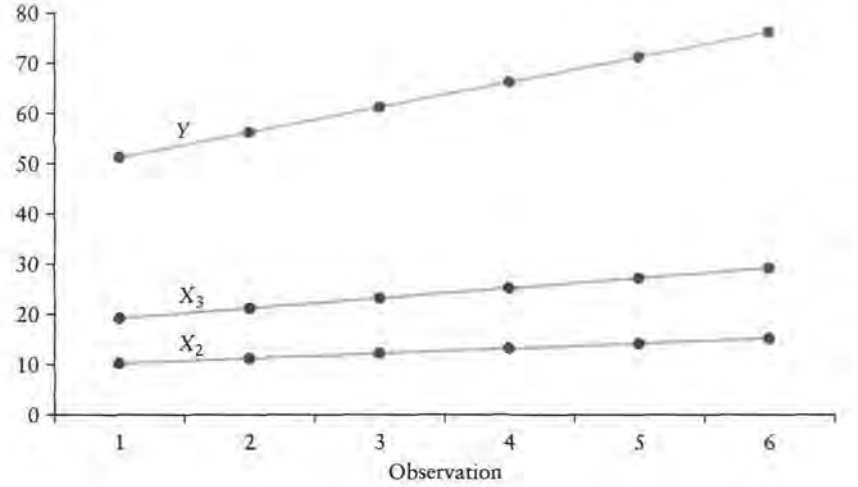

Figure 3.3 Example of exact multicollinearity

Looking at the data, you could come to any of the following conclusions:

1. Y is determined by (3.38).
2. X_3 is irrelevant and Y is determined by the relationship

$$Y = 1 + 5X_2 + u. \tag{3.40}$$

3. X_2 is irrelevant and Y is determined by the relationship

$$Y = 3.5 + 2.5X_3 + u. \tag{3.41}$$

In fact these are not the only possibilities. Any relationship that is a weighted average of (3.40) and (3.41) would also fit the data. For example, (3.38) may be regarded as such a weighted average, being (3.40) multiplied by 0.6 plus (3.41) multiplied by 0.4.

In such a situation, it is impossible for regression analysis, or any other technique for that matter, to distinguish between these possibilities. You would not even be able to calculate the regression coefficients because both the numerator and the denominator of the regression coefficients would collapse to zero. This will be demonstrated with the general two-variable case. Suppose

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u \tag{3.42}$$

and

$$X_3 = \lambda + \mu X_2. \tag{3.43}$$

First note that, given (3.43),

$$\left(X_{3i} - \bar{X}_3\right) = \left([\lambda + \mu X_{2i}] - [\lambda + \mu \bar{X}_2]\right) = \mu\left(X_{2i} - \bar{X}_2\right). \tag{3.44}$$

Hence,

$$\sum_{i=1}^{n}(X_{3i}-\bar{X}_3)^2=\mu^2\sum_{i=1}^{n}(X_{2i}-\bar{X}_2)^2 \tag{3.45}$$

$$\sum_{i=1}^{n}(X_{3i}-\bar{X}_3)(Y_i-\bar{Y})=\mu\sum_{i=1}^{n}(X_{2i}-\bar{X}_2)(Y_i-\bar{Y}) \tag{3.46}$$

$$\sum_{i=1}^{n}(X_{2i}-\bar{X}_2)(X_{3i}-\bar{X}_3)=\mu\sum_{i=1}^{n}(X_{2i}-\bar{X}_2)^2. \tag{3.47}$$

Substituting for X_3 in (3.11), one obtains

$$\begin{aligned}b_2&=\frac{\sum_{i=1}^{n}(X_{2i}-\bar{X}_2)(Y_i-\bar{Y})\sum_{i=1}^{n}(X_{3i}-\bar{X}_3)^2-\sum_{i=1}^{n}(X_{3i}-\bar{X}_3)(Y_i-\bar{Y})\sum_{i=1}^{n}(X_{2i}-\bar{X}_2)(X_{3i}-\bar{X}_3)}{\sum_{i=1}^{n}(X_{2i}-\bar{X}_2)^2\sum_{i=1}^{n}(X_{3i}-\bar{X}_3)^2-\left(\sum_{i=1}^{n}(X_{2i}-\bar{X}_2)(X_{3i}-\bar{X}_3)\right)^2}\\&=\frac{\sum_{i=1}^{n}(X_{2i}-\bar{X}_2)(Y_i-\bar{Y})\left(\mu^2\sum_{i=1}^{n}(X_{2i}-\bar{X}_2)^2\right)-\left(\mu\sum_{i=1}^{n}(X_{2i}-\bar{X}_2)(Y_i-\bar{Y})\right)\left(\mu\sum_{i=1}^{n}(X_{2i}-\bar{X}_2)^2\right)}{\sum_{i=1}^{n}(X_{2i}-\bar{X}_2)^2\left(\mu^2\sum_{i=1}^{n}(X_{2i}-\bar{X}_2)^2\right)-\left(\mu\sum_{i=1}^{n}(X_{2i}-\bar{X}_2)^2\right)^2}\\&=\frac{0}{0}.\end{aligned} \tag{3.48}$$

It is unusual for there to be an exact relationship between the explanatory variables in a regression. When this occurs, it is typically because there is a logical error in the specification. An example is provided by Exercise 3.16. However, it often happens that there is an approximate relationship, and typically this takes the form of high correlations among two or more explanatory variables.

Multicollinearity in models with more than two explanatory variables

The discussion of multicollinearity so far has been restricted to the case where there are two explanatory variables. In models with a greater number of explanatory variables, multicollinearity may be caused by an approximate linear relationship among them. It may be difficult to discriminate between the effects of one variable and those of a linear combination of the remainder. In the model with two explanatory variables, an approximate linear relationship automatically means a high correlation, but when there are three or more, this is not necessarily the case because a linear relationship does not inevitably imply high pairwise correlations between any of the variables. The effects of multicollinearity are the same as in the case with two explanatory variables. As in that case, the problem may not be serious if the population variance of the disturbance term is small, the number of observations large, and the mean square deviations of the explanatory variables large.

Example of multicollinearity

Suppose that one is investigating the determinants of educational attainment. A basic specification, proposed in Exercise 3.1, is

$$S = \beta_1 + \beta_2 SM + \beta_3 SF + \beta_4 ASVABC + u, \tag{3.49}$$

where S is highest grade completed by the respondent, SM and SF are highest grades completed by the respondent's mother and father, and $ASVABC$ is a measure of cognitive ability. $ASVABC$ is a composite of the scores on three tests: $ASVAB02$, arithmetical reasoning; $ASVAB03$, word knowledge; and $ASVAB04$, paragraph comprehension. If we include these separately instead of the composite, the regression results using Data Set 21 are as shown in Table 3.7.

All the regression coefficients are positive, as expected, but those for SM and $ASVAB04$ are not significant, even at the 5 percent level, and that for $ASVAB03$ is significant only at that level. Since maternal education may reasonably be assumed to be a powerful determinant of educational attainment, the lack of significance of the coefficient of SM is surprising. Multicollinearity is the most likely explanation, given that, on account of assortative mating, the correlation between SM and SF is 0.62. The insignificant coefficient of $ASVAB04$ might simply indicate that that variable was not a determinant of educational attainment, but here also multicollinearity must be suspected, for the correlation between $ASVAB03$ and $ASVAB04$, 0.80, is even higher.

What can you do about multicollinearity?

The various ways of trying to alleviate multicollinearity fall into two categories: direct attempts to improve the conditions responsible for the variances of the regression coefficients, and indirect methods.

