1

Review of the general linear model

This chapter reviews some of the standard theory of estimation of econometric relationships which makes up many econometric courses. We show how a set of assumptions regarding the structure of the model lead ordinary least squares (OLS) to be an optimal estimator and how the failure of these assumptions can produce highly misleading results. This chapter sets the scene for much of the rest of the book as later chapters focus both on the problems which arise when these assumptions are violated and more importantly on the range of new techniques which have been developed for dealing with these problems.

1.1 Economic and statistical models

We may define an *economic* model as one that has some basis in economic theory. Economic theory usually (but not exclusively) yields static, or 'long-run' relationships. For example, in the simple Keynesian consumption function, consumption at time t, y_t say, is assumed proportional to income, x_t say. If we assume instantaneous adjustment of y to x, we may write

$$y_t = \beta_1 + \beta_2 x_t + \varepsilon_t \tag{1.1}$$

where ε_t is a random error term which encapsulates deviations from the model; we discuss various possible properties of ε_t below. It may be possible, however, to obtain a good approximation to the behaviour of y without recourse to *any* economic theory. A simple, pure time series model of consumption might be a univariate autoregressive model of order one – an AR(1):

$$y_t = \alpha_1 + \alpha_2 y_{t-1} + \varepsilon_t \tag{1.2}$$

It may be the case that *some* economic theory is consistent with equation (1.2), but a pure time series modeller need not be concerned with this. Although an 'economic modeller' and a 'time series modeller' may end up with similar statistical models, their aims will usually be subtly different. The time series modeller is aiming for a succinct summary of the time series behaviour of the variable. Typically, the applied economist will want to go further than this, to test some kind of *restrictions* on the time series model, as a test of an economic hypothesis. For example, the life-cycle theory of consumption under rational expectations would suggest that a univariate model of consumption of the form (1.2) should hold, with $\alpha_2 = 1$.

1.2 Time series and stochastic processes

A *stochastic process* is a sequence of random variables – any one element of the sequence may take on any of a range of values in any particular realisation. Thus, if I plan to roll a fair, six-sided die every morning before breakfast next week, then I can imagine seven random variables (each morning's score) associated with this activity which together form a stochastic process. If I denote the number of dots uppermost on the *i*th day as d_i then the sequence $(d_i)_{i=1\ldots7}$ denotes a stochastic process. If a stochastic process has one element for each of a set of points in time, then any realisation of the stochastic process is a *time series*. Thus, if the number of dots uppermost on the die each morning was as follows: Monday 1, Tuesday 3, Wednesday 5, Thursday 5, Friday 2, Saturday 4, Sunday 4; then the sequence (1,3,5,5,2,4,4) denotes a time series. Any element of a stochastic process is a *random variable*. Any element of a time series is a *number* which is referred to as an *observation*. In general, when econometricians speak of *modelling* a time series, they mean the act of postulating a *stochastic process* which may have generated the observed time series. Following standard practice, we shall, where there is no possibility of confusion, use the same notation to denote a stochastic process, a time series or an element of either.

1.3 Properties of stochastic processes

In the early morning die-rolling example given above, the *sample* mean is just the mean of the observed time series (which is 24/7); the

population mean is the expected value of any element of the stochastic process (which is 21/6). Roughly speaking, if a process is *ergodic*, then its *moments* (i.e. mean, variance, etc.) can be estimated 'well' (or, to be precise, *consistently* – see below) by the corresponding moments of the observed time series over a long period of time. Consider the following AR(1) model for y:

$$y_t = \beta y_{t-1} + \varepsilon_t \tag{1.3}$$

where ε_t is a zero–mean random variable with constant variance σ^2, which is uncorrelated with any other variable in the sequence $(\varepsilon_t)_{t=-\infty}^{+\infty}$, i.e.

$$E(\varepsilon_t) = 0 \tag{1.4a}$$

$$\mathrm{Var}(\varepsilon_t) = E(\varepsilon_t^2) = \sigma_\varepsilon^2 \tag{1.4b}$$

$$\mathrm{Cov}(\varepsilon_t, \varepsilon_{t-j}) = 0, \text{ for all } j \neq 0 \tag{1.4c}$$

A stochastic process displaying these properties is often referred to as *white noise*. A white noise process is a special case of a more general class of stochastic processes, namely those which are *stationary*. A *covariance stationary* stochastic process, y_t say, has a constant mean and variance and the covariation between any two elements in the sequence is a function only of the distance in time between the two elements:

$$E(y_t) = \mu \tag{1.5a}$$

$$\mathrm{Var}(y_t) = E[(y_t - \mu)^2] = \gamma(0) < \infty \tag{1.5b}$$

$$\mathrm{Cov}(y_t, y_{t-j}) = \gamma(j) \quad \text{for all } j \tag{1.5c}$$

A stochastic process is *strictly stationary* if the joint probability of any consecutive r observations is always the same, for any integer r. In this book we shall generally use the term *stationary* to refer to *weak* or *covariance* stationarity. Note that, if a process is both covariance stationary and normally distributed, then it is also strictly stationary.

Equation (1.3) can be written in the form

$$(1 - \beta L)y_t = \varepsilon_t \tag{1.6}$$

where L is the lag operator, which has the property:

$$L^m y_t = y_{t-m}$$

and $(1 - \beta L)$ is thus a polynomial of order one in the lag operator. If we lag (1.3) by one period, we have

$$y_{t-1} = \beta y_{t-2} + \varepsilon_{t-1} \tag{1.7}$$

Substituting equation (1.7) into (1.3):

$$y_t = \beta^2 y_{t-2} + \varepsilon_t + \beta\varepsilon_{t-1} \quad \textbf{(1.8)}$$

If we now lag (1.3) twice [i.e. lag (1.7) once] and substitute into equation (1.8) we have an expression in y_{t-3}, ε_t, ε_{t-1} and ε_{t-2}. Continually substituting for lagged values of y in this fashion we have, after $n - 1$ substitutions:

$$y_t = \beta^n y_{t-n} + \varepsilon_t + \beta\varepsilon_{t-1} + \beta^2\varepsilon_{t-2} + \beta^3\varepsilon_{t-3} + \ldots + \beta^{n-1}\varepsilon_{t-n+1} \quad \textbf{(1.9)}$$

If β is less than one in absolute value, $|\beta| < 1$, then as n gets bigger and bigger (tends towards infinity), β^n gets smaller and smaller (tends towards zero). Thus, for large n we can write:

$$y_t = \varepsilon_t + \beta\varepsilon_{t-1} + \beta^2\varepsilon_{t-2} + \beta^3\varepsilon_{t-3} + \ldots \quad \textbf{(1.10a)}$$

or

$$y_t = [1 + \beta L + (\beta L)^2 + (\beta L)^3 + \ldots]\varepsilon_t \quad \textbf{(1.10b)}$$

where we have again used the lag operator. Multiplying both sides of equation (1.10b) by βL and subtracting the resulting expression from (1.10a) gives

$$y_t(1 - \beta L) = \varepsilon_t \quad \textbf{(1.11a)}$$

or

$$y_t = (1 - \beta L)^{-1}\varepsilon_t \quad \textbf{(1.11b)}$$

Since ε_t is a white noise process, (1.10a) implies the following:

$$E(y_t) = E(\varepsilon_t) + \beta E(\varepsilon_{t-1}) + \beta^2 E(\varepsilon_{t-2}) + \ldots = 0 \quad \textbf{(1.12a)}$$

$$\mathrm{Var}(y_t) = E(y_t^2) = (1 + \beta^2 + \beta^4 + \beta^6 + \ldots)\sigma_\varepsilon^2$$
$$= (1 - \beta^2)^{-1}\sigma_\varepsilon^2 \quad \textbf{(1.12b)}$$

$$\mathrm{Cov}(y_t, y_{t-j}) = E[(\varepsilon_t + \beta\varepsilon_{t-1} + \beta^2\varepsilon_{t-2} + \ldots) \times (\varepsilon_{t-j} + \beta\varepsilon_{t-j-1} + \beta^2\varepsilon_{t-j-2} + \ldots)]$$
$$= \beta^{j2} E(y_t^2) \quad \textbf{(1.12c)}$$

Comparing (1.12) with (1.5), we can see that the AR(1) process (1.3) is stationary for $|\beta| < 1$.

Virtually the whole of standard econometric theory is based on the assumption that the processes under examination are stationary. However many economic time series – particularly macroeconomic and financial time series – appear to be generated by non-stationary processes. Recently, however, a body of literature has developed which deals with non-stationary processes directly. This will be the subject matter of Chapter 5.

1.4 Properties of estimators

Econometrics is largely to do with estimating the parameters of economic relationships and testing hypotheses with respect to those parameters. For example, consider again the simple, linear Keynesian consumption function relating consumption, y, to income, x:

$$y_t = \beta_1 + \beta_2 x_t + \varepsilon_t \quad \textbf{(1.13)}$$

Economic theory suggests the form of the consumption function (Keynes's 'fundamental psychological law' – Keynes 1936), and may even suggest qualitative restrictions on the parameters. For example, since β_1 is equal to autonomous consumption and β_2 is the marginal propensity to consume, we can infer:

$$\beta_1 \geq 0,\ 0 \leq \beta_2 \leq 1 \quad \textbf{(1.14a)}$$

In general, however, economic theory will be silent on the *exact* values of the parameters of a model. Moreover, even when an exact value of a parameter *is* suggested by economic theory, an econometrician may still want to estimate it to see if the data is in accordance with the theory. Econometrics can thus be used to obtain estimates of unknown parameters in empirical economic models and to test hypotheses with respect to them.

For example, Davis (1952) uses annual data for the United States for the period 1929–40 (deflated for price and population changes) and estimates the parameters in (1.13) as:

$$y_t = 11.45 + 0.78x_t \quad \textbf{(1.14b)}$$

Thus, Davis's estimate of β_1 ('autonomous consumption') is 11.45 and of β_2 (the 'marginal propensity to consume') is 0.78. These are numbers. To obtain these *estimates*, Davis used formulae suggested by econometric theory. These formulae are *estimators*.

There is an infinite number of estimators, all but a few of which are unacceptable to an econometrician. For example, a particularly silly estimator could be obtained simply by writing down the number of the day of the month. Whilst it may be *obvious* that such an estimator is silly, there are other estimators which are not obviously so. Thus, we need a formal set of criteria by which to judge an estimator.

Sampling distributions

Consider the model

$$y_t = \beta x_t + \varepsilon_t \tag{1.15}$$

where ε_t is assumed to be white noise. Equation (1.15) defines an assumed data-generating process for y_t. Suppose we have observed time series for y and x. For any given estimator of β, β^* say, we can construct an estimate using these observed time series. But since time series are realisations of stochastic processes, it is equally possible that different realisations, i.e. time series, could have been obtained. Theoretically, we can consider how the estimate given by the estimator will vary according to different realisations – this is the basis for the *sampling distribution* of an econometric estimator. The sampling distribution simply allows us to calculate the probability of observing an estimate within a given interval, i.e. it is the frequency distribution of the estimator.

For concreteness, suppose that x is in fact non-stochastic – for example, it may be a time trend. Then we could carry out a *Monte Carlo* experiment whereby say, 2000 series for ε were generated using a random number generator. Given the series for x, equation (1.15) then implies 2000 time series for y – we simply *fix* β at a number, e.g. 2.5. Since the true value of β is known in the experiment, we can then see how the estimator behaves with respect to it in *repeated samples* by constructing, say, 2000 estimates of β (i.e. realisations of β^*). The manner in which these estimates differ is called the *empirical sampling distribution*, which could be approximated by constructing a histogram of the estimates. Monte Carlo studies are often used to construct empirical sampling distributions where the model or the estimator is particularly complex, or its behaviour is known only in very large samples. Often, however, we can deduce the properties of the sampling distribution from the assumptions we have made concerning the model.

Econometricians normally judge the quality of an estimator by considering the properties of its sampling distribution. In particular, an estimator will clearly be more attractive if there is a high, rather than a low probability that it will yield an estimate that is close to the true (but unknown) value of the parameter which is being estimated.

Unbiasedness

The first property we consider is *unbiasedness*. An estimator is unbiased if the mean of its sampling distribution is in fact the true value of the parameter being estimated. This *does not* mean that, 'on average' we should expect an unbiased estimator to yield the true value of the parameter vector, since the sampling distribution is continuous, the probability of this happening is in fact zero.

