5 Univariate time series modelling and forecasting

Learning Outcomes

In this chapter, you will learn how to

- Explain the defining characteristics of various types of stochastic processes
- Identify the appropriate time series model for a given data series
- Produce forecasts for ARMA and exponential smoothing models
- Evaluate the accuracy of predictions using various metrics
- Estimate time series models and produce forecasts from them in EViews

5.1 Introduction

Univariate time series models are a class of specifications where one attempts to model and to predict financial variables using only information contained in their own past values and possibly current and past values of an error term. This practice can be contrasted with *structural models*, which are multivariate in nature, and attempt to explain changes in a variable by reference to the movements in the current or past values of other (explanatory) variables. Time series models are usually a-theoretical, implying that their construction and use is not based upon any underlying theoretical model of the behaviour of a variable. Instead, time series models are an attempt to capture empirically relevant features of the observed data that may have arisen from a variety of different (but unspecified) structural models. An important class of time series models is the family of AutoRegressive Integrated Moving Average (ARIMA) models, usually associated with Box and Jenkins (1976). Time series models may be useful

when a structural model is inappropriate. For example, suppose that there is some variable y_t whose movements a researcher wishes to explain. It may be that the variables thought to drive movements of y_t are not observable or not measurable, or that these forcing variables are measured at a lower frequency of observation than y_t. For example, y_t might be a series of daily stock returns, where possible explanatory variables could be macroeconomic indicators that are available monthly. Additionally, as will be examined later in this chapter, structural models are often not useful for out-of-sample forecasting. These observations motivate the consideration of pure time series models, which are the focus of this chapter.

The approach adopted for this topic is as follows. In order to define, estimate and use ARIMA models, one first needs to specify the notation and to define several important concepts. The chapter will then consider the properties and characteristics of a number of specific models from the ARIMA family. The book endeavours to answer the following question: 'For a specified time series model with given parameter values, what will be its defining characteristics?' Following this, the problem will be reversed, so that the reverse question is asked: 'Given a set of data, with characteristics that have been determined, what is a plausible model to describe that data?'

5.2 Some notation and concepts

The following sub-sections define and describe several important concepts in time series analysis. Each will be elucidated and drawn upon later in the chapter. The first of these concepts is the notion of whether a series is *stationary* or not. Determining whether a series is stationary or not is very important, for the stationarity or otherwise of a series can strongly influence its behaviour and properties. Further detailed discussion of stationarity, testing for it, and implications of it not being present, are covered in chapter 7.

5.2.1 A strictly stationary process

A strictly stationary process is one where, for any $t_1, t_2, \ldots, t_T \in Z$, any $k \in Z$ and $T = 1, 2, \ldots$

$$F_{y_{t_1}, y_{t_2}, \ldots, y_{t_T}}(y_1, \ldots, y_T) = F_{y_{t_1+k}, y_{t_2+k}, \ldots, y_{t_T+k}}(y_1, \ldots, y_T) \qquad (5.1)$$

where F denotes the joint distribution function of the set of random variables (Tong, 1990, p.3). It can also be stated that the probability measure for the sequence $\{y_t\}$ is the same as that for $\{y_{t+k}\} \forall k$ (where '$\forall k$' means

'for all values of k'). In other words, a series is strictly stationary if the distribution of its values remains the same as time progresses, implying that the probability that y falls within a particular interval is the same now as at any time in the past or the future.

5.2.2 A weakly stationary process

If a series satisfies (5.2)–(5.4) for $t = 1, 2, \ldots, \infty$, it is said to be weakly or covariance stationary

$$(1)\ E(y_t) = \mu \tag{5.2}$$
$$(2)\ E(y_t - \mu)(y_t - \mu) = \sigma^2 < \infty \tag{5.3}$$
$$(3)\ E(y_{t_1} - \mu)(y_{t_2} - \mu) = \gamma_{t_2 - t_1} \quad \forall t_1, t_2 \tag{5.4}$$

These three equations state that a stationary process should have a constant mean, a constant variance and a constant autocovariance structure, respectively. Definitions of the mean and variance of a random variable are probably well known to readers, but the autocovariances may not be.

The autocovariances determine how y is related to its previous values, and for a stationary series they depend only on the difference between t_1 and t_2, so that the covariance between y_t and y_{t-1} is the same as the covariance between y_{t-10} and y_{t-11}, etc. The moment

$$E(y_t - E(y_t))(y_{t-s} - E(y_{t-s})) = \gamma_s, s = 0, 1, 2, \ldots \tag{5.5}$$

is known as the *autocovariance function*. When $s = 0$, the autocovariance at lag zero is obtained, which is the autocovariance of y_t with y_t, i.e. the variance of y. These covariances, γ_s, are also known as autocovariances since they are the covariances of y with its own previous values. The autocovariances are not a particularly useful measure of the relationship between y and its previous values, however, since the values of the autocovariances depend on the units of measurement of y_t, and hence the values that they take have no immediate interpretation.

It is thus more convenient to use the autocorrelations, which are the autocovariances normalised by dividing by the variance

$$\tau_s = \frac{\gamma_s}{\gamma_0}, \quad s = 0, 1, 2, \ldots \tag{5.6}$$

The series τ_s now has the standard property of correlation coefficients that the values are bounded to lie between ± 1. In the case that $s = 0$, the autocorrelation at lag zero is obtained, i.e. the correlation of y_t with y_t, which is of course 1. If τ_s is plotted against $s = 0, 1, 2, \ldots$, a graph known as the *autocorrelation function* (acf) or *correlogram* is obtained.

5.2.3 A white noise process

Roughly speaking, a white noise process is one with no discernible structure. A definition of a white noise process is

$$E(y_t) = \mu \tag{5.7}$$

$$\text{var}(y_t) = \sigma^2 \tag{5.8}$$

$$\gamma_{t-r} = \begin{cases} \sigma^2 & \text{if} \quad t = r \\ 0 & \quad otherwise \end{cases} \tag{5.9}$$

Thus a white noise process has constant mean and variance, and zero autocovariances, except at lag zero. Another way to state this last condition would be to say that each observation is uncorrelated with all other values in the sequence. Hence the autocorrelation function for a white noise process will be zero apart from a single peak of 1 at $s = 0$. If $\mu = 0$, and the three conditions hold, the process is known as zero mean white noise.

If it is further assumed that y_t is distributed normally, then the sample autocorrelation coefficients are also approximately normally distributed

$$\hat{\tau}_s \sim approx.\ N(0, 1/T)$$

where T is the sample size, and $\hat{\tau}_s$ denotes the autocorrelation coefficient at lag s estimated from a sample. This result can be used to conduct significance tests for the autocorrelation coefficients by constructing a non-rejection region (like a confidence interval) for an estimated autocorrelation coefficient to determine whether it is significantly different from zero. For example, a 95% non-rejection region would be given by

$$\pm 1.96 \times \frac{1}{\sqrt{T}}$$

for $s \neq 0$. If the sample autocorrelation coefficient, $\hat{\tau}_s$, falls outside this region for a given value of s, then the null hypothesis that the true value of the coefficient at that lag s is zero is rejected.

It is also possible to test the joint hypothesis that all m of the τ_k correlation coefficients are simultaneously equal to zero using the Q-statistic developed by Box and Pierce (1970)

$$Q = T \sum_{k=1}^{m} \hat{\tau}_k^2 \tag{5.10}$$

where T = sample size, m = maximum lag length.

The correlation coefficients are squared so that the positive and negative coefficients do not cancel each other out. Since the sum of squares of independent standard normal variates is itself a χ^2 variate with degrees

of freedom equal to the number of squares in the sum, it can be stated that the Q-statistic is asymptotically distributed as a χ^2_m under the null hypothesis that all m autocorrelation coefficients are zero. As for any joint hypothesis test, only one autocorrelation coefficient needs to be statistically significant for the test to result in a rejection.

However, the Box–Pierce test has poor small sample properties, implying that it leads to the wrong decision too frequently for small samples. A variant of the Box–Pierce test, having better small sample properties, has been developed. The modified statistic is known as the Ljung–Box (1978) statistic

$$Q^* = T(T+2)\sum_{k=1}^{m}\frac{\hat{\tau}_k^2}{T-k} \sim \chi^2_m \tag{5.11}$$

It should be clear from the form of the statistic that asymptotically (that is, as the sample size increases towards infinity), the $(T+2)$ and $(T-k)$ terms in the Ljung–Box formulation will cancel out, so that the statistic is equivalent to the Box–Pierce test. This statistic is very useful as a portmanteau (general) test of linear dependence in time series.

Example 5.1

Suppose that a researcher had estimated the first five autocorrelation coefficients using a series of length 100 observations, and found them to be

Lag	1	2	3	4	5
Autocorrelation coefficient	0.207	−0.013	0.086	0.005	−0.022

Test each of the individual correlation coefficients for significance, and test all five jointly using the Box–Pierce and Ljung–Box tests.

A 95% confidence interval can be constructed for each coefficient using

$$\pm 1.96 \times \frac{1}{\sqrt{T}}$$

where $T = 100$ in this case. The decision rule is thus to reject the null hypothesis that a given coefficient is zero in the cases where the coefficient lies outside the range (−0.196, +0.196). For this example, it would be concluded that only the first autocorrelation coefficient is significantly different from zero at the 5% level.

Now, turning to the joint tests, the null hypothesis is that all of the first five autocorrelation coefficients are jointly zero, i.e.

$$H_0 : \tau_1 = 0, \tau_2 = 0, \tau_3 = 0, \tau_4 = 0, \tau_5 = 0$$

The test statistics for the Box–Pierce and Ljung–Box tests are given respectively as

$$\begin{aligned} Q &= 100 \times (0.207^2 + -0.013^2 + 0.086^2 + 0.005^2 + -0.022^2) \\ &= 5.09 \end{aligned} \tag{5.12}$$

$$\begin{aligned} Q^* = 100 \times 102 \times \Bigg(&\frac{0.207^2}{100-1} + \frac{-0.013^2}{100-2} + \frac{0.086^2}{100-3} \\ &+ \frac{0.005^2}{100-4} + \frac{-0.022^2}{100-5} \Bigg) = 5.26 \end{aligned} \tag{5.13}$$

The relevant critical values are from a χ^2 distribution with 5 degrees of freedom, which are 11.1 at the 5% level, and 15.1 at the 1% level. Clearly, in both cases, the joint null hypothesis that all of the first five autocorrelation coefficients are zero cannot be rejected. Note that, in this instance, the individual test caused a rejection while the joint test did not. This is an unexpected result that may have arisen as a result of the low power of the joint test when four of the five individual autocorrelation coefficients are insignificant. Thus the effect of the significant autocorrelation coefficient is diluted in the joint test by the insignificant coefficients. The sample size used in this example is also modest relative to those commonly available in finance.