Table 3.7

```
. reg S SM SF ASVAB02 ASVAB03 ASVAB04
```

Source	SS	df	MS	Number of obs	= 540
				F(5, 534)	= 62.50
Model	1183.14727	5	236.629455	Prob > F	= 0.0000
Residual	2021.83606	534	3.78620985	R-squared	= 0.3692
				Adj R-squared	= 0.3633
Total	3204.98333	539	5.94616574	Root MSE	= 1.9458

S	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
SM	.04978	.0392495	1.27	0.205	-.0273223	.1268823
SF	.1072127	.0310202	3.46	0.001	.0462762	.1681493
ASVAB02	.0749955	.0127422	5.89	0.000	.0499645	.1000265
ASVAB03	.0354954	.016693	2.13	0.034	.0027034	.0682874
ASVAB04	.0259897	.0149738	1.74	0.083	-.0034251	.0554045
_cons	4.968064	.5140692	9.66	0.000	3.958218	5.97791

Direct methods

The four factors responsible for the variances of the coefficients are the variance of the disturbance term, the number of observations in the sample, the mean square deviations of the explanatory variables, and the correlations among the explanatory variables.

We will start with the number of observations. Table 3.8 shows the result of running the regression with all 2,714 observations in the *EAEF* data set. Comparing this result with that using Data Set 21, we see that the standard errors are much smaller, as expected. As a consequence, the *t* statistics of all of the variables are large and the coefficients significantly different from zero at the 0.1 percent level. The problem of multicollinearity has disappeared.

This example is artificial because the output in Table 3.7 used a subset of the data base. In practice, there would never be any reason to use a subset, unless the entire data base were so enormous, as in the case of a national census of population, that the costs of processing it fully might not be justified.

If you are working with cross-sectional data (individuals, households, enterprises, etc.) and you are undertaking a survey, you could increase the size of the sample by negotiating a bigger budget. Alternatively, you could make a fixed budget go further by using a technique known as clustering. You divide the country geographically into localities. For example, the National Longitudinal Survey of Youth, from which the *EAEF* data are drawn, divides the country into counties, independent cities, and standard metropolitan statistical areas. You select a number of localities randomly, perhaps using stratified random sampling to make sure that metropolitan, other urban, and rural areas are properly represented. You then confine the survey to the localities selected. This reduces the travel time of the fieldworkers, allowing them to interview a greater number of respondents.

Table 3.8

```
. reg S SM SF ASVAB02 ASVAB03 ASVAB04

      Source |       SS       df       MS              Number of obs =    2714
-------------+------------------------------           F(5, 2708)    =  330.75
       Model |  6294.19353     5  1258.83871            Prob > F      =  0.0000
    Residual |   10306.567  2708  3.80597008            R-squared     =  0.3792
-------------+------------------------------           Adj R-squared =  0.3780
       Total |  16600.7605  2713  6.11896812            Root MSE      =  1.9509

------------------------------------------------------------------------------
           S |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          SM |   .1275844   .0186513     6.84   0.000     .0910122    .1641566
          SF |   .1187439   .0142069     8.36   0.000     .0908864    .1466014
    ASVAB02  |   .0671771   .0056405    11.91   0.000      .056117    .0782372
    ASVAB03  |   .0283495    .006953     4.08   0.000     .0147157    .0419833
    ASVAB04  |   .0292889   .0064412     4.55   0.000     .0166588    .0419191
       _cons |   4.549561   .2350241    19.36   0.000     4.088716    5.010406
------------------------------------------------------------------------------
```

If you are working with time series data, you may be able to increase the sample by working with shorter time intervals for the data, for example quarterly or even monthly data instead of annual data. This is such an obvious thing to do that most researchers working with time series almost automatically use quarterly data, if they are available, instead of annual data, even if there does not appear to be a problem of multicollinearity, simply to minimize the population variances of the regression coefficients. There are, however, potential problems. You may introduce, or aggravate, autocorrelation (see Chapter 12), but this can be neutralized. Also you may introduce, or aggravate, measurement error bias (see Chapter 8) if the quarterly data are less accurately measured than the corresponding annual data. This problem is not so easily overcome, but it may be a minor one.

The next direct method that we will consider is the reduction of σ_u^2. The disturbance term is the joint effect of all the variables influencing Y that have not been included explicitly in the regression equation. If you can think of an important variable that you have omitted, and is therefore contributing to u, you will reduce the population variance of the disturbance term if you add it to the regression equation, and hence the variances of the regression coefficients.

In the case of the educational attainment function, we might consider adding *AGE*, the age of the respondent, to the specification. Educational attainment in the population is gradually rising, each new generation tending to be better educated than the last, and so one would anticipate a negative relationship between age and attainment.

Table 3.9 shows the regression output with *AGE* included. Looking at the t statistic, we see that the coefficient of *AGE* is significant at the 5 percent level (p value 0.014), so its inclusion appears to have improved the specification. Comparing Tables 3.7 and 3.9, *RSS* has fallen from 2021.8 to 1999.1.

Table 3.9

```
. reg S SM SF ASVAB02 ASVAB03 ASVAB04 AGE

      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(6, 533)     =   53.58
       Model |  1205.87487     6  200.979145            Prob > F      =  0.0000
    Residual |  1999.10846   533  3.75067254            R-squared     =  0.3762
-------------+------------------------------           Adj R-squared =  0.3692
       Total |  3204.98333   539  5.94616574            Root MSE      =  1.9367

------------------------------------------------------------------------------
           S |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          SM |   .0426916   .0391708     1.09   0.276    -.0342566    .1196397
          SF |   .1039566   .0309026     3.36   0.001     .0432509    .1646624
     ASVAB02 |   .0733302   .0127003     5.77   0.000     .0483814     .098279
     ASVAB03 |   .0415514   .0167956     2.47   0.014     .0085577    .0745452
     ASVAB04 |   .0279754   .0149252     1.87   0.061    -.001344     .0572948
         AGE |  -.0916652   .0372376    -2.46   0.014    -.1648157   -.0185146
       _cons |   8.522519   1.531916     5.56   0.000     5.513186    11.53185
------------------------------------------------------------------------------
```

As a consequence, s_u^2, the estimator of σ_u^2, has fallen from 3.79 to 3.75 (see the calculation of the residual sum of squares divided by the number of degrees of freedom in the top left quarter of the regression output). Although this is a step in the right direction, it is very small. As a consequence, the reduction of the standard errors of the coefficients of *SM* and *ASVAB04* is very small and those coefficients remain statistically insignificant. This outcome is actually fairly typical. You will probably already have included all of the major variables in your original specification, so those remaining in your data set are likely to be marginal. The approach can even have an effect opposite to that intended. The standard errors of the existing variables in the specification may actually increase if the new variables are correlated with them.

A third possible way of reducing the problem of multicollinearity might be to increase the mean square deviation of the explanatory variables. This is possible only at the design stage of a survey. For example, if you were planning a household survey with the aim of investigating how expenditure patterns vary with income, you should make sure that the sample included relatively rich and relatively poor households as well as middle-income households by stratifying the sample. (For a discussion of sampling theory and techniques, see, for example, Moser and Kalton, 1985, or Fowler, 2009.)

The fourth direct method is the most direct of all. If you are still at the design stage of a survey, you should do your best to obtain a sample where the explanatory variables are less related (more easily said than done, of course).

Indirect methods

Next, we will consider indirect methods. If the correlated variables are similar conceptually, it may be reasonable to combine them into some overall index. That is precisely what has been done with the three cognitive variables in the Armed Services Vocational Aptitude Battery. *ASVABC* has been calculated as a weighted average of *ASVAB02* (arithmetic reasoning), *ASVAB03* (word knowledge), and *ASVAB04* (paragraph comprehension). The three components are highly correlated and by combining them rather than using them individually, one avoids a potential problem of multicollinearity.