An alternative way of thinking about this property is to consider the *bias* of an estimator. The bias is the difference between the mean of the sampling distribution – the expected value of the estimator – and the true value:

$$B = E(\beta^*) - \beta \tag{1.16}$$

Consider the sampling distributions of two univariate estimators: β^* which is unbiased but which has a large variance and $\tilde{\beta}$ which has a small degree of bias but with a very small variance. Because the variance of the sampling distribution of $\tilde{\beta}$ is smaller than the sampling distribution of β^*, $\tilde{\beta}$ is more *efficient* than β^*. Thus it is probable that $\tilde{\beta}$ will yield an estimate closer to β than β^* in any particular realisation. This example shows very clearly that the variance, as well as the mean of the sampling distribution should be considered when assessing the quality of an estimator.

Best unbiased

The preceding discussion illustrated the importance of considering the variance as well as the mean of the sampling distribution. In general, we should choose estimators which have 'low' variance. It is, however, almost meaningless to speak of a 'minimum variance' estimator. Suppose, for example that whenever a model with one parameter was being considered, we used the estimator $\beta^* = 103.9$ – *regardless* of the context, or the data, or whatever. Because this estimator never varies, its variance is zero, the smallest possible, notwithstanding its patent silliness. For this reason, it is necessary to qualify the search for low variance. Normally, this is done by considering only estimators which are unbiased. Consider the sampling distribution of two unbiased estimators, one of which, β^*, has lower variance than the other, $\tilde{\beta}$. Clearly, β^*, the more efficient estimator, is more likely to yield an estimate closer to the true value of the parameter than is $\tilde{\beta}$.

An estimator which has the lowest variance – is the most efficient – within a certain class of estimators is said to be the *best* estimator in that class. As we shall see in Chapter 2, there is a general principle for choosing estimators, the maximum likelihood principle, which will always give the best unbiased estimator, if it exists. Often, however, econometricians will want to restrict the analysis to consider only estimators which are linear functions of the errors. An estimator which is linear, unbiased and minimum variance among all linear unbiased estimators is termed the best linear unbiased estimator (BLUE).

Where we are considering estimating a parameter vector with more than one element, the discussion of efficiency has to be qualified somewhat. In general, if we are considering two $k \times 1$ estimators β^* and $\tilde{\beta}$, then we will be comparing the $k \times k$ covariance matrices of these estimators. If the matrix

$$\mathrm{Var}(\tilde{\beta}) - \mathrm{Var}(\beta^*)$$

is a positive semidefinite matrix, then β^* is said to be more efficient than $\tilde{\beta}$.

Asymptotic properties of estimators

The properties discussed above relate to an estimator's sampling distribution, regardless of the number of observations in the time series employed by the estimator. An unbiased estimator, for example, has an expected value equal to the true parameter, independently of how many data points, or observations are available. In many situations, however, an estimator with these desirable properties does not exist, and it is then necessary to inspect an estimator's *asymptotic* properties, i.e. to see how it behaves when very large samples of data are used. Sometimes, where an estimator's properties are known only asymptotically, Monte Carlo experiments are performed to try to simulate the behaviour of the estimator in small samples.

To get an intuitive idea of what asymptotic theory is about, consider again the Monte Carlo experiment with reference to equation (1.15). The independent variable, x, is assumed non-stochastic (e.g. a time trend) and the Monte Carlo procedure consists of generating a time series for the disturbance term and so, for a given value of β, of y. The estimator is then applied to this data to produce an estimate. Repeating this a large number of times then produces an estimate of the sampling distribution of the estimator.

Now, this will be for a given sample size – i.e. we generate series for ε and y which are a certain length, say 100 observations. Let us denote the sample size or number of observations by T, so initially $T = 100$. We could then repeat the Monte Carlo experiment for $T = 101$, then for $T = 102$, then for $T = 103$ and so on, letting T get bigger and bigger. For each value of T we would have a different empirical sampling distribution. If the estimator's properties do not depend on sample size, then the empirical sampling distribution will look very similar, regardless of the value of T. If, on the other hand, sample size does affect the estimator's behaviour, then the shape and/or the location of the empirical sampling distribution will tend to alter as T gets bigger and bigger. For many estimators, we do not in fact have to carry out such experiments to find out what its properties are when T is very large – we can work out mathematically how it behaves as T tends in the limit to infinity. The properties of an estimator as T tends to infinity are termed its *asymptotic properties*.

As we mentioned previously, however, the shape and location of the empirical sampling distribution for small values of T may be examined in order to assess the small-sample properties of the estimator if these cannot be determined mathematically. Note that the sequence $(\beta_T^*)_{T=k}^{\infty}$, where β_T^* denotes the estimator applied to a sample of size T, is itself a stochastic process since each element in the sequence is a random variable which can take on any of a range of values depending on the particular time series used.

The sampling distribution of an estimator as T tends to infinity is termed the *asymptotic distribution*. If the asymptotic distribution has a mean equal to the true value of the parameter being estimated, the estimator is said to be asymptotically unbiased. Often, however, we are more concerned with another asymptotic property – *consistency*. If the asymptotic distribution is concentrated on the true value of the parameter, then the estimator is said to be consistent. Formally, consistency requires that the probability of an estimate generated from an estimator being an arbitrarily small distance from the true value should be unity as the sample size tends to infinity:

$$\lim_{T\to\infty} \Pr\{|\beta_T^* - \beta| < \delta\} = 1 \qquad \textbf{(1.17)}$$

If an estimator is consistent, then its probability limit is equal to the true value of the parameter. If we are considering estimating a parameter vector then the estimator is said to be consistent if each element converges in probability to the corresponding element of the true parameter vector.

A shorthand way of writing equation (1.17) is:

$$\operatorname*{plim}_{T\to\infty} \beta_T^* = \beta \qquad \textbf{(1.18)}$$

Suppose we have two estimators applied to a sample of size T, α_T^* and β_T^* such that equation (1.18) holds and

$$\operatorname*{plim}_{T\to\infty} \alpha_T^* = \alpha \qquad \textbf{(1.19)}$$

Then the following properties of probability limits can be established:

$$\operatorname*{plim}_{T\to\infty}(\alpha_T^* \pm \beta_T^*) = \operatorname*{plim}_{T\to\infty}\alpha_T^* \pm \operatorname*{plim}_{T\to\infty}\beta_T^* = \alpha \pm \beta \qquad \textbf{(1.20a)}$$

$$\operatorname*{plim}_{T\to\infty}(\alpha_T^*\beta_T^*) = \{\operatorname*{plim}_{T\to\infty}\alpha_T^*\}\{\operatorname*{plim}_{T\to\infty}\beta_T^*\} = \alpha\beta \qquad \textbf{(1.20b)}$$

If $\beta_T^* \neq 0$ and $\beta \neq 0$:

$$\operatorname*{plim}_{T\to\infty}(\alpha_T^*/\beta_T^*) = \{\operatorname*{plim}_{T\to\infty}\alpha_T^*\}/\{\operatorname*{plim}_{T\to\infty}\beta_T^*\} = \alpha/\beta \qquad \textbf{(1.20c)}$$

If $\beta_T^* \geqslant 0$ and $\beta \geqslant 0$:

$$\operatorname*{plim}_{T\to\infty}\sqrt{\beta_T^*} = \sqrt{(\operatorname*{plim}_{T\to\infty}\beta_T^*)} = \sqrt{\beta} \qquad \textbf{(1.20d)}$$

If γ is a constant:

$$\operatorname*{plim}_{T\to\infty}\gamma = \gamma \qquad \textbf{(1.20e)}$$

If $\phi(\ )$ is a continuous function:

$$\operatorname*{plim}_{T\to\infty}\phi(\beta_T^*) = \phi(\beta) \qquad \textbf{(1.20f)}$$

The last expression, (1.20f), is sometimes referred to as the Slutsky theorem.

A common source of confusion in econometrics concerns the relationship between the mean and variance of the asymptotic distribution, the asymptotic mean and variance and the probability limit of an estimator. The asymptotic mean and variance are the limits of the first and second moments of the sampling distribution:

$$\text{Asymptotic mean} = \lim_{T\to\infty} E(\beta_T^*)$$

$$\text{Asymptotic variance} = \lim_{T\to\infty}\operatorname{Var}(\beta_T^*)$$

$$= \lim_{T\to\infty} E\{[\beta_T^* - \lim_{T\to\infty} E(\beta_T^*)]^2\}$$

There are circumstances in which the asymptotic mean and variance do not exist while the mean and variance of the asymptotic distribution do, so that the latter are often thought of as the more useful concepts. A sufficient condition for an estimator to be consistent is that the mean of the asymptotic distribution be equal to the true parameter value and that the variance of the asymptotic distribution be zero. The following example, however, demonstrates very clearly that this is not a necessary condition.

Suppose the sampling distribution of the estimator β_T^* is described as:

$$\Pr(|\beta_T^* - \beta| < \delta) = 1 - 1/T$$

$$\Pr(|\beta_T^* - T| < \delta) = 1/T$$

where δ is an arbitrarily small number. Clearly, such an estimator would be consistent since $1/T$ tends to zero as T tends to infinity. The asymptotic mean and variance, however, can be calculated as:

$$\lim_{T\to\infty} E(\beta_T^*) = \lim_{T\to\infty}[\beta(1 - 1/T) + T(1/T)]$$

$$= \beta + 1$$

and

$$\lim_{T\to\infty} E\{[\beta_T^* - \lim_{T\to\infty} E(\beta_T^*)]^2\}$$

$$= \lim_{T\to\infty} E[\beta^2(1 - 1/T) + T^2(1/T) - \{\beta(1 - 1/T) + 1\}^2]$$

$$= \infty$$

Thus, the asymptotic mean is not equal to the true value of the parameter and, moreover, its asymptotic variance is infinite. Nevertheless, it is still a consistent estimator.

1.5 The general linear model

In this section we begin to develop the core of econometrics – the general linear model. Starting from a well-defined set of assumptions we can develop the basic econometric estimator – the ordinary least squares estimator. Much of standard econometric theory can be viewed as adapting this estimator to deal with circumstances in which one or more of these so-called classical assumptions break down. In order to keep the discussion as general as possible, much of the discussion in the remainder of this chapter is in matrix notation.

The classical assumptions, the OLS estimator and the Gauss–Markov theorem

The starting point in our review of standard econometric theory is the general linear regression model. At its most basic, this asserts that the data generating process for an observed variable y_t is a linear combination of K known explanatory variables, x_{kt}, $k = 1, \ldots, K$, plus a stochastic disturbance term u_t:

$$Y_t = \beta_1 x_{1t} + \beta_2 x_{2t} + \beta_3 x_{3t} + \ldots + \beta_K x_{Kt} + u_t \qquad \textbf{(1.21)}$$

where the β_is are unknown. A basic objective of econometrics is to provide 'optimal' estimates of the unknown parameters in relationships such as (1.21). If we have available T observations on y_t and the x_{kt}, we can write them all in matrix notation as

$$Y = X\beta + u \tag{1.22}$$

where

$$Y = (y_1 y_2 \dots y_T)'$$

$$X = \begin{bmatrix} x_{11} & x_{21} & x_{31} & \dots & x_{k1} \\ x_{12} & x_{22} & x_{31} & \dots & x_{k2} \\ \vdots & \vdots & \vdots & & \vdots \\ x_{1T} & x_{2T} & x_{3T} & \dots & x_{kT} \end{bmatrix}$$

$$\beta = (\beta_1 \beta_2 \dots \beta_k)'$$

$$u = (u_1 u_2 \dots u_T)'$$

Thus, Y is a $(T \times 1)$ vector of observations on the dependent variable (the 'regressand'), X is a $(T \times K)$ matrix of observations on the explanatory variables ('regressors'), β is a $(K \times 1)$ vector of unknown parameters and u is a $(T \times 1)$ vector of unobservable stochastic disturbances. X is sometimes termed the 'design matrix'.

The classical linear regression model makes certain assumptions in order to establish various properties of econometric estimators. These are:

1. The disturbances are uncorrelated with one another and each has mean zero and finite variance σ^2:

$$E(u) = 0, \text{Var}(u) = \sigma^2 I$$

2. The explanatory variables are non-stochastic and are thus independent of the disturbances:

$$E(X'u) = 0$$

3. The explanatory variables are linearly independent:

$$\text{rank}(X'X) = \text{rank}(X)$$
$$= K$$

and hence $(X'X)^{-1}$ exists.

