3 Further development and analysis of the classical linear regression model

Learning Outcomes

In this chapter, you will learn how to

- Construct models with more than one explanatory variable
- Test multiple hypotheses using an F-test
- Determine how well a model fits the data
- Form a restricted regression
- Derive the OLS parameter and standard error estimators using matrix algebra
- Estimate multiple regression models and test multiple hypotheses in EViews

3.1 Generalising the simple model to multiple linear regression

Previously, a model of the following form has been used:

$$y_t = \alpha + \beta x_t + u_t \quad t = 1, 2, \ldots, T \tag{3.1}$$

Equation (3.1) is a simple bivariate regression model. That is, changes in the dependent variable are explained by reference to changes in one single explanatory variable x. But what if the financial theory or idea that is sought to be tested suggests that the dependent variable is influenced by more than one independent variable? For example, simple estimation and tests of the CAPM can be conducted using an equation of the form of (3.1), but arbitrage pricing theory does not pre-suppose that there is only a single factor affecting stock returns. So, to give one illustration, stock returns might be purported to depend on their sensitivity to unexpected changes in:

(1) inflation
(2) the differences in returns on short- and long-dated bonds
(3) industrial production
(4) default risks.

Having just one independent variable would be no good in this case. It would of course be possible to use each of the four proposed explanatory factors in separate regressions. But it is of greater interest and it is more valid to have more than one explanatory variable in the regression equation at the same time, and therefore to examine the effect of all of the explanatory variables together on the explained variable.

It is very easy to generalise the simple model to one with k regressors (independent variables). Equation (3.1) becomes

$$y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \cdots + \beta_k x_{kt} + u_t, \quad t = 1, 2, \ldots, T \tag{3.2}$$

So the variables x_{2t}, $x_{3t}, \ldots,$ x_{kt} are a set of $k-1$ explanatory variables which are thought to influence y, and the coefficient estimates β_1, $\beta_2, \ldots,$ β_k are the parameters which quantify the effect of each of these explanatory variables on y. The coefficient interpretations are slightly altered in the multiple regression context. Each coefficient is now known as a partial regression coefficient, interpreted as representing the partial effect of the given explanatory variable on the explained variable, after holding constant, or eliminating the effect of, all other explanatory variables. For example, $\hat{\beta}_2$ measures the effect of x_2 on y after eliminating the effects of x_3, $x_4, \ldots,$ x_k. Stating this in other words, each coefficient measures the average change in the dependent variable per unit change in a given independent variable, holding all other independent variables constant at their average values.

3.2 The constant term

In (3.2) above, astute readers will have noticed that the explanatory variables are numbered x_2, $x_3, \ldots$ i.e. the list starts with x_2 and not x_1. So, where is x_1? In fact, it is the constant term, usually represented by a column of ones of length T:

$$x_1 = \begin{bmatrix} 1 \\ 1 \\ \cdot \\ \cdot \\ \cdot \\ 1 \end{bmatrix} \tag{3.3}$$

Thus there is a variable implicitly hiding next to β_1, which is a column vector of ones, the length of which is the number of observations in the sample. The x_1 in the regression equation is not usually written, in the same way that one unit of p and 2 units of q would be written as '$p + 2q$' and not '$1p + 2q$'. β_1 is the coefficient attached to the constant term (which was called α in the previous chapter). This coefficient can still be referred to as the *intercept*, which can be interpreted as the average value which y would take if all of the explanatory variables took a value of zero.

A tighter definition of k, the number of explanatory variables, is probably now necessary. Throughout this book, k is defined as the number of 'explanatory variables' or 'regressors' including the constant term. This is equivalent to the number of parameters that are estimated in the regression equation. Strictly speaking, it is not sensible to call the constant an explanatory variable, since it does not explain anything and it always takes the same values. However, this definition of k will be employed for notational convenience.

Equation (3.2) can be expressed even more compactly by writing it in matrix form

$$y = X\beta + u \tag{3.4}$$

where: y is of dimension $T \times 1$
X is of dimension $T \times k$
β is of dimension $k \times 1$
u is of dimension $T \times 1$

The difference between (3.2) and (3.4) is that all of the time observations have been stacked up in a vector, and also that all of the different explanatory variables have been squashed together so that there is a column for each in the X matrix. Such a notation may seem unnecessarily complex, but in fact, the matrix notation is usually more compact and convenient. So, for example, if k is 2, i.e. there are two regressors, one of which is the constant term (equivalent to a simple bivariate regression $y_t = \alpha + \beta x_t + u_t$), it is possible to write

$$\underset{T \times 1}{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{bmatrix}} = \underset{T \times 2}{\begin{bmatrix} 1 & x_{21} \\ 1 & x_{22} \\ \vdots & \vdots \\ 1 & x_{2T} \end{bmatrix}} \underset{2 \times 1}{\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}} + \underset{T \times 1}{\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_T \end{bmatrix}} \tag{3.5}$$

so that the x_{ij} element of the matrix X represents the jth time observation on the ith variable. Notice that the matrices written in this way are *conformable* – in other words, there is a valid matrix multiplication and addition on the RHS.

The above presentation is the standard way to express matrices in the time series econometrics literature, although the ordering of the indices is different to that used in the mathematics of matrix algebra (as presented in the mathematical appendix at the end of this book). In the latter case, x_{ij} would represent the element in row i and column j, although in the notation used in the body of this book it is the other way around.

3.3 How are the parameters (the elements of the β vector) calculated in the generalised case?

Previously, the residual sum of squares, $\sum \hat{u}_i^2$ was minimised with respect to α and β. In the multiple regression context, in order to obtain estimates of the parameters, $\beta_1, \beta_2, \ldots, \beta_k$, the *RSS* would be minimised with respect to all the elements of β. Now, the residuals can be stacked in a vector:

$$\hat{u} = \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \vdots \\ \hat{u}_T \end{bmatrix} \tag{3.6}$$

The *RSS* is still the relevant loss function, and would be given in a matrix notation by

$$L = \hat{u}'\hat{u} = [\hat{u}_1 \hat{u}_2 \cdots \hat{u}_T] \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \vdots \\ \hat{u}_T \end{bmatrix} = \hat{u}_1^2 + \hat{u}_2^2 + \cdots + \hat{u}_T^2 = \sum \hat{u}_t^2 \tag{3.7}$$

Using a similar procedure to that employed in the bivariate regression case, i.e. substituting into (3.7), and denoting the vector of estimated parameters as $\hat{\beta}$, it can be shown (see the appendix to this chapter) that the coefficient estimates will be given by the elements of the expression

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = (X'X)^{-1}X'y \tag{3.8}$$

If one were to check the dimensions of the RHS of (3.8), it would be observed to be $k \times 1$. This is as required since there are k parameters to be estimated by the formula for $\hat{\beta}$.

But how are the standard errors of the coefficient estimates calculated? Previously, to estimate the variance of the errors, σ^2, an estimator denoted by s^2 was used

$$s^2 = \frac{\sum \hat{u}_t^2}{T-2} \tag{3.9}$$

The denominator of (3.9) is given by $T-2$, which is the number of degrees of freedom for the bivariate regression model (i.e. the number of observations minus two). This essentially applies since two observations are effectively 'lost' in estimating the two model parameters (i.e. in deriving estimates for α and β). In the case where there is more than one explanatory variable plus a constant, and using the matrix notation, (3.9) would be modified to

$$s^2 = \frac{\hat{u}'\hat{u}}{T-k} \tag{3.10}$$

where k = number of regressors including a constant. In this case, k observations are 'lost' as k parameters are estimated, leaving $T-k$ degrees of freedom. It can also be shown (see the appendix to this chapter) that the parameter variance–covariance matrix is given by

$$\operatorname{var}(\hat{\beta}) = s^2(X'X)^{-1} \tag{3.11}$$

The leading diagonal terms give the coefficient variances while the off-diagonal terms give the covariances between the parameter estimates, so that the variance of $\hat{\beta}_1$ is the first diagonal element, the variance of $\hat{\beta}_2$ is the second element on the leading diagonal, and the variance of $\hat{\beta}_k$ is the kth diagonal element. The coefficient standard errors are thus simply given by taking the square roots of each of the terms on the leading diagonal.

Example 3.1

The following model with 3 regressors (including the constant) is estimated over 15 observations

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + u \tag{3.12}$$

and the following data have been calculated from the original xs

$$(X'X)^{-1} = \begin{bmatrix} 2.0 & 3.5 & -1.0 \\ 3.5 & 1.0 & 6.5 \\ -1.0 & 6.5 & 4.3 \end{bmatrix}, \quad (X'y) = \begin{bmatrix} -3.0 \\ 2.2 \\ 0.6 \end{bmatrix}, \quad \hat{u}'\hat{u} = 10.96$$

Calculate the coefficient estimates and their standard errors.

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = (X'X)^{-1}X'y = \begin{bmatrix} 2.0 & 3.5 & -1.0 \\ 3.5 & 1.0 & 6.5 \\ -1.0 & 6.5 & 4.3 \end{bmatrix} \times \begin{bmatrix} -3.0 \\ 2.2 \\ 0.6 \end{bmatrix} = \begin{bmatrix} 1.10 \\ -4.40 \\ 19.88 \end{bmatrix} \tag{3.13}$$

To calculate the standard errors, an estimate of σ^2 is required

$$s^2 = \frac{RSS}{T-k} = \frac{10.96}{15-3} = 0.91 \tag{3.14}$$

The variance–covariance matrix of $\hat{\beta}$ is given by

$$s^2(X'X)^{-1} = 0.91(X'X)^{-1} = \begin{bmatrix} 1.82 & 3.19 & -0.91 \\ 3.19 & 0.91 & 5.92 \\ -0.91 & 5.92 & 3.91 \end{bmatrix} \tag{3.15}$$

The coefficient variances are on the diagonals, and the standard errors are found by taking the square roots of each of the coefficient variances

$$\text{var}(\hat{\beta}_1) = 1.82 \qquad SE(\hat{\beta}_1) = 1.35 \tag{3.16}$$

$$\text{var}(\hat{\beta}_2) = 0.91 \Leftrightarrow SE(\hat{\beta}_2) = 0.95 \tag{3.17}$$

$$\text{var}(\hat{\beta}_3) = 3.91 \qquad SE(\hat{\beta}_3) = 1.98 \tag{3.18}$$

The estimated equation would be written

$$\begin{aligned} \hat{y} = \; & 1.10 - 4.40x_2 + 19.88x_3 \\ & (1.35) \; (0.95) \qquad (1.98) \end{aligned} \tag{3.19}$$

Fortunately, in practice all econometrics software packages will estimate the cofficient values and their standard errors. Clearly, though, it is still useful to understand where these estimates came from.