However, the *ASVABC* composite is not helpful if one is concerned with determining whether verbal skills or numerical skills are more important for educational attainment. For this purpose, we might leave *ASVAB02* as the indicator of numerical skills and combine *ASVAB03* and *ASVAB04* into an index of verbal skills. In the output shown in Table 3.10, *VERBAL* is defined as the sum of *ASVAB03* and *ASVAB04*. The standard error of its coefficient, 0.00697, is less than half of those of *ASVAB03* and *ASVAB04* in Table 3.7, indicating a large improvement in precision. If one is really interested in the effects of verbal skills, the difference between word knowledge and paragraph comprehension may be unimportant, and using a composite is helpful.

Another possible solution to the problem of multicollinearity is to drop some of the correlated variables if they have insignificant coefficients. Table 3.11 shows

Table 3.10

```
. gen VERBAL = ASVAB03 + ASVAB04
. reg S SM SF ASVAB02 VERBAL
```

Source	SS	df	MS	Number of obs =	540
				F(4, 535) =	78.22
Model	1182.72356	4	295.68089	Prob > F =	0.0000
Residual	2022.25977	535	3.77992481	R-squared =	0.3690
				Adj R-squared =	0.3643
Total	3204.98333	539	5.94616574	Root MSE =	1.9442

S	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
SM	.0504512	.0391656	1.29	0.198	-.026486	.1273884
SF	.1076162	.030971	3.37	0.001	.0467766	.1684558
ASVAB02	.0754085	.0126717	5.95	0.000	.050516	.1003009
VERBAL	.0304221	.00697	4.36	0.000	.0167301	.0441141
_cons	4.963665	.5134743	9.67	0.000	3.954992	5.972338

Table 3.11

```
. reg S SM SF ASVAB02 ASVAB03
```

Source	SS	df	MS	Number of obs =	540
				F(4, 535) =	77.08
Model	1171.74101	4	292.935253	Prob > F =	0.0000
Residual	2033.24232	535	3.80045293	R-squared =	0.3656
				Adj R-squared =	0.3609
Total	3204.98333	539	5.94616574	Root MSE =	1.9495

S	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
SM	.0510298	.0393166	1.30	0.195	-.0262041	.1282636
SF	.1072792	.0310784	3.45	0.001	.0462284	.1683299
ASVAB02	.07999	.0124364	6.43	0.000	.0555599	.1044201
ASVAB03	.0531464	.0132626	4.01	0.000	.0270933	.0791996
_cons	5.132012	.5062662	10.14	0.000	4.137499	6.126525

the output when *ASVAB04* has been dropped. The coefficient of *ASVAB03* is now highly significant. However, the increase in the *t* statistic provides a somewhat misleading indication of the effectiveness of this measure. The increase in precision in the estimation of the coefficient is not dramatic. The standard error does fall, from 0.0167 to 0.0133. But most of the increase in the *t* statistic is attributable to an increase in the size of the estimated coefficient, from 0.0331 to 0.0531. There is always a risk that a variable with an insignificant coefficient does truly belong in the model and that dropping it may distort the results by giving rise to omitted variable bias. We will discuss this in Chapter 6.

A further way of dealing with the problem of multicollinearity is to use extraneous information, if available, concerning the coefficient of one of the variables. For example, suppose that one is relating aggregate demand for a

category of consumer expenditure, Y, to aggregate disposable personal income, X, and a price index for the category, P.

$$Y = \beta_1 + \beta_2 X + \beta_3 P + u. \tag{3.50}$$

To fit a model of this type you would use time series data. If X and P possess strong time trends and are therefore highly correlated, which is often the case with time series variables, multicollinearity is likely to be a problem. Suppose, however, that you also have cross-sectional data on Y and X derived from a separate household survey. These variables will be denoted Y' and X' to indicate that the data are household data, not aggregate data. Assuming that all the households in the survey were paying roughly the same price for the commodity, one would fit the simple regression

$$\hat{Y}' = b_1' + b_2' X'. \tag{3.51}$$

Now substitute b_2' for β_2 in the time series model,

$$Y = \beta_1 + b_2' X + \beta_3 P + u, \tag{3.52}$$

subtract $b_2' X$ from both sides,

$$Y - b_2' X = \beta_1 + \beta_3 P + u, \tag{3.53}$$

and regress $Z = Y - b_2' X$ on price. This is a simple regression, so multicollinearity has been eliminated.

There are, however, two possible problems with this technique. First, the estimate of β_3 in (3.53) depends on the accuracy of the estimate of b_2', and this of course is subject to sampling error. Second, you are assuming that the income coefficient has the same meaning in time series and cross-sectional contexts, and this may not be the case. For many commodities, the short-run and long-run effects of changes in income may differ because expenditure patterns are subject to inertia. A change in income can affect expenditure both directly, by altering the budget constraint, and indirectly, through causing a change in lifestyle, and the indirect effect is much slower than the direct one. As a first approximation, it is commonly argued that time series regressions, particularly those using short sample periods, estimate short-run effects, while cross-sectional regressions estimate long-run ones. For a discussion of this and related issues, see Kuh and Meyer (1957).

Last, but by no means least among the indirect methods of alleviating multicollinearity, is the use of a theoretical restriction, which is defined as a hypothetical relationship among the parameters of a regression model. Returning to the educational attainment model,

$$S = \beta_1 + \beta_2 SM + \beta_3 SF + \beta_4 ASVAB02 + \beta_5 ASVAB03 + \beta_6 ASVAB04 + u \tag{3.54}$$

suppose that we hypothesize that word knowledge, *ASVAB03*, and paragraph comprehension, *ASVAB04*, are equally important. We can then impose the restriction $\beta_5 = \beta_6$. This allows us to write the equation as

$$\begin{aligned} S &= \beta_1 + \beta_2 SM + \beta_3 SF + \beta_4 ASVAB02 + \beta_5 (ASVAB03 + ASVAB04) + \\ &= \beta_1 + \beta_2 SM + \beta_3 SF + \beta_4 ASVAB02 + \beta_5 VERBAL + u. \end{aligned} \tag{3.55}$$

Thus, we have returned to the first of the indirect methods that we have considered, the use of a composite. Since the specification is the same, the consequences are the same. The only difference is in the justification for the procedure. Earlier, we decided to use the composite *VERBAL* instead of *ASVAB03* and *ASVAB04* individually as a pragmatic measure. Here we are doing so as a consequence of the use of a restriction.

In the same way, we might hypothesize that mother's and father's education are equally important for educational attainment. We can then impose another restriction, $\beta_2 = \beta_3$, and write the model as

$$\begin{aligned} S &= \beta_1 + \beta_2 (SM + SF) + \beta_4 ASVAB02 + \beta_5 VERBAL + \\ &= \beta_1 + \beta_2 SP + \beta_4 ASVAB02 + \beta_5 VERBAL + u, \end{aligned} \tag{3.56}$$

where $SP = SM + SF$ is total parental education. Table 3.12 gives the corresponding regression output.

The estimate of β_2 is now 0.083. Not surprisingly, this is a compromise between the coefficients of *SM* and *SF* in the previous specification. The standard error of *SP* is much smaller than those of *SM* and *SF*, indicating that, in this case also, the use of the restriction has led to a gain in efficiency.