5.3 Moving average processes

The simplest class of time series model that one could entertain is that of the moving average process. Let u_t ($t = 1, 2, 3, \ldots$) be a white noise process with $\mathrm{E}(u_t) = 0$ and $\mathrm{var}(u_t) = \sigma^2$. Then

$$y_t = \mu + u_t + \theta_1 u_{t-1} + \theta_2 u_{t-2} + \cdots + \theta_q u_{t-q} \tag{5.14}$$

is a qth order moving average mode, denoted MA(q). This can be expressed using sigma notation as

$$y_t = \mu + \sum_{i=1}^{q} \theta_i u_{t-i} + u_t \tag{5.15}$$

A moving average model is simply a linear combination of white noise processes, so that y_t depends on the current and previous values of a white noise disturbance term. Equation (5.15) will later have to be manipulated, and such a process is most easily achieved by introducing the lag operator notation. This would be written $Ly_t = y_{t-1}$ to denote that y_t is lagged once. In order to show that the ith lag of y_t is being taken (that is, the value that y_t took i periods ago), the notation would be $L^i y_t = y_{t-i}$. Note that in

some books and studies, the lag operator is referred to as the 'backshift operator', denoted by B. Using the lag operator notation, (5.15) would be written as

$$y_t = \mu + \sum_{i=1}^{q} \theta_i L^i u_t + u_t \tag{5.16}$$

or as

$$y_t = \mu + \theta(L)u_t \tag{5.17}$$

where: $\theta(L) = 1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q$.

In much of what follows, the constant (μ) is dropped from the equations. Removing μ considerably eases the complexity of algebra involved, and is inconsequential for it can be achieved without loss of generality. To see this, consider a sample of observations on a series, z_t that has a mean $\bar{z}$. A zero-mean series, y_t can be constructed by simply subtracting $\bar{z}$ from each observation z_t.

The distinguishing properties of the moving average process of order q given above are

$$(1)\ \mathrm{E}(y_t) = \mu \tag{5.18}$$

$$(2)\ \mathrm{var}(y_t) = \gamma_0 = \left(1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2\right)\sigma^2 \tag{5.19}$$

(3) covariances γ_s

$$= \begin{cases} (\theta_s + \theta_{s+1}\theta_1 + \theta_{s+2}\theta_2 + \cdots + \theta_q\theta_{q-s})\,\sigma^2 & \textit{for}\ \ s = 1, 2, \ldots, q \\ 0 & \textit{for}\ \ s > q \end{cases} \tag{5.20}$$

So, a moving average process has constant mean, constant variance, and autocovariances which may be non-zero to lag q and will always be zero thereafter. Each of these results will be derived below.

Example 5.2

Consider the following MA(2) process

$$y_t = u_t + \theta_1 u_{t-1} + \theta_2 u_{t-2} \tag{5.21}$$

where u_t is a zero mean white noise process with variance σ^2.

(1) Calculate the mean and variance of y_t
(2) Derive the autocorrelation function for this process (i.e. express the autocorrelations, $\tau_1, \tau_2, \ldots$ as functions of the parameters θ_1 and θ_2)
(3) If $\theta_1 = -0.5$ and $\theta_2 = 0.25$, sketch the acf of y_t.

Solution

(1) If $E(u_t) = 0$, then $\mathrm{E}(u_{t-i}) = 0 \; \forall \; i$ (5.22)

So the expected value of the error term is zero for all time periods. Taking expectations of both sides of (5.21) gives

$$\begin{aligned} \mathrm{E}(y_t) &= \mathrm{E}(u_t + \theta_1 u_{t-1} + \theta_2 u_{t-2}) \\ &= \mathrm{E}(u_t) + \theta_1 \mathrm{E}(u_{t-1}) + \theta_2 \mathrm{E}(u_{t-2}) = 0 \end{aligned} \tag{5.23}$$

$$\mathrm{var}(y_t) = \mathrm{E}[y_t - \mathrm{E}(y_t)][y_t - \mathrm{E}(y_t)] \tag{5.24}$$

but $\mathrm{E}(y_t) = 0$, so that the last component in each set of square brackets in (5.24) is zero and this reduces to

$$\mathrm{var}(y_t) = \mathrm{E}[(y_t)(y_t)] \tag{5.25}$$

Replacing y_t in (5.25) with the RHS of (5.21)

$$\mathrm{var}(y_t) = \mathrm{E}[(u_t + \theta_1 u_{t-1} + \theta_2 u_{t-2})(u_t + \theta_1 u_{t-1} + \theta_2 u_{t-2})] \tag{5.26}$$

$$\mathrm{var}(y_t) = \mathrm{E}\left[u_t^2 + \theta_1^2 u_{t-1}^2 + \theta_2^2 u_{t-2}^2 + \textit{cross-products}\right] \tag{5.27}$$

But E[*cross-products*] = 0 since $\mathrm{cov}(u_t, u_{t-s}) = 0$ for $s \neq 0$. '*Cross-products*' is thus a catchall expression for all of the terms in u which have different time subscripts, such as $u_{t-1}u_{t-2}$ or $u_{t-5}u_{t-20}$, etc. Again, one does not need to worry about these cross-product terms, since these are effectively the autocovariances of u_t, which will all be zero by definition since u_t is a random error process, which will have zero autocovariances (except at lag zero). So

$$\mathrm{var}(y_t) = \gamma_0 = E\left[u_t^2 + \theta_1^2 u_{t-1}^2 + \theta_2^2 u_{t-2}^2\right] \tag{5.28}$$

$$\mathrm{var}(y_t) = \gamma_0 = \sigma^2 + \theta_1^2 \sigma^2 + \theta_2^2 \sigma^2 \tag{5.29}$$

$$\mathrm{var}(y_t) = \gamma_0 = \left(1 + \theta_1^2 + \theta_2^2\right) \sigma^2 \tag{5.30}$$

γ_0 can also be interpreted as the autocovariance at lag zero.

(2) Calculating now the acf of y_t, first determine the autocovariances and then the autocorrelations by dividing the autocovariances by the variance.

The autocovariance at lag 1 is given by

$$\gamma_1 = \mathrm{E}[y_t - \mathrm{E}(y_t)][y_{t-1} - \mathrm{E}(y_{t-1})] \tag{5.31}$$

$$\gamma_1 = \mathrm{E}[y_t][y_{t-1}] \tag{5.32}$$

$$\gamma_1 = \mathrm{E}[(u_t + \theta_1 u_{t-1} + \theta_2 u_{t-2})(u_{t-1} + \theta_1 u_{t-2} + \theta_2 u_{t-3})] \tag{5.33}$$

Again, ignoring the cross-products, (5.33) can be written as

$$\gamma_1 = \mathrm{E}\left[\left(\theta_1 u_{t-1}^2 + \theta_1\theta_2 u_{t-2}^2\right)\right] \tag{5.34}$$

$$\gamma_1 = \theta_1\sigma^2 + \theta_1\theta_2\sigma^2 \tag{5.35}$$

$$\gamma_1 = (\theta_1 + \theta_1\theta_2)\sigma^2 \tag{5.36}$$

The autocovariance at lag 2 is given by

$$\gamma_2 = \mathrm{E}[y_t - \mathrm{E}(y_t)][y_{t-2} - \mathrm{E}(y_{t-2})] \tag{5.37}$$

$$\gamma_2 = \mathrm{E}[y_t][y_{t-2}] \tag{5.38}$$

$$\gamma_2 = \mathrm{E}[(u_t + \theta_1 u_{t-1} + \theta_2 u_{t-2})(u_{t-2} + \theta_1 u_{t-3} + \theta_2 u_{t-4})] \tag{5.39}$$

$$\gamma_2 = \mathrm{E}\left[\left(\theta_2 u_{t-2}^2\right)\right] \tag{5.40}$$

$$\gamma_2 = \theta_2\sigma^2 \tag{5.41}$$

The autocovariance at lag 3 is given by

$$\gamma_3 = \mathrm{E}[y_t - \mathrm{E}(y_t)][y_{t-3} - \mathrm{E}(y_{t-3})] \tag{5.42}$$

$$\gamma_3 = \mathrm{E}[y_t][y_{t-3}] \tag{5.43}$$

$$\gamma_3 = \mathrm{E}[(u_t + \theta_1 u_{t-1} + \theta_2 u_{t-2})(u_{t-3} + \theta_1 u_{t-4} + \theta_2 u_{t-5})] \tag{5.44}$$

$$\gamma_3 = 0 \tag{5.45}$$

So $\gamma_s = 0$ for s 2. All autocovariances for the MA(2) process will be zero for any lag length, s, greater than 2.

The autocorrelation at lag 0 is given by

$$\tau_0 = \frac{\gamma_0}{\gamma_0} = 1 \tag{5.46}$$

The autocorrelation at lag 1 is given by

$$\tau_1 = \frac{\gamma_1}{\gamma_0} = \frac{(\theta_1 + \theta_1\theta_2)\sigma^2}{\left(1 + \theta_1^2 + \theta_2^2\right)\sigma^2} = \frac{(\theta_1 + \theta_1\theta_2)}{\left(1 + \theta_1^2 + \theta_2^2\right)} \tag{5.47}$$

The autocorrelation at lag 2 is given by

$$\tau_2 = \frac{\gamma_2}{\gamma_0} = \frac{(\theta_2)\sigma^2}{\left(1 + \theta_1^2 + \theta_2^2\right)\sigma^2} = \frac{\theta_2}{\left(1 + \theta_1^2 + \theta_2^2\right)} \tag{5.48}$$

The autocorrelation at lag 3 is given by

$$\tau_3 = \frac{\gamma_3}{\gamma_0} = 0 \tag{5.49}$$

The autocorrelation at lag s is given by

$$\tau_s = \frac{\gamma_s}{\gamma_0} = 0 \ \forall \ s > 2 \tag{5.50}$$

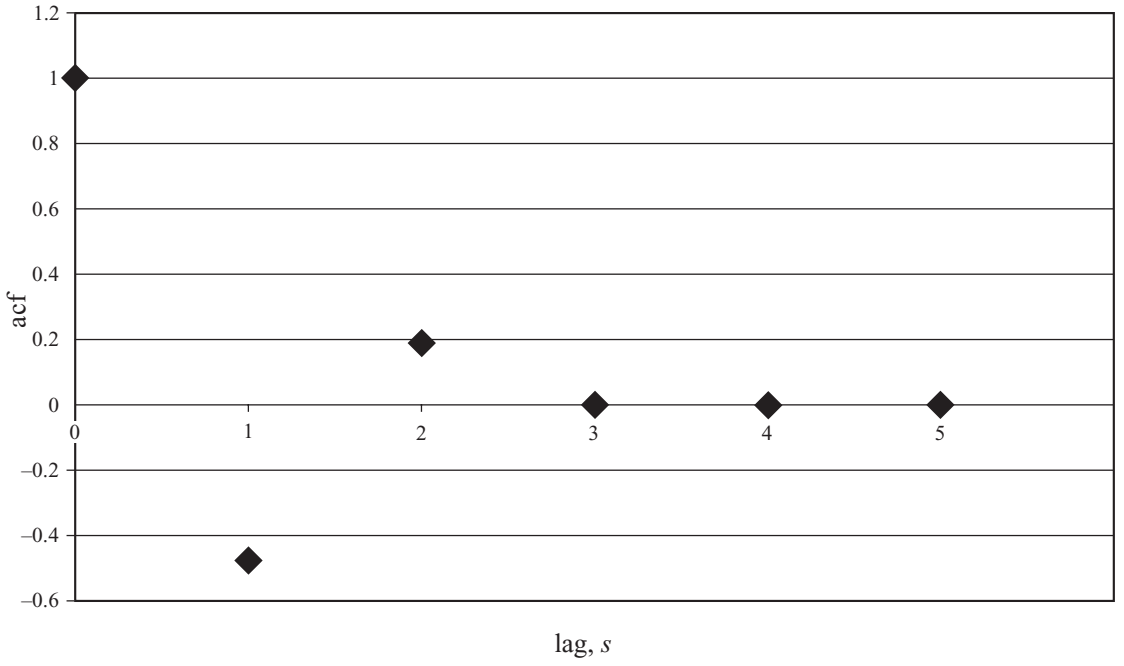


Figure 5.1 Autocorrelation function for sample MA(2) process

(3) For $\theta_1 = -0.5$ and $\theta_2 = 0.25$, substituting these into the formulae above gives the first two autocorrelation coefficients as $\tau_1 = -0.476$, $\tau_2 = 0.190$. Autocorrelation coefficients for lags greater than 2 will all be zero for an MA(2) model. Thus the acf plot will appear as in figure 5.1.