Note that we have not yet made any assertions concerning the statistical distribution of the disturbances. Nor have we assumed that the disturbances are independently distributed (i.e. that their joint density function is just the product of their individual density functions), although this property follows from the zero correlation property under normality. Under assumptions 1–3, the best (i.e. minimum sampling distribution variance) linear unbiased estimator of β, $\hat{\beta}$ say, is given by minimising the sum of squared estimated disturbances, or residuals:

$$\min_{\beta} S = (Y - X\hat{\beta})'(Y - X\hat{\beta}) \tag{1.23}$$

The first order conditions for equation (1.23) are:

$$\frac{\partial S}{\partial \hat{\beta}} = -2X'(Y - X\hat{\beta}) = 0$$

which can be expressed as the 'normal equations':

$$X'Y = X'X\hat{\beta} \tag{1.24}$$

and since we know by assumption 3 that $(X'X)$ is non-singular, we have the ordinary least squares (OLS) estimator:

$$\hat{\beta} = (X'X)^{-1} X'Y \tag{1.25}$$

That equation (1.25) solves (1.23) is clear since the second order conditions are satisfied:

$$\frac{\partial^2 S}{\partial \hat{\beta} \partial \hat{\beta}'} = 2X'X$$

which is positive definite.

Since the elements of X are fixed, $(X'X)^{-1}X'$ can be interpreted as a linear function which maps ('projects') any vector in T-dimensional space (Y) into a vector in K-dimensional space (β):

$$(X'X)^{-1}X'\colon R^T \to R^K$$

Thus the matrix $(X'X)^{-1}X'$ is often referred to as the *projection matrix* P_X, with the useful result that $P_X X = I$ and $\hat{\beta} = P_X Y$. Since $\hat{\beta}$ is a linear function of Y, it is a linear estimator. It is also unbiased in the sense that the expected value of $\hat{\beta}$ is the true parameter vector β:

$$\begin{aligned} E(\hat{\beta}) &= E[(X'X)^{-1}X'Y] \\ &= E[(X\ X)^{-1}\ X'(X\beta + u)] \\ &= \beta + (X'X)^{-1}E(X'u) \\ &= \beta \end{aligned} \tag{1.26}$$

where we have used $P_X X = I$ and assumption 2 (non-stochastic regressors). It is clear from equation (1.26) that $\hat{\beta}$ is also a linear function of the errors u.

The variance–covariance matrix for $\hat{\beta}$ is easily established using, $\hat{\beta} = \beta + P_X u$:

$$\begin{aligned}\mathrm{Var}(\beta) &= E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] \\ &= E[P_X uu' P_X] \\ &= (X'X)^{-1}X'(\sigma^2 I)X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1} \end{aligned} \quad \textbf{(1.27)}$$

where assumptions 1 and 3 have been used.

If β^* is any other linear unbiased estimator of β, it is straightforward to show that the variance of β^* exceeds that of $\hat{\beta}$ in the sense that $[\mathrm{Var}(\beta^*) - \mathrm{Var}(\hat{\beta})]$ is a positive semidefinite matrix (the Gauss–Markov theorem). Since β^* is a linear estimator, we can write it as:

$$\beta^* = AY$$

where A is a $K \times T$ matrix of constants. If we define

$$C = A - (X'X)^{-1}X'$$

then clearly

$$\begin{aligned}\beta^* &= [(X'X)^{-1}X' + C]Y \\ &= [(X'X)^{-1}X' + C](X\beta + u) \\ &= \beta + CX\beta + [(X'X)^{-1}X' + C]u\end{aligned}$$

Thus

$$E(\beta^*) = \beta + CX\beta$$

Hence, if β^* is to be unbiased, $CX = 0$. Thus,

$$\begin{aligned}\mathrm{Var}(\beta^*) &= E(\beta^* - \beta)(\beta^* - \beta)' \\ &= E[(X'X)^{-1}X' + C]uu'[(X'X)^{-1}X' + C]' \\ &= \sigma^2[(X'X)^{-1} + CC']\end{aligned}$$

where the property $CX = 0$ has been used. Hence,

$$\mathrm{Var}(\beta^*) - \mathrm{Var}(\hat{\beta}) = \sigma^2 CC'$$

So $\mathrm{Var}(\beta^*)$ exceeds $\mathrm{Var}(\hat{\beta})$ by a positive semidefinite matrix. In particular, note that the diagonal elements of $\sigma^2 CC'$ must be non-negative, so that

$$\mathrm{Var}(\beta_i^*) - \mathrm{Var}(\hat{\beta}_i) > 0, \; i = 1, \ldots, K.$$

Thus, under assumptions 1–3, the OLS estimator $\hat{\beta}$ is the best (minimum variance) linear unbiased estimator (BLUE).

Goodness of fit: coefficient of determination and error variance

Given $\hat{\beta}$, we can divide the Y vector into the sum of an 'explained' part $\hat{Y}$ and an unexplained part $\hat{u}$:

$$Y = X\hat{\beta} + \hat{u} = \hat{Y} + \hat{u} \quad \textbf{(1.28)}$$

One way of determining how well an estimated model fits is to calculate the proportion of the variation in Y which is 'explained' by variation in $\hat{Y}$, and how much is unexplained, due to variation in $\hat{u}$. One measure of variability is the sum of squared y_t's, $Y'Y$. Using equation (1.28):

$$Y'Y = \hat{\beta}'X'X\hat{\beta} + \hat{u}'\hat{u} + 2\hat{\beta}'X'\hat{u} \quad \textbf{(1.29)}$$

The OLS estimator constructs the residual vector $\hat{u}$ so that it is orthogonal to the regressors:

$$\begin{aligned}X'\hat{u} &= X'(Y - X\hat{\beta}) \\ &= X'[I - X(X'X)^{-1}X']Y \\ &= 0\end{aligned}$$

so that the last term in equation (1.29) is zero. Hence, $Y'Y$ is partitioned into two components, one due to the explanatory variables and one unexplained by the model:

$$\begin{aligned}Y'Y &= \hat{\beta}'X'X\hat{\beta} + \hat{u}'\hat{u} \\ &= \hat{Y}'\hat{Y} + \hat{u}'\hat{u}\end{aligned} \quad \textbf{(1.30)}$$

It is, however, more usual to measure variation in a variable around its mean. If we denote the total sum of squares (TSS):

$$\text{TSS} = \sum_{t=1}^{T} (y_t - \bar{y})^2$$

where

$$\bar{y} = T^{-1} \sum_{t=1}^{T} y_t$$

or

$$\text{TSS} = Y'Y - T\bar{y}^2$$

Thus, subtracting $T\bar{y}^2$ from (1.30):

$$\text{TSS} = (\hat{Y}'\hat{Y} - T\bar{y}^2) + \hat{u}'\hat{u} \quad \textbf{(1.31)}$$

If the model contains an intercept, then $x_{1t} = 1$ for all t (see Note 1) and so the first row of the normal equations, (1.24), is:

$$x_1'Y = x_1'X\hat{\beta}$$
$$= \bar{T}\hat{y}$$

where

$$x_1 = (x_{11}, x_{12}, \ldots, x_{1T})'$$
$$= (1, 1, \ldots, 1)'.$$

Thus,

$$\sum y = T^{-1}x_1'X\hat{\beta}$$
$$= T^{-1}\sum_{t=1}^{T}\hat{y}_t$$

Thus, the first bracketed term in equation (1.31) measures the variation in the 'explained' part of Y, that is $\hat{Y}$, around its mean, or the explained sum of squares (ESS). It is trivial to demonstrate that the OLS residuals have mean zero, hence the second term in (1.31) measures the unexplained (or residual) sum of squares (USS):

$$\text{TSS} = \text{ESS} + \text{USS} \tag{1.32}$$

The coefficient of determination, or R^2, measures ESS as a proportion of TSS:

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{USS}}{\text{TSS}} \tag{1.33}$$

Clearly, $0 \leqslant R^2 \leqslant 1$. The closer R^2 is to unity, the better the fit of the regression. Since the R^2 cannot fall, and will usually rise, as the number of regressors is expanded, an allowance is sometimes made for the degrees of freedom lost in constructing the R^2. K degrees of freedom are used up in constructing ESS (corresponding to the K estimated parameters) and 1 in constructing TSS (corresponding to $\bar{y}$). Hence, the degrees-of-freedom corrected R^2, $\bar{R}^2$, is

$$\bar{R}^2 = 1 - \frac{\hat{u}'\hat{u}/(T-K)}{(Y'Y - T\bar{y}^2)/(T-1)} \tag{1.34}$$

or

$$\bar{R}^2 = 1 - [(T-1)/(T-K)](1-R^2)$$

Error variance

An unbiased *estimator* of the error variance σ^2 is often written as s^2 given by:

$$s^2 = \hat{u}'\hat{u}/(T-K)$$

We can demonstrate that s^2 is an unbiased estimator of σ^2 as follows. Note that if we define $M = I - X(X'X)^{-1}X'$, then

$$\hat{u} = Y - X\hat{\beta} = (X\beta + u) - (X\beta + XP_Xu) = (I - XP_X)u$$

So $\hat{u} = Mu$.

It is easily seen that M is symmetric ($M = M'$) and idempotent ($M'M = MM' = M$) hence using $\hat{u} = Mu$:

$$s^2 = u'[M'M]u/(T-K) \tag{1.35}$$

Since s^2 is a scalar, it is trivially equal to its own trace. The properties of trace can be exploited usefully on the right-hand side of equation (1.35), however, in determining the expected value of s^2 (see Notes 2 and 3):

$$E(s^2) = E[\text{trace}[u'(I - X(X'X)^{-1}X')u]/(T-K)]$$
$$= \text{trace}[[I - X(X'X)^{-1}X']E(uu')]/(T-K)$$
$$= \text{trace}[[I - X(X'X)^{-1}]I\sigma^2]/(T-K)$$
$$= \sigma^2(T-K)/(T-K)$$
$$= \sigma^2 \tag{1.36}$$

hence s^2 is an unbiased estimator of σ^2.

It can be shown that, given two regression models, one of which is assumed to be true, the expected value of s^2 for the true model is less than or equal to the expected value for the alternative model. To see this, let

$$Y = X\beta + u$$

be the true model and

$$Y = Z\gamma + u$$

be the alternative, where X and Z are $T \times K_x$ and $T \times K_z$ matrices and X contains at least one variable not included in Z. Then we can write the s^2 for the two models using:

$$\hat{u} = Y - X\hat{\beta} = Y - X(P_xY) = (I - XP_X)Y = M_xY$$

Hence $s_x^2 = Y'M_xY/(T-K_x)$ $s_z^2 = Y'M_zY/(T-K_z)$

where $M_x = I - X(X'X)^{-1}X'$; $M_z = I - Z(Z'Z)^{-1}Z'$. It follows that

$$\begin{aligned}(T - K_z)E(s_z^2) &= E(Y'M_zY)\\ &= E[(X\beta + u)'M_z(X\beta + u)]\\ &= \beta'X'M_zX\beta + E(u'M_zu)\\ &= \beta'X'M_zX\beta + (T - K_z)\sigma^2\\ &> (T - K_z)\sigma^2\end{aligned}$$

Thus $E(s_z^2) > \sigma^2$ or, using the unbiasedness result (1.36) for the 'true model' s_x^2, we then have

$$E(s_z^2) > E(s_x^2) \tag{1.37}$$

Relation (1.37) is sometimes used to justify specification search strategies which maximise the R^2, since, from (1.33):

$$R^2 = 1 - (T - K)s^2/(T - 1)s_y^2 \tag{1.38}$$

where s_y^2 is the sample variance of y_t. Hence, loosely speaking, searching over alternative variables to minimise s^2 also maximises R^2.