3.4 Testing multiple hypotheses: the *F*-test

The *t*-test was used to test single hypotheses, i.e. hypotheses involving only one coefficient. But what if it is of interest to test more than one coefficient simultaneously? For example, what if a researcher wanted to determine whether a restriction that the coefficient values for β_2 and β_3 are both unity could be imposed, so that an increase in either one of the two variables x_2 or x_3 would cause y to rise by one unit? The *t*-testing framework is not sufficiently general to cope with this sort of hypothesis test. Instead, a more general framework is employed, centring on an F-test. Under the F-test framework, two regressions are required, known as the unrestricted and the restricted regressions. The unrestricted regression is the one in which the coefficients are freely determined by the data, as has been constructed previously. The restricted regression is the one in which the coefficients are restricted, i.e. the restrictions are imposed on some βs. Thus the F-test approach to hypothesis testing is also termed restricted least squares, for obvious reasons.

The residual sums of squares from each regression are determined, and the two residual sums of squares are 'compared' in the test statistic. The F-test statistic for testing multiple hypotheses about the coefficient estimates is given by

$$\textit{test statistic} = \frac{RRSS - URSS}{URSS} \times \frac{T-k}{m} \tag{3.20}$$

where the following notation applies:

$$\begin{aligned} URSS &= \text{residual sum of squares from unrestricted regression} \\ RRSS &= \text{residual sum of squares from restricted regression} \\ m &= \text{number of restrictions} \\ T &= \text{number of observations} \\ k &= \text{number of regressors in unrestricted regression} \end{aligned}$$

The most important part of the test statistic to understand is the numerator expression $RRSS - URSS$. To see why the test centres around a comparison of the residual sums of squares from the restricted and unrestricted regressions, recall that OLS estimation involved choosing the model that minimised the residual sum of squares, with no constraints imposed. Now if, after imposing constraints on the model, a residual sum of squares results that is not much higher than the unconstrained model's residual sum of squares, it would be concluded that the restrictions were supported by the data. On the other hand, if the residual sum of squares increased considerably after the restrictions were imposed, it would be concluded that the restrictions were not supported by the data and therefore that the hypothesis should be rejected.

It can be further stated that $RRSS \geq URSS$. Only under a particular set of very extreme circumstances will the residual sums of squares for the restricted and unrestricted models be exactly equal. This would be the case when the restriction was already present in the data, so that it is not really a restriction at all (it would be said that the restriction is 'not binding', i.e. it does not make any difference to the parameter estimates). So, for example, if the null hypothesis is H$_0$: $\beta_2 = 1$ and $\beta_3 = 1$, then $RRSS = URSS$ only in the case where the coefficient estimates for the unrestricted regression had been $\hat{\beta}_2 = 1$ and $\hat{\beta}_3 = 1$. Of course, such an event is extremely unlikely to occur in practice.

Example 3.2

Dropping the time subscripts for simplicity, suppose that the general regression is

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + u \tag{3.21}$$

and that the restriction $\beta_3 + \beta_4 = 1$ is under test (there exists some hypothesis from theory which suggests that this would be an interesting hypothesis to study). The unrestricted regression is (3.21) above, but what is the restricted regression? It could be expressed as

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + u \;\; \text{s.t. (subject to) } \beta_3 + \beta_4 = 1 \tag{3.22}$$

The restriction ($\beta_3 + \beta_4 = 1$) is substituted into the regression so that it is automatically imposed on the data. The way that this would be achieved would be to make either β_3 or β_4 the subject of (3.22), e.g.

$$\beta_3 + \beta_4 = 1 \Rightarrow \beta_4 = 1 - \beta_3 \tag{3.23}$$

and then substitute into (3.21) for β_4

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + (1 - \beta_3) x_4 + u \tag{3.24}$$

Equation (3.24) is already a restricted form of the regression, but it is not yet in the form that is required to estimate it using a computer package. In order to be able to estimate a model using OLS, software packages usually require each RHS variable to be multiplied by one coefficient only. Therefore, a little more algebraic manipulation is required. First, expanding the brackets around $(1 - \beta_3)$

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + x_4 - \beta_3 x_4 + u \tag{3.25}$$

Then, gathering all of the terms in each β_i together and rearranging

$$(y - x_4) = \beta_1 + \beta_2 x_2 + \beta_3 (x_3 - x_4) + u \tag{3.26}$$

Note that any variables without coefficients attached (e.g. x_4 in (3.25)) are taken over to the LHS and are then combined with y. Equation (3.26) is the restricted regression. It is actually estimated by creating two new variables – call them, say, P and Q, where $P = y - x_4$ and $Q = x_3 - x_4$ – so the regression that is actually estimated is

$$P = \beta_1 + \beta_2 x_2 + \beta_3 Q + u \tag{3.27}$$

What would have happened if instead β_3 had been made the subject of (3.23) and β_3 had therefore been removed from the equation? Although the equation that would have been estimated would have been different from (3.27), the value of the residual sum of squares for these two models (both of which have imposed upon them the same restriction) would be the same.

The test statistic follows the F-distribution under the null hypothesis. The F-distribution has 2 degrees of freedom parameters (recall that the t-distribution had only 1 degree of freedom parameter, equal to $T - k$). The value of the degrees of freedom parameters for the F-test are m, the number of restrictions imposed on the model, and $(T - k)$, the number of observations less the number of regressors for the unrestricted regression, respectively. Note that the order of the degree of freedom parameters is important. The appropriate critical value will be in column m, row $(T - k)$ of the F-distribution tables.

3.4.1 The relationship between the t- and the F-distributions

Any hypothesis that could be tested with a t-test could also have been tested using an F-test, but not the other way around. So, single hypotheses involving one coefficient can be tested using a t- or an F-test, but multiple hypotheses can be tested only using an F-test. For example, consider the hypothesis

$$H_0 : \beta_2 = 0.5$$
$$H_1 : \beta_2 \neq 0.5$$

This hypothesis could have been tested using the usual t-test

$$\textit{test stat} = \frac{\hat{\beta}_2 - 0.5}{SE(\hat{\beta}_2)} \tag{3.28}$$

or it could be tested in the framework above for the F-test. Note that the two tests always give the same conclusion since the t-distribution is just a special case of the F-distribution. For example, consider any random variable Z that follows a t-distribution with $T - k$ degrees of freedom, and square it. The square of the t is equivalent to a particular form of the F-distribution

$$Z^2 \sim t^2\ (T - k) \text{ then also } Z^2 \sim F(1, T - k)$$

Thus the square of a t-distributed random variable with $T - k$ degrees of freedom also follows an F-distribution with 1 and $T - k$ degrees of freedom. This relationship between the t and the F-distributions will always hold – take some examples from the statistical tables and try it!

The F-distribution has only positive values and is not symmetrical. Therefore, the null is rejected only if the test statistic exceeds the critical F-value, although the test is a two-sided one in the sense that rejection will occur if $\hat{\beta}_2$ is significantly bigger or significantly smaller than 0.5.

3.4.2 Determining the number of restrictions, m

How is the appropriate value of m decided in each case? Informally, the number of restrictions can be seen as 'the number of equality signs under the null hypothesis'. To give some examples

H_0 : *hypothesis*	*No. of restrictions*, m
$\beta_1 + \beta_2 = 2$	1
$\beta_2 = 1$ and $\beta_3 = -1$	2
$\beta_2 = 0$, $\beta_3 = 0$ and $\beta_4 = 0$	3

At first glance, you may have thought that in the first of these cases, the number of restrictions was two. In fact, there is only one restriction that involves two coefficients. The number of restrictions in the second two examples is obvious, as they involve two and three separate component restrictions, respectively.

The last of these three examples is particularly important. If the model is

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + u \tag{3.29}$$

then the null hypothesis of

$$\mathrm{H}_0 : \beta_2 = 0 \quad \text{and} \quad \beta_3 = 0 \quad \text{and} \quad \beta_4 = 0$$

is tested by 'THE' regression F-statistic. It tests the null hypothesis that all of the coefficients except the intercept coefficient are zero. This test is sometimes called a test for 'junk regressions', since if this null hypothesis cannot be rejected, it would imply that none of the independent variables in the model was able to explain variations in y.

Note the form of the alternative hypothesis for all tests when more than one restriction is involved

$$\mathrm{H}_1 : \beta_2 \neq 0 \quad \text{or} \quad \beta_3 \neq 0 \quad \text{or} \quad \beta_4 \neq 0$$

In other words, 'and' occurs under the null hypothesis and 'or' under the alternative, so that it takes only one part of a joint null hypothesis to be wrong for the null hypothesis as a whole to be rejected.

3.4.3 Hypotheses that cannot be tested with either an F- or a t-test

It is not possible to test hypotheses that are not linear or that are multiplicative using this framework – for example, $H_0 : \beta_2\beta_3 = 2$, or $H_0 : \beta_2^2 = 1$ cannot be tested.

Example 3.3

Suppose that a researcher wants to test whether the returns on a company stock (y) show unit sensitivity to two factors (factor x_2 and factor x_3) among three considered. The regression is carried out on 144 monthly observations. The regression is

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + u \tag{3.30}$$

(1) What are the restricted and unrestricted regressions?
(2) If the two *RSS* are 436.1 and 397.2, respectively, perform the test.

Unit sensitivity to factors x_2 and x_3 implies the restriction that the coefficients on these two variables should be unity, so H_0: $\beta_2 = 1$ and $\beta_3 = 1$. The unrestricted regression will be the one given by (3.30) above. To derive the restricted regression, first impose the restriction:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + u \quad \text{s.t.} \quad \beta_2 = 1 \quad \text{and} \quad \beta_3 = 1 \tag{3.31}$$

Replacing β_2 and β_3 by their values under the null hypothesis

$$y = \beta_1 + x_2 + x_3 + \beta_4 x_4 + u \tag{3.32}$$

Rearranging

$$y - x_2 - x_3 = \beta_1 + \beta_4 x_4 + u \tag{3.33}$$

Defining $z = y - x_2 - x_3$, the restricted regression is one of z on a constant and x_4

$$z = \beta_1 + \beta_4 x_4 + u \tag{3.34}$$

The formula for the F-test statistic is given in (3.20) above. For this application, the following inputs to the formula are available: $T = 144$, $k = 4$, $m = 2$, $RRSS = 436.1$, $URSS = 397.2$. Plugging these into the formula gives an F-test statistic value of 6.86. This statistic should be compared with an $F(m, T - k)$, which in this case is an $F(2, 140)$. The critical values are 3.07 at the 5% level and 4.79 at the 1% level. The test statistic clearly exceeds the critical values at both the 5% and 1% levels, and hence the null hypothesis is rejected. It would thus be concluded that the restriction is not supported by the data.

The following sections will now re-examine the CAPM model as an illustration of how to conduct multiple hypothesis tests using EViews.