Two final notes are in order. First, in this example, the two restrictions were of the same type: two parameters were hypothesized to be equal to one another. However, this was a coincidence. We will encounter other types of restriction in

Table 3.12

```
.gen SP = SM + SF
.reg S SP ASVAB02 VERBAL
```

Source	SS	df	MS	Number of obs =	540
				F(3, 536) =	104.04
Model	1179.48715	3	393.162385	Prob > F =	0.0000
Residual	2025.49618	536	3.77891078	R-squared =	0.3680
				Adj R-squared =	0.3645
Total	3204.98333	539	5.94616574	Root MSE =	1.9439

S	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
SP	.0833379	.0164531	5.07	0.000	.0510174	.1156585
ASVAB02	.0755626	.0126689	5.96	0.000	.0506757	.1004495
VERBAL	.0301108	.006961	4.33	0.000	.0164367	.0437849
_cons	4.8936	.5077924	9.64	0.000	3.896093	5.891108

the remainder of this text. Second, restrictions are hypothetical and we cannot assume that they are valid. We need to subject them to formal tests. We shall see how to do this in Chapter 6.

EXERCISES

3.14 Using your *EAEF* data set, regress *S* on *SM*, *SF*, *ASVAB02*, *ASVAB03*, and *ASVAB04*, the three components of the *ASVABC* composite score. Compare the coefficients and their standard errors with those of *ASVABC* in a regression of *S* on *SM*, *SF*, and *ASVABC*. Calculate correlation coefficients for the three *ASVAB* components.

3.15 In Exercise 1.9, the number of children in the respondent's family was regressed on mother's years of schooling. Fit an extended version of the model adding father's years of schooling using your *EAEF* data set. Define *CHILDREN* = *SIBLINGS* + 1 and regress *CHILDREN* on *SM* and *SF*. *SM* and *SF* are likely to be highly correlated (find the correlation in your data set) and the regression may be subject to multicollinearity. Introduce the restriction that the theoretical coefficients of *SM* and *SF* are equal and run the regression a second time, replacing *SM* and *SF* by their sum, *SP*. Evaluate the regression results.

3.16* A researcher investigating the determinants of the demand for public transport in a certain city has the following data for 100 residents for the previous calendar year: expenditure on public transport, *E*, measured in dollars; number of days worked, *W*; and number of days not worked, *NW*. By definition *NW* is equal to 365 – *W*. He attempts to fit the following model:

$$E = \beta_1 + \beta_2 W + \beta_3 NW + u.$$

Explain why he is unable to fit this equation. (Give both intuitive and technical explanations.) How might he resolve the problem?

3.17 Work experience is generally found to be an important determinant of earnings. If a direct measure is lacking in a data set, it is standard practice to use potential work experience, *PWE*, defined by

$$PWE = AGE - S - 5$$

as a proxy. This is the maximum number of years since the completion of full-time education, assuming that an individual enters first grade at the age of six. Using your *EAEF* data set, first regress *EARNINGS* on *S* and *PWE*, and then run the regression a second time adding *AGE* as well. Comment on the regression results.

3.5 Goodness of fit: R^2

As in simple regression analysis, the coefficient of determination, R^2, measures the proportion of the variation in Y explained by the regression and is defined equivalently by

$$R^2 = \frac{\sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2}{\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2}, \tag{3.57}$$

by

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}, \tag{3.58}$$

or by the square of the correlation coefficient for Y and $\hat{Y}$. It can never decrease, and generally will increase, if you add another variable to a regression equation, provided that you retain all the previous explanatory variables. To see this, suppose that you regress Y on X_2 and X_3 and fit the equation

$$\hat{Y}_i = b_1 + b_2 X_{2i} + b_3 X_{3i}. \tag{3.59}$$

Next suppose that you regress Y on X_2 only and the result is

$$\hat{Y}_i = b_1^* + b_2^* X_{2i}. \tag{3.60}$$

This can be rewritten

$$\hat{Y}_i = b_1^* + b_2^* X_{2i} + 0 X_{3i}. \tag{3.61}$$

Comparing (3.59) and (3.61), the coefficients in (3.59) have been determined freely by the OLS technique using the data for Y, X_2, and X_3 to give the best possible fit. In (3.61), however, the coefficient of X_3 has arbitrarily been set at zero, and the fit will be suboptimal unless, by coincidence, b_3 happens to be zero, in which case the fit will be the same. (b_1^* will then be equal to b_1, and b_2^* will be equal to b_2). Hence, in general, the level of R^2 will be higher in (3.59) than in (3.61), and it cannot be lower. Of course, if the new variable does not genuinely belong in the equation, the increase in R^2 is likely to be negligible.

You might think that, because R^2 measures the proportion of the variation jointly explained by the explanatory variables, it should be possible to deduce the individual contribution of each explanatory variable and thus obtain a measure of its relative importance. At least it would be very convenient if one could. Unfortunately, such a decomposition is impossible if the explanatory variables are correlated because their explanatory power will overlap. The problem will be discussed further in Section 6.2.

F tests

We saw in Section 2.7 that we could perform an F test of the explanatory power of the simple regression model

$$Y_i = \beta_1 + \beta_2 X_i + u_i, \tag{3.62}$$

the null hypothesis being H_0: $\beta_2 = 0$ and the alternative being H_1: $\beta_2 \neq 0$. The null hypothesis was the same as that for a t test on the slope coefficient and it turned out that the F test was equivalent to a (two-sided) t test. However, in the case of

the multiple regression model, the tests have different roles. The t tests test the significance of the coefficient of each variable individually, while the F test tests their joint explanatory power. The null hypothesis, which we hope to reject, is that the model has no explanatory power. The model will have no explanatory power if it turns out that Y is unrelated to any of the explanatory variables. Mathematically, therefore, if the model is

$$Y_i = \beta_1 + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i, \tag{3.63}$$

the null hypothesis for the F test is that all the slope coefficients $\beta_2, \ldots, \beta_k$ are zero:

$$H_0: \beta_2 = \ldots = \beta_k = 0. \tag{3.64}$$

The alternative hypothesis H_1 is that at least one of the slope coefficients $\beta_2, \ldots, \beta_k$ is different from zero. The F statistic is defined as

$$F(k-1, n-k) = \frac{ESS/(k-1)}{RSS/(n-k)} \tag{3.65}$$

and the test is performed by comparing this with the critical level of F in the column corresponding to $k - 1$ degrees of freedom and the row corresponding to $n - k$ degrees of freedom in the appropriate part of Table A.3 in Appendix A.

This F statistic may also be expressed in terms of R^2 by dividing both the numerator and denominator of (3.65) by TSS, the total sum of squares, and noting that ESS/TSS is R^2 and RSS/TSS is $(1 - R^2)$:

$$F(k-1, n-k) = \frac{R^2/(k-1)}{(1-R^2)/(n-k)}. \tag{3.66}$$

Example

The educational attainment model will be used as an illustration. For simplicity, we will use the composite cognitive ability measure *ASVABC* instead of any of its components discussed in the previous section. We will suppose that S depends on *ASVABC*, *SM*, and *SF*:

$$S = \beta_1 + \beta_2 SM + \beta_3 SF + \beta_4 ASVABC + u. \tag{3.67}$$

The null hypothesis for the F test of goodness of fit is that all three slope coefficients are equal to zero:

$$H_0: \beta_2 = \beta_3 = \beta_4 = 0. \tag{3.68}$$

The alternative hypothesis is that at least one of them is nonzero. The regression output using *EAEF* Data Set 21 is shown in Table 3.13.