Imposing linear restrictions

Suppose that we wished to impose a set of linear restrictions on our estimate of the parameter vector β, in accordance with some underlying economic theory for example. Linear restrictions can always be written in the form

$$R\hat{\beta} = r \tag{1.39}$$

where R is a $(q \times K)$ matrix, q being the number of restrictions, and r is a $K \times 1$ vector. Suppose, for example, that the vector of parameter estimates was 3×1:

$$\hat{\beta} = (\hat{\beta}_1\hat{\beta}_2\hat{\beta}_3)'$$

and we wished to impose the restrictions

$$\hat{\beta}_1 = 1;\ \hat{\beta}_2 + \hat{\beta}_3 = 1 \tag{1.40}$$

To write the restrictions (1.40) in the form of (1.39) let

$$R = \begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 1\end{bmatrix}$$

and

$$r = (1\ 1)'$$

The method of obtaining the restricted least squares estimator should be familiar to an economist: constrained optimisation. We minimise the sum of squared residuals subject to the restrictions. One way of doing this is by unconstrained minimisation of a Lagrangean:

$$\min_{\beta} \pounds = (Y - X\beta)'(Y - X\beta) + 2\lambda'(R\beta - r) \tag{1.41}$$

where 2λ is a $q \times 1$ vector of Lagrange multipliers, scaled by 2 in order to simplify some of the following algebra. The first-order conditions for expression (1.41) are:

$$\frac{\partial \pounds}{\partial \beta} = -2X'Y + 2X'X\beta + 2R'\lambda = O \tag{1.42}$$

$$\frac{\partial \pounds}{\partial \lambda'} = R\beta - r = O \tag{1.43}$$

Premultiply (1.42) by $R(X'X)^{-1}$:

$$\begin{aligned}[R(X'X)^{-1}R']\lambda &= R(X'X)^{-1}X'Y - R\beta\\ &= R\hat{\beta} - r\end{aligned}$$

using (1.43). Thus:

$$\lambda = [R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r) \tag{1.44}$$

where

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{1.45}$$

i.e. $\hat{\beta}$ is the unconstrained OLS estimator. Substituting (1.44) back into (1.42), premultiplying by $(X'X)^{-1}$ and using (1.45), we derive the restricted least squares (RLS) estimator:

$$\tilde{\beta} = \hat{\beta} - (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r) \tag{1.46}$$

If the restrictions were true and all the other classical assumptions were satisfied, then the vector $(R\hat{\beta} - r)$ should be small – the OLS estimates should be close to satisfying the restrictions. From equation (1.46), the RLS estimates will then be close to the OLS estimates. Moreover, the bigger the difference between the OLS and RLS estimates, the less faith we might have in the restrictions. In order to formalise this intuition, however, we need to make some further assumptions.

The distribution of the OLS estimator and linear hypothesis testing

In an earlier section we stated the three classical assumptions which have been used up until now to establish certain properties of the OLS estimator. The first of these assumptions was that the disturbance vector has a zero mean $[E(u) = 0]$ and a scalar covariance matrix $[\text{Var}(u) = \sigma^2 I]$. In order to go further, for example to establish the distribution of the OLS estimator and discuss hypothesis testing, we now need to make some assumptions concerning the statistical distribution of the disturbances.

It is usual to assume that u has a multivariate normal distribution as well as being mean zero and having a scalar covariance matrix:

$$u \sim N(0, \sigma^2 I) \tag{1.47}$$

that is u_t is a Gaussian white noise process. Since, by classical assumption 2, the elements of X are non-stochastic, we can also infer the distribution of Y from (1.47):

$$Y \sim N(X\beta, \sigma^2 I) \tag{1.48}$$

Since the OLS estimator $\hat{\beta}$ is a linear function of Y, it too must be normally distributed, with mean and variance as given by (1.26) and (1.27):

$$\hat{\beta} \sim N[\beta, \sigma^2(X'X)^{-1}] \tag{1.49}$$

Hence, under the classical assumptions plus the assumption of normally distributed disturbances, the OLS estimator is normally distributed with mean β, the true parameter vector, and covariance matrix $\sigma^2(X'X)^{-1}$. Although the error variance σ^2 will usually be unknown, we derived above an unbiased estimator of this quantity, s^2 in equation (1.35), which can be used to construct an unbiased estimate of the covariance matrix:

$$\text{Var}(\hat{\beta}) = s^2(X'X)^{-1} \tag{1.50}$$

We can now apply this framework to derive statistical tests of linear restrictions of the kind considered above. In particular, suppose we wished to test the null hypothesis

$$H_0\colon R\beta - r = O \tag{1.51}$$

where R is an $q \times K$ matrix, r is an $q \times 1$ vector and O is an $q \times 1$ null vector. As we suggested above, if the restrictions (1.51) are correct, then we should expect the vector $(R\hat{\beta} - r)$ to be close to the origin. Given the distribution of the OLS estimator, (1.49), we can infer the distribution of $(R\hat{\beta} - r)$:

$$(R\hat{\beta} - r) \underset{H_0}{\sim} N(0, \sigma^2 R(X'X)^{-1}R') \tag{1.52}$$

where '$\underset{H_0}{\sim}$' is to be read 'is distributed under the null hypothesis as'. If $R\hat{\beta} - r$ is close to the origin, then the following quadratic form should be close to zero:

$$F' = (R\hat{\beta} - r)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)\sigma^{-2} \tag{1.53}$$

Now,

$$\hat{\beta} = (X'X)^{-1}X'Y = P_X(X\beta + u) = \beta + P_X u$$

where we have used $P_X X = I$ and $P_X = (X'X)^{-1}X'$ is a $K \times T$ 'projection matrix' of constants. Hence, under the null hypothesis,

$$R\hat{\beta} - r = RP_x u$$

Since u is, by assumption, a vector of independent, normally distributed random variables, $RP_x u$ is a vector of q independent, normally distributed random variables and so, given (1.52) and (1.53), F' is the sum of $-q$ squared independent standard normal variates; it is therefore a chi-square variate with q degrees of freedom:

$$F' \underset{H_0}{\sim} \chi^2(q) \tag{1.54}$$

Since, however, we do not, in general, know the value of σ^2, expression (1.53) is non-operational. Intuitively, one might be tempted to use an unbiased estimator of σ^2, such as s^2 in (1.53). Since s^2 is itself an estimator then (1.54) would no longer be true. However we can obtain the distribution of s^2 as follows. We have seen that

$$\hat{u} = [I_T - X(X'X)^{-1}X']u = Mu$$

where M is symmetric and idempotent and the subscript in 'I_T' is to make clear the dimensions of this identity matrix. From expression (1.47), we know

$$u\sigma^{-1} \sim N(0, I)$$

A standard result in statistics is that:

$$\frac{\hat{u}'\hat{u}}{\sigma^2} = \frac{u'Mu}{\sigma^2} \sim \chi^2(\text{rank } M)$$

Moreover, by the properties of idempotent matrices:

$$\begin{aligned}\text{rank } M &= \text{trace } M \\ &= \text{trace } I_T - \text{trace } X(X'X)^{-1}X' \\ &= T - \text{trace}\,(X'X)^{-1}X'X \\ &= T - \text{trace } I_K \\ &= T - K\end{aligned}$$

Thus, given $s^2 = \hat{u}'\hat{u}/(T-K)$ we have:

$$(T-K)s^2/\sigma^2 \sim \chi^2(T-K) \qquad (1.55)$$

From expressions (1.53), (1.54) and (1.55) we can therefore write:

$$\frac{(R\hat{\beta}-r)'\,[R(X'X)R']^{-1}(R\hat{\beta}-r)/q}{s^2} \underset{H_0}{\sim} F(q,\,T-K) \qquad (1.56)$$

Expression (1.56) contains no unknown quantities; it can therefore be used to test linear restrictions on the model under the relevant assumptions. Although (1.56) may appear rather cumbersome, it can in fact be computed in a relatively straightforward fashion, as the following demonstrates.

From the definition of the RLS estimator $\tilde{\beta}$, (1.46), we have:

$$X'X(\hat{\beta}-\tilde{\beta}) = R'[R(X'X)^{-1}R^1]^{-1}(R\hat{\beta}-r) \qquad (1.57)$$

Now, $\tilde{\beta}$ must satisfy the restrictions, so that $R\tilde{\beta}=r$, hence:

$$(R\hat{\beta}-r)' = (R\hat{\beta}-R\tilde{\beta})'$$
$$= (\hat{\beta}-\tilde{\beta})'R'$$

So, premultiplying (1.57) by $(\hat{\beta}-\tilde{\beta})'$:

$$(\hat{\beta}-\tilde{\beta})'X'X(\hat{\beta}-\tilde{\beta})$$
$$= (R\hat{\beta}-r)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta}-r) \qquad (1.58)$$

Now consider the restricted sum of squared residuals ($e_r'e_r$) and the unrestricted sum of squared residuals ($e_u'e_u$):

$$e_r'e_r = (Y-X\tilde{\beta})'(Y-X\tilde{\beta}) \qquad (1.59)$$

$$e_u'e_u = (Y-X\hat{\beta})'(Y-X\hat{\beta})$$
$$= (T-K)s^2 \qquad (1.60)$$

Developing equation (1.59):

$$e_r'e_r = (Y-X\hat{\beta}+X\hat{\beta}-X\tilde{\beta})'(Y-X\hat{\beta}+X\hat{\beta}+X\tilde{\beta})$$
$$= (Y-X\hat{\beta})'(Y-X\hat{\beta}) + (\hat{\beta}-\tilde{\beta})'X'X(\hat{\beta}-\tilde{\beta}) \qquad (1.61)$$

where we have used the orthogonality property $X'\hat{u} = X'(Y-X\hat{\beta}) = 0$ to eliminate some terms. From (1.60) and (1.61) we then have:

$$e_r'e_r - e_u'e_u = (\hat{\beta}-\tilde{\beta})'X'X(\hat{\beta}-\tilde{\beta})$$

or, using (1.58):

$$e_r'e_r - e_u'e_u = (R\hat{\beta}-r)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta}-r)$$

Hence, (1.56) may be expressed alternatively:

$$\frac{(e_r'e_r - e_u'e_u)/q}{e_r'e_r/(T-K)} \underset{H_0}{\sim} F(q,\,T-K) \qquad (1.62)$$

The formulation (1.62) is quite intuitive. Since the unrestricted OLS estimator minimises the sum of squared residuals, imposing the restrictions must increase the sum of squares. The left-hand side of (1.62) thus gives the increase in the sum of squares per restriction. We would want to reject restrictions that led to a 'large' increase in the sum of squares; exactly how large 'large' is can be determined from the tables for the F distribution once we choose a specific probability of making an error.

Confidence intervals

Consider expression (1.56) again. Under the null hypothesis, (1.51), $r = R\beta$ (where β is the true parameter vector), so that (1.56) may be expressed alternatively:

$$\frac{(\hat{\beta}-\beta)'R'[R(X'X)^{-1}R']^{-1}R(\hat{\beta}-\beta)}{s^2} \underset{H_0}{\sim} F(q,\,T-K) \qquad (1.63)$$

Now let $F_\alpha(q,\,T-K)$ denote the critical value for the upper 100α per cent of the distribution (or 'test size'), i.e. it is the point on the horizontal axis such that the area under a graph of the central $F(q,\,T-K)$ distribution to the right of this point is α (or, since the total area under the graph must sum to unity, 100α per cent). This allows us to construct a $100(1-\alpha)$ per cent confidence ellipsoid:

$$\Pr\left\{\frac{(\hat{\beta}-\beta)'R'[R(X'X)^{-1}R']^{-1}R(\hat{\beta}-\beta)}{s^2} \leqslant F_\alpha(q,\,T-K)\right\}$$
$$= 1-\alpha \qquad (1.64)$$

What is the interpretation of expression (1.64)? Suppose we were given repeated samples of the data – that is to say, given the true model (1.22), suppose that we generated many Y vectors using the same values for the design matrix, X, and the same coefficients, β, but a different disturbance vector, u, for every case. This would allow us to derive a sampling distribution for $\hat{\beta}$, since there will generally be a different $\hat{\beta}$ for each sample. The statements concerning the unbiasedness and efficiency of the OLS estimator discussed earlier in this chapter are in fact statements about the mean and variance of this sampling distribution. Now suppose that, for each repeated sample we constructed the region in m-dimensional Euclidean space described by the term inside the braces on the left-hand side of (1.64). Expression (1.64) tells us that in $100(1-\alpha)$ per cent of repeated samples the region considered will contain the true value of $R\beta$.