3.5 Sample EViews output for multiple hypothesis tests

Reload the 'capm.wk1' workfile constructed in the previous chapter. As a reminder, the results are included again below.

```
Dependent Variable: ERFORD
Method: Least Squares
Date: 08/21/07 Time: 15:02
Sample (adjusted): 2002M02 2007M04
Included observations: 63 after adjustments

                         Coefficient   Std. Error   t-Statistic      Prob.

           C                2.020219     2.801382      0.721151     0.4736
       ERSANDP              0.359726     0.794443      0.452803     0.6523

R-squared                   0.003350   Mean dependent var         2.097445
Adjusted R-squared         -0.012989   S.D. dependent var         22.05129
S.E. of regression          22.19404   Akaike info criterion      9.068756
Sum squared resid           30047.09   Schwarz criterion          9.136792
Log likelihood             -283.6658   Hannan-Quinn criter.       9.095514
F-statistic                 0.205031   Durbin-Watson stat         1.785699
Prob(F-statistic)           0.652297
```

If we examine the regression F-test, this also shows that the regression slope coefficient is not significantly different from zero, which in this case is exactly the same result as the t-test for the beta coefficient (since there is only one slope coefficient). Thus, in this instance, the F-test statistic is equal to the square of the slope t-ratio.

Now suppose that we wish to conduct a joint test that both the intercept and slope parameters are 1. We would perform this test exactly as for a test involving only one coefficient. Select **View/Coefficient Tests/Wald - Coefficient Restrictions**... and then in the box that appears, type **C(1)=1, C(2)=1**. There are two versions of the test given: an F-version and a χ^2-version. The F-version is adjusted for small sample bias and should be used when the regression is estimated using a small sample (see chapter 4). Both statistics asymptotically yield the same result, and in this case the p-values are very similar. The conclusion is that the joint null hypothesis, $H_0 : \beta_1 = 1$ and $\beta_2 = 1$, is not rejected.

3.6 Multiple regression in EViews using an APT-style model

In the spirit of arbitrage pricing theory (APT), the following example will examine regressions that seek to determine whether the monthly returns on Microsoft stock can be explained by reference to unexpected changes in a set of macroeconomic and financial variables. **Open a new EViews workfile** to store the data. There are 254 monthly observations in the file 'macro.xls', starting in March 1986 and ending in April 2007. There are 13 series plus a column of dates. The series in the Excel file are the Microsoft stock price, the S&P500 index value, the consumer price index, an industrial production index, Treasury bill yields for the following maturities: three months, six months, one year, three years, five years and ten years, a measure of 'narrow' money supply, a consumer credit series, and a 'credit spread' series. The latter is defined as the difference in annualised average yields between a portfolio of bonds rated AAA and a portfolio of bonds rated BAA.

Import the data from the Excel file and save the resulting workfile as 'macro.wf1'.

The first stage is to generate a set of changes or *differences* for each of the variables, since the APT posits that the stock returns can be explained by reference to the *unexpected changes* in the macroeconomic variables rather than their levels. The unexpected value of a variable can be defined as the difference between the actual (realised) value of the variable and its expected value. The question then arises about how we believe that investors might have formed their expectations, and while there are many ways to construct measures of expectations, the easiest is to assume that investors have naive expectations that the next period value of the variable is equal to the current value. This being the case, the entire change in the variable from one period to the next is the unexpected change (because investors are assumed to expect no change).[1]

Transforming the variables can be done as described above. Press **Genr** and then enter the following in the 'Enter equation' box:

dspread = baa_aaa_spread - baa_aaa_spread(-1)

Repeat these steps to conduct all of the following transformations:

dcredit = consumer_credit - consumer_credit(-1)
dprod = industrial_production - industrial_production(-1)
rmsoft = 100*dlog(microsoft)
rsandp = 100*dlog(sandp)
dmoney = m1money_supply - m1money_supply(-1)

[1] It is an interesting question as to whether the differences should be taken on the levels of the variables or their logarithms. If the former, we have absolute changes in the variables, whereas the latter would lead to proportionate changes. The choice between the two is essentially an empirical one, and this example assumes that the former is chosen, apart from for the stock price series themselves and the consumer price series.

inflation = 100*dlog(cpi)
term = ustb10y - ustb3m

and then click **OK**. Next, we need to apply further transformations to some of the transformed series, so **repeat the above steps** to generate

dinflation = inflation - inflation(-1)
mustb3m = ustb3m/12
rterm = term - term(-1)
ermsoft = rmsoft - mustb3m
ersandp = rsandp - mustb3m

The final two of these calculate excess returns for the stock and for the index.

We can now run the regression. So click **Object/New Object/Equation** and name the object '**msoftreg**'. Type the following variables in the Equation specification window

ERMSOFT C ERSANDP DPROD DCREDIT DINFLATION DMONEY DSPREAD RTERM

and use **Least Squares** over the whole sample period. The table of results will appear as follows.

```
Dependent Variable: ERMSOFT
Method: Least Squares
Date: 08/21/07 Time: 21:45
Sample (adjusted): 1986M05 2007M04
Included observations: 252 after adjustments

                     Coefficient   Std. Error   t-Statistic      Prob.

        C              −0.587603     1.457898     −0.403048     0.6873
     ERSANDP            1.489434     0.203276      7.327137     0.0000
      DPROD             0.289322     0.500919      0.577583     0.5641
     DCREDIT           −5.58E-05     0.000160     −0.347925     0.7282
    DINFLATION          4.247809     2.977342      1.426712     0.1549
      DMONEY           −1.161526     0.713974     −1.626847     0.1051
     DSPREAD            12.15775     13.55097      0.897187     0.3705
      RTERM             6.067609     3.321363      1.826843     0.0689

R-squared               0.203545   Mean dependent var        −0.420803
Adjusted R-squared      0.180696   S.D. dependent var         15.41135
S.E. of regression      13.94965   Akaike info criterion      8.140017
Sum squared resid       47480.62   Schwarz criterion          8.252062
Log likelihood         −1017.642   Hannan-Quinn criter.       8.185102
F-statistic             8.908218   Durbin-Watson stat         2.156221
Prob(F-statistic)       0.000000
```

Take a few minutes to examine the main regression results. Which of the variables has a statistically significant impact on the Microsoft excess returns? Using your knowledge of the effects of the financial and macroeconomic environment on stock returns, examine whether the coefficients have their expected signs and whether the sizes of the parameters are plausible.

The regression *F*-statistic takes a value 8.908. Remember that this tests the null hypothesis that all of the slope parameters are jointly zero. The *p*-value of zero attached to the test statistic shows that this null hypothesis should be rejected. However, there are a number of parameter estimates that are not significantly different from zero – specifically those on the DPROD, DCREDIT and DSPREAD variables. Let us test the null hypothesis that the parameters on these three variables are jointly zero using an *F*-test. To test this, Click on **View/Coefficient Tests/Wald – Coefficient Restrictions**. . . and in the box that appears type **C(3)=0, C(4)=0, C(7)=0 and click OK.** The resulting *F*-test statistic follows an *F*(3, 244) distribution as there are three restrictions, 252 usable observations and eight parameters to estimate in the unrestricted regression. The *F*-statistic value is 0.402 with *p*-value 0.752, suggesting that the null hypothesis cannot be rejected. The parameters on DINLATION and DMONEY are almost significant at the 10% level and so the associated parameters are not included in this *F*-test and the variables are retained.

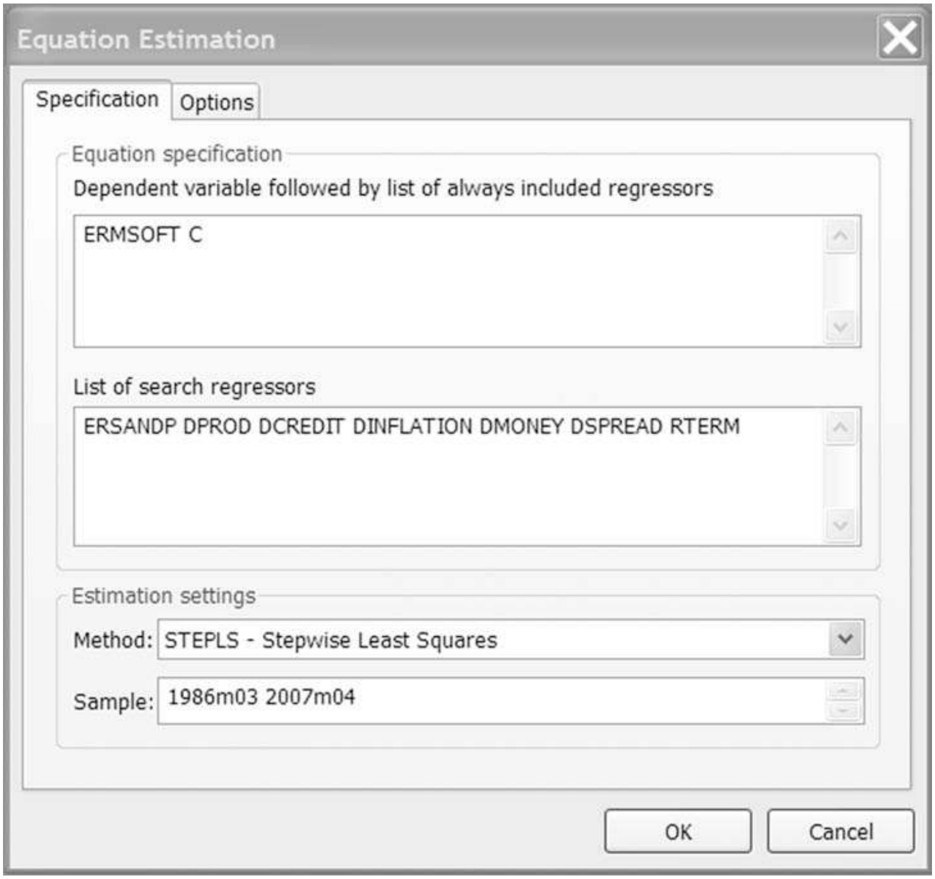
There is a procedure known as a *stepwise regression* that is now available in EViews 6. Stepwise regression is an automatic variable selection procedure which chooses the jointly most ‘important’ (variously defined) explanatory variables from a set of candidate variables. There are a number of different stepwise regression procedures, but the simplest is the uni-directional forwards method. This starts with no variables in the regression (or only those variables that are always required by the researcher to be in the regression) and then it selects first the variable with the lowest *p*-value (largest *t*-ratio) if it were included, then the variable with the second lowest *p*-value conditional upon the first variable already being included, and so on. The procedure continues until the next lowest *p*-value relative to those already included variables is larger than some specified threshold value, then the selection stops, with no more variables being incorporated into the model.