In this example, $k - 1$, the number of explanatory variables, is equal to 3 and $n - k$, the number of degrees of freedom, is equal to 536. The numerator of the F statistic is the explained sum of squares divided by $k - 1$. In the Stata output,

Table 3.13

```
. reg S SM SF ASVABC

      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(3, 536)     =  104.30
       Model |  1181.36981     3  393.789935            Prob > F      =  0.0000
    Residual |  2023.61353   536  3.77539837            R-squared     =  0.3686
-------------+------------------------------           Adj R-squared =  0.3651
       Total |  3204.98333   539  5.94616574            Root MSE      =   1.943

------------------------------------------------------------------------------
           S |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          SM |   .0492424   .0390901     1.26   0.208    -.027546    .1260309
          SF |   .1076825   .0309522     3.48   0.001     .04688     .1684851
      ASVABC |   .1257087   .0098533    12.76   0.000     .1063528    .1450646
       _cons |   5.370631   .4882155    11.00   0.000     4.41158     6.329681
------------------------------------------------------------------------------
```

these numbers, 1181.4 and 3, respectively, are given in the Model row. The denominator is the residual sum of squares divided by the number of degrees of freedom remaining, 2023.6 and 536, respectively. Hence the F statistic is

$$F(3,536) = \frac{1181.4/3}{2023.6/536} = 104.3 \tag{3.69}$$

as in the printed output. All serious regression applications compute this F statistic for you as part of the diagnostics in the regression output.

The critical value for $F(3,536)$ is not given in the F tables, but we know it must be lower than $F(3,500)$, which is given. At the 0.1 percent level, this is 5.51. Hence we reject H_0 at that significance level. This result could have been anticipated because both *ASVABC* and *SF* have highly significant t statistics. So we knew in advance that both β_2 and β_3 were nonzero.

In practice, the F statistic will almost always be significant if any t statistic is. In principle, however, it might not be. Suppose that you ran a nonsense regression with 40 explanatory variables, none being a true determinant of the dependent variable. Then the F statistic should be low enough for H_0 not to be rejected. However, if you are performing t tests on the slope coefficients at the 5 percent level, with a 5 percent chance of a Type I error, on average 2 of the 40 variables could be expected to have 'significant' coefficients.

On the other hand, it can easily happen that the F statistic is significant while the t statistics are not. Suppose you have a multiple regression model that is correctly specified and R^2 is high. You would be likely to have a highly significant F statistic. However, if the explanatory variables are highly correlated and the model is subject to severe multicollinearity, the standard errors of the slope coefficients could all be so large that none of the t statistics is significant. In this situation, you would know that your model has high explanatory power, but you are not in a position to pinpoint the contributions made by the explanatory variables individually.

Further analysis of variance

Besides testing the equation as a whole, you can use an F test to see whether or not the joint marginal contribution of a group of variables is significant. Suppose that you first fit the model

$$Y = \beta_1 + \beta_2 X_2 + \cdots + \beta_k X_k + u, \tag{3.70}$$

with explained sum of squares ESS_k. Next you add $m - k$ variables and fit the model

$$Y = \beta_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \beta_{k+1} X_{k+1} + \cdots + \beta_m X_m + u, \tag{3.71}$$

with explained sum of squares ESS_m. You have then explained an additional sum of squares equal to $ESS_m - ESS_k$ using up an additional $m - k$ degrees of freedom, and you want to see whether the increase is greater than is likely to have arisen by chance.

Again an F test is used and the appropriate F statistic may be expressed in verbal terms as

$$F = \frac{\text{improvement in fit/extra degrees of freedom used up}}{\text{residual sum of squares remaining/degrees of freedom remaining}}. \tag{3.72}$$

Since RSS_m, the unexplained sum of squares in the second model, is equal to $TSS - ESS_m$, and RSS_k, the residual sum of squares in the first model, is equal to $TSS - ESS_k$, the improvement in the fit when the extra variables are added, $ESS_m - ESS_k$, is equal to $RSS_k - RSS_m$. Hence, the appropriate F statistic is

$$F(m-k, n-m) = \frac{(RSS_k - RSS_m)/(m-k)}{RSS_m/(n-m)}. \tag{3.73}$$

Under the null hypothesis that the additional variables contribute nothing to the equation,

$$H_0: \beta_{k+1} = \beta_{k+2} = \ldots = \beta_m = 0, \tag{3.74}$$

this F statistic is distributed with $m - k$ and $n - m$ degrees of freedom. The upper half of Table 3.14 gives the analysis of variance for the explanatory power of the original $k - 1$ variables. The lower half gives it for the joint marginal contribution of the new variables.

Example

We will illustrate the test with the educational attainment example. Table 3.15 shows the output from a regression of *S* on *ASVABC* using *EAEF* Data Set 21. We make a note of the residual sum of squares, 2123.0.

Now we add a group of two variables, the years of schooling of each parent, with the output shown in Table 3.16. Do the two new variables jointly make a significant contribution to the explanatory power of the model? Well, we can see

Table 3.14 Analysis of variance, original variables and a group of additional variables

	Sum of squares	Degrees of freedom	Sum of squares divided by degrees of freedom	F statistic
Explained by original variables	ESS_k	$k-1$	$ESS_k/(k-1)$	$\frac{ESS_k/(k-1)}{RSS_k/(n-k)}$
Residual	$RSS_k = TSS - ESS_k$	$n-k$	$RSS_k/(n-k)$	
Explained by new variables	$ESS_m - ESS_k = RSS_k - RSS_m$	$m-k$	$(RSS_k - RSS_m)/(m-k)$	$\frac{(RSS_k - RSS_m)/(m-k)}{RSS_m/(n-m)}$
Residual	$RSS_m = TSS - ESS_m$	$n-m$	$RSS_m/(n-m)$	

Table 3.15

```
. reg S ASVABC

      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(1, 538)     =  274.19
       Model |  1081.97059     1  1081.97059            Prob > F      =  0.0000
    Residual |  2123.01275   538  3.94612035           R-squared     =  0.3376
-------------+------------------------------           Adj R-squared =  0.3364
       Total |  3204.98333   539  5.94616574           Root MSE      =  1.9865

------------------------------------------------------------------------------
           S |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      ASVABC |    .148084   .0089431    16.56   0.000     .1305165    .1656516
       _cons |    6.066225   .4672261    12.98   0.000     5.148413    6.984036
------------------------------------------------------------------------------
```

that a t test would show that SF has a highly significant coefficient, but we will perform the F test anyway. We make a note of RSS, 2023.6.

The improvement in the fit on adding the parental schooling variables is the reduction in the residual sum of squares, 2123.0 – 2023.6. The cost is two degrees of freedom because two additional parameters have been estimated. The residual sum of squares remaining unexplained after adding SM and SF is 2023.6. The number of degrees of freedom remaining after adding the new variables is 540 – 4 = 536.

$$F(2,536) = \frac{(2123.0 - 2023.6)/2}{2023.6/536} = 13.16. \tag{3.75}$$

Thus, the F statistic is 13.16. The critical value of $F(2,500)$ at the 0.1 percent level is 7.00. The critical value of $F(2,536)$ must be lower, so we reject H_0 and

Table 3.16

```
. reg S ASVABC SM SF

      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(3, 536)     =  104.30
       Model |  1181.36981     3  393.789935            Prob > F      =  0.0000
    Residual |  2023.61353   536  3.77539837           R-squared     =  0.3686
-------------+------------------------------           Adj R-squared =  0.3651
       Total |  3204.98333   539  5.94616574           Root MSE      =   1.943

------------------------------------------------------------------------------
           S |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      ASVABC |   .1257087   .0098533    12.76   0.000     .1063528   .1450646
          SM |   .0492424   .0390901     1.26   0.208    -.027546    .1260309
          SF |   .1076825   .0309522     3.48   0.001     .04688     .1684851
       _cons |   5.370631   .4882155    11.00   0.000     4.41158    6.329681
------------------------------------------------------------------------------
```

conclude that the parental education variables do have significant joint explanatory power.