5.4 Autoregressive processes

An autoregressive model is one where the current value of a variable, y, depends upon only the values that the variable took in previous periods plus an error term. An autoregressive model of order p, denoted as AR(p), can be expressed as

$$y_t = \mu + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + u_t \tag{5.51}$$

where u_t is a white noise disturbance term. A manipulation of expression (5.51) will be required to demonstrate the properties of an autoregressive model. This expression can be written more compactly using sigma notation

$$y_t = \mu + \sum_{i=1}^{p} \phi_i y_{t-i} + u_t \tag{5.52}$$

or using the lag operator, as

$$y_t = \mu + \sum_{i=1}^{p} \phi_i L^i y_t + u_t \tag{5.53}$$

or

$$\phi(L) y_t = \mu + u_t \tag{5.54}$$

where $\phi(L) = (1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p)$.

5.4.1 The stationarity condition

Stationarity is a desirable property of an estimated AR model, for several reasons. One important reason is that a model whose coefficients are non-stationary will exhibit the unfortunate property that previous values of the error term will have a non-declining effect on the current value of y_t as time progresses. This is arguably counter-intuitive and empirically implausible in many cases. More discussion on this issue will be presented in chapter 7. Box 5.1 defines the stationarity condition algebraically.

Box 5.1 The stationarity condition for an AR(p) model

Setting μ to zero in (5.54), for a zero mean AR (p) process, y_t, given by

$$\phi(L) y_t = u_t \tag{5.55}$$

it would be stated that the process is stationary if it is possible to write

$$y_t = \phi(L)^{-1} u_t \tag{5.56}$$

with $\phi(L)^{-1}$ converging to zero. This means that the autocorrelations will decline eventually as the lag length is increased. When the expansion $\phi(L)^{-1}$ is calculated, it will contain an infinite number of terms, and can be written as an MA(∞), e.g. $a_1 u_{t-1} + a_2 u_{t-2} + a_3 u_{t-3} + \cdots + u_t$. If the process given by (5.54) is stationary, the coefficients in the MA(∞) representation will decline eventually with lag length. On the other hand, if the process is non-stationary, the coefficients in the MA(∞) representation would not converge to zero as the lag length increases.

The condition for testing for the stationarity of a general AR(p) model is that the roots of the 'characteristic equation'

$$1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p = 0 \tag{5.57}$$

all lie outside the unit circle. The notion of a characteristic equation is so-called because its roots determine the characteristics of the process y_t – for example, the acf for an AR process will depend on the roots of this characteristic equation, which is a polynomial in z.

Example 5.3

Is the following model stationary?

$$y_t = y_{t-1} + u_t \tag{5.58}$$

In order to test this, first write y_{t-1} in lag operator notation (i.e. as Ly_t), and take this term over to the LHS of (5.58), and factorise

$$y_t = Ly_t + u_t \tag{5.59}$$
$$y_t - Ly_t = u_t \tag{5.60}$$
$$y_t(1 - L) = u_t \tag{5.61}$$

Then the characteristic equation is

$$1 - z = 0, \tag{5.62}$$

having the root $z = 1$, which lies on, not outside, the unit circle. In fact, the particular AR(p) model given by (5.58) is a non-stationary process known as a random walk (see chapter 7).

This procedure can also be adopted for autoregressive models with longer lag lengths and where the stationarity or otherwise of the process is less obvious. For example, is the following process for y_t stationary?

$$y_t = 3y_{t-1} - 2.75y_{t-2} + 0.75y_{t-3} + u_t \tag{5.63}$$

Again, the first stage is to express this equation using the lag operator notation, and then taking all the terms in y over to the LHS

$$y_t = 3Ly_t - 2.75L^2y_t + 0.75L^3y_t + u_t \tag{5.64}$$
$$(1 - 3L + 2.75L^2 - 0.75L^3)y_t = u_t \tag{5.65}$$

The characteristic equation is

$$1 - 3z + 2.75z^2 - 0.75z^3 = 0 \tag{5.66}$$

which fortunately factorises to

$$(1 - z)(1 - 1.5z)(1 - 0.5z) = 0 \tag{5.67}$$

so that the roots are $z = 1$, $z = 2/3$, and $z = 2$. Only one of these lies outside the unit circle and hence the process for y_t described by (5.63) is not stationary.

5.4.2 *Wold's decomposition theorem*

Wold's decomposition theorem states that any stationary series can be decomposed into the sum of two unrelated processes, a purely deterministic

part and a purely stochastic part, which will be an MA(∞). A simpler way of stating this in the context of AR modelling is that any stationary autoregressive process of order p with no constant and no other terms can be expressed as an infinite order moving average model. This result is important for deriving the autocorrelation function for an autoregressive process.

For the AR(p) model, given in, for example, (5.51) (with μ set to zero for simplicity) and expressed using the lag polynomial notation, $\phi(L)y_t = u_t$, the Wold decomposition is

$$y_t = \psi(L)u_t \tag{5.68}$$

where $\psi(L) = \phi(L)^{-1} = (1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p)^{-1}$

The characteristics of an autoregressive process are as follows. The (unconditional) mean of y is given by

$$E(y_t) = \frac{\mu}{1 - \phi_1 - \phi_2 - \cdots - \phi_p} \tag{5.69}$$

The autocovariances and autocorrelation functions can be obtained by solving a set of simultaneous equations known as the Yule–Walker equations. The Yule–Walker equations express the correlogram (the τs) as a function of the autoregressive coefficients (the ϕs)

$$\begin{aligned} \tau_1 &= \phi_1 + \tau_1\phi_2 + \cdots + \tau_{p-1}\phi_p \\ \tau_2 &= \tau_1\phi_1 + \phi_2 + \cdots + \tau_{p-2}\phi_p \\ &\vdots\ \vdots\ \vdots \\ \tau_p &= \tau_{p-1}\phi_1 + \tau_{p-2}\phi_2 + \cdots + \phi_p \end{aligned} \tag{5.70}$$

For any AR model that is stationary, the autocorrelation function will decay geometrically to zero.[1] These characteristics of an autoregressive process will be derived from first principles below using an illustrative example.

Example 5.4

Consider the following simple AR(1) model

$$y_t = \mu + \phi_1 y_{t-1} + u_t \tag{5.71}$$

(i) Calculate the (unconditional) mean y_t.
For the remainder of the question, set the constant to zero ($\mu = 0$) for simplicity.

[1] Note that the τ_s will not follow an exact geometric sequence, but rather the absolute value of the τ_s is bounded by a geometric series. This means that the autocorrelation function does not have to be monotonically decreasing and may change sign.

(ii) Calculate the (unconditional) variance of y_t.
(iii) Derive the autocorrelation function for this process.

Solution

(i) The unconditional mean will be given by the expected value of expression (5.71)

$$\mathrm{E}(y_t) = \mathrm{E}(\mu + \phi_1 y_{t-1}) \tag{5.72}$$
$$\mathrm{E}(y_t) = \mu + \phi_1 \mathrm{E}(y_{t-1}) \tag{5.73}$$

But also

$$y_{t-1} = \mu + \phi_1 y_{t-2} + u_{t-1} \tag{5.74}$$

So, replacing y_{t-1} in (5.73) with the RHS of (5.74)

$$\mathrm{E}(y_t) = \mu + \phi_1(\mu + \phi_1 \mathrm{E}(y_{t-2})) \tag{5.75}$$
$$\mathrm{E}(y_t) = \mu + \phi_1\mu + \phi_1^2 \mathrm{E}(y_{t-2}) \tag{5.76}$$

Lagging (5.74) by a further one period

$$y_{t-2} = \mu + \phi_1 y_{t-3} + u_{t-2} \tag{5.77}$$

Repeating the steps given above one more time

$$\mathrm{E}(y_t) = \mu + \phi_1\mu + \phi_1^2(\mu + \phi_1 \mathrm{E}(y_{t-3})) \tag{5.78}$$
$$\mathrm{E}(y_t) = \mu + \phi_1\mu + \phi_1^2\mu + \phi_1^3 \mathrm{E}(y_{t-3}) \tag{5.79}$$

Hopefully, readers will by now be able to see a pattern emerging. Making n such substitutions would give

$$\mathrm{E}(y_t) = \mu\left(1 + \phi_1 + \phi_1^2 + \cdots + \phi_1^{n-1}\right) + \phi_1^t \mathrm{E}(y_{t-n}) \tag{5.80}$$

So long as the model is stationary, i.e. $|\phi_1| < 1$, then $\phi_1^\infty = 0$. Therefore, taking limits as $n \to \infty$, then $lim_{n\to\infty}\phi_1^t \mathrm{E}(y_{t-n}) = 0$, and so

$$\mathrm{E}(y_t) = \mu\left(1 + \phi_1 + \phi_1^2 + \cdots\right) \tag{5.81}$$

Recall the rule of algebra that the finite sum of an infinite number of geometrically declining terms in a series is given by 'first term in series divided by (1 minus common difference)', where the common difference is the quantity that each term in the series is multiplied by to arrive at the next term. It can thus be stated from (5.81) that

$$E(y_t) = \frac{\mu}{1 - \phi_1} \tag{5.82}$$

Thus the expected or mean value of an autoregressive process of order one is given by the intercept parameter divided by one minus the autoregressive coefficient.

(ii) Calculating now the variance of y_t, with μ set to zero

$$y_t = \phi_1 y_{t-1} + u_t \tag{5.83}$$

This can be written equivalently as

$$y_t(1 - \phi_1 L) = u_t \tag{5.84}$$

From Wold's decomposition theorem, the AR(p) can be expressed as an MA(∞)

$$y_t = (1 - \phi_1 L)^{-1} u_t \tag{5.85}$$

$$y_t = \left(1 + \phi_1 L + \phi_1^2 L^2 + \cdots\right) u_t \tag{5.86}$$

or

$$y_t = u_t + \phi_1 u_{t-1} + \phi_1^2 u_{t-2} + \phi_1^3 u_{t-3} + \cdots \tag{5.87}$$

So long as $|\phi_1| < 1$, i.e. so long as the process for y_t is stationary, this sum will converge.