A special case of interest is where R is a $K \times K$ identity matrix. Expression (1.64) then becomes

$$\Pr\left\{\frac{(\hat{\beta}-\beta)'(X'X)(\hat{\beta}-\beta)/K}{s^2} \leqslant F_\alpha(K, T-K)\right\} = 1-\alpha \tag{1.65}$$

This says that in $100(1-\alpha)$ per cent of repeated samples, the ellipsoid in K-dimensional Euclidean space described by the term in braces will contain the true parameter vector β.

Another interesting case is where R is a K-dimensional row vector with unity in the ith element and zeros elsewhere. Expression (1.64) then becomes

$$\Pr\left\{\frac{(\hat{\beta}_i-\beta_i)^2}{s^2(X'X)^{-1}_{ii}} \leqslant F_\alpha(1, T-K)\right\} = 1-\alpha \tag{1.66}$$

where '$[\ ]_{ii}$' denotes the (i, i)th (i.e. ith diagonal) element of the matrix inside the brackets. Using the fact that the square root of an $F(1, T-K)$ variate is distributed as $t(T-K)$, this can be written

$$\Pr\left\{-t_{\alpha/2}(T-K) \leqslant \frac{(\hat{\beta}-\beta)}{s[(X'X)^{-1}_{ii}]^{1/2}} \leqslant t_{\alpha/2}(T-K)\right\}$$
$$= 1-\alpha$$

or

$$\Pr\{\hat{\beta}_i - t_{\alpha/2}(T-K)se(\hat{\beta}_i) \leqslant \beta_i \leqslant \hat{\beta}_i + t_{\alpha/2}(T-K)se(\hat{\beta}_i)\}$$
$$= 1-\alpha \tag{1.67}$$

In moving from (1.66) to (1.67) we have used the notation '$se(\hat{\beta}_i)$' to denote the square root of the ith diagonal element of $s^2(X'X)^{-1}$, the estimated standard error of $\hat{\beta}_i$ and have moved to a two-sided confidence region because the square root may be either positive or negative. Suppose, for example, that $T-K=60$; then since $t_{0.025}(60)=2$, equation (1.67) would mean that in 95% of repeated samples, a region consisting of the point estimate of β_i plus or minus two estimated standard errors would contain the true value of β_i.

Note that a distinction should be made between testing a number of restrictions individually, and testing all of them jointly. Consider, for example, the individual null hypotheses

$$H_a{:}\beta_i = 0$$
$$H_b{:}\beta_j = 0$$

and the *joint* null hypothesis

$$H_c{:}(\beta_i, \beta_j) = (0, 0)$$

Suppose that the $100(1-\alpha)$ per cent confidence regions for β_i and β_j each contain zero. Then we would not be able to reject either H_a or H_b at the 100α per cent significance level. It may be, however, that the two-dimensional $100(1-\alpha)$ per cent joint confidence ellipse for β_i and β_j does not contain the origin, so that H_c may be rejected at the 100α per cent level (i.e. although we cannot reject a hypothesis that one of these coefficients is zero, we can reject the joint hypothesis that they are both zero).

In order to construct the joint confidence ellipse for β_i and β_j, let R be a $2 \times k$ matrix with unity in the $(1, i)$-th and $(2, j)$-th elements and zeros elsewhere. Let the estimated covariance of $\hat{\beta}_i$ and $\hat{\beta}_j$, the (i, j)-th element of the (symmetric) matrix $s^2(X'X)^{-1}$, be denoted $\operatorname{Cov}(\hat{\beta}_i, \hat{\beta}_j)$ and let $se(\hat{\beta}_i)$ and $se(\hat{\beta}_j)$ denote the positive square roots of the ith and jth diagonal elements of this matrix. Then, with R as just defined, expression (1.64) becomes:

$$\Pr\{(1/2)(\hat{\beta}_i-\beta_i)^2 se(\beta_i)^2 + (1/2)(\hat{\beta}_j-\beta_j)se(\beta_j)^2$$
$$+ (\hat{\beta}_i-\beta_i)(\hat{\beta}_j-\beta_j)\operatorname{Cov}(\hat{\beta}_i, \hat{\beta}_j)$$
$$\leqslant F_\alpha(2, T-K)\} = 1-\alpha \tag{1.68}$$

The region described by the term in braces on the left-hand side of (1.68) is an ellipse with centre $(\hat{\beta}_i, \hat{\beta}_j)$. If we make the assumption that $\operatorname{Cov}(\hat{\beta}_i, \hat{\beta}_j)=0$, then the region defined by (1.68) would be a rectangle centred on $(\hat{\beta}_i, \hat{\beta}_j)$, with sides equal to the individual $100(1-\alpha)$ per cent confidence intervals for $\hat{\beta}_i$ and $\hat{\beta}_j$ derived from expressions analogous to (1.67). However, if we know that $\hat{\beta}_i$ and $\hat{\beta}_j$ have positive covariance, then we know that an over-estimate (under-estimate) of β_i is likely to be accompanied by an over-estimate (under-estimate) of β_j. This allows us to rule out the corner areas of the rectangle and so we derive an ellipse which is appropriate in forming a joint $100(1-\alpha)$ per cent confidence region.

It is also straightforward to construct a confidence region for σ^2, the (constant) variance of the disturbance. From expression (1.55), we know that $(T-K)s^2/\sigma^2$ has a χ^2 distribution with $(T-K)$ degrees of freedom. Let $\chi^2(T-K, 1-\alpha/2)$ and $\chi^2(T-K, \alpha/2)$ denote, respectively, the lower and upper $100\alpha/2$ per cent critical values of the $\chi^2(T-K)$ distribution. Then

$$\Pr[\chi^2(T-K, 1-\alpha/2) < (T-K)s^2/\sigma^2 < \chi^2(T-K, \alpha/2)]$$
$$= 1-\alpha$$

which implies:

$$\Pr\left\{\frac{(T-K)s^2}{\chi^2(T-K, 1-\alpha/2)} \leqslant \sigma^2 \leqslant \frac{(T-K)s^2}{\chi^2(T-K, \alpha/2)}\right\} \quad \textbf{(1.69)}$$

1.6 Departures from the classical assumptions

In this section we consider the consequences of and possible remedies to various breakdowns in the classical assumptions.

Omitted variables

So far, our analysis has been conducted under the assumption that the assumed model is correctly specified as in section 1.5. Suppose, however, that we have omitted some important explanatory variables, so that the true model is in fact as in equation (1.70):

$$Y = X\beta + Z\gamma + u \quad \textbf{(1.70)}$$

where Z is a $T \times r$ matrix of observations and γ is an $r \times 1$ parameter vector. Thus, we have omitted r explanatory variables.

The residual vector obtained from the regression $Y = X\hat{\beta} + \hat{v}$, that is excluding Z, may be written

$$\hat{v} = M_X Y$$

Substituting the true expression for Y

$$\hat{v} = M_x(X\beta + Z\gamma + u)$$
$$= M_X Z\gamma + M_X u$$

where we have used

$$M_X X = (I - X(X'X)^{-1}X')X = 0 \quad \text{Thus,}$$

$$E(\hat{v}) = M_x Z\gamma \quad \textbf{(1.71)}$$

Expression (1.71) means that the residual vector $\hat{v}$, where we have omitted the variables Z will have an expected value equal to the residual vector obtained by regressing $Z\gamma$ on to X. Thus, the residuals $\hat{v}$ should be of use in checking for misspecification – this will be developed further in Chapter 4.

Now consider the bias in the OLS estimator, $Y = X\hat{\beta} + \hat{v}$

$$\hat{\beta} = (X'X)^{-1}X'Y = P_x Y$$

Given that the true model is (1.70), and $P_x X = I$, it is easy to show that

$$E(\hat{\beta}) = \beta + P_x Z\gamma \neq \beta$$

and therefore 'omitted variables' will generally lead to biased estimates. However there is no omitted variable bias when X and Z are orthogonal, i.e.

$$X'Z = 0$$

Non-scalar covariance matrix

The first of the classical assumptions which we listed in section 1.5 was that the disturbance terms in the regression model were mean zero and uncorrelated with one another and that each has a constant, finite variance:

$$E(u) = 0, \operatorname{Var}(u) = \sigma^2 I$$

The violation of the assumption of zero-mean disturbances causes no major problems – this effect will simply be picked up by including an intercept term among the regressors. The violation of the assumption that the variance–covariance matrix is a diagonal matrix with a constant term on the main diagonal (i.e. a 'scalar matrix') is, however, quite important.

Each element on the main diagonal of the variance–covariance matrix gives the variance of the distribution from which that element is assumed to be drawn. If the elements of the main *diagonal* of the disturbance variance–covariance matrix differ from observation to observation, then the series is said to be *heteroscedastic* – as opposed to the case of *homoscedastic disturbances*, where the variance is constant. This means that each element in the disturbance vector can be thought of as being drawn from a different distribution.

Each off-diagonal element of the variance–covariance matrix gives the covariance between the disturbances associated with two of the sample observations (for example, the element in the third row, second column gives the covariance between the third observation and the second observation). If all of the off-diagonal terms are zero, the disturbances are said to be uncorrelated; otherwise they are serially correlated.

If the disturbance vector is characterised by either heteroscedasticity or serial correlation, or both, then the variance–covariance matrix will no longer be a scalar matrix:

$$E(uu') = \Omega \neq \sigma^2 I \tag{1.72}$$

Since the proof of unbiasedness of the OLS estimator relied only on the first-moment properties of the model, the OLS estimator will still be unbiased and consistent in this case. The distribution of the OLS estimator will, however, be affected:

$$\begin{aligned}\text{Var}(\beta) &= E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] \\ &= E[(X'X)^{-1}X'uu'X(X'X)^{-1}] \\ &= (X'X)^{-1}X'\Omega X(X'X)^{-1}\end{aligned} \tag{1.73}$$

Thus, the standard formula for the variance–covariance matrix of the estimator, $s^2(X'X)^{-1}$ is incorrect and hence is biased and inconsistent.

There are two possible ways of remedying this problem. One way is to transform the model so that the disturbance variance–covariance matrix is transformed to a scalar matrix and then to apply OLS to the transformed equation. This is generalised least squares (GLS). Note that this method assumes an exact knowledge of the changing covariance structure of the model. Another method, which is becoming increasingly popular, is to use the OLS point estimates for β, since they are unbiased and consistent, but to use a consistent estimate of Ω in equation (1.73) to obtain a consistent estimate of the variance–covariance matrix of the OLS estimator. Since information with respect to Ω is not used in the latter approach, it will result in a less efficient estimator than if the transformation approach is taken. Recently, however, authors have developed methods of estimating Ω consistently without specifying in detail the form of the heteroscedasticity or serial correlation, so that the latter approach can be seen as more general.

Generalised least squares

The 'generalised' linear regression model is

$$Y = X\beta + u \tag{1.74}$$

$$E(u) = 0,\ E(uu') = \sigma^2\Omega$$

where we assume that β and σ^2 are unknown and Ω is known. We have scaled the covariance matrix by the unknown σ^2 in order to reinforce the idea that GLS requires that the form of the covariance matrix need only be known up to a scalar multiple.

Since Ω is a positive definite matrix, it can be shown that there exists a non-singular matrix P which has the property that

$$P\Omega P' = I$$

from which it follows that

$$P'P = \Omega^{-1}$$

Premultiplying (1.74) by P, we have:

$$PY = PX\beta + Pu$$

or

$$Y^* = X^*\beta + u^* \tag{1.75}$$

where $Y^* = PY$, $X^* = PX$, $u^* = Pu$. The covariance matrix of u^* is given by

$$\begin{aligned}E(u^*u^{*\prime}) &= E(Puu'P') \\ &= \sigma^2 P\Omega P' \\ &= \sigma^2 I\end{aligned}$$

Thus, applying OLS to (1.75) will yield the best, linear, unbiased estimator of β:

$$\begin{aligned}\hat{\beta}_{\text{GLS}} &= (X^{*\prime}X^*)^{-1}X^{*\prime}Y^* \\ &= (X'P'PX)^{-1}X'P'PY \\ &= (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y\end{aligned} \tag{1.76}$$

Equation (1.76) gives the GLS estimator. Intuitively, the GLS estimator works because it *weights* the data – for example, in the case of heteroscedasticity, an observation associated with a disturbance whose variance is thought to be especially large would receive less weight than one whose disturbance variance was thought to be small.