Relationship between *F* statistic and *t* statistic

Suppose that you are considering the following alternative model specifications:

$$Y = \beta_1 + \beta_2 X_2 + \cdots + \beta_{k-1} X_{k-1} + u \tag{3.76}$$

$$Y = \beta_1 + \beta_2 X_2 + \cdots + \beta_{k-1} X_{k-1} + \beta_k X_k + u, \tag{3.77}$$

the only difference being the addition of X_k as an explanatory variable in (3.77). You now have two ways to test whether X_k belongs in the model. You could perform a t test on its coefficient when (3.77) is fitted. Alternatively, you could perform an F test of the type just discussed, treating X_k as a 'group' of just one variable, to test its marginal explanatory power. For the F test the null hypothesis will be H_0: $\beta_k = 0$, since only X_k has been added and this is the same null hypothesis as that for the t test. Thus, it might appear that there is a risk that the outcomes of the two tests might conflict with each other.

Fortunately, this is impossible, since it can be shown that the F statistic for this test must be equal to the square of the t statistic and that the critical value of F is equal to the square of the critical value of t (two-sided test). This result means that the t test of the coefficient of a variable is in effect a test of its marginal explanatory power, *after all the other variables have been included in the equation.*

If the variable is correlated with one or more of the other variables, its marginal explanatory power may be quite low, even if it genuinely belongs in the model. If all the variables are correlated, it is possible for all of them to have low marginal explanatory power and for none of the t tests to be significant, even

Table 3.17

```
.reg S ASVABC SM

      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(2, 537)     =  147.36
       Model |  1135.67473     2  567.837363            Prob > F      =  0.0000
    Residual |  2069.30861   537  3.85346109            R-squared     =  0.3543
-------------+------------------------------           Adj R-squared =  0.3519
       Total |  3204.98333   539  5.94616574            Root MSE      =   1.963

------------------------------------------------------------------------------
           S |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      ASVABC |   .1328069   .0097389    13.64   0.000     .1136758    .151938
          SM |   .1235071   .0330837     3.73   0.000     .0585178    .1884963
       _cons |   5.420733   .4930224    10.99   0.000     4.452244    6.389222
------------------------------------------------------------------------------
```

though the F test for their joint explanatory power is highly significant. If this is the case, the model is said to be suffering from the problem of multicollinearity discussed earlier in this chapter.

No proof of the equivalence will be offered here, but it will be illustrated with the educational attainment model. In Table 3.17, it is hypothesized that S depends on *ASVABC* and *SM*. In Table 3.16, it is hypothesized that it depends on *SF* as well.

Comparing Tables 3.16 and 3.17, the improvement on adding *SF* is the reduction in the residual sum of squares, 2069.3 – 2023.6. The cost is just the single degree of freedom lost when estimating the coefficient of *SF*. The residual sum of squares remaining after adding *SF* is 2023.6. The number of degrees of freedom remaining after adding *SF* is 540 – 4 = 536. Hence the F statistic is 12.10:

$$F(1,536) = \frac{(2069.3 - 2023.6)/1}{2023.6/536} = 12.10. \tag{3.78}$$

The critical value of F at the 0.1 percent significance level with 500 degrees of freedom is 10.96. The critical value with 536 degrees of freedom must be lower, so we reject H_0 at the 0.1 percent level. The t statistic for the coefficient of *SF* in the regression with both *SM* and *SF* is 3.48. The critical value of t at the 0.1 percent level with 500 degrees of freedom is 3.31. The critical value with 536 degrees of freedom must be lower, so we also reject H_0 with the t test. The square of 3.48 is 12.11, equal to the F statistic, except for rounding error, and the square of 3.31 is 10.96, equal to the critical value of $F(1,500)$. Hence, the conclusions of the two tests must coincide.

'Adjusted' R^2

If you look at regression output, you will almost certainly find near the R^2 statistic something called the 'adjusted' R^2. Sometimes it is called the 'corrected'

R^2. 'Corrected' makes it sound as if it is better than the ordinary one, but this is questionable.

As was noted earlier in this section, R^2 cannot fall, and generally increases, if you add another variable to a regression equation. The adjusted R^2, usually denoted $\bar{R}^2$, attempts to compensate for this automatic upward shift by imposing a penalty for increasing the number of explanatory variables. It is defined as

$$\begin{aligned}\bar{R}^2 &= 1-(1-R^2)\frac{n-1}{n-k} = \frac{n-1}{n-k}R^2 - \frac{k-1}{n-k} \\ &= R^2 - \frac{k-1}{n-k}(1-R^2),\end{aligned} \tag{3.79}$$

where $k-1$ is the number of explanatory variables. As k increases, $(k-1)/(n-k)$ increases, and so the negative adjustment to R^2 increases.

It can be shown that the addition of a new variable to a regression will cause $\bar{R}^2$ to rise if and only if the absolute value of its t statistic is greater than 1. Hence, a rise in $\bar{R}^2$ when a new variable is added does not necessarily mean that the coefficient of the new variable is significantly different from zero. It therefore does not follow, as is sometimes suggested, that a rise in $\bar{R}^2$ implies that the specification of an equation has improved.

This is one reason why $\bar{R}^2$ is not widely used as a diagnostic statistic. Another is that small variations in R^2 are not all that critical anyway. Initially, it may seem that R^2 should be regarded as a key indicator of the success of model specification. In practice, however, as will be seen in the following chapters, even a very badly specified regression model may yield a high R^2, and recognition of this fact leads to a reduction in the perceived importance of R^2. R^2 should be regarded as just one of a whole set of diagnostic statistics that should be examined when evaluating a regression model. Consequently, there is little to be gained by fine tuning it with a 'correction' of dubious value.

EXERCISES

3.18 Using your *EAEF* data set, fit an educational attainment function, regressing *S* on *ASVABC*, *SM*, and *SF*. Calculate the *F* statistic using the explained sum of squares and the residual sum of squares in the regression output, verify that it matches the *F* statistic in the output, and perform a test of the explanatory power of the equation as a whole. Also calculate the *F* statistic using R^2 and verify that it is the same.

3.19 Fit an educational attainment function using the specification in Exercise 3.18, adding the ASVAB speed test scores *ASVAB05* and *ASVAB06*. Perform an *F* test of the joint explanatory power of *ASVAB05* and *ASVAB06*, using the results of this regression and that in Exercise 3.18.

3.20 Fit an educational attainment function, regressing *S* on *ASVABC*, *SM*, *SF*, and *ASVAB05*. Perform an *F* test of the explanatory power of *ASVAB06*, using the results of this regression and that in Exercise 3.19. Verify that it leads to the same conclusion as a two-sided *t* test.

3.21* The researcher in Exercise 3.16 decides to divide the number of days not worked into the number of days not worked because of illness, I, and the number of days not worked for other reasons, O. The mean value of I in the sample is 2.1 and the mean value of O is 120.2. He fits the regression (standard errors in parentheses):

$$\hat{E} = -9.6 + 2.10W + 0.45O \qquad R^2 = 0.72.$$
$$(8.3)\ (1.98)\quad (1.77)$$

Perform t tests on the regression coefficients and an F test on the goodness of fit of the equation. Explain why the t tests and the F test have different outcomes.

3.6 Prediction

The word prediction is often related to the forecasting of time series, but it is also of practical relevance in a cross-sectional context. One example is hedonic pricing, which we shall use to illustrate the topic.