From the definition of the variance of any random variable y, it is possible to write

$$\text{var}(y_t) = \text{E}[y_t - \text{E}(y_t)][y_t - \text{E}(y_t)] \tag{5.88}$$

but $\text{E}(y_t) = 0$, since μ is set to zero to obtain (5.83) above. Thus

$$\text{var}(y_t) = \text{E}[(y_t)(y_t)] \tag{5.89}$$

$$\text{var}(y_t) = \text{E}\left[\left(u_t + \phi_1 u_{t-1} + \phi_1^2 u_{t-2} + \cdots\right)\left(u_t + \phi_1 u_{t-1} + \phi_1^2 u_{t-2} + \cdots\right)\right] \tag{5.90}$$

$$\text{var}(y_t) = \text{E}\left[u_t^2 + \phi_1^2 u_{t-1}^2 + \phi_1^4 u_{t-2}^2 + \cdots + \textit{cross-products}\right] \tag{5.91}$$

As discussed above, the 'cross-products' can be set to zero.

$$\text{var}(y_t) = \gamma_0 = \text{E}\left[u_t^2 + \phi_1^2 u_{t-1}^2 + \phi_1^4 u_{t-2}^2 + \cdots\right] \tag{5.92}$$

$$\text{var}(y_t) = \sigma^2 + \phi_1^2 \sigma^2 + \phi_1^4 \sigma^2 + \cdots \tag{5.93}$$

$$\text{var}(y_t) = \sigma^2 \left(1 + \phi_1^2 + \phi_1^4 + \cdots\right) \tag{5.94}$$

Provided that $|\phi_1| < 1$, the infinite sum in (5.94) can be written as

$$\text{var}(y_t) = \frac{\sigma^2}{\left(1 - \phi_1^2\right)} \tag{5.95}$$

(iii) Turning now to the calculation of the autocorrelation function, the autocovariances must first be calculated. This is achieved by following

similar algebraic manipulations as for the variance above, starting with the definition of the autocovariances for a random variable. The autocovariances for lags 1, 2, 3, ..., s, will be denoted by $\gamma_1, \gamma_2, \gamma_3, \ldots, \gamma_s$, as previously.

$$\gamma_1 = \text{cov}\,(y_t, y_{t-1}) = \text{E}[y_t - E(y_t)][y_{t-1} - E(y_{t-1})] \tag{5.96}$$

Since μ has been set to zero, $\text{E}(y_t) = 0$ and $\text{E}(y_{t-1}) = 0$, so

$$\gamma_1 = \text{E}[y_t y_{t-1}] \tag{5.97}$$

under the result above that $\text{E}(y_t) = \text{E}(y_{t-1}) = 0$. Thus

$$\gamma_1 = \text{E}\big[\big(u_t + \phi_1 u_{t-1} + \phi_1^2 u_{t-2} + \cdots\big)\big(u_{t-1} + \phi_1 u_{t-2} + \phi_1^2 u_{t-3} + \cdots\big)\big] \tag{5.98}$$

$$\gamma_1 = \text{E}\big[\phi_1 u_{t-1}^2 + \phi_1^3 u_{t-2}^2 + \cdots + cross - products\big] \tag{5.99}$$

Again, the cross-products can be ignored so that

$$\gamma_1 = \phi_1\sigma^2 + \phi_1^3\sigma^2 + \phi_1^5\sigma^2 + \cdots \tag{5.100}$$

$$\gamma_1 = \phi_1\sigma^2\big(1 + \phi_1^2 + \phi_1^4 + \cdots\big) \tag{5.101}$$

$$\gamma_1 = \frac{\phi_1\sigma^2}{\big(1 - \phi_1^2\big)} \tag{5.102}$$

For the second autocovariance,

$$\gamma_2 = \text{cov}(y_t, y_{t-2}) = \text{E}[y_t - \text{E}(y_t)][y_{t-2} - \text{E}(y_{t-2})] \tag{5.103}$$

Using the same rules as applied above for the lag 1 covariance

$$\gamma_2 = \text{E}[y_t y_{t-2}] \tag{5.104}$$

$$\gamma_2 = \text{E}\big[\big(u_t + \phi_1 u_{t-1} + \phi_1^2 u_{t-2} + \cdots\big)\big(u_{t-2} + \phi_1 u_{t-3} + \phi_1^2 u_{t-4} + \cdots\big)\big] \tag{5.105}$$

$$\gamma_2 = \text{E}\big[\phi_1^2 u_{t-2}^2 + \phi_1^4 u_{t-3}^2 + \cdots + cross\text{-}products\big] \tag{5.106}$$

$$\gamma_2 = \phi_1^2\sigma^2 + \phi_1^4\sigma^2 + \cdots \tag{5.107}$$

$$\gamma_2 = \phi_1^2\sigma^2\big(1 + \phi_1^2 + \phi_1^4 + \cdots\big) \tag{5.108}$$

$$\gamma_2 = \frac{\phi_1^2\sigma^2}{\big(1 - \phi_1^2\big)} \tag{5.109}$$

By now it should be possible to see a pattern emerging. If these steps were repeated for γ_3, the following expression would be obtained

$$\gamma_3 = \frac{\phi_1^3\sigma^2}{\big(1 - \phi_1^2\big)} \tag{5.110}$$

and for any lag s, the autocovariance would be given by

$$\gamma_s = \frac{\phi_1^s \sigma^2}{\left(1 - \phi_1^2\right)} \tag{5.111}$$

The acf can now be obtained by dividing the covariances by the variance, so that

$$\tau_0 = \frac{\gamma_0}{\gamma_0} = 1 \tag{5.112}$$

$$\tau_1 = \frac{\gamma_1}{\gamma_0} = \frac{\left(\frac{\phi_1 \sigma^2}{\left(1 - \phi_1^2\right)}\right)}{\left(\frac{\sigma^2}{\left(1 - \phi_1^2\right)}\right)} = \phi_1 \tag{5.113}$$

$$\tau_2 = \frac{\gamma_2}{\gamma_0} = \frac{\left(\frac{\phi_1^2 \sigma^2}{\left(1 - \phi_1^2\right)}\right)}{\left(\frac{\sigma^2}{\left(1 - \phi_1^2\right)}\right)} = \phi_1^2 \tag{5.114}$$

$$\tau_3 = \phi_1^3 \tag{5.115}$$

The autocorrelation at lag s is given by

$$\tau_s = \phi_1^s \tag{5.116}$$

which means that $\text{corr}(y_t, y_{t-s}) = \phi_1^s$. Note that use of the Yule–Walker equations would have given the same answer.

5.5 The partial autocorrelation function

The partial autocorrelation function, or pacf (denoted τ_{kk}), measures the correlation between an observation k periods ago and the current observation, after controlling for observations at intermediate lags (i.e. all lags $< k$) – i.e. the correlation between y_t and y_{t-k}, after removing the effects of $y_{t-k+1}, y_{t-k+2}, \ldots, y_{t-1}$. For example, the pacf for lag 3 would measure the correlation between y_t and y_{t-3} after controlling for the effects of y_{t-1} and y_{t-2}.

At lag 1, the autocorrelation and partial autocorrelation coefficients are equal, since there are no intermediate lag effects to eliminate. Thus, $\tau_{11} = \tau_1$, where τ_1 is the autocorrelation coefficient at lag 1.

At lag 2

$$\tau_{22} = \left(\tau_2 - \tau_1^2\right) / \left(1 - \tau_1^2\right) \tag{5.117}$$

where τ_1 and τ_2 are the autocorrelation coefficients at lags 1 and 2, respectively. For lags greater than two, the formulae are more complex and hence a presentation of these is beyond the scope of this book. There now proceeds, however, an intuitive explanation of the characteristic shape of the pacf for a moving average and for an autoregressive process.

In the case of an autoregressive process of order p, there will be direct connections between y_t and y_{t-s} for $s \leq p$, but no direct connections for $s > p$. For example, consider the following AR(3) model

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \phi_3 y_{t-3} + u_t \tag{5.118}$$

There is a direct connection through the model between y_t and y_{t-1}, and between y_t and y_{t-2}, and between y_t and y_{t-3}, but not between y_t and y_{t-s}, for $s > 3$. Hence the pacf will usually have non-zero partial autocorrelation coefficients for lags up to the order of the model, but will have zero partial autocorrelation coefficients thereafter. In the case of the AR(3), only the first three partial autocorrelation coefficients will be non-zero.

What shape would the partial autocorrelation function take for a moving average process? One would need to think about the MA model as being transformed into an AR in order to consider whether y_t and y_{t-k}, $k = 1, 2, \ldots$, are directly connected. In fact, so long as the MA(q) process is invertible, it can be expressed as an AR(∞). Thus a definition of invertibility is now required.

5.5.1 *The invertibility condition*

An MA(q) model is typically required to have roots of the characteristic equation $\theta(z) = 0$ greater than one in absolute value. The invertibility condition is mathematically the same as the stationarity condition, but is different in the sense that the former refers to MA rather than AR processes. This condition prevents the model from exploding under an AR(∞) representation, so that $\theta^{-1}(L)$ converges to zero. Box 5.2 shows the invertibility condition for an MA(2) model.

5.6 ARMA processes

By combining the AR(p) and MA(q) models, an ARMA(p, q) model is obtained. Such a model states that the current value of some series y depends linearly on its own previous values plus a combination of current and previous values of a white noise error term. The model could be

Box 5.2 The invertibility condition for an MA(2) model

In order to examine the shape of the pacf for moving average processes, consider the following MA(2) process for y_t

$$y_t = u_t + \theta_1 u_{t-1} + \theta_2 u_{t-2} = \theta(L)u_t \tag{5.119}$$

Provided that this process is invertible, this MA(2) can be expressed as an AR(∞)

$$y_t = \sum_{i=1}^{\infty} c_i L^i y_{t-i} + u_t \tag{5.120}$$

$$y_t = c_1 y_{t-1} + c_2 y_{t-2} + c_3 y_{t-3} + \cdots + u_t \tag{5.121}$$

It is now evident when expressed in this way that for a moving average model, there are direct connections between the current value of y and all of its previous values. Thus, the partial autocorrelation function for an MA(q) model will decline geometrically, rather than dropping off to zero after q lags, as is the case for its autocorrelation function. It could thus be stated that the acf for an AR has the same basic shape as the pacf for an MA, and the acf for an MA has the same shape as the pacf for an AR.

written

$$\phi(L)y_t = \mu + \theta(L)u_t \tag{5.122}$$

where

$$\phi(L) = 1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p \quad \text{and}$$
$$\theta(L) = 1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q$$

or

$$y_t = \mu + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \theta_1 u_{t-1} + \theta_2 u_{t-2} + \cdots + \theta_q u_{t-q} + u_t \tag{5.123}$$

with

$$\mathrm{E}(u_t) = 0; E\left(u_t^2\right) = \sigma^2; E(u_t u_s) = 0, t \neq s$$

The characteristics of an ARMA process will be a combination of those from the autoregressive (AR) and moving average (MA) parts. Note that the pacf is particularly useful in this context. The acf alone can distinguish between a pure autoregressive and a pure moving average process. However, an ARMA process will have a geometrically declining acf, as will a pure AR process. So, the pacf is useful for distinguishing between an AR(p) process and an ARMA(p, q) process – the former will have a geometrically declining autocorrelation function, but a partial autocorrelation function which cuts off to zero after p lags, while the latter will have

both autocorrelation and partial autocorrelation functions which decline geometrically.