To conduct a stepwise regression which will automatically select from among these variables the most important ones for explaining the variations in Microsoft stock returns, click **Proc** and then **Equation**. Name the equation **Msoftstepwise** and then in the ‘Estimation settings/Method’ box, change *LS – Least Squares (NLS and ARMA)* to ***STEPLS – Stepwise Least***

Squares and then in the top box that appears, 'Dependent variable followed by list of always included regressors', enter

ERMSOFT C

This shows that the dependent variable will be the excess returns on Microsoft stock and that an intercept will always be included in the regression. If the researcher had a strong prior view that a particular explanatory variable must always be included in the regression, it should be listed in this first box. In the second box, 'List of search regressors', type the list of all of the explanatory variables used above: **ERSANDP DPROD DCREDIT DINFLATION DMONEY DSPREAD RTERM**. The window will appear as in screenshot 3.1.



Dependent variable followed by list of always included regressors: ERMSOFT C

List of search regressors: ERSANDP DPROD DCREDIT DINFLATION DMONEY DSPREAD RTERM

Method: STEPLS - Stepwise Least Squares

Sample: 1986m03 2007m04

Screenshot 3.1
Stepwise procedure equation estimation window

Clicking on the 'Options' tab gives a number of ways to conduct the regression. For example, 'Forwards' will start with the list of required regressors (the intercept only in this case) and will sequentially add to them, while ‘Backwards’ will start by including all of the variables and will sequentially delete variables from the regression. The default criterion is to include variables if the *p*-value is less than 0.5, but this seems high and could potentially result in the inclusion of some very insignificant variables, so **modify this to 0.2** and then click **OK** to see the results.

As can be seen, the excess market return, the term structure, money supply and unexpected inflation variables have all been included, while the default spread and credit variables have been omitted.

```
Dependent Variable: ERMSOFT
Method: Stepwise Regression
Date: 08/27/07 Time: 10:21
Sample (adjusted): 1986M05 2007M04
Included observations: 252 after adjustments
Number of always included regressors: 1
Number of search regressors: 7
Selection method: Stepwise forwards
Stopping criterion: p-value forwards/backwards = 0.2/0.2

                      Coefficient   Std. Error   t-Statistic        Prob.*

           C            -0.947198       0.8787     -1.077954        0.2821

       ERSANDP           1.471400     0.201459      7.303725        0.0000
        RTERM            6.121657     3.292863      1.859068        0.0642
        DMONEY          -1.171273     0.702523     -1.667238        0.0967
      DINFLATION         4.013512     2.876986      1.395040        0.1643

R-squared                0.199612   Mean dependent var           -0.420803
Adjusted R-squared       0.186650   S.D. dependent var            15.41135
S.E. of regression       13.89887   Akaike info criterion         8.121133
Sum squared resid        47715.09   Schwarz criterion             8.191162
Log likelihood          -1018.263   Hannan-Quinn criter.          8.149311
F-statistic              15.40008   Durbin-Watson stat            2.150604
Prob(F-statistic)        0.000000

Selection Summary

  Added ERSANDP
   Added RTERM
  Added DMONEY
 Added DINFLATION
```

*Note: p-values and subsequent tests do not account for stepwise selection.

Stepwise procedures have been strongly criticised by statistical purists. At the most basic level, they are sometimes argued to be no better than automated procedures for data mining, in particular if the list of potential candidate variables is long and results from a ‘fishing trip’ rather than a strong prior financial theory. More subtly, the iterative nature of the variable selection process implies that the size of the tests on parameters attached to variables in the final model will not be the nominal values (e.g. 5%) that would have applied had this model been the only one estimated. Thus the p-values for tests involving parameters in the final regression should really be modified to take into account that the model results from a sequential procedure, although they are usually not in statistical packages such as EViews.

3.6.1 A note on sample sizes and asymptotic theory

A question that is often asked by those new to econometrics is 'what is an appropriate sample size for model estimation?' While there is no definitive answer to this question, it should be noted that most testing procedures in econometrics rely on asymptotic theory. That is, the results in theory hold only if there are an *infinite number of observations*. In practice, an infinite number of observations will never be available and fortunately, an infinite number of observations are not usually required to invoke the asymptotic theory! An approximation to the asymptotic behaviour of the test statistics can be obtained using finite samples, provided that they are large enough. In general, as many observations as possible should be used (although there are important caveats to this statement relating to 'structural stability', discussed in chapter 4). The reason is that all the researcher has at his disposal is a sample of data from which to estimate parameter values and to infer their likely population counterparts. A sample may fail to deliver something close to the exact population values owing to sampling error. Even if the sample is randomly drawn from the population, some samples will be more representative of the behaviour of the population than others, purely owing to 'luck of the draw'. Sampling error is minimised by increasing the size of the sample, since the larger the sample, the less likely it is that all of the data drawn will be unrepresentative of the population.

3.7 Data mining and the true size of the test

Recall that the probability of rejecting a correct null hypothesis is equal to the size of the test, denoted α. The possibility of rejecting a correct null hypothesis arises from the fact that test statistics are assumed to follow a random distribution and hence they will take on extreme values that fall in the rejection region some of the time by chance alone. A consequence of this is that it will almost always be possible to find significant relationships between variables if enough variables are examined. For example, suppose that a dependent variable y_t and 20 explanatory variables $x_{2t}, \ldots, x_{21t}$ (excluding a constant term) are generated separately as independent normally distributed random variables. Then y is regressed separately on each of the 20 explanatory variables plus a constant, and the significance of each explanatory variable in the regressions is examined. If this experiment is repeated many times, on average one of the 20 regressions will have a slope coefficient that is significant at the 5% level for each experiment. The implication is that for any regression, if enough explanatory variables are employed in a regression, often one or more will be significant by chance alone. More concretely, it could be stated that if an α% size of test is used, on average one in every $(100/\alpha)$ regressions will have a significant slope coefficient by chance alone.

Trying many variables in a regression without basing the selection of the candidate variables on a financial or economic theory is known as 'data mining' or 'data snooping'. The result in such cases is that the true significance level will be considerably greater than the nominal significance level assumed. For example, suppose that 20 separate regressions are conducted, of which three contain a significant regressor, and a 5% nominal significance level is assumed, then the true significance level would be much higher (e.g. 25%). Therefore, if the researcher then shows only the results for the regression containing the final three equations and states that they are significant at the 5% level, inappropriate conclusions concerning the significance of the variables would result.

As well as ensuring that the selection of candidate regressors for inclusion in a model is made on the basis of financial or economic theory, another way to avoid data mining is by examining the forecast performance of the model in an 'out-of-sample' data set (see chapter 5). The idea is essentially that a proportion of the data is not used in model estimation, but is retained for model testing. A relationship observed in the estimation period that is purely the result of data mining, and is therefore spurious, is very unlikely to be repeated for the out-of-sample period. Therefore, models that are the product of data mining are likely to fit very poorly and to give very inaccurate forecasts for the out-of-sample period.

3.8 Goodness of fit statistics

3.8.1 R^2

It is desirable to have some measure of how well the regression model actually fits the data. In other words, it is desirable to have an answer to the question, 'how well does the model containing the explanatory variables that was proposed actually explain variations in the dependent variable?' Quantities known as *goodness of fit statistics* are available to test how well the sample regression function (SRF) fits the data – that is, how 'close' the fitted regression line is to all of the data points taken together. Note that it is not possible to say how well the sample regression function fits the population regression function – i.e. how the estimated model compares with the true relationship between the variables, since the latter is never known.

But what measures might make plausible candidates to be goodness of fit statistics? A first response to this might be to look at the residual sum of squares (*RSS*). Recall that OLS selected the coefficient estimates that minimised this quantity, so the lower was the minimised value of the *RSS*, the better the model fitted the data. Consideration of the *RSS* is certainly one possibility, but *RSS* is unbounded from above (strictly, *RSS* is bounded from above by the total sum of squares – see below) – i.e. it can take any (non-negative) value. So, for example, if the value of the *RSS* under OLS estimation was 136.4, what does this actually mean? It would therefore be very difficult, by looking at this number alone, to tell whether the regression line fitted the data closely or not. The value of *RSS* depends to a great extent on the scale of the dependent variable. Thus, one way to pointlessly reduce the *RSS* would be to divide all of the observations on y by 10!

In fact, a *scaled version* of the residual sum of squares is usually employed. The most common goodness of fit statistic is known as R^2. One way to define R^2 is to say that it is the square of the correlation coefficient between y and $\hat{y}$ – that is, the square of the correlation between the values of the dependent variable and the corresponding fitted values from the model. A correlation coefficient must lie between -1 and $+1$ by definition. Since R^2 defined in this way is the square of a correlation coefficient, it must lie between 0 and 1. If this correlation is high, the model fits the data well, while if the correlation is low (close to zero), the model is not providing a good fit to the data.

Another definition of R^2 requires a consideration of what the model is attempting to explain. What the model is trying to do in effect is to explain variability of y about its mean value, $\bar{y}$. This quantity, $\bar{y}$, which is more specifically known as the unconditional mean of y, acts like a benchmark since, if the researcher had no model for y, he could do no worse than to regress y on a constant only. In fact, the coefficient estimate for this regression would be the mean of y. So, from the regression

$$y_t = \beta_1 + u_t \tag{3.35}$$

the coefficient estimate $\hat{\beta}_1$, will be the mean of y, i.e. $\bar{y}$. The total variation across all observations of the dependent variable about its mean value is known as the total sum of squares, *TSS*, which is given by:

$$TSS = \sum_t (y_t - \bar{y})^2 \tag{3.36}$$

The *TSS* can be split into two parts: the part that has been explained by the model (known as the explained sum of squares, *ESS*) and the part that the model was not able to explain (the *RSS*). That is

$$TSS = ESS + RSS \tag{3.37}$$

$$\sum_t (y_t - \bar{y})^2 = \sum_t (\hat{y}_t - \bar{y})^2 + \sum_t \hat{u}_t^2 \tag{3.38}$$

Recall also that the residual sum of squares can also be expressed as

$$\sum_t (y_t - \hat{y}_t)^2$$

since a residual for observation t is defined as the difference between the actual and fitted values for that observation. The goodness of fit statistic is given by the ratio of the explained sum of squares to the total sum of squares:

$$R^2 = \frac{ESS}{TSS} \tag{3.39}$$

but since $TSS = ESS + RSS$, it is also possible to write

$$R^2 = \frac{ESS}{TSS} = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} \tag{3.40}$$

R^2 must always lie between zero and one (provided that there is a constant term in the regression). This is intuitive from the correlation interpretation of R^2 given above, but for another explanation, consider two extreme cases

$$\begin{array}{llllll} RSS = TSS & \text{i.e.} & ESS = 0 & \text{so} & R^2 = ESS/TSS = 0 \\ ESS = TSS & \text{i.e.} & RSS = 0 & \text{so} & R^2 = ESS/TSS = 1 \end{array}$$

In the first case, the model has not succeeded in explaining any of the variability of y about its mean value, and hence the residual and total sums of squares are equal. This would happen only where the estimated values of all of the coefficients were exactly zero. In the second case, the model has explained all of the variability of y about its mean value, which implies that the residual sum of squares will be zero. This would happen only in the case where all of the observation points lie exactly on the fitted line. Neither of these two extremes is likely in practice, of course, but they do show that R^2 is bounded to lie between zero and one, with a higher R^2 implying, everything else being equal, that the model fits the data better.