Hedonic pricing supposes that a good or service has a number of characteristics that individually give it value to the purchaser. The market price of the good is then assumed to be a function, typically a linear combination, of the prices of the characteristics. Thus, one has

$$P_i = \beta_1 + \sum_{j=2}^{k} \beta_j X_{ji} + u_i, \tag{3.80}$$

where P_i is the price of the good, the X_j are the characteristics, and the β_j are their prices. In principle, the β_j may themselves be market prices, but more often they are implicit.

A common example is the pricing of houses, with the value of a house being related to plot size, floor space, the number of bedrooms, proximity to a metropolitan area, and other details. Another example, responsible for much of the growth of the early literature, is the pricing of automobiles, with value being related to size, weight, engine power, etc. Another is the pricing of computers, value being related to the speed of the processor, the size of the hard drive, etc.

Since a computer is just an assembly of traded components, it might be possible to determine its price from the component prices, but typically the prices of the characteristics have to be inferred, and of course multiple regression analysis is an appropriate tool. Given the prices of automobiles of a roughly similar nature with differing specifications, multiple regression analysis may be used to infer the prices of the latter. In the case of houses, the environment is so important that a hedonic pricing function is never going to explain all the variance, but even so some intuitive kind of multiple regression analysis underlies an assertion that, in a given location, an extra bedroom adds so many extra dollars to the

value. More generally, hedonic pricing underlies, if only subjectively, the notion of what is a fair price for a good or a service.

Besides being useful in the marketplace, hedonic pricing is widely used in national accounting. Changes in the money value of output and other aggregates are separated into real changes and changes attributable to inflation. The separation is especially tricky when improvements in technology are causing specifications to change rapidly. In the case of laptop computers, for example, the average price may not seem to change much from year to year, and as a consequence it might seem that changes in volume could be measured in terms of the number of units sold. However, one needs to recognize that the underlying technology, in terms of processor speed, size of hard drive, size of DRAM, etc., is continually being upgraded. Failure to take account of this would lead to underestimation of both the real growth of output and the reduction in prices.

Example

The term 'hedonic price index' was coined in Court (1939), one of the earliest such studies. The study relates to the pricing of automobiles in the 1920s and 1930s and was prompted by the paradox that the price index for automobiles published by the Bureau of Labor Statistics showed an increase of 45 percent from 1925 to 1935, when it was obvious that, in reality, prices had fallen dramatically. The cause of the contradiction was the fact that the Bureau was using a very broad definition of passenger automobile that did not take account of the great improvement in average specification during the period. When one controlled for specification, the conclusion was exactly the opposite.

Table 3.18 presents representative data for eight manufacturers for their cheapest four-passenger vehicles in 1920. Court did not publish his full data set,

Table 3.18 Characteristics and factory price of cheapest four-passenger car, 1920

	Weight (lbs)	Wheelbase (inches)	Brake horsepower	Price ($)
Chrysler	3100	117	45	3170
General Motors	2739	112	44	2435
Graham–Paige	3150	119	43	3260
Hudson	2955	109	55	3010
Hupp	3400	123	45	3400
Nash–Kelvinator	3455	121	35	3285
Studebaker	2900	112	45	2780
Willys–Overland	2152	100	35	1675

Source: Court (1939), p. 106.

but these observations will suffice for illustration. A regression of price on the characteristics yields the following results (standard errors in parentheses):

$$\hat{price} = -2441 + 1.13\ \text{weight} + 10.11\ \text{wheelbase} + 18.28\ \text{horsepower}\quad R^2 = 0.97. \quad (3.81)$$
$$(1667)\ (0.43)\qquad (23.64)\qquad (8.50)$$

The coefficients indicate that an extra pound of weight adds \$1.13 to the value of a car, an extra inch of wheelbase \$10.11, and one extra horsepower \$18.28, with the weight and horsepower coefficients being significant at the 1 and 5 percent levels, despite the tiny size of the sample. Of course, the regression specification is crude and should be viewed only as a conceptual starting point. For an automobile in this category with the average specification of 2,981 pounds, 114 inch wheelbase, and 43 horsepower, the regression specification indicates a price of \$2,869. By 1939, for an average specification, horsepower had doubled to 85, with wheelbase unchanged and weight increased marginally to 2,934 pounds. The regression indicates a price of \$3,583. In the interval, with the Great Depression, there had been general deflation of 28 percent. Taking this into account, one would anticipate a price of \$2,573. The actual average price was \$795, lower by 70 percent, achieved by improvements in production technology and the exploitation of economies of scale, especially between 1920 and 1925.

Properties of least squares predictors

Suppose that one has fitted a hedonic pricing model

$$\hat{P}_i = b_1 + \sum_{j=2}^{k} b_j X_{ji} \quad (3.82)$$

and one encounters a new variety of the good with characteristics $\{X_2^*, X_3^*, \ldots, X_k^*\}$. Given the sample regression result, it is natural to predict that the price of the new variety should be given by

$$\hat{P}^* = b_1 + \sum_{j=2}^{k} b_j X_j^*. \quad (3.83)$$

What can one say about the properties of this prediction? First, it is natural to ask whether it is fair, in the sense of not systematically overestimating or underestimating the actual price. (It is tempting to talk about an 'unbiased' prediction, but we are reserving this term for the estimation of parameters.) Second, we will be concerned about the likely accuracy of the prediction.

We will start by supposing that the good has only one relevant characteristic and that we have fitted the simple regression model

$$\hat{P}_i = b_1 + b_2 X_i \quad (3.84)$$

and that, given a new variety of the good with characteristic $X = X^*$, the predicted price is

$$\hat{P}^* = b_1 + b_2 X^* \tag{3.85}$$

We will define the prediction error of the model, *PE*, as the difference between the actual price, P^*, and the predicted price, $\hat{P}^*$:

$$PE = P^* - \hat{P}^* \tag{3.86}$$

We will assume that the model applies to the new good and therefore the actual price is generated as

$$P^* = \beta_1 + \beta_2 X^* + u^* \tag{3.87}$$

where u^* is the value of the disturbance term for the new good. Then

$$PE = P^* - \hat{P}^* = (\beta_1 + \beta_2 X^* + u^*) - (b_1 + b_2 X^*) \tag{3.88}$$

and the expected value of the prediction error is given by

$$\begin{aligned} E(PE) &= E(\beta_1 + \beta_2 X^* + u^*) - E(b_1 + b_2 X^*) \\ &= \beta_1 + \beta_2 X^* + E(u^*) - E(b_1) - X^* E(b_2) \\ &= \beta_1 + \beta_2 X^* - \beta_1 - X^* \beta_2 = 0 \end{aligned} \tag{3.89}$$

Thus, the expected prediction error is zero. Note that we have assumed that the regression model assumptions are satisfied for the sample period, so that $E(b_1) = \beta_1$ and $E(b_2) = \beta_2$, and that the disturbance term of the new good satisfies the assumption that the expected value of the disturbance term be zero. The result generalizes to the case where there are multiple characteristics and the new good embodies a new combination of them. The proof is left as an exercise.

The population variance of the prediction error is given by

$$\sigma_{PE}^2 = \left\{ 1 + \frac{1}{n} + \frac{(X^* - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \right\} \sigma_u^2, \tag{3.90}$$

where $\bar{X}$ and $\sum (X_i - \bar{X})^2$ are the sample mean and sum of squared deviations of *X*. Unsurprisingly, this implies that, the further is the value of X^* from the sample mean, the larger will be the population variance of the prediction error. It also implies, again unsurprisingly, that, the larger is the sample, the smaller will be the population variance of the prediction error, with a lower limit of σ_u^2. Provided that the regression model assumptions are valid, b_1 and b_2 will tend to

their true values as the sample becomes large, so the only source of error in the prediction will be u^*, and by definition this has population variance σ_u^2.