We can now summarise the defining characteristics of AR, MA and ARMA processes.

An autoregressive process has:

- a geometrically decaying acf
- a number of non-zero points of pacf = AR order.

A moving average process has:

- number of non-zero points of acf = MA order
- a geometrically decaying pacf.

A combination autoregressive moving average process has:

- a geometrically decaying acf
- a geometrically decaying pacf.

In fact, the mean of an ARMA series is given by

$$E(y_t) = \frac{\mu}{1 - \phi_1 - \phi_2 - \cdots - \phi_p} \tag{5.124}$$

The autocorrelation function will display combinations of behaviour derived from the AR and MA parts, but for lags beyond q, the acf will simply be identical to the individual AR(p) model, so that the AR part will dominate in the long term. Deriving the acf and pacf for an ARMA process requires no new algebra, but is tedious and hence is left as an exercise for interested readers.

5.6.1 *Sample acf and pacf plots for standard processes*

Figures 5.2–5.8 give some examples of typical processes from the ARMA family with their characteristic autocorrelation and partial autocorrelation functions. The acf and pacf are not produced analytically from the relevant formulae for a model of that type, but rather are estimated using 100,000 simulated observations with disturbances drawn from a normal distribution. Each figure also has 5% (two-sided) rejection bands represented by dotted lines. These are based on $(\pm 1.96/\sqrt{100000}) = \pm 0.0062$, calculated in the same way as given above. Notice how, in each case, the acf and pacf are identical for the first lag.

In figure 5.2, the MA(1) has an acf that is significant for only lag 1, while the pacf declines geometrically, and is significant until lag 7. The acf at lag 1 and all of the pacfs are negative as a result of the negative coefficient in the MA generating process.

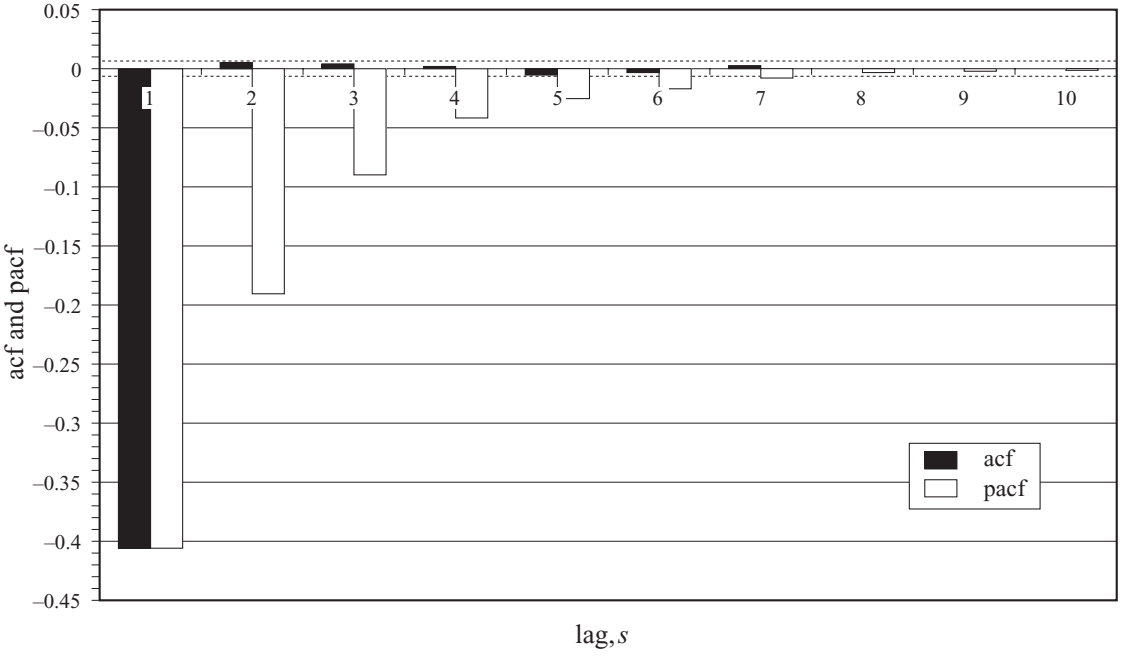


Figure 5.2 Sample autocorrelation and partial autocorrelation functions for an MA(1) model: $y_t = -0.5u_{t-1} + u_t$

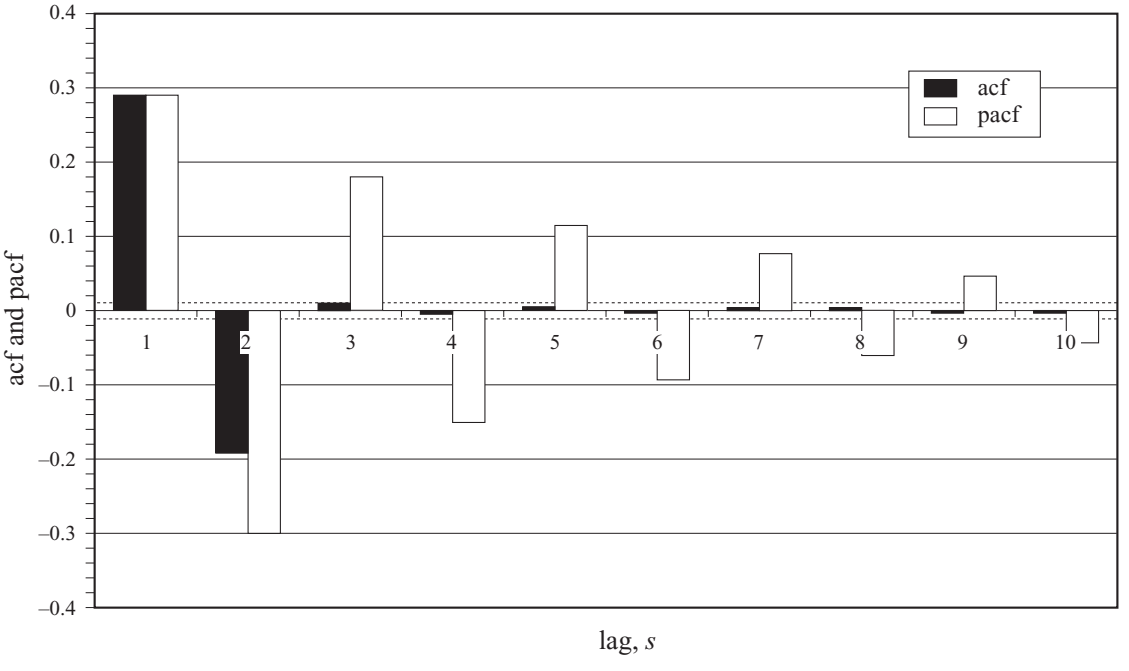


Figure 5.3 Sample autocorrelation and partial autocorrelation functions for an MA(2) model: $y_t = 0.5u_{t-1} - 0.25u_{t-2} + u_t$

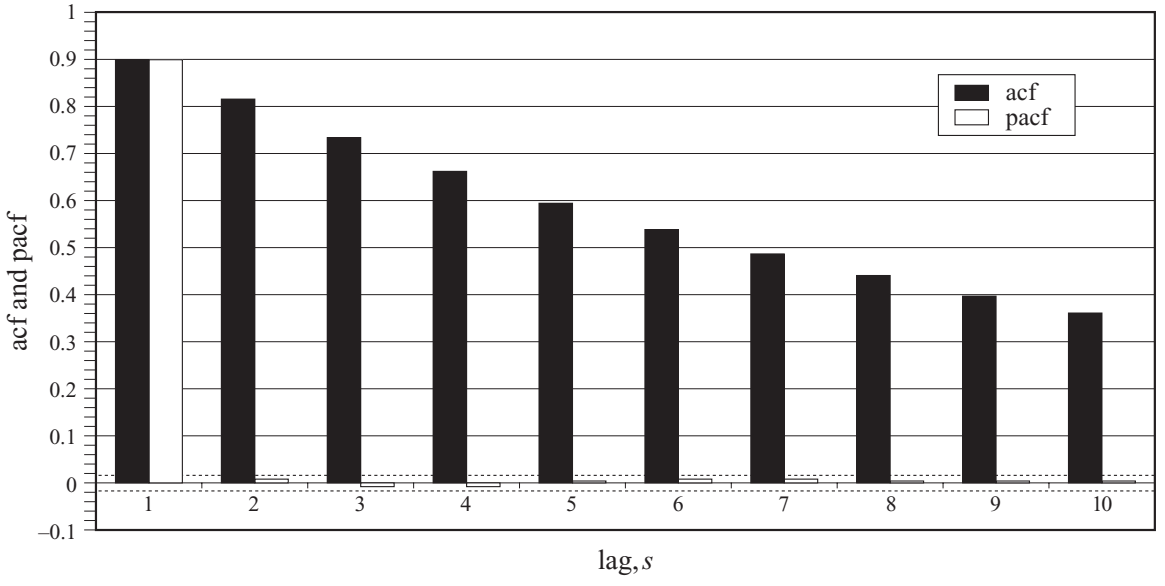


Figure 5.4 Sample autocorrelation and partial autocorrelation functions for a slowly decaying AR(1) model: $y_t = 0.9y_{t-1} + u_t$

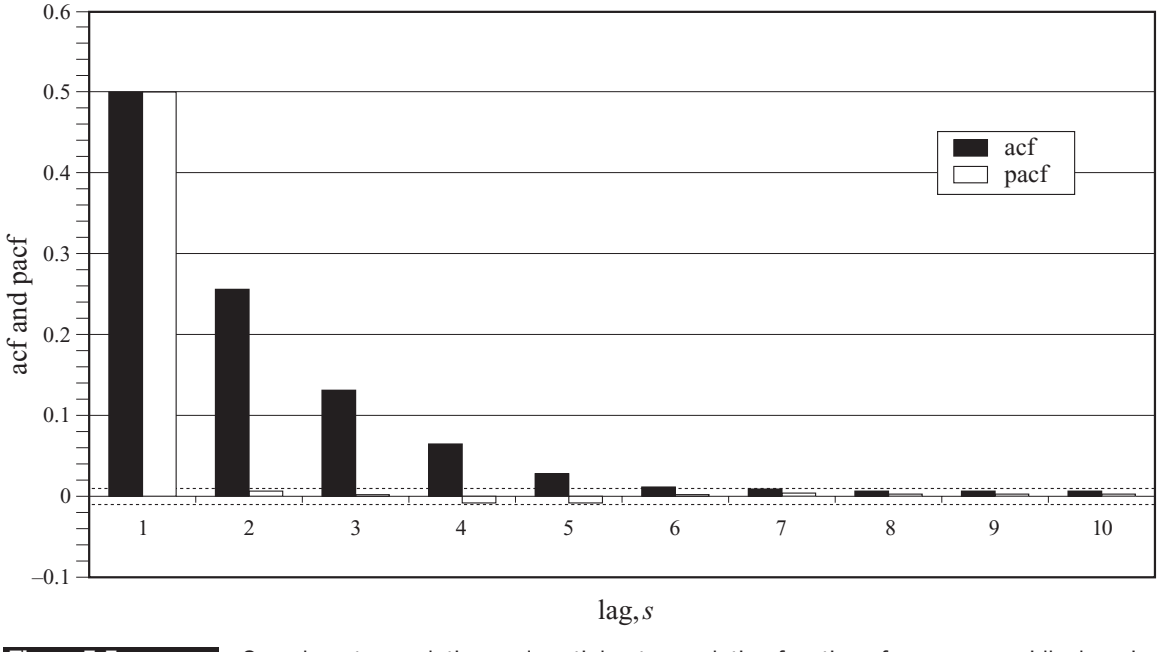


Figure 5.5 Sample autocorrelation and partial autocorrelation functions for a more rapidly decaying AR(1) model: $y_t = 0.5y_{t-1} + u_t$