Note that, as it stands, the GLS estimator is non-operational because it requires that Ω be known. In general, researchers have to estimate Ω in advance before substituting it in to an equation such as (1.76). This results in the *feasible generalised least squares estimator*. The way in which it is estimated will depend on whether or not the researcher is assuming heteroscedasticity, or autocorrelation, or both.

Heteroscedasticity

Consider again the simple Keynesian consumption function where consumption, y_t, is assumed to depend on current income, x_t:

$$y_t = \beta_1 + \beta_2 x_t + \varepsilon_t \tag{1.77}$$

It may well be the case that the variance of the disturbance term varies as income rises, since the bigger one's income, the more room there is for acts of caprice in consumption, rather than sticking fairly closely to a basic consumption bundle. Suppose, for example, that the variance of the disturbance was assumed to vary with the *square* of income:

$$\text{Var}(\varepsilon_t) = \sigma_t^2 = \alpha x_t^2 \tag{1.78}$$

Deflating (1.77) by x_t yields:

$$y_t^* = \beta_1 z_t^* + \beta_2 + \varepsilon_t^* \tag{1.79}$$

where $y_t^* = y_t/x_t$, $z_t^* = 1/x_t$, $\varepsilon_t^* = \varepsilon_t/x_t$. The variance of the tth transformed disturbance is

$$\text{Var}(\varepsilon_t^*) = \text{Var}(\varepsilon_t)/x_t^2 = \alpha x_t^2/x_t^2 = \alpha$$

which demonstrates that it is homoscedastic, so that OLS can be applied to the *transformed equation* (1.79) to yield an optimal estimator.

In terms of the more general discussion of the previous subsection, we can write, in matrix notation:

$$Y^* = X^*\beta + \varepsilon^* \tag{1.80}$$

$$\Omega = \alpha \begin{bmatrix} x_1^2 & & & & \\ & x_2^2 & & 0 & \\ & & x_3^2 & & \\ & 0 & & \ddots & \\ & & & & x_T^2 \end{bmatrix} \tag{1.81}$$

$$P = \begin{bmatrix} 1/x_1 & & & & \\ & 1/x_2 & & 0 & \\ & & 1/x_3 & & \\ & 0 & & \ddots & \\ & & & & 1/x_T \end{bmatrix}$$

It is easily seen that $P\Omega P' = \alpha I$ – a scalar matrix – hence the GLS estimator (i.e. OLS applied to (1.79)) will have the desired properties.

A more general method, suggested by White (1980) is to obtain unbiased point estimates of β using OLS and then to estimate Ω as a diagonal matrix with the ith squared OLS residual as the (i,i)th element in Ω:

$$\hat{\Omega} = \begin{bmatrix} \hat{\varepsilon}_1^2 & & & & \\ & \hat{\varepsilon}_2^2 & & 0 & \\ & & \hat{\varepsilon}_3^2 & & \\ & 0 & & \ddots & \\ & & & & \hat{\varepsilon}_T^2 \end{bmatrix} \tag{1.81a}$$

White then shows that

$$\operatorname*{plim}_{T\to\infty}(X'X)^{-1}X'\hat{\Omega}X(X'X) = (X'X)^{-1}X'\Omega X(X'X)^{-1}$$

so that the formula

$$\text{Var}(\hat{\beta}_{\text{OLS}}) = (X'X)^{-1}X'\hat{\Omega}X(X'X)^{-1}$$

can be used as a consistent estimator of the variance–covariance matrix of the OLS estimator – regardless of the precise form of the heteroscedasticity. Many regression packages will now calculate heteroscedasticity-consistent, or 'robust' estimated standard errors using this formula or some variant of it. They have also been widely used in estimating equations containing expectations terms (see Chapter 6).

Autocorrelation

A particularly simple case of serially correlated disturbances is where the disturbance is assumed to follow a first-order autoregressive, or AR(1) process. For example:

$$y_t = \beta_1 + \beta_2 x_t + \varepsilon_t \tag{1.82}$$

$$\varepsilon_t = \rho\varepsilon_{t-1} + v_t \tag{1.83}$$

where v_t is assumed to be a white noise process and, in order to ensure the stationarity of ε_t, $|\rho| < 1$:

$$E(v_t) = 0, \tag{1.84a}$$

$$E(v_t^2) = \sigma_v^2, \tag{1.84b}$$

$$E(v_t v_{t-j}) = 0, \text{ for all } j \neq 0 \tag{1.84c}$$

If we lag (1.82) once, multiply it by ρ and subtract the result from (1.82), we have:

$$y_t^* = \beta_1(1-\rho) + \beta_2 x_t^* + v_t \tag{1.85}$$

where $y_t^* = (y_t - \rho y_{t-1})$, $x_t^* = (x_t - \rho x_{t-1})$. Since v_t is white noise, OLS applied to (1.85) will be optimal. (See Note 4.)

To see that this is equivalent to the general form for the GLS estimator discussed above, we need to derive the variance–covariance matrix of the autoregressive disturbance term.

We have already discussed the AR(1) model. In particular, equation (1.83) can be written as:

$$\varepsilon_t = v_t + \rho v_{t-1} + \rho^2 v_{t-2} + \rho^3 v_{t-3} + \rho^4 v_{t-4} + \ldots$$

$$= \sum_{i=0}^{\infty} \rho^i v_{t-i}$$

By substituting (1.85) into (1.82) we can see that y_t is influenced by past error terms – with geometrically declining weights. Thus, the data-generating process for y is *dynamic*; a fact which is not obvious in (1.82).

From (1.83) we have

$$E(\varepsilon_t) = \sum_{i=0}^{\infty} \rho^i E(v_{t-1}) = 0$$

which follows from the assumption that v is a white noise process (1.84a). Thus, the assumption of zero-mean disturbances is unaffected.

Now construct the variance–covariance matrix for $\varepsilon = (\varepsilon_1\varepsilon_2\varepsilon_3 \ldots \varepsilon_T)'$. Using (1.84b):

$$\begin{aligned} E(\varepsilon_t^2) &= E(v_t^2 + \rho^2 v_{t-1}^2 + \rho^4 v_{t-2}^2 + \rho^6 v_{t-3}^2 + \ldots \\ &\quad + \rho v_t v_{t-1} + \rho^2 v_t v_{t-2} \ldots) \\ &= E(v_t^2) + \rho^2 E(v_{t-1}^2) + \rho^4 E(v_{t-2}^2) + \rho^6 E(v_{t-3}^2) + \ldots \\ &= \sigma_v^2 [1 + \rho^2 + \rho^4 + \rho^6 + \ldots] \\ &= \sigma_v^2/(1-\rho^2) \end{aligned} \tag{1.86}$$

Note that the cross-product terms in (1.86) disappear because v is an uncorrelated process (1.84c).

By a similar procedure used to derive (1.86), the covariance between two disturbances j periods apart is given by:

$$E(v_t v_{t-j}) = E(v_t v_{t+j}) = \rho^j \sigma_v^2/(1-\rho^2) \tag{1.87}$$

Thus, the variance–covariance matrix can be written:

$$\Omega = E(\varepsilon\varepsilon') = \frac{\sigma_v^2}{1-\rho^2}\begin{bmatrix} 1 & \rho & \rho^2 & \ldots & \rho^{T-1} \\ \rho & 1 & \rho & \ldots & \rho^{T-2} \\ \rho^2 & \rho & 1 & \ldots & \rho^{T-3} \\ \rho^3 & \rho^2 & \rho & \ldots & \rho^{T-4} \\ \vdots & & & & \vdots \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \ldots & 1 \end{bmatrix} \tag{1.88}$$

We now need to find a matrix P such that $P'P = \Omega^{-1}$. It can be shown that this matrix is given by:

$$P = \begin{bmatrix} \sqrt{(1-\rho^2)} & 0 & 0 & \ldots & 0 \\ -\rho & 1 & 0 & \ldots & 0 \\ 0 & -\rho & 1 & \ldots & 0 \\ 0 & 0 & -\rho & \ldots & \vdots \\ \vdots & & & & \\ 0 & 0 & 0 & \ldots & 1 \end{bmatrix} \tag{1.89}$$

If (1.82) is written in matrix form as

$$Y = X\beta + \varepsilon$$

then premultiplying by P yields

$$Y^* = X^*\beta + \varepsilon^*$$

where

$$Y^* = \begin{bmatrix} \sqrt{(1-\rho^2)}y_1 \\ y_2 - \rho y_1 \\ \vdots \\ y_T - \rho y_{t-1} \end{bmatrix} \tag{1.90a}$$

$$X^* = \begin{bmatrix} \sqrt{(1-\rho^2)} & \sqrt{(1-\rho^2)}x_1 \\ 1-\rho & x_2 - \rho x_1 \\ \vdots & \vdots \\ 1-\rho & x_T - \rho x_{T-1} \end{bmatrix} \tag{1.90b}$$

$$\varepsilon^* = \begin{bmatrix} \sqrt{(1-\rho^2)}\varepsilon_1 \\ v_2 \\ v_3 \\ \vdots \\ v_T \end{bmatrix} \tag{1.90c}$$

It is straightforward to show that $\sqrt{(1-\rho^2)}\varepsilon_1$ has variance σ_v^2 and is uncorrelated with v_t for $t \geq 2$, so:

$$E(\varepsilon^*\varepsilon^{*\prime}) = \sigma_v^2 I$$

Thus, applying OLS to the transformed data, i.e. GLS, will yield the optimal GLS estimator. This method only differs from the intuitive procedure for correcting for AR(1) disturbances in the treatment of the first observation.

In practice, of course, ρ is not known *a priori*, and hence Ω is not (see Note 5). In practice, therefore, this parameter must be estimated. There exists a number of techniques for doing this. This first, the so called Hildreth–Liu technique, involves carrying out an exhaustive grid search for ρ over its admissible range, i.e. -1 to $+1$. For each value of ρ, the GLS estimator is calculated and the sum of squared residuals $(Y^* - X^*\beta)'(Y^* - X^*\beta)$ is computed. The value of ρ is then chosen which minimises this sum of squares.

The second algorithm is due to Cochrane and Orcutt (1949). The Cochrane–Orcutt technique starts by exploiting the fact that OLS will provide an unbiased and consistent estimate of the parameter vector in the presence of autocorrelation (see Note 6), and thus of the disturbance vector. The resulting estimates of ε can then be substituted into (1.83) and OLS applied to yield an estimate of ρ. This then can be used to find a GLS estimate of β, which yields a more efficient estimate of ε. OLS is then applied to the new set of residuals to find a more efficient estimate of ρ, which is again used to construct the GLS estimate, and so on. The procedure stops when successive estimated values of ρ are deemed to be sufficiently close – i.e. until the algorithm converges.

Finally, we should sound a note of caution in 'correcting' for autocorrelation in this fashion indiscriminately. In particular, it is important to distinguish between autocorrelation in the 'true' errors and autocorrelated regression residuals. The latter may be indicative of the former, but they may also indicate dynamic misspecification. One way of attempting to discriminate between the two is to apply a common factor test. For example, consider the dynamic model

$$y_t = \alpha_1 + \alpha_2 x_t + \alpha_3 x_{t-1} + \rho y_{t-1} + v_t \tag{1.91}$$

Using the lag operator, (1.91) can be written:

$$(1 - \rho L)y_t = \alpha_1 + \alpha_2[1 + (\alpha_3/\alpha_2)L]x_t + v_t \tag{1.92}$$

If the restriction $-\rho = (\alpha_3/\alpha_2)$ holds, or equivalently:

$$\rho\alpha_2 + \alpha_3 = 0 \tag{1.93}$$

then we can divide (1.92) through by the common factor $(1 - \rho L)$ to obtain the AR(1) disturbance model (1.82), (1.83). Thus, estimating a static model with an AR(1) disturbance term is tantamount to imposing the common factor restrictions (1.93) on the dynamic model with

a white noise error, (1.92). If there is sign of serial correlation in the static regression residuals (such as from the Durbin–Watson statistic) but the AR(1) common factor restrictions are rejected, then the remedy is not to 'correct' for serially correlated residuals, but to improve the dynamic specification of the model. Testing non-linear restrictions of the kind (1.93) lies outside the scope of this introductory chapter but will be discussed in Chapter 2. The topic of dynamic specification is discussed at length in Chapter 4.