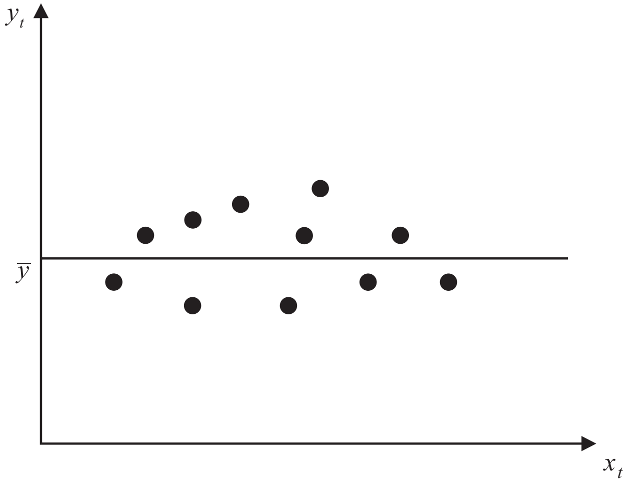


Figure 3.1
$R^2 = 0$ demonstrated by a flat estimated line, i.e. a zero slope coefficient

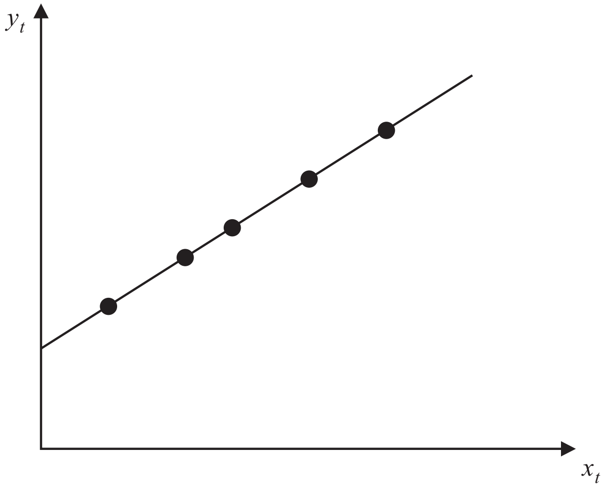


Figure 3.2
$R^2 = 1$ when all data points lie exactly on the estimated line

To sum up, a simple way (but crude, as explained next) to tell whether the regression line fits the data well is to look at the value of R^2. A value of R^2 close to 1 indicates that the model explains nearly all of the variability of the dependent variable about its mean value, while a value close to zero indicates that the model fits the data poorly. The two extreme cases, where $R^2 = 0$ and $R^2 = 1$, are indicated in figures 3.1 and 3.2 in the context of a simple bivariate regression.

3.8.2 *Problems with R^2 as a goodness of fit measure*

R^2 is simple to calculate, intuitive to understand, and provides a broad indication of the fit of the model to the data. However, there are a number of problems with R^2 as a goodness of fit measure:

(1) R^2 is defined in terms of variation about the mean of y so that if a model is reparameterised (rearranged) and the dependent variable changes, R^2 will change, even if the second model was a simple rearrangement of the first, with identical *RSS*. Thus it is not sensible to compare the value of R^2 across models with different dependent variables.

(2) R^2 never falls if more regressors are added to the regression. For example, consider the following two models:

$$\text{Regression 1: } y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + u \tag{3.41}$$

$$\text{Regression 2: } y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + u \tag{3.42}$$

R^2 will always be at least as high for regression 2 relative to regression 1. The R^2 from regression 2 would be exactly the same as that for regression 1 only if the estimated value of the coefficient on the new variable were exactly zero, i.e. $\hat{\beta}_4 = 0$. In practice, $\hat{\beta}_4$ will always be nonzero, even if not significantly so, and thus in practice R^2 always rises as more variables are added to a model. This feature of R^2 essentially makes it impossible to use as a determinant of whether a given variable should be present in the model or not.

(3) R^2 can take values of 0.9 or higher for time series regressions, and hence it is not good at discriminating between models, since a wide array of models will frequently have broadly similar (and high) values of R^2.

3.8.3 Adjusted R^2

In order to get around the second of these three problems, a modification to R^2 is often made which takes into account the loss of degrees of freedom associated with adding extra variables. This is known as $\bar{R}^2$, or adjusted R^2, which is defined as

$$\bar{R}^2 = 1 - \left[\frac{T-1}{T-k}(1 - R^2)\right] \tag{3.43}$$

So if an extra regressor (variable) is added to the model, k increases and unless R^2 increases by a more than off-setting amount, $\bar{R}^2$ will actually fall. Hence $\bar{R}^2$ can be used as a decision-making tool for determining whether a given variable should be included in a regression model or not, with the rule being: include the variable if $\bar{R}^2$ rises and do not include it if $\bar{R}^2$ falls.

However, there are still problems with the maximisation of $\bar{R}^2$ as criterion for model selection, and principal among these is that it is a ‘soft’ rule, implying that by following it, the researcher will typically end up with a large model, containing a lot of marginally significant or insignificant variables. Also, while R^2 must be at least zero if an intercept is included in the regression, its adjusted counterpart may take negative values, even with an intercept in the regression, if the model fits the data very poorly.

Now reconsider the results from the previous exercises using EViews in the previous chapter and earlier in this chapter. If we first consider the hedging model from chapter 2, the R^2 value for the returns regression was only 0.01, indicating that a mere 1% of the variation in spot returns is explained by the futures returns – a very poor model fit indeed.

The fit is no better for the Ford stock CAPM regression described in chapter 2, where the R^2 is less than 1% and the adjusted R^2 is actually negative. The conclusion here would be that for this stock and this sample period, almost none of the monthly movement in the excess returns can be attributed to movements in the market as a whole, as measured by the S&P500.

Finally, if we look at the results from the recent regressions for Microsoft, we find a considerably better fit. It is of interest to compare the model fit for the original regression that included all of the variables with the results of the stepwise procedure. We can see that the raw R^2 is slightly higher for the original regression (0.204 versus 0.200 for the stepwise regression, to three decimal places), exactly as we would expect. Since the original regression contains more variables, the R^2-value must be at least as high. But comparing the $\bar{R}^2$s, the stepwise regression value (0.187) is slightly higher than for the full regression (0.181), indicating that the additional regressors in the full regression do not justify their presence, at least according to this criterion.

Box 3.1 The relationship between the regression F-statistic and R^2

There is a particular relationship between a regression's R^2 value and the regression F-statistic. Recall that the regression F-statistic tests the null hypothesis that all of the regression slope parameters are simultaneously zero. Let us call the residual sum of squares for the unrestricted regression including all of the explanatory variables RSS, while the restricted regression will simply be one of y_t on a constant

$$y_t = \beta_1 + u_t \tag{3.44}$$

Since there are no slope parameters in this model, none of the variability of y_t about its mean value would have been explained. Thus the residual sum of squares for equation (3.44) will actually be the total sum of squares of y_t, TSS. We could write the usual F-statistic formula for testing this null that all of the slope parameters are jointly zero as

$$F - stat = \frac{TSS - RSS}{RSS} \times \frac{T - k}{k - 1} \tag{3.45}$$

In this case, the number of restrictions ('m') is equal to the number of slope parameters, $k - 1$. Recall that $TSS - RSS = ESS$ and dividing the numerator and denominator of equation (3.45) by TSS, we obtain

$$F - stat = \frac{ESS/TSS}{RSS/TSS} \times \frac{T - k}{k - 1} \tag{3.46}$$

Now the numerator of equation (3.46) is R^2, while the denominator is $1 - R^2$, so that the F-statistic can be written

$$F - stat = \frac{R^2(T - k)}{1 - R^2(k - 1)} \tag{3.47}$$

This relationship between the F-statistic and R^2 holds only for a test of this null hypothesis and not for any others.

There now follows another case study of the application of the OLS method of regression estimation, including interpretation of t-ratios and R^2.

3.9 Hedonic pricing models

One application of econometric techniques where the coefficients have a particularly intuitively appealing interpretation is in the area of hedonic pricing models. *Hedonic models* are used to value real assets, especially housing, and view the asset as representing a bundle of characteristics, each of which gives either utility or disutility to its consumer. Hedonic models are often used to produce appraisals or valuations of properties, given their characteristics (e.g. size of dwelling, number of bedrooms, location, number of bathrooms, etc). In these models, the coefficient estimates represent 'prices of the characteristics'.

One such application of a hedonic pricing model is given by Des Rosiers and Thérialt (1996), who consider the effect of various amenities on rental values for buildings and apartments in five sub-markets in the Quebec area of Canada. After accounting for the effect of 'contract-specific' features which will affect rental values (such as whether furnishings, lighting, or hot water are included in the rental price), they arrive at a model where the rental value in Canadian dollars per month (the dependent variable) is a function of 9–14 variables (depending on the area under consideration). The paper employs 1990 data for the Quebec City region, and there are 13,378 observations. The 12 explanatory variables are:

LnAGE	log of the apparent age of the property
NBROOMS	number of bedrooms
AREABYRM	area per room (in square metres)
ELEVATOR	a dummy variable = 1 if the building has an elevator; 0 otherwise
BASEMENT	a dummy variable = 1 if the unit is located in a basement; 0 otherwise
OUTPARK	number of outdoor parking spaces
INDPARK	number of indoor parking spaces
NOLEASE	a dummy variable = 1 if the unit has no lease attached to it; 0 otherwise
LnDISTCBD	log of the distance in kilometres to the central business district (CBD)
SINGLPAR	percentage of single parent families in the area where the building stands
DSHOPCNTR	distance in kilometres to the nearest shopping centre
VACDIFF1	vacancy difference between the building and the census figure

This list includes several variables that are dummy variables. Dummy variables are also known as *qualitative variables* because they are often used to numerically represent a qualitative entity. Dummy variables are usually specified to take on one of a narrow range of integer values, and in most instances only zero and one are used.

Dummy variables can be used in the context of cross-sectional or time series regressions. The latter case will be discussed extensively below. Examples of the use of dummy variables as cross-sectional regressors would be for sex in the context of starting salaries for new traders (e.g. male = 0, female = 1) or in the context of sovereign credit ratings (e.g. developing country = 0, developed country = 1), and so on. In each case, the dummy variables are used in the same way as other explanatory variables and the coefficients on the dummy variables can be interpreted as the average differences in the values of the dependent variable for each category, given all of the other factors in the model.