We can obtain the standard error of the prediction error by replacing σ_u^2 in (3.90) by s_u^2 and taking the square root. Then $(P^* - \hat{P}^*)$/standard error follows a t distribution with the number of degrees of freedom when fitting the equation in the sample period, $n - k$. Hence we can derive a confidence interval for the actual outcome, P^*

$$\hat{P}^* - t_{\text{crit}} \times \text{s.e.} < P^* < \hat{P}^* + t_{\text{crit}} \times \text{s.e.}, \tag{3.91}$$

where t_{crit} is the critical level of t, given the significance level selected and the number of degrees of freedom, and s.e. is the standard error of the prediction. Figure 3.4 depicts in general terms the relationship between the confidence interval for prediction and the value of the explanatory variable.

Naturally, when there are multiple explanatory variables, the expression for the prediction variance becomes complex. One point to note is that multicollinearity may not have an adverse effect on prediction precision, even though the estimates of the coefficients have large variances. The intuitive reason for this is easily explained. For simplicity, suppose that there are two explanatory variables, that both have positive true coefficients, and that they are positively correlated, the model being

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u \tag{3.92}$$

and that we are predicting the value Y^*, given values X_2^* and X_3^*. Then if the effect of X_2 is overestimated, so that $b_2 > \beta_2$, the effect of X_3 will almost certainly be underestimated, with $b_3 < \beta_3$. As a consequence, the effects of the

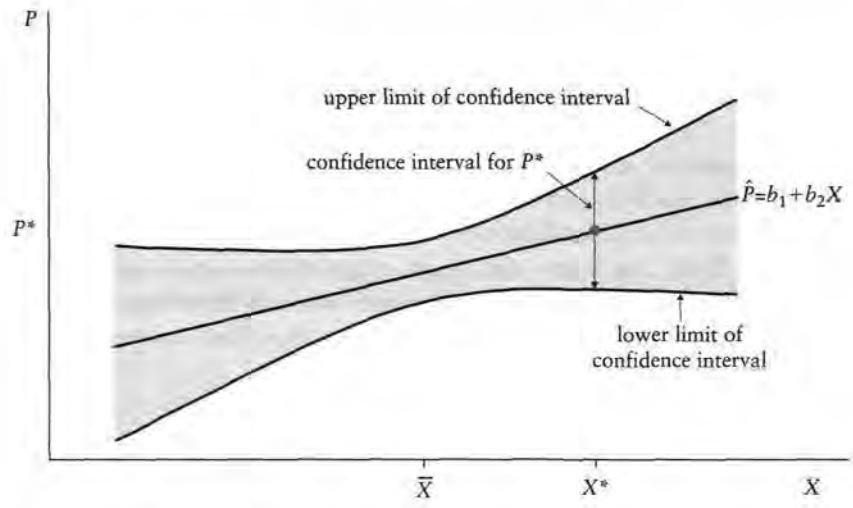

Figure 3.4 Confidence interval for a prediction

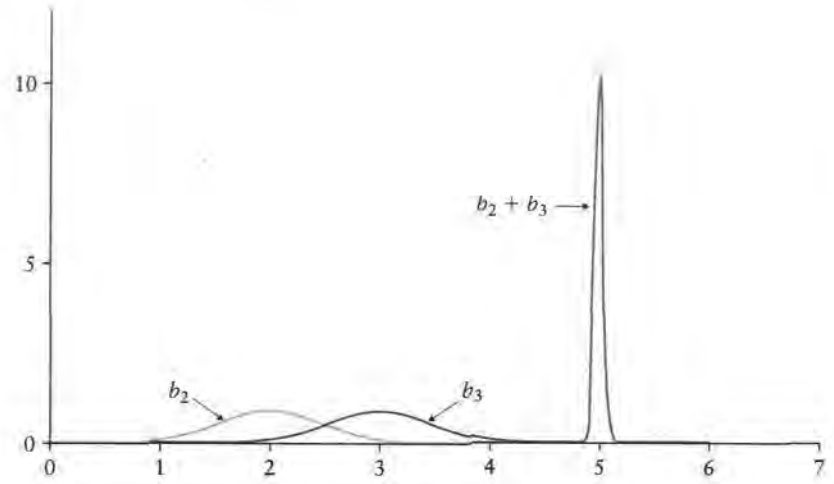

Figure 3.5 Distributions of b_2, b_3, and $b_2 + b_3$

errors may to some extent cancel out, with the result that the linear combination $(b_2X_2^* + b_3X_3^*)$ may be close to $(\beta_2X_2^* + \beta_3X_3^*)$.

This will be illustrated with a simulation, with the model

$$Y = 10 + 2X_2 + 3X_3 + u. \tag{3.93}$$

X_2 is assigned the numbers 1 to 20, and X_3 is almost identical, being assigned the numbers {2,2,4,4,6,6, ...,18,18,20,20}. The correlation between X_2 and X_3 is 0.9962. The disturbance term u is generated as a normal random variable with zero mean and unit variance. We fit the model and make the prediction

$$Y^* = b_1 + b_2X_2^* + b_3X_3^*. \tag{3.94}$$

Since X_2 and X_3 are virtually identical, this may be approximated as

$$Y^* = b_1 + (b_2 + b_3)X_2^*. \tag{3.95}$$

Thus, the predictive accuracy depends on how close $(b_2 + b_3)$ is to $(\beta_2 + \beta_3)$, that is, to 5.

Figure 3.5 shows the distributions of b_2 and b_3 for 10 million samples. Their distributions have relatively wide variances around their true values, as should be expected, given the multicollinearity. The actual standard deviation of their distributions is 0.45. The figure also shows the distribution of their sum. As anticipated, it is distributed around 5, but with a much lower standard deviation, 0.04, despite the multicollinearity affecting the point estimates of the individual coefficients.

Key terms

- adjusted R^2
- Frisch–Waugh–Lovell theorem
- hedonic pricing
- multicollinearity
- multiple regression analysis
- restriction

EXERCISES

3.22 It seems that the first study in hedonic pricing was Waugh's investigation of how the prices of vegetables were determined in the Boston wholesale market (Waugh, 1929). Waugh was an economist working for the Bureau of Agricultural Economics and he was surprised to find that one box of cucumbers might sell for \$7 while another for only \$1. Being told that thinner cucumbers had better texture and taste than fat ones, he fitted the following regression (standard errors in parentheses, data from 1925):

$$\hat{P}_i = 508 + 32.3\ L_i - 8.80\ D_i \qquad R^2 = 0.35$$
$$(272)\ \ (20.1) \qquad (4.45) \qquad F(2,47) = 12.43$$

where P_i is the price, in cents, of a box of cucumbers and L_i and D_i are the length in inches and the diameter/length ratio, as a percentage, of the cucumbers in the box. The boxes in the market were carefully sorted so that their contents were uniform in terms of these characteristics. Give an interpretation of the regression results.

3.23 The standard deviations of the distributions of b_2 and b_3 for the 10 million samples in Figure 3.5 are both 0.45. Verify that this is what you would expect theoretically, given that the correlation between X_2 and X_3 is 0.9962 and that $\sum_{i=1}^{20}(X_2 - \bar{X}_2)^2 = 665$ and $\sum_{i=1}^{20}(X_3 - \bar{X}_3)^2 = 660$.