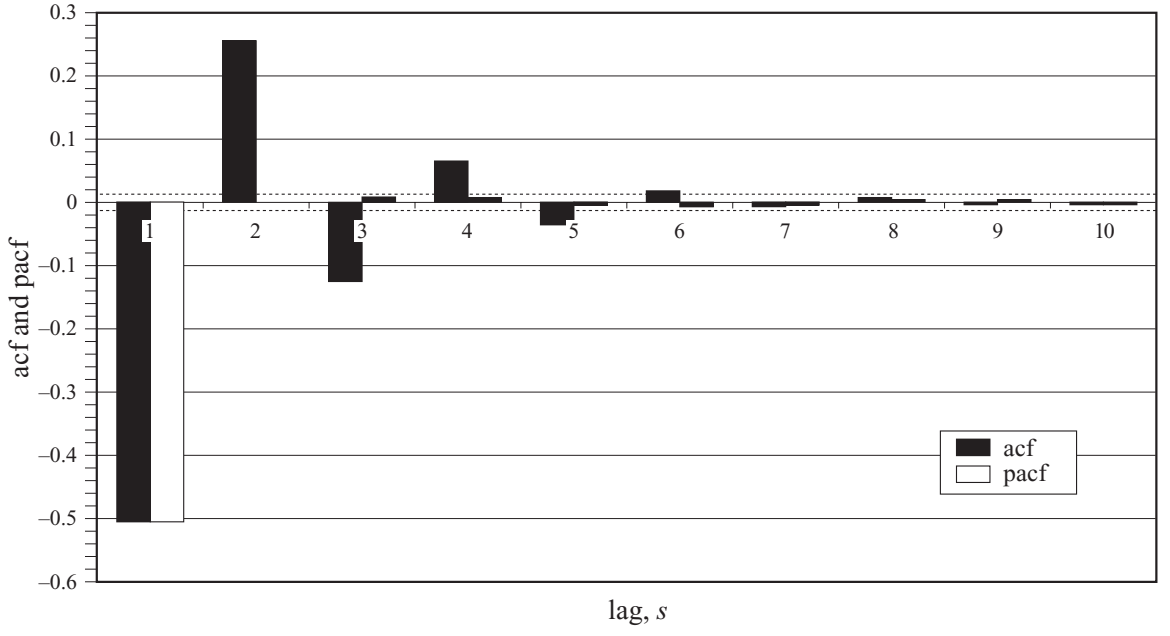


Figure 5.6 Sample autocorrelation and partial autocorrelation functions for a more rapidly decaying AR(1) model with negative coefficient: $y_t = -0.5y_{t-1} + u_t$

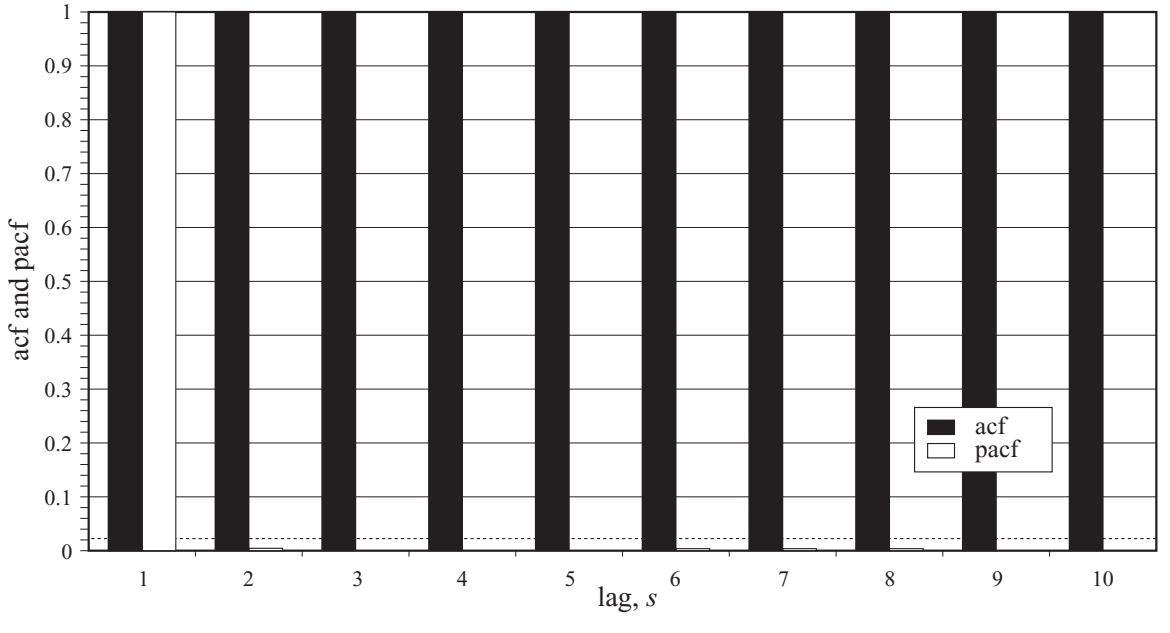


Figure 5.7 Sample autocorrelation and partial autocorrelation functions for a non-stationary model (i.e. a unit coefficient): $y_t = y_{t-1} + u_t$

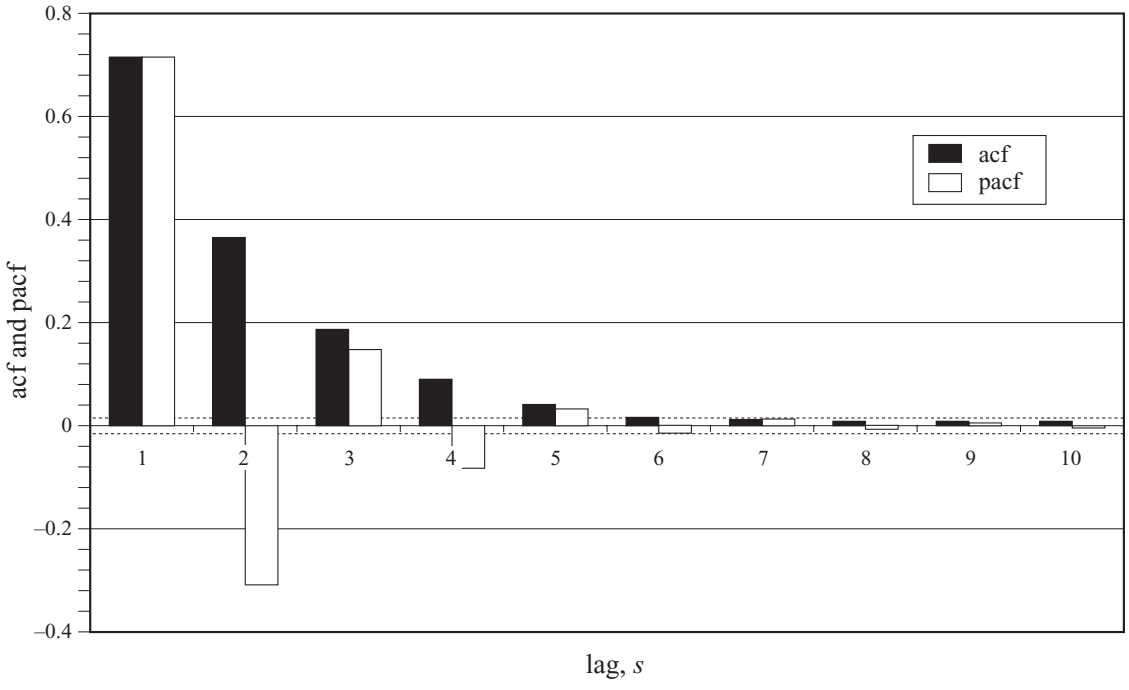


Figure 5.8 Sample autocorrelation and partial autocorrelation functions for an ARMA(1, 1) model: $y_t = 0.5y_{t-1} + 0.5u_{t-1} + u_t$

Again, the structures of the acf and pacf in figure 5.3 are as anticipated. The first two autocorrelation coefficients only are significant, while the partial autocorrelation coefficients are geometrically declining. Note also that, since the second coefficient on the lagged error term in the MA is negative, the acf and pacf alternate between positive and negative. In the case of the pacf, we term this alternating and declining function a 'damped sine wave' or 'damped sinusoid'.

For the autoregressive model of order 1 with a fairly high coefficient – i.e. relatively close to 1 – the autocorrelation function would be expected to die away relatively slowly, and this is exactly what is observed here in figure 5.4. Again, as expected for an AR(1), only the first pacf coefficient is significant, while all others are virtually zero and are not significant.

Figure 5.5 plots an AR(1), which was generated using identical error terms, but a much smaller autoregressive coefficient. In this case, the autocorrelation function dies away much more quickly than in the previous example, and in fact becomes insignificant after around 5 lags.

Figure 5.6 shows the acf and pacf for an identical AR(1) process to that used for figure 5.5, except that the autoregressive coefficient is now negative. This results in a damped sinusoidal pattern for the acf, which again

becomes insignificant after around lag 5. Recalling that the autocorrelation coefficient for this AR(1) at lag s is equal to $(-0.5)^s$, this will be positive for even s, and negative for odd s. Only the first pacf coefficient is significant (and negative).

Figure 5.7 plots the acf and pacf for a non-stationary series (see chapter 7 for an extensive discussion) that has a unit coefficient on the lagged dependent variable. The result is that shocks to y never die away, and persist indefinitely in the system. Consequently, the acf function remains relatively flat at unity, even up to lag 10. In fact, even by lag 10, the autocorrelation coefficient has fallen only to 0.9989. Note also that on some occasions, the acf does die away, rather than looking like figure 5.7, even for such a non-stationary process, owing to its inherent instability combined with finite computer precision. The pacf, however, is significant only for lag 1, correctly suggesting that an autoregressive model with no moving average term is most appropriate.

Finally, figure 5.8 plots the acf and pacf for a mixed ARMA process. As one would expect of such a process, both the acf and the pacf decline geometrically – the acf as a result of the AR part and the pacf as a result of the MA part. The coefficients on the AR and MA are, however, sufficiently small that both acf and pacf coefficients have become insignificant by lag 6.

5.7 Building ARMA models: the Box–Jenkins approach

Although the existence of ARMA models predates them, Box and Jenkins (1976) were the first to approach the task of estimating an ARMA model in a systematic manner. Their approach was a practical and pragmatic one, involving three steps:

(1) Identification
(2) Estimation
(3) Diagnostic checking.