Des Rosiers and Thérialt (1996) report several specifications for five different regions, and they present results for the model with variables as

Table 3.1 Hedonic model of rental values in Quebec City, 1990. Dependent variable: Canadian dollars per month

Variable	Coefficient	t-ratio	A priori sign expected
Intercept	282.21	56.09	+
LnAGE	−53.10	−59.71	−
NBROOMS	48.47	104.81	+
AREABYRM	3.97	29.99	+
ELEVATOR	88.51	45.04	+
BASEMENT	−15.90	−11.32	−
OUTPARK	7.17	7.07	+
INDPARK	73.76	31.25	+
NOLEASE	−16.99	−7.62	−
LnDISTCBD	5.84	4.60	−
SINGLPAR	−4.27	−38.88	−
DSHOPCNTR	−10.04	−5.97	−
VACDIFF1	0.29	5.98	−

Notes: Adjusted $R^2 = 0.651$; regression *F*-statistic = 2082.27.
Source: Des Rosiers and Thérialt (1996). Reprinted with permission of American Real Estate Society.

discussed here in their exhibit 4, which is adapted and reported here as table 3.1.

The adjusted R^2 value indicates that 65% of the total variability of rental prices about their mean value is explained by the model. For a cross-sectional regression, this is quite high. Also, all variables are significant at the 0.01% level or lower and consequently, the regression *F*-statistic rejects very strongly the null hypothesis that all coefficient values on explanatory variables are zero.

As stated above, one way to evaluate an econometric model is to determine whether it is consistent with theory. In this instance, no real theory is available, but instead there is a notion that each variable will affect rental values in a given direction. The actual signs of the coefficients can be compared with their expected values, given in the last column of table 3.1 (as determined by this author). It can be seen that all coefficients except two (the log of the distance to the CBD and the vacancy differential) have their predicted signs. It is argued by Des Rosiers and Thérialt that the ‘distance to the CBD’ coefficient may be expected to have a positive sign since, while it is usually viewed as desirable to live close to a town centre, everything else being equal, in this instance most of the least desirable neighbourhoods are located towards the centre.

The coefficient estimates themselves show the Canadian dollar rental price per month of each feature of the dwelling. To offer a few illustrations, the NBROOMS value of 48 (rounded) shows that, everything else being equal, one additional bedroom will lead to an average increase in the rental price of the property by $48 per month at 1990 prices. A basement coefficient of −16 suggests that an apartment located in a basement commands a rental $16 less than an identical apartment above ground. Finally the coefficients for parking suggest that on average each outdoor parking space adds $7 to the rent while each indoor parking space adds $74, and so on. The intercept shows, in theory, the rental that would be required of a property that had zero values on all the attributes. This case demonstrates, as stated previously, that the coefficient on the constant term often has little useful interpretation, as it would refer to a dwelling that has just been built, has no bedrooms each of zero size, no parking spaces, no lease, right in the CBD and shopping centre, etc.

One limitation of such studies that is worth mentioning at this stage is their assumption that the implicit price of each characteristic is identical across types of property, and that these characteristics do not become saturated. In other words, it is implicitly assumed that if more and more bedrooms or allocated parking spaces are added to a dwelling indefinitely, the monthly rental price will rise each time by $48 and $7, respectively. This assumption is very unlikely to be upheld in practice, and will result in the estimated model being appropriate for only an ‘average’ dwelling. For example, an additional indoor parking space is likely to add far more value to a luxury apartment than a basic one. Similarly, the marginal value of an additional bedroom is likely to be bigger if the dwelling currently has one bedroom than if it already has ten. One potential remedy for this would be to use dummy variables with fixed effects in the regressions; see, for example, chapter 10 for an explanation of these.

3.10 Tests of non-nested hypotheses

All of the hypothesis tests conducted thus far in this book have been in the context of ‘nested’ models. This means that, in each case, the test involved imposing restrictions on the original model to arrive at a restricted formulation that would be a sub-set of, or nested within, the original specification.

However, it is sometimes of interest to compare between non-nested models. For example, suppose that there are two researchers working independently, each with a separate financial theory for explaining the variation in some variable, y_t. The models selected by the researchers respectively could be

$$y_t = \alpha_1 + \alpha_2 x_{2t} + u_t \tag{3.48}$$
$$y_t = \beta_1 + \beta_2 x_{3t} + v_t \tag{3.49}$$

where u_t and v_t are iid error terms. Model (3.48) includes variable x_2 but not x_3, while model (3.49) includes x_3 but not x_2. In this case, neither model can be viewed as a restriction of the other, so how then can the two models be compared as to which better represents the data, y_t? Given the discussion in section 3.8, an obvious answer would be to compare the values of R^2 or adjusted R^2 between the models. Either would be equally applicable in this case since the two specifications have the same number of RHS variables. Adjusted R^2 could be used even in cases where the number of variables was different across the two models, since it employs a penalty term that makes an allowance for the number of explanatory variables. However, adjusted R^2 is based upon a particular penalty function (that is, $T - k$ appears in a specific way in the formula). This form of penalty term may not necessarily be optimal. Also, given the statement above that adjusted R^2 is a soft rule, it is likely on balance that use of it to choose between models will imply that models with more explanatory variables are favoured. Several other similar rules are available, each having more or less strict penalty terms; these are collectively known as 'information criteria'. These are explained in some detail in chapter 5, but suffice to say for now that a different strictness of the penalty term will in many cases lead to a different preferred model.

An alternative approach to comparing between non-nested models would be to estimate an encompassing or hybrid model. In the case of (3.48) and (3.49), the relevant encompassing model would be

$$y_t = \gamma_1 + \gamma_2 x_{2t} + \gamma_3 x_{3t} + w_t \tag{3.50}$$

where w_t is an error term. Formulation (3.50) contains both (3.48) and (3.49) as special cases when γ_3 and γ_2 are zero, respectively. Therefore, a test for the best model would be conducted via an examination of the significances of γ_2 and γ_3 in model (3.50). There will be four possible outcomes (box 3.2).

However, there are several limitations to the use of encompassing regressions to select between non-nested models. Most importantly, even if models (3.48) and (3.49) have a strong theoretical basis for including the RHS variables that they do, the hybrid model may be meaningless. For example, it could be the case that financial theory suggests that y could either follow model (3.48) or model (3.49), but model (3.50) is implausible.

Box 3.2 Selecting between models

(1) γ_2 is statistically significant but γ_3 is not. In this case, (3.50) collapses to (3.48), and the latter is the preferred model.
(2) γ_3 is statistically significant but γ_2 is not. In this case, (3.50) collapses to (3.49), and the latter is the preferred model.
(3) γ_2 and γ_3 are both statistically significant. This would imply that both x_2 and x_3 have incremental explanatory power for y, in which case both variables should be retained. Models (3.48) and (3.49) are both ditched and (3.50) is the preferred model.
(4) Neither γ_2 nor γ_3 are statistically significant. In this case, none of the models can be dropped, and some other method for choosing between them must be employed.

Also, if the competing explanatory variables x_2 and x_3 are highly related (i.e. they are near collinear), it could be the case that if they are both included, neither γ_2 nor γ_3 are statistically significant, while each is significant in their separate regressions (3.48) and (3.49); see the section on multicollinearity in chapter 4.

An alternative approach is via the J-encompassing test due to Davidson and MacKinnon (1981). Interested readers are referred to their work or to Gujarati (2003, pp. 533–6) for further details.

Key concepts

The key terms to be able to define and explain from this chapter are

- multiple regression model
- restricted regression
- R^2
- hedonic model
- data mining
- variance-covariance matrix
- F-distribution
- $\bar{R}^2$
- encompassing regression

Appendix 3.1 Mathematical derivations of CLRM results

Derivation of the OLS coefficient estimator in the multiple regression context

In the multiple regression context, in order to obtain the parameter estimates, $\beta_1, \beta_2, \ldots, \beta_k$, the RSS would be minimised with respect to all the elements of β. Now the residuals are expressed in a vector:

$$\hat{u} = \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \vdots \\ \hat{u}_T \end{bmatrix} \tag{3A.1}$$

The *RSS* is still the relevant loss function, and would be given in a matrix notation by expression (3A.2)

$$L = \hat{u}'\hat{u} = [\hat{u}_1\hat{u}_2 \dots \hat{u}_T] \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \vdots \\ \hat{u}_T \end{bmatrix} = \hat{u}_1^2 + \hat{u}_2^2 + \cdots + \hat{u}_T^2 = \sum \hat{u}_t^2 \tag{3A.2}$$

Denoting the vector of estimated parameters as $\hat{\beta}$, it is also possible to write

$$L = \hat{u}'\hat{u} = (y - X\hat{\beta})'(y - X\hat{\beta}) = y'y - \hat{\beta}'X'y - y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta} \tag{3A.3}$$

It turns out that $\hat{\beta}'X'y$ is $(1 \times k) \times (k \times T) \times (T \times 1) = 1 \times 1$, and also that $y'X\hat{\beta}$ is $(1 \times T) \times (T \times k) \times (k \times 1) = 1 \times 1$, so in fact $\hat{\beta}'X'y = y'X\hat{\beta}$. Thus (3A.3) can be written

$$L = \hat{u}'\hat{u} = (y - X\hat{\beta})'(y - X\hat{\beta}) = y'y - 2\hat{\beta}'X'y + \hat{\beta}'X'X\hat{\beta} \tag{3A.4}$$

Differentiating this expression with respect to $\hat{\beta}$ and setting it to zero in order to find the parameter values that minimise the residual sum of squares would yield

$$\frac{\partial L}{\partial \hat{\beta}} = -2X'y + 2X'X\hat{\beta} = 0 \tag{3A.5}$$

This expression arises since the derivative of $y'y$ is zero with respect to $\hat{\beta}$, and $\hat{\beta}'X'X\hat{\beta}$ acts like a square of $X\hat{\beta}$, which is differentiated to $2X'X\hat{\beta}$. Rearranging (3A.5)

$$2X'y = 2X'X\hat{\beta} \tag{3A.6}$$

$$X'y = X'X\hat{\beta} \tag{3A.7}$$

Pre-multiplying both sides of (3A.7) by the inverse of $X'X$

$$\hat{\beta} = (X'X)^{-1}X'y \tag{3A.8}$$

Thus, the vector of OLS coefficient estimates for a set of k parameters is given by

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = (X'X)^{-1}X'y \tag{3A.9}$$

Derivation of the OLS standard error estimator in the multiple regression context