Stochastic regressors

The second classical assumption which we listed in section 1.5 was that the regressors are non-stochastic and are thus independent of the errors.

$$E(X'u) = 0$$

This assumption was required to derive the unbiasedness property of the OLS estimator. In general, however, the assumption that the regressors are non-stochastic – or fixed in repeated samples – can be seen to be quite restrictive. For example, we may have a lagged dependent variable, representing some degree of inertia in the behaviour of a variable. More generally, there seems to be little sense in asserting that some economic time series such as consumption are stochastic, while others such as income are not. Moreover, there may be random errors in the measurement of the regressors, i.e. 'errors in variables', or the equation we are considering may be part of a larger simultaneous system involving stochastic feedback between variables.

If the regressors are not considered to be non-stochastic, but it is considered safe to assume that they are distributed independently of the disturbance, then most of the desirable characteristics of the OLS estimator can in fact be recovered, although the algebra becomes considerably more complicated. This is the *conditional regression model* which is discussed briefly in Chapter 4. The small-sample properties of the OLS estimator under the assumption of stochastic regressors cannot, however, be retrieved, although consistency holds if the regressors are *contemporaneously* uncorrelated with the disturbances, i.e. the nth observation of the regressor is uncorrelated with the nth observation of the disturbance.

Consider the general linear model, where the design matrix is not assumed to be non-stochastic:

$$Y = X\beta + u \tag{1.94}$$

The covariance matrix is assumed to be scalar. If the following conditions hold:

$$\operatorname*{plim}_{T\to\infty} T^{-1}X'X = \Sigma \tag{1.95a}$$

$$\operatorname*{plim}_{T\to\infty} T^{-1}X'u = 0 \tag{1.95b}$$

where Σ is a non-singular matrix, then the OLS estimator $\hat{\beta}$ is consistent:

$$\begin{aligned}
\operatorname*{plim}_{T\to\infty} \hat{\beta} &= \operatorname*{plim}_{T\to\infty} (X'X)^{-1}X'Y \\
&= \operatorname*{plim}_{T\to\infty} (X'X)^{-1}X'(X\beta + u) \\
&= \beta + \operatorname*{plim}_{T\to\infty} (X'X)^{-1}X'u \\
&= \beta + \operatorname*{plim}_{T\to\infty} (T^{-1}X'X)^{-1} \operatorname*{plim}_{T\to\infty} (T^{-1}X'u) \\
&= \beta + \Sigma 0 \\
&= \beta
\end{aligned}$$

Assumptions (1.95a) and (1.95b) thus replace the second classical assumption. Assumption (1.95a) will hold if X consists of realisations from a stationary multivariate stochastic process with a non-singular contemporaneous variance–covariance matrix. It can also be shown that the standard estimators of the disturbance variance and of the variance–covariance matrix of the OLS parameters, $\hat{\beta}$, will also be consistent.

Errors in variables

Another reason that regressors may be stochastic is where there is stochastic measurement error in one or more of the regressors. In this case, however, the OLS estimator is no longer even consistent.

Say, for example, we believe Y and X are related by an exact linear relationship:

$$\tilde{Y} = \tilde{X}\beta \tag{1.96}$$

but instead of observing $\tilde{X}$ and $\tilde{Y}$ directly, we observe only measured data X and Y which may be contaminated by measurement error:

$$X = \tilde{X} + \zeta \tag{1.97a}$$

$$Y = \tilde{Y} + \mu \tag{1.97b}$$

where ζ and μ represent the measurement error. Often, there is no reason why measurement error should not be autocorrelated; for example, where x represents a stock (such as money supply for example), it may make sense to propose a first-order moving average representation for the measurement error:

$$\zeta_t = v_t - \partial v_{t-1} \tag{1.98}$$

This would imply that a proportion ∂ of measurement error tends to be reversed in the following period. For our purposes, however, we need only assume that the measurement errors are white noise stochastic processes.

Substituting from (1.97) and (1.96), we have:

$$Y = X\beta + \omega \tag{1.99a}$$

$$\omega = \mu - \beta\zeta \tag{1.99b}$$

If the measurement errors are assumed uncorrelated, then the covariance matrix for ω is:

$$E(\omega\omega') = \sigma_\mu^2 I + \beta^2\sigma_\zeta^2 I = \sigma_\omega^2 I \tag{1.100}$$

where σ_μ^2 and σ_ζ^2 denote the variance of μ and ζ respectively. From equation (1.100) it is clear that ω has a scalar covariance matrix.

From equations (1.99a) and (1.99b), however, it is clear that the disturbance term in (1.99a), ω, is correlated with the regressor, X, which violates one of the classical assumptions which we discussed in section 1.5 and which was needed to derive the unbiasedness property of the OLS estimator. Moreover, the OLS estimator is no longer even consistent since, although condition (1.95a) may still be assumed to hold, condition (1.95b) is violated:

$$\operatorname*{plim}_{T\to\infty} T^{-1}X'X = \Sigma$$

$$\begin{aligned}
\operatorname*{plim}_{T\to\infty} T^{-1}X'\omega &= \operatorname*{plim}_{T\to\infty} T^{-1}(\tilde{X} + \zeta)'(\mu - \beta\zeta) \\
&= -\beta\sigma_\zeta^2 I \neq 0
\end{aligned}$$

Thus:

$$\begin{aligned}
\operatorname*{plim}_{T\to\infty} \hat{\beta} &= \operatorname*{plim}_{T\to\infty} (X'X)^{-1}X'Y \\
&= \operatorname*{plim}_{T\to\infty} (X'X)^{-1}X'(X\beta + \omega) \\
&= \beta + \operatorname*{plim}_{T\to\infty} T(X'X)^{-1} \operatorname*{plim}_{T\to\infty} T^{-1}X'\omega \\
&= \beta - \beta\sigma_\zeta^2\Sigma^{-1} \\
&\neq \beta
\end{aligned}$$

Simultaneous equations and instrumental variables

Another standard case where some of the independent variables in a regression are correlated with the errors is in a simultaneous equation system. In such a system there is a *contemporaneous* feedback between the endogenous variables of the system. OLS on any single equation therefore gives biased and inconsistent parameter estimates. Although it is possible to estimate the full system 'at one go' (see Chapter 2), frequently, applied economists only wish to estimate a single structural equation. Nevertheless they are aware that the equation of interest may be part of a larger simultaneous system and hence OLS is inappropriate. A consistent estimator is provided by the method of instrumental variables (IV) on a single equation although it is not always obvious how one chooses a particular set of instruments and whether they are independent of the error term. IV is a single equation estimation technique and does not consider all of the information in the rest of the system of equations, (although it may be generalised to a system estimator, three stage least squares, 3SLS, is a limited information estimator which can be compared to a full information estimator such as maximum likelihood (see Chapter 2).

The IV approach is very general and there is a wide variety of estimators within this general class. The approach is to take a set of variables ('instruments') which satisfy the classical assumption and use them to construct a 'proxy' for the variable which is endogenous. To delineate members of the class, one's choice of instrument set may determine the name given to a particular IV estimator. For example, the two stage least squares (2SLS) estimator is a specific form of IV estimator. To complicate matters the 2SLS estimator may also be interpreted as a two-step estimator; it is equivalent to doing two (particular) OLS regressions. However there are some subtle differences between 2SLS viewed as a special form of IV estimator and the two-step procedure.

The IV estimator is derived as follows. Suppose in the general linear model we have a subset of variables $X_1(1 \times k_1)$ that are uncorrelated with the error term u in large samples. But the subset of variables $X_2(1 \times k_2)$ are correlated with u

$$Y = X\beta + u = (X_1, X_2)\beta + u \tag{1.101}$$

where

$$\operatorname{plim} T^{-1}(X_1'u) = 0 \tag{1.102a}$$

$$\operatorname{plim} T^{-1}(X_2'u) \neq 0 \tag{1.102b}$$

Without loss of generality assume $u \sim N(0, \sigma^2 I)$. Suppose there exists a set of k_2 variables denoted W_1 (the 'instruments') which have the properties:

$$\operatorname*{plim}_{T\to\infty} T^{-1}(W_1'u) = 0 \tag{1.103}$$

$$\operatorname*{plim}_{T\to\infty} T^{-1}(W_1'X_2) \neq 0$$

Hence W_1 is uncorrelated in the limit with u and there is a non-zero correlation between W_1 and X_2 (with a constant asymptotic moment matrix, W'_1X_2).

The full matrix of instruments is

$$W = (W_1, X_1) \tag{1.104}$$

where X_1 acts, in effect, as its own instrument and W_1 has 'replaced' the variables X_2. Now we premultiply (1.101) by W' and take probability limits:

$$\operatorname*{plim}_{T\to\infty} T^{-1}(W'Y) = \operatorname*{plim}_{T\to\infty} T^{-1}(W'X)\beta + \operatorname*{plim}_{T\to\infty} T^{-1}(W'u) \tag{1.105}$$

Taking the sample moments as estimates of their population values (which we assume throughout this section) and using equation (1.103) above, it is easily seen that $\hat{\beta}_{IV}$ the *instrumental variable estimator* is

$$\hat{\beta}_{IV} = (W'X)^{-1}(W'Y) \tag{1.106}$$

Note that if all the X variables satisfy the classical assumptions then W is the same as X and this is simply the OLS estimator. The asymptotic covariance matrix of the IV estimator (which we simply denote as $\operatorname{Var}(\hat{\beta}_{IV})$) may be derived as follows. Substituting (1.101) in (1.106) produces

$$\hat{\beta}_{IV} - \beta = (W'X)^{-1}(W'u) \tag{1.107}$$

Hence

$$\operatorname{Var}(\hat{\beta}_{IV}) = \operatorname*{plim}_{T\to\infty} T(W'X)^{-1} \operatorname*{plim}_{T\to\infty} T^{-2}(W'uu'W) \times \operatorname*{plim}_{T\to\infty} T(X'W)^{-1}$$

$$= \sigma^2(W'X)^{-1}(W'W)(X'W)^{-1} \tag{1.108}$$

Since $\hat{\beta}_{IV}$ is consistent (to see this take 'plims' of (1.107)) and using (1.102), (1.103) and (1.104) the residuals

$$\hat{u}_{IV} = Y - X\hat{\beta}_{IV} \tag{1.109}$$

can be used to obtain a consistent estimator for σ^2:

$$s^2_{IV} = (\hat{u}_{IV}'\hat{u}_{IV})/T \tag{1.110}$$

Note that X and not W is used in (1.109) and that (1.108) would again be the OLS formula when X and W are identical.

We now turn to a simple two-equation simultaneous equation system to demonstrate the relationship between IV and 2SLS. (Because of space constraints we do not discuss the identification problem in simultaneous models, although we make the implicit assumption that the systems we are discussing are identified. This means that the order condition, that the number of predetermined variables excluded from any equation must at least equal the number of endogenous variables included on the *right-hand side*, and the rank condition, are met. We also require that there are at least as many instruments as endogenous variables.)

A simultaneous system

Our simple illustrative system is:

$$y_1 = \alpha y_2 + \beta x_1 + \varepsilon_1 \quad \textbf{(1.111a)}$$

$$= Q\delta + \varepsilon_1$$

$$y_2 = \gamma y_1 + \theta x_2 + \varepsilon_2 \quad \textbf{(1.111b)}$$

where $\varepsilon_i \sim N(0, \sigma_i^2 I)$

and $\operatorname{plim}_{T\to\infty} (x_i'\varepsilon_j)/T = 0 \ (i, j = 1, 2)$

$$E(\varepsilon_{1t}\varepsilon_{2t}) = E(\varepsilon_{1t}\varepsilon_{2t-j}) = 0$$

and we define $Q = (y_2, x_1)$, $\delta = (\alpha, \beta)$

Because we wish to isolate the issues that arise solely from simultaneity between the endogenous variables y_{1t} and y_{2t} we assume white noise errors in each equation and no contemporaneous correlation between the errors in different equations.