These steps are now explained in greater detail.

Step 1

This involves *determining the order of the model required* to capture the dynamic features of the data. Graphical procedures are used (plotting the data over time and plotting the acf and pacf) to determine the most appropriate specification.

Step 2

This involves *estimation of the parameters of the model* specified in step 1. This can be done using least squares or another technique, known as maximum likelihood, depending on the model.

Step 3

This involves *model checking* – i.e. determining whether the model specified and estimated is adequate. Box and Jenkins suggest two methods: overfitting and residual diagnostics. *Overfitting* involves deliberately fitting a larger model than that required to capture the dynamics of the data as identified in stage 1. If the model specified at step 1 is adequate, any extra terms added to the ARMA model would be insignificant. *Residual diagnostics* imply checking the residuals for evidence of linear dependence which, if present, would suggest that the model originally specified was inadequate to capture the features of the data. The acf, pacf or Ljung–Box tests could be used.

It is worth noting that 'diagnostic testing' in the Box–Jenkins world essentially involves only autocorrelation tests rather than the whole barrage of tests outlined in chapter 4. Also, such approaches to determining the adequacy of the model could only reveal a model that is underparameterised ('too small') and would not reveal a model that is overparameterised ('too big').

Examining whether the residuals are free from autocorrelation is much more commonly used than overfitting, and this may partly have arisen since for ARMA models, it can give rise to common factors in the overfitted model that make estimation of this model difficult and the statistical tests ill behaved. For example, if the true model is an ARMA(1,1) and we deliberately then fit an ARMA(2,2) there will be a common factor so that not all of the parameters in the latter model can be identified. This problem does not arise with pure AR or MA models, only with mixed processes.

It is usually the objective to form a *parsimonious model*, which is one that describes all of the features of data of interest using as few parameters (i.e. as simple a model) as possible. A parsimonious model is desirable because:

- The residual sum of squares is *inversely proportional* to the number of degrees of freedom. A model which contains irrelevant lags of the variable or of the error term (and therefore unnecessary parameters) will usually lead to increased coefficient standard errors, implying that it will be more difficult to find significant relationships in the data. Whether an increase in the number of variables (i.e. a reduction in

the number of degrees of freedom) will actually cause the estimated parameter standard errors to rise or fall will obviously depend on how much the RSS falls, and on the relative sizes of T and k. If T is very large relative to k, then the decrease in RSS is likely to outweigh the reduction in $T - k$ so that the standard errors fall. Hence 'large' models with many parameters are more often chosen when the sample size is large.

- Models that are profligate might be inclined to fit to data specific features, which would not be replicated out-of-sample. This means that the models may appear to fit the data very well, with perhaps a high value of R^2, but would give very inaccurate forecasts. Another interpretation of this concept, borrowed from physics, is that of the distinction between 'signal' and 'noise'. The idea is to fit a model which *captures the signal* (the important features of the data, or the underlying trends or patterns), but which does not try to fit a spurious model to the noise (the completely random aspect of the series).

5.7.1 Information criteria for ARMA model selection

The identification stage would now typically not be done using graphical plots of the acf and pacf. The reason is that when 'messy' real data is used, it unfortunately rarely exhibits the simple patterns of figures 5.2–5.8. This makes the acf and pacf very hard to interpret, and thus it is difficult to specify a model for the data. Another technique, which removes some of the subjectivity involved in interpreting the acf and pacf, is to use what are known as *information criteria*. Information criteria embody two factors: a term which is a function of the residual sum of squares (*RSS*), and some penalty for the loss of degrees of freedom from adding extra parameters. So, adding a new variable or an additional lag to a model will have two competing effects on the information criteria: the residual sum of squares will fall but the value of the penalty term will increase.

The object is to choose the number of parameters which minimises the value of the information criteria. So, adding an extra term will reduce the value of the criteria only if the fall in the residual sum of squares is sufficient to more than outweigh the increased value of the penalty term. There are several different criteria, which vary according to how stiff the penalty term is. The three most popular information criteria are Akaike's (1974) information criterion (*AIC*), Schwarz's (1978) Bayesian information criterion (*SBIC*), and the Hannan–Quinn criterion (*HQIC*).

Algebraically, these are expressed, respectively, as

$$AIC = \ln(\hat{\sigma}^2) + \frac{2k}{T} \tag{5.125}$$

$$SBIC = \ln(\hat{\sigma}^2) + \frac{k}{T}\ln T \tag{5.126}$$

$$HQIC = \ln(\hat{\sigma}^2) + \frac{2k}{T}\ln(\ln(T)) \tag{5.127}$$

where $\hat{\sigma}^2$ is the residual variance (also equivalent to the residual sum of squares divided by the number of observations, T), $k = p + q + 1$ is the total number of parameters estimated and T is the sample size. The information criteria are actually minimised subject to $p \leq \bar{p}, q \leq \bar{q}$, i.e. an upper limit is specified on the number of moving average ($\bar{q}$) and/or autoregressive ($\bar{p}$) terms that will be considered.

It is worth noting that *SBIC* embodies a much stiffer penalty term than *AIC*, while *HQIC* is somewhere in between. The adjusted R^2 measure can also be viewed as an information criterion, although it is a very soft one, which would typically select the largest models of all.

5.7.2 Which criterion should be preferred if they suggest different model orders?

SBIC is strongly consistent (but inefficient) and *AIC* is not consistent, but is generally more efficient. In other words, *SBIC* will asymptotically deliver the correct model order, while *AIC* will deliver on average too large a model, even with an infinite amount of data. On the other hand, the average variation in selected model orders from different samples within a given population will be greater in the context of *SBIC* than *AIC*. Overall, then, no criterion is definitely superior to others.

5.7.3 ARIMA modelling

ARIMA modelling, as distinct from ARMA modelling, has the additional letter 'I' in the acronym, standing for 'integrated'. An *integrated autoregressive process* is one whose characteristic equation has a root on the unit circle. Typically researchers difference the variable as necessary and then build an ARMA model on those differenced variables. An ARMA(p, q) model in the variable differenced d times is equivalent to an ARIMA(p, d, q) model on the original data – see chapter 7 for further details. For the remainder of this chapter, it is assumed that the data used in model construction are stationary, or have been suitably transformed to make them stationary. Thus only ARMA models will be considered further.

5.8 Constructing ARMA models in EViews

5.8.1 Getting started

This example uses the monthly UK house price series which was already incorporated in an EViews workfile in chapter 1. There were a total of 196 monthly observations running from February 1991 (recall that the January observation was 'lost' in constructing the lagged value) to May 2007 for the percentage change in house price series.

The objective of this exercise is to build an ARMA model for the house price changes. Recall that there are three stages involved: identification, estimation and diagnostic checking. The first stage is carried out by looking at the autocorrelation and partial autocorrelation coefficients to identify any structure in the data.

5.8.2 Estimating the autocorrelation coefficients for up to 12 lags

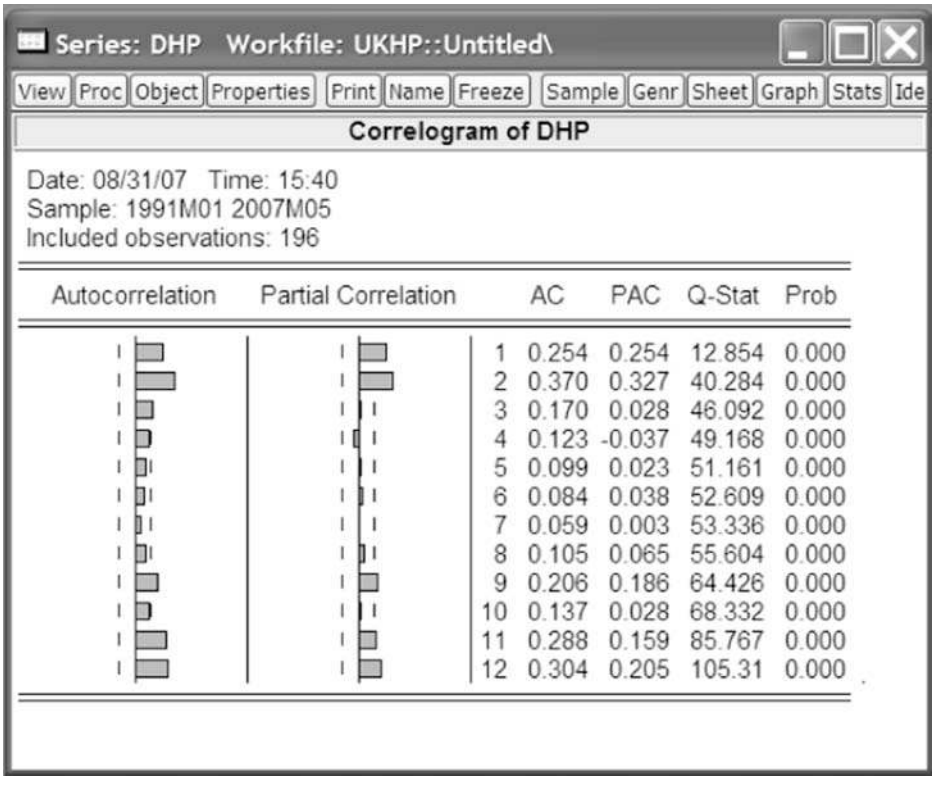
Double click on the **DHP series** and then click **View** and choose **Correlogram** .... In the 'Correlogram Specification' window, choose **Level** (since the series we are investigating has already been transformed into percentage returns or percentage changes) and in the 'Lags to include' box, type **12**. Click on **OK**. The output, including relevant test statistics, is given in screenshot 5.1.

It is clearly evident from the first columns that the series is quite persistent given that it is already in percentage change form. The autocorrelation function dies away quite slowly. Only the first partial autocorrelation coefficient appears strongly significant. The numerical values of the autocorrelation and partial autocorrelation coefficients at lags 1–12 are given in the fourth and fifth columns of the output, with the lag length given in the third column.

The penultimate column of output gives the statistic resulting from a Ljung–Box test with number of lags in the sum equal to the row number (i.e. the number in the third column). The test statistics will follow a $\chi^2(1)$ for the first row, a $\chi^2(2)$ for the second row, and so on. p-values associated with these test statistics are given in the last column.

Remember that as a rule of thumb, a given autocorrelation coefficient is classed as significant if it is outside a $\pm 1.96 \times 1/(T)^{1/2}$ band, where T is the number of observations. In this case, it would imply that a correlation coefficient is classed as significant if it is bigger than approximately 0.14 or smaller than -0.14. The band is of course wider when the sampling frequency is monthly, as it is here, rather than daily where there would be more observations. It can be deduced that the first three

Screenshot 5.1
Estimating the correlogram



autocorrelation coefficients and the first two partial autocorrelation coefficients are significant under this rule. Since the first acf coefficient is highly significant, the Ljung–Box joint test statistic rejects the null hypothesis of no autocorrelation at the 1% level for all numbers of lags considered. It could be concluded that a mixed ARMA process could be appropriate, although it is hard to precisely determine the appropriate order given these results. In order to investigate this issue further, the information criteria are now employed.