The variance of a vector of random variables $\hat{\beta}$ is given by the formula $E[(\hat{\beta}-\beta)(\hat{\beta}-\beta)']$. Since $y = X\beta + u$, it can also be stated, given (3A.9), that

$$\hat{\beta} = (X'X)^{-1}X'(X\beta + u) \tag{3A.10}$$

Expanding the parentheses

$$\hat{\beta} = (X'X)^{-1}X'X\beta + (X'X)^{-1}X'u \tag{3A.11}$$
$$\hat{\beta} = \beta + (X'X)^{-1}X'u \tag{3A.12}$$

Thus, it is possible to express the variance of $\hat{\beta}$ as

$$E[(\hat{\beta}-\beta)(\hat{\beta}-\beta)'] = E[(\beta + (X'X)^{-1}X'u - \beta)(\beta + (X'X)^{-1}X'u - \beta)'] \tag{3A.13}$$

Cancelling the β terms in each set of parentheses

$$E[(\hat{\beta}-\beta)(\hat{\beta}-\beta)'] = E[((X'X)^{-1}X'u)((X'X)^{-1}X'u)'] \tag{3A.14}$$

Expanding the parentheses on the RHS of (3A.14) gives

$$E[(\hat{\beta}-\beta)(\hat{\beta}-\beta)'] = E[(X'X)^{-1}X'uu'X(X'X)^{-1}] \tag{3A.15}$$
$$E[(\hat{\beta}-\beta)(\hat{\beta}-\beta)'] = (X'X)^{-1}X'E[uu']X(X'X)^{-1} \tag{3A.16}$$

Now $E[uu']$ is estimated by s^2I, so that

$$E[(\hat{\beta}-\beta)(\hat{\beta}-\beta)'] = (X'X)^{-1}X's^2IX(X'X)^{-1} \tag{3A.17}$$

where I is a $k \times k$ identity matrix. Rearranging further,

$$E[(\hat{\beta}-\beta)(\hat{\beta}-\beta)'] = s^2(X'X)^{-1}X'X(X'X)^{-1} \tag{3A.18}$$

The $X'X$ and the last $(X'X)^{-1}$ term cancel out to leave

$$\text{var}(\hat{\beta}) = s^2(X'X)^{-1} \tag{3A.19}$$

as the expression for the parameter variance–covariance matrix. This quantity, $s^2(X'X)^{-1}$, is known as the estimated variance–covariance matrix of the coefficients. The leading diagonal terms give the estimated coefficient variances while the off-diagonal terms give the estimated covariances between the parameter estimates. The variance of $\hat{\beta}_1$ is the first diagonal element, the variance of $\hat{\beta}_2$ is the second element on the leading diagonal, $\ldots$, and the variance of $\hat{\beta}_k$ is the kth diagonal element, etc. as discussed in the body of the chapter.

Appendix 3.2 A brief introduction to factor models and principal components analysis

Factor models are employed primarily as dimensionality reduction techniques in situations where we have a large number of closely related variables and where we wish to allow for the most important influences from all of these variables at the same time. Factor models decompose the structure of a set of series into factors that are common to all series and a proportion that is specific to each series (idiosyncratic variation). There are broadly two types of such models, which can be loosely characterised as either macroeconomic or mathematical factor models. The key distinction between the two is that the factors are observable for the former but are latent (unobservable) for the latter. Observable factor models include the APT model of Ross (1976). The most common mathematical factor model is principal components analysis (PCA). PCA is a technique that may be useful where explanatory variables are closely related – for example, in the context of near multicollinearity. Specifically, if there are k explanatory variables in the regression model, PCA will transform them into k uncorrelated new variables. To elucidate, suppose that the original explanatory variables are denoted $x_1, x_2, \ldots, x_k$, and denote the principal components by $p_1, p_2, \ldots, p_k$. These principal components are independent linear combinations of the original data

$$\begin{aligned} p_1 &= \alpha_{11}x_1 + \alpha_{12}x_2 + \cdots + \alpha_{1k}x_k \\ p_2 &= \alpha_{21}x_1 + \alpha_{22}x_2 + \cdots + \alpha_{2k}x_k \\ \ldots \quad &\ldots \qquad \ldots \qquad\qquad \ldots \\ p_k &= \alpha_{k1}x_1 + \alpha_{k2}x_2 + \cdots + \alpha_{kk}x_k \end{aligned} \tag{3A.20}$$

where α_{ij} are coefficients to be calculated, representing the coefficient on the jth explanatory variable in the ith principal component. These coefficients are also known as factor loadings. Note that there will be T observations on each principal component if there were T observations on each explanatory variable.

It is also required that the sum of the squares of the coefficients for each component is one, i.e.

$$\begin{aligned} &\alpha_{11}^2 + \alpha_{12}^2 + \cdots + \alpha_{1k}^2 = 1 \\ &\qquad\quad \vdots \quad\; \vdots \\ &\alpha_{k1}^2 + \alpha_{k2}^2 + \cdots + \alpha_{kk}^2 = 1 \end{aligned} \tag{3A.21}$$

This requirement could also be expressed using sigma notation

$$\sum_{j=1}^{k} \alpha_{ij}^2 = 1 \quad \forall \quad i = 1, \ldots, k \tag{3A.22}$$

Constructing the components is a purely mathematical exercise in constrained optimisation, and thus no assumption is made concerning the structure, distribution, or other properties of the variables.

The principal components are derived in such a way that they are in descending order of importance. Although there are k principal components, the same as the number of explanatory variables, if there is some collinearity between these original explanatory variables, it is likely that some of the (last few) principal components will account for so little of the variation that they can be discarded. However, if all of the original explanatory variables were already essentially uncorrelated, all of the components would be required, although in such a case there would have been little motivation for using PCA in the first place.

The principal components can also be understood as the eigenvalues of $(X'X)$, where X is the matrix of observations on the original variables. Thus the number of eigenvalues will be equal to the number of variables, k. If the ordered eigenvalues are denoted λ_i $(i = 1, \ldots, k)$, the ratio

$$\phi_i = \frac{\lambda_i}{\sum_{i=1}^{k} \lambda_i}$$

gives the proportion of the total variation in the original data explained by the principal component i. Suppose that only the first $r\,(0 < r < k)$ principal components are deemed sufficiently useful in explaining the variation of $(X'X)$, and that they are to be retained, with the remaining $k - r$ components being discarded. The regression finally estimated, after the principal components have been formed, would be one of y on the r principal components

$$y_t = \gamma_0 + \gamma_1 p_{1t} + \cdots + \gamma_r p_{rt} + u_t \tag{3A.23}$$

In this way, the principal components are argued to keep most of the important information contained in the original explanatory variables, but are orthogonal. This may be particularly useful for independent variables that are very closely related. The principal component estimates $(\hat{\gamma}_i, i = 1, \ldots, r)$ will be biased estimates, although they will be more efficient than the OLS estimators since redundant information has been removed. In fact, if the OLS estimator for the original regression of y on x is denoted $\hat{\beta}$, it can be shown that

$$\hat{\gamma}_r = P_r'\hat{\beta} \tag{3A.24}$$

where $\hat{\gamma}_r$ are the coefficient estimates for the principal components, and P_r is a matrix of the first r principal components. The principal component coefficient estimates are thus simply linear combinations of the original OLS estimates.

An application of principal components to interest rates

Many economic and financial models make use of interest rates in some form or another as independent variables. Researchers may wish to include interest rates on a large number of different assets in order to reflect the variety of investment opportunities open to investors. However, market interest rates could be argued to be not sufficiently independent of one another to make the inclusion of several interest rate series in an econometric model statistically sensible. One approach to examining this issue would be to use PCA on several related interest rate series to determine whether they did move independently of one another over some historical time period or not.

Fase (1973) conducted such a study in the context of monthly Dutch market interest rates from January 1962 until December 1970 (108 months). Fase examined both ‘money market’ and ‘capital market’ rates, although only the money market results will be discussed here in the interests of brevity. The money market instruments investigated were:

- Call money
- Three-month Treasury paper
- One-year Treasury paper
- Two-year Treasury paper
- Three-year Treasury paper
- Five-year Treasury paper
- Loans to local authorities: three-month
- Loans to local authorities: one-year
- Eurodollar deposits
- Netherlands Bank official discount rate.

Prior to analysis, each series was standardised to have zero mean and unit variance by subtracting the mean and dividing by the standard deviation in each case. The three largest of the ten eigenvalues are given in table 3A.1.

Table 3A.1 Principal component ordered eigenvalues for Dutch interest rates, 1962–1970

	Monthly data			*Quarterly data*
	Jan 62–Dec 70	Jan 62–Jun 66	Jul 66–Dec 70	Jan 62–Dec 70
λ_1	9.57	9.31	9.32	9.67
λ_2	0.20	0.31	0.40	0.16
λ_3	0.09	0.20	0.17	0.07
ϕ_1	95.7%	93.1%	93.2%	96.7%

Source: Fase (1973). Reprinted with the permission of Elsevier Science.

Table 3A.2 Factor loadings of the first and second principal components for Dutch interest rates, 1962–1970

j	Debt instrument	α_{j1}	α_{j2}
1	Call money	0.95	−0.22
2	3-month Treasury paper	0.98	0.12
3	1-year Treasury paper	0.99	0.15
4	2-year Treasury paper	0.99	0.13
5	3-year Treasury paper	0.99	0.11
6	5-year Treasury paper	0.99	0.09
7	Loans to local authorities: 3-month	0.99	−0.08
8	Loans to local authorities: 1-year	0.99	−0.04
9	Eurodollar deposits	0.96	−0.26
10	Netherlands Bank official discount rate	0.96	−0.03
	Eigenvalue, λ_i	9.57	0.20
	Proportion of variability explained by eigenvalue i, $\phi_i(\%)$	95.7	2.0

Source: Fase (1973). Reprinted with the permission of Elsevier Science.

The results in table 3A.1 are presented for the whole period using the monthly data, for two monthly sub-samples, and for the whole period using data sampled quarterly instead of monthly. The results show clearly that the first principal component is sufficient to describe the common variation in these Dutch interest rate series. The first component is able to explain over 90% of the variation in all four cases, as given in the last row of table 3A.1. Clearly, the estimated eigenvalues are fairly stable across the sample periods and are relatively invariant to the frequency of sampling of the data. The factor loadings (coefficient estimates) for the first two ordered components are given in table 3A.2.

As table 3A.2 shows, the loadings on each factor making up the first principal component are all positive. Since each series has been standardised to have zero mean and unit variance, the coefficients α_{j1} and α_{j2} can be interpreted as the correlations between the interest rate j and the first and second principal components, respectively. The factor loadings for each interest rate series on the first component are all very close to one. Fase (1973) therefore argues that the first component can be interpreted simply as an equally weighted combination of all of the market interest rates. The second component, which explains much less of the variability of the rates, shows a factor loading pattern of positive coefficients for the Treasury paper series and negative or almost zero values for the other series. Fase (1973) argues that this is owing to the characteristics of the Dutch Treasury instruments that they rarely change hands and have low transactions costs, and therefore have less sensitivity to general interest rate movements. Also, they are not subject to default risks in the same way as, for example Eurodollar deposits. Therefore, the second principal component is broadly interpreted as relating to default risk and transactions costs.