The reduced form equations of the system are:

$$y_{1t} = x_{1t}\pi_{11} + x_{2t}\pi_{12} + v_{1t} \quad \textbf{(1.112a)}$$

$$y_{2t} = x_{1t}\pi_{21} + x_{2t}\pi_{22} + v_{2t} \quad \textbf{(1.112b)}$$

where $\pi_{11} = (1-\alpha\gamma)^{-1}\beta$, $\pi_{12} = (1-\alpha\gamma)^{-1}\alpha\theta$, $\pi_{21} = (1-\alpha\gamma)^{-1}\gamma\beta$,

$\pi_{22} = (1-\alpha\gamma)^{-1}\theta$, $v_{1t} = (1-\alpha\gamma)^{-1}(\varepsilon_{1t} + \alpha\varepsilon_{2t})$

$v_{2t} = (1-\alpha\gamma)^{-1}(\varepsilon_{2t} + \gamma\varepsilon_{1t})$

In what follows, of crucial importance in (1.112a) and (1.112b) is that y_{1t} and y_{2t} depend on a linear combination of the structural errors $\varepsilon_{it} (i = 1, 2)$. This arises because of the simultaneity of the system and

so the classical assumptions outlined in section 1.5 do not hold, also condition (1.95b) is violated and so by the proof given in section 1.6 OLS is neither unbiased nor consistent. An instrumental variable estimator may be used to provide consistent parameter estimates of either a part of the system or the complete system.

In most of what follows we assume the econometrician is only interested in estimating the structural equation (1.111a) but is aware that this equation is embedded in a simultaneous system which could consist of a number of additional equations. Also we could easily extend the analysis to consider y_2 to be a vector of endogenous variables, just as there could be many more than two x variables. However, for pedagogic reasons we assume for the moment the simple model outlined above.

The IV estimator of (1.111a) is consistent. We require an instrument for y_2, that is both independent of ε_1 in large samples and has some non-zero correlation with y_2. Call this variable w_1. The instrument matrix is then

$$W_1 = (w_1, x_1)$$

where x_1 may be thought of as acting as its own instrument. The IV estimator is,

$$\delta = (W_1'Q)^{-1}W_1'y_1 \quad \textbf{(1.113)}$$

An obvious question is how do we choose a particular variable to act as an instrument for y_{2t}? An obvious candidate is x_{2t} since by assumption this is independent of ε_{1t} and from (1.111b) is correlated with y_{2t}. But if we know the system we can do better than this.

Two-stage least squares (2SLS)

An alternative is to use a linear combination of *all* the predetermined variables in the system. To obtain our linear combination we perform the OLS regression

$$y_2 = x_1\hat{\pi}_{21} + x_2\hat{\pi}_{22} + \hat{v}_2 \quad \textbf{(1.114)}$$

and form $\hat{y}_2$ as the fitted values from this model, note that this is the true reduced form equation for y_2. The instrument matrix is then

$$W = (\hat{y}_2, x_1)$$

and

$$\delta^*(\text{2SLS}) = (W'Q)^{-1}(W'y_1) \quad \textbf{(1.115)}$$

When we use $\hat{y}_2$ as the instrument and apply IV then δ^* is known as the 2SLS estimator (a particular form of IV). The name originates because we obtain $\hat{y}_2$ in the 'first-stage' regression and use this in the IV (second stage) formula. For the moment we assume that $\operatorname{plim} T^{-1}(\hat{y}_2'\varepsilon_1) = 0$. Intuitively we might expect $\hat{y}_2$ to be uncorrelated with ε_1 because it is a linear combination of x_1 and x_2 which are both assumed to be independent of ε_1.

Other IV *estimators*

More often than not applied economists are interested only in estimating one structural equation although they are aware that this equation may form part of a larger simultaneous system. If we are interested only in (1.111a) we will still obtain biased parameter estimates because of the existence of (1.111b) even if we do not explicitly formulate this second equation. So whenever the possibility of simultaneity (or the failure of weak exogeneity, see Chapter 4) exists we must be prepared to consider an IV estimation strategy. However we may have only a hazy idea of the form of the rest of the model and of the set of weakly endogenous variables in the complete system. We may therefore try a number of alternative instruments sets for y_{2t}. We could choose any one x_i variable from the potentially large set of X $(X = (x_1, \ldots, x_k))$. Alternatively we can choose one of many essentially arbitrary sub-sets of X, $X^j \subset X$ and perform the OLS regression of y_2 on X^j. We could then form

$$\hat{y}_2^j = X^{j\prime}\hat{\Pi} \tag{1.116}$$

where $\hat{y}_2^j$ may be used as an instrument for y_2. Clearly many such instruments may be constructed depending on the choice of X^j and they will all differ and give somewhat different parameter estimates in finite samples. By assumption, however, all of these IV estimators are consistent. This is a practical problem with IV estimation: results are not invariant to the choice of instrument set. Ideally one should report a sensitivity analysis with respect to alternative instrument sets. As a general guideline in small samples there is also a trade-off between efficiency and consistency. Choosing a very small set of instruments will ensure consistency but may yield very inefficient estimates but, in a small sample, as the number of instruments grows the IV parameter estimate will converge on the OLS estimator and will be inconsistent. One way of checking the validity of the instrument set is to use a test due to Sargan which tests for the orthogonality of the instrument set and the structural residual; this test is discussed in Chapter 4.

Two-step and two-stage least squares

We will now consider the 2SLS estimator from a slightly different angle, namely as *two applications of* OLS. Suppose we have a fixed set of instruments $X = \{x_1, x_2\}$ where x_1 are the k^* predetermined variables in the structural equation of interest and x_2 are the k^{**} predetermined variables excluded from the equation. The OLS regression of y_2 on X is the first stage regression and we may construct the instrument:

$$\hat{y}_2 = x_1\hat{\pi}_1 + x_2\hat{\pi}_2 = X\hat{\Pi} \tag{1.117}$$

Now let us replace the endogenous variable y_2 with $\hat{y}_2$ in (1.111a) and then estimate the resulting equation by OLS

$$\begin{aligned} y_1 &= \alpha\hat{y}_{2t} + \beta x_1 + \omega_1 \\ &= \hat{Q}\delta + \omega_1 \end{aligned} \tag{1.118}$$

Then the '2-step least squares estimator' (OLS done twice) is

$$\hat{\delta}_p = (\hat{Q}'\hat{Q})^{-1}(\hat{Q}'y_1) \tag{1.119}$$

The OLS formula for the covariance matrix $\operatorname{Var}(\hat{\delta}_p)$ and the variance of the equation s_p^2 produced by standard regression packages on (1.118) will be:

$$\operatorname{Var}(\hat{\delta}_p) = s_p^2(\hat{Q}'\hat{Q}) \tag{1.120}$$

$$s_p^2 = (y_1 - \hat{Q}\hat{\delta}_p)'(y_1 - \hat{Q}\hat{\delta}_p)/(T - K) \tag{1.121}$$

How do these formulae compare with the ones given for the 2SLS estimator? For 2SLS the instrument matrix is:

$$W = (\hat{y}_2, x_1) = \hat{Q} \tag{1.122}$$

which is the same as that used above in the second stage of the two-stage estimation. So (1.115) reduces to δ^* (2SLS) $= (\hat{Q}'Q)^{-1}(\hat{Q}'y_1)$. However, it may be shown that

$$(\hat{Q}'Q) = \hat{Q}'\hat{Q} \tag{1.123}$$

and therefore the 2SLS estimator gives exactly the same numerical value as the two-step estimator for the estimated parameters. Also, using $(\hat{Q}'Q) = (\hat{Q}'\hat{Q})$, $W = \hat{Q}$, and noting that $X \equiv Q$ then (1.108) gives:

$$\operatorname{Var}(\delta_{2SLS}) = s_{2SLS}^2(\hat{Q}'\hat{Q})^{-1} \tag{1.124}$$

The difference in the two formulae (1.120) and (1.124) lies in the estimate of s^2. Equation (1.124) constructs s^2 from the IV/2SLS residuals defined by $u_{2SLS} = y_1 - Q\hat{\delta}_{2SLS}$ where $Q = (y_2, x_1)$ while s_p^2 is constructed using the residuals defined in (1.121) using $\hat{Q} = (\hat{y}_2, x_1)$; these are not the same. Thus the two-step procedure constructs the residuals using $\hat{y}_2$ while the 2SLS procedure uses y_2, the actual value of y_2. Hence, while the two-step procedure provides consistent parameter estimates it does not calculate correctly the variance of the equation or the covariance matrix of the parameters; for these the IV/2SLS formulae must be used.

Consistency of the two-step estimator $\hat{\delta}_p$

We have already implicitly established the consistency of the two-step procedure by appealing to its numerical equivalence with the IV estimator but given the use we will make of the two-step procedure in Chapter 6 on rational expectations it is useful to establish this result and outline a complication. If we take (1.111a) and add and subtract $\alpha\hat{y}_2$ from it we may restate it as

$$y_1 = \alpha\hat{y}_2 + \beta x_1 + \omega_1 \tag{1.125}$$

where

$$\omega_1 = \varepsilon_1 + \alpha(y_2 - \hat{y}_2) \tag{1.126}$$

Consistency of the two-step estimator then requires:

$$\operatorname*{plim}_{T\to\infty} T^{-1}(x_1'\omega_1) = \operatorname*{plim}_{T\to\infty} T^{-1}(\hat{y}_2'\omega_1) = 0$$

We have:

$$\operatorname*{plim}_{T\to\infty} T^{-1}(x_1'\varepsilon_1) = 0 \quad \text{by assumption}$$

$$\operatorname*{plim}_{T\to\infty} T^{-1}(x_1'(y_2 - \hat{y}_2)) = \operatorname*{plim}_{T\to\infty} T^{-1}(x_1'\hat{v}_2) = 0 \quad \text{by OLS}$$

$$\operatorname*{plim}_{T\to\infty} T^{-1}(\hat{y}_2'\varepsilon_1) = 0 \quad \text{by equation (1.117)}$$

$$\operatorname*{plim}_{T\to\infty} T^{-1}(\hat{y}_2'(y_2 - \hat{y}_2)) = \operatorname*{plim}_{T\to\infty} T^{-1}(\hat{y}_2'\hat{v}_2) = 0 \quad \text{by OLS}$$

which establishes the consistency of the estimator. Note that a complication is that this proof rests on the assumption that x_1 and $\hat{y}_2$ are uncorrelated with $\hat{v}_2$; this is correct by construction given our specification of (1.117), but if we had omitted x_1 from the specification this would no longer be valid. So if we define $\hat{y}_2^*$ from OLS on:

$$\hat{y}_2^* = x_2\hat{\pi}^* \tag{1.127}$$

then

$$x_1'(y_2 - \hat{y}_2^*) \neq 0$$

if x_1 has any influence on y_2. In this case the two-step estimator using $\hat{y}_2^*$ is inconsistent. This arises in expectations models (see Chapter 6). However if $\hat{y}_2^*$ is used *as an instrument* for y_2,

$$W = (\hat{y}_2^*, x_1)$$

then the IV formulae yield consistent estimates of δ, s^2 and $\mathrm{Var}(\alpha)$, $\mathrm{Var}(\beta)$.

1.7 Conclusion

In this chapter we have given an account of the standard econometric results which underlie single equation estimation by OLS, we have shown that under a fairly stringent set of assumptions OLS is an optimal estimator and we have outlined how the failure of these assumptions leads to a poor performance on the part of this technique. So far we have said little about systems estimation and ways in which we can deal with the problems which arise when the classical assumptions are violated. Much of the rest of the book is aimed at dealing with these problems. Chapter 2 introduces the notion of maximum likelihood and this allows systems of equations to be treated effectively. The failure of the assumption of stationarity is the central issue of Chapter 5 on cointegration and correct conditioning and testing of the underlying dynamic specification is the heart of dynamic modelling, treated in Chapter 4.

Notes

1. Of course, the intercept need not be placed first in the regression. If it were placed in the ith position, then we would examine the ith row of the normal equations.
2. The *trace* of a matrix is defined as the sum of the elements on the leading diagonal.
3. An idempotent matrix, M, has the property that $MM = M$. If M is non-singular (i.e. if its inverse exists), then it follows immediately that M is the identity matrix. In general, an idempotent matrix is singular.
4. Note that we have not stated explicitly what should be done with the first observation – y_t^* is defined only for $t \geq 2$. One option is simply to drop this observation. A more satisfactory alternative will become clear below.
5. The variance of v_t need not be known because, as we noted above, the variance–covariance matrix of the disturbance need be known only up to a scalar multiple.
6. Provided there are no lagged dependent variables. If there were, this would in any case violate the assumption of non-stochastic regressors, which we consider below.