Principal components can be useful in some circumstances, although the technique has limited applicability for the following reasons:

- A change in the units of measurement of x will change the principal components. It is thus usual to transform all of the variables to have zero mean and unit variance prior to applying PCA.
- The principal components usually have no theoretical motivation or interpretation whatsoever.
- The r principal components retained from the original k are the ones that explain most of the variation in x, but these components might not be the most useful as explanations for y.

Calculating principal components in EViews

In order to calculate the principal components of a set of series with EViews, the first stage is to compile the series concerned into a group. **Re-open the 'macro.wf1' file** which contains US Treasury bill and bond series of various maturities. Select **New Object/Group** but do not name the object. When EViews prompts you to give a 'List of series, groups and/or series expressions', enter

USTB3M USTB6M USTB1Y USTB3Y USTB5Y USTB10Y

and click **OK**, then name the group **Interest** by clicking the **Name** tab. The group will now appear as a set of series in a spreadsheet format. From within this window, click **View/Principal Components**. Screenshot 3.2 will appear.

There are many features of principal components that can be examined, but for now keep the defaults and click **OK**. The results will appear as in the following table.

Principal Components Analysis
Date: 08/31/07 Time: 14:45
Sample: 1986M03 2007M04
Included observations: 254
Computed using: Ordinary correlations
Extracting 6 of 6 possible components

Eigenvalues: (Sum = 6, Average = 1)

Number	Value	Difference	Proportion	Cumulative Value	Cumulative Proportion
1	5.645020	5.307297	0.9408	5.645020	0.9408
2	0.337724	0.323663	0.0563	5.982744	0.9971
3	0.014061	0.011660	0.0023	5.996805	0.9995
4	0.002400	0.001928	0.0004	5.999205	0.9999
5	0.000473	0.000150	0.0001	5.999678	0.9999
6	0.000322	–	0.0001	6.000000	1.0000

Eigenvectors (loadings):

Variable	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6
USTB3M	0.405126	−0.450928	0.556508	−0.407061	0.393026	−0.051647
USTB6M	0.409611	−0.393843	0.084066	0.204579	−0.746089	0.267466
USTB1Y	0.415240	−0.265576	−0.370498	0.577827	0.335650	−0.416211
USTB3Y	0.418939	0.118972	−0.540272	−0.295318	0.243919	0.609699
USTB5Y	0.410743	0.371439	−0.159996	−0.461981	−0.326636	−0.589582
USTB10Y	0.389162	0.647225	0.477986	0.3973990	0.100167	0.182274

Ordinary correlations:

	USTB3M	USTB6M	USTB1Y	USTB3Y	USTB5Y	USTB10Y
USTB3M	1.000000					
USTB6M	0.997052	1.000000				
USTB1Y	0.986682	0.995161	1.000000			
USTB3Y	0.936070	0.952056	0.973701	1.000000		
USTB5Y	0.881930	0.899989	0.929703	0.987689	1.000000	
USTB10Y	0.794794	0.814497	0.852213	0.942477	0.981955	1.000000

It is evident that there is a great deal of common variation in the series, since the first principal component captures 94% of the variation in the series and the first two components capture 99.7%. Consequently, if we wished, we could reduce the dimensionality of the system by using two components rather than the entire six interest rate series. Interestingly,

Screenshot 3.2 Conducting PCA in EViews

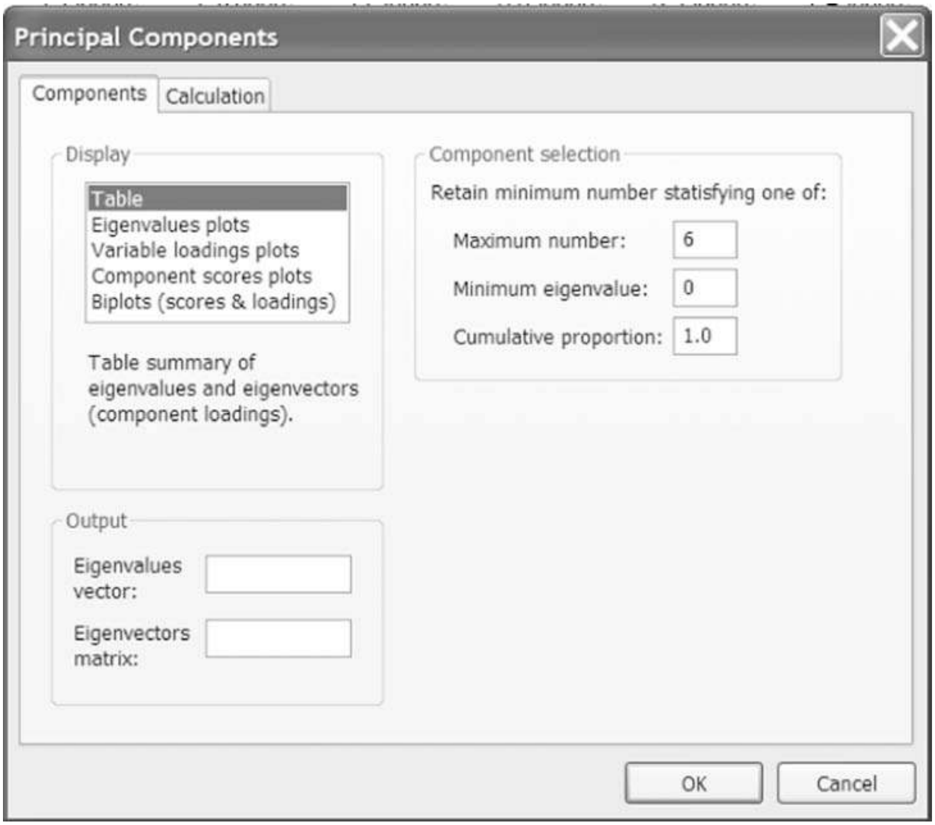


Table summary of eigenvalues and eigenvectors (component loadings).

Retain minimum number statisfying one of:

Maximum number: 6

Minimum eigenvalue: 0

Cumulative proportion: 1.0

Eigenvalues vector:

Eigenvectors matrix:

the first component comprises almost exactly equal weights in all six series.

Then **Minimise this group** and you will see that the ‘Interest’ group has been added to the list of objects.

Review questions

1. By using examples from the relevant statistical tables, explain the relationship between the t- and the F-distributions.

For questions 2–5, assume that the econometric model is of the form

$$y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \beta_4 x_{4t} + \beta_5 x_{5t} + u_t \tag{3.51}$$

2. Which of the following hypotheses about the coefficients can be tested using a t-test? Which of them can be tested using an F-test? In each case, state the number of restrictions.

   (a) $H_0 : \beta_3 = 2$
   (b) $H_0 : \beta_3 + \beta_4 = 1$

(c) $H_0 : \beta_3 + \beta_4 = 1 \text{ and } \beta_5 = 1$
(d) $H_0 : \beta_2 = 0 \text{ and } \beta_3 = 0 \text{ and } \beta_4 = 0 \text{ and } \beta_5 = 0$
(e) $H_0 : \beta_2\beta_3 = 1$

3. Which of the above null hypotheses constitutes 'THE' regression F-statistic in the context of (3.51)? Why is this null hypothesis always of interest whatever the regression relationship under study? What exactly would constitute the alternative hypothesis in this case?
4. Which would you expect to be bigger – the unrestricted residual sum of squares or the restricted residual sum of squares, and why?
5. You decide to investigate the relationship given in the null hypothesis of question 2, part (c). What would constitute the restricted regression? The regressions are carried out on a sample of 96 quarterly observations, and the residual sums of squares for the restricted and unrestricted regressions are 102.87 and 91.41, respectively. Perform the test. What is your conclusion?
6. You estimate a regression of the form given by (3.52) below in order to evaluate the effect of various firm-specific factors on the returns of a sample of firms. You run a cross-sectional regression with 200 firms

$$r_i = \beta_0 + \beta_1 S_i + \beta_2 MB_i + \beta_3 PE_i + \beta_4 BETA_i + u_i \tag{3.52}$$

where: r_i is the percentage annual return for the stock
S_i is the size of firm i measured in terms of sales revenue
MB_i is the market to book ratio of the firm
PE_i is the price/earnings (P/E) ratio of the firm
$BETA_i$ is the stock's CAPM beta coefficient

You obtain the following results (with standard errors in parentheses)

$$\begin{array}{llllll} \hat{r}_i = & 0.080 & + 0.801S_i & + 0.321MB_i & + 0.164PE_i & - 0.084BETA_i \\ & (0.064) & (0.147) & (0.136) & (0.420) & (0.120) \end{array} \tag{3.53}$$

Calculate the t-ratios. What do you conclude about the effect of each variable on the returns of the security? On the basis of your results, what variables would you consider deleting from the regression? If a stock's beta increased from 1 to 1.2, what would be the expected effect on the stock's return? Is the sign on beta as you would have expected? Explain your answers in each case.

7. A researcher estimates the following econometric models including a lagged dependent variable

$$y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \beta_4 y_{t-1} + u_t \tag{3.54}$$

$$\Delta y_t = \gamma_1 + \gamma_2 x_{2t} + \gamma_3 x_{3t} + \gamma_4 y_{t-1} + v_t \tag{3.55}$$

where u_t and v_t are iid disturbances.

Will these models have the same value of (a) The residual sum of squares (*RSS*), (b) R^2, (c) Adjusted R^2? Explain your answers in each case.

8. A researcher estimates the following two econometric models

$$y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + u_t \tag{3.56}$$

$$y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \beta_4 x_{4t} + v_t \tag{3.57}$$

where u_t and v_t are iid disturbances and x_{3t} is an irrelevant variable which does not enter into the data generating process for y_t. Will the value of (a) R^2, (b) Adjusted R^2, be higher for the second model than the first? Explain your answers.

9. Re-open the CAPM Eviews file and estimate CAPM betas for each of the other stocks in the file.
   (a) Which of the stocks, on the basis of the parameter estimates you obtain, would you class as defensive stocks and which as aggressive stocks? Explain your answer.
   (b) Is the CAPM able to provide any reasonable explanation of the overall variability of the returns to each of the stocks over the sample period? Why or why not?

10. Re-open the Macro file and apply the same APT-type model to some of the other time-series of stock returns contained in the CAPM-file.
    (a) Run the stepwise procedure in each case. Is the same sub-set of variables selected for each stock? Can you rationalise the differences between the series chosen?
    (b) Examine the sizes and signs of the parameters in the regressions in each case – do these make sense?

11. What are the units of R^2?