5.8.3 *Using information criteria to decide on model orders*

As demonstrated above, deciding on the appropriate model orders from autocorrelation functions could be very difficult in practice. An easier way is to choose the model order that minimises the value of an information criterion.

An important point to note is that books and statistical packages often differ in their construction of the test statistic. For example, the formulae given earlier in this chapter for Akaike's and Schwarz's Information

Criteria were

$$AIC = \ln(\hat{\sigma}^2) + \frac{2k}{T} \tag{5.128}$$

$$SBIC = \ln(\hat{\sigma}^2) + \frac{k}{T}(\ln T) \tag{5.129}$$

where $\hat{\sigma}^2$ is the estimator of the variance of regressions disturbances u_t, k is the number of parameters and T is the sample size. When using the criterion based on the estimated standard errors, the model with the lowest value of *AIC* and *SBIC* should be chosen. However, EViews uses a formulation of the test statistic derived from the log-likelihood function value based on a maximum likelihood estimation (see chapter 8). The corresponding EViews formulae are

$$AIC_\ell = -2\ell/T + \frac{2k}{T} \tag{5.130}$$

$$SBIC_\ell = -2\ell/T + \frac{k}{T}(\ln T) \tag{5.131}$$

where $l = -\frac{T}{2}(1 + \ln(2\pi) + \ln(\hat{u}'\hat{u}/T))$

Unfortunately, this modification is not benign, since it affects the relative strength of the penalty term compared with the error variance, sometimes leading different packages to select different model orders for the same data and criterion!

Suppose that it is thought that ARMA models from order (0,0) to (5,5) are plausible for the house price changes. This would entail considering 36 models (ARMA(0,0), ARMA(1,0), ARMA(2,0), ... ARMA(5,5)), i.e. up to five lags in both the autoregressive and moving average terms.

In EViews, this can be done by separately estimating each of the models and noting down the value of the information criteria in each case.[2] This would be done in the following way. On the EViews main menu, click on **Quick** and choose **Estimate Equation** .... EViews will open an Equation Specification window. In the Equation Specification editor, type, for example

dhp c ar(1) ma(1)

For the estimation settings, select **LS – Least Squares (NLS and ARMA)**, select the whole sample, and click **OK** – this will specify an ARMA(1,1). The output is given in the table below.

[2] Alternatively, any reader who knows how to write programs in EViews could set up a structure to loop over the model orders and calculate all the values of the information criteria together – see chapter 12.

Dependent Variable: DHP
Method: Least Squares
Date: 08/31/07 Time: 16:09
Sample (adjusted): 1991M03 2007M05
Included observations: 195 after adjustments
Convergence achieved after 19 iterations
MA Backcast: 1991M02

	Coefficient	Std. Error	t-Statistic	Prob.
C	0.868177	0.334573	2.594884	0.0102
AR(1)	0.975461	0.019471	50.09854	0.0000
MA(1)	−0.909851	0.039596	−22.9784	0.0000

R-squared	0.144695	Mean dependent var	0.635212
Adjusted R-squared	0.135786	S.D. dependent var	1.149146
S.E. of regression	1.068282	Akaike info criterion	2.985245
Sum squared resid	219.1154	Schwarz criterion	3.035599
Log likelihood	−288.0614	Hannan-Quinn criter.	3.005633
F-statistic	16.24067	Durbin-Watson stat	1.842823
Prob(F-statistic)	0.000000		

Inverted AR Roots	.98
Inverted MA Roots	.91

In theory, the output would then be interpreted in a similar way to that discussed in chapter 3. However, in reality it is very difficult to interpret the parameter estimates in the sense of, for example, saying, 'a 1 unit increase in x leads to a β unit increase in y'. In part because the construction of ARMA models is not based on any economic or financial theory, it is often best not to even try to interpret the individual parameter estimates, but rather to examine the plausibility of the model as a whole and to determine whether it describes the data well and produces accurate forecasts (if this is the objective of the exercise, which it often is).

The inverses of the AR and MA roots of the characteristic equation are also shown. These can be used to check whether the process implied by the model is stationary and invertible. For the AR and MA parts of the process to be stationary and invertible, respectively, the inverted roots in each case must be smaller than 1 in absolute value, which they are in this case, although only just. Note also that the header for the EViews output for ARMA models states the number of iterations that have been used in the model estimation process. This shows that, in fact, an iterative numerical optimisation procedure has been employed to estimate the coefficients (see chapter 8 for further details).

Repeating these steps for the other ARMA models would give all of the required values for the information criteria. To give just one more example, in the case of an ARMA(5,5), the following would be typed in the Equation Specification editor box:

dhp c ar(1) ar(2) ar(3) ar(4) ar(5) ma(1) ma(2) ma(3) ma(4) ma(5)

Note that, in order to estimate an ARMA(5,5) model, it is necessary to write out the whole list of terms as above rather than to simply write, for example, 'dhp c ar(5) ma(5)', which would give a model with a fifth lag of the dependent variable and a fifth lag of the error term but no other variables. The values of all of the information criteria, calculated using EViews, are as follows:

Information criteria for ARMA models of the percentage changes in UK house prices

AIC						
p/q	0	1	2	3	4	5
0	3.116	3.086	2.973	2.973	2.977	2.977
1	3.065	2.985	2.965	2.935	2.931	2.938
2	2.951	2.961	2.968	2.924	2.941	2.957
3	2.960	2.968	2.970	2.980	2.937	2.914
4	2.969	2.979	2.931	2.940	**2.862**	2.924
5	2.984	2.932	2.955	2.986	2.937	2.936
SBIC						
p/q	0	1	2	3	4	5
0	3.133	3.120	3.023	3.040	3.061	3.078
1	3.098	3.036	3.032	3.019	3.032	3.056
2	**3.002**	3.029	3.053	3.025	3.059	3.091
3	3.028	3.053	3.072	3.098	3.072	3.066
4	3.054	3.081	3.049	3.076	3.015	3.094
5	3.086	3.052	3.092	3.049	3.108	3.123

So which model actually minimises the two information criteria? In this case, the criteria choose different models: *AIC* selects an ARMA(4,4), while *SBIC* selects the smaller ARMA(2,0) model – i.e. an AR(2). These chosen models are highlighted in bold in the table. It will always be the case that *SBIC* selects a model that is at least as small (i.e. with fewer or the same number of parameters) as *AIC*, because the former criterion has a stricter penalty term. This means that *SBIC* penalises the incorporation of additional terms more heavily. Many different models provide almost

identical values of the information criteria, suggesting that the chosen models do not provide particularly sharp characterisations of the data and that a number of other specifications would fit the data almost as well.

5.9 Examples of time series modelling in finance

5.9.1 Covered and uncovered interest parity

The determination of the price of one currency in terms of another (i.e. the exchange rate) has received a great deal of empirical examination in the international finance literature. Of these, three hypotheses in particular are studied – covered interest parity (CIP), uncovered interest parity (UIP) and purchasing power parity (PPP). The first two of these will be considered as illustrative examples in this chapter, while PPP will be discussed in chapter 7. All three relations are relevant for students of finance, for violation of one or more of the parities may offer the potential for arbitrage, or at least will offer further insights into how financial markets operate. All are discussed briefly here; for a more comprehensive treatment, see Cuthbertson and Nitsche (2004) or the many references therein.

5.9.2 Covered interest parity

Stated in its simplest terms, CIP implies that, if financial markets are efficient, it should not be possible to make a riskless profit by borrowing at a risk-free rate of interest in a domestic currency, switching the funds borrowed into another (foreign) currency, investing them there at a risk-free rate and locking in a forward sale to guarantee the rate of exchange back to the domestic currency. Thus, if CIP holds, it is possible to write

$$f_t - s_t = (r - r^*)_t \tag{5.132}$$

where f_t and s_t are the log of the forward and spot prices of the domestic in terms of the foreign currency at time t, r is the domestic interest rate and r^* is the foreign interest rate. This is an equilibrium condition which must hold otherwise there would exist riskless arbitrage opportunities, and the existence of such arbitrage would ensure that any deviation from the condition cannot hold indefinitely. It is worth noting that, underlying CIP are the assumptions that the risk-free rates are truly risk-free – that is, there is no possibility for default risk. It is also assumed that there are no transactions costs, such as broker's fees, bid–ask spreads, stamp duty, etc., and that there are no capital controls, so that funds can be moved without restriction from one currency to another.

5.9.3 Uncovered interest parity

UIP takes CIP and adds to it a further condition known as 'forward rate unbiasedness' (FRU). Forward rate unbiasedness states that the forward rate of foreign exchange should be an unbiased predictor of the future value of the spot rate. If this condition does not hold, again in theory riskless arbitrage opportunities could exist. UIP, in essence, states that the expected change in the exchange rate should be equal to the interest rate differential between that available risk-free in each of the currencies. Algebraically, this may be stated as

$$s_{t+1}^{e} - s_t = (r - r^*)_t \tag{5.133}$$

where the notation is as above and s_{t+1}^{e} is the expectation, made at time t of the spot exchange rate that will prevail at time $t+1$.

The literature testing CIP and UIP is huge with literally hundreds of published papers. Tests of CIP unsurprisingly (for it is a pure arbitrage condition) tend not to reject the hypothesis that the condition holds. Taylor (1987, 1989) has conducted extensive examinations of CIP, and concluded that there were historical periods when arbitrage was profitable, particularly during periods where the exchange rates were under management.

Relatively simple tests of UIP and FRU take equations of the form (5.133) and add intuitively relevant additional terms. If UIP holds, these additional terms should be insignificant. Ito (1988) tests UIP for the yen/dollar exchange rate with the three-month forward rate for January 1973 until February 1985. The sample period is split into three as a consequence of perceived structural breaks in the series. Strict controls on capital movements were in force in Japan until 1977, when some were relaxed and finally removed in 1980. A Chow test confirms Ito's intuition and suggests that the three sample periods should be analysed separately. Two separate regressions are estimated for each of the three sample sub-periods

$$s_{t+3} - f_{t,3} = a + b_1(s_t - f_{t-3,3}) + b_2(s_{t-1} - f_{t-4,3}) + u_t \tag{5.134}$$

where s_{t+3} is the spot interest rate prevailing at time $t+3$, $f_{t,3}$ is the forward rate for three periods ahead available at time t, and so on, and u_t is an error term. A natural joint hypothesis to test is H_0: $a = 0$ and $b_1 = 0$ and $b_2 = 0$. This hypothesis represents the restriction that the deviation of the forward rate from the realised rate should have a mean value insignificantly different from zero ($a = 0$) and it should be independent of any information available at time t ($b_1 = 0$ and $b_2 = 0$). All three of these conditions must be fulfilled for UIP to hold. The second equation that Ito

Table 5.1 Uncovered interest parity test results

Sample period	1973M1–1977M3	1977M4–1980M12	1981M1–1985M2
Panel A: Estimates and hypothesis tests for $S_{t+3} - f_{t,3} = a + b_1(s_t - f_{t-3,3}) + b_2(s_{t-1} - f_{t-4,3}) + u_t$			
Estimate of a	0.0099	0.0031	0.027
Estimate of b_1	0.020	0.24	0.077
Estimate of b_2	−0.37	0.16	−0.21
Joint test $\chi^2(3)$	23.388	5.248	6.022
P-value for joint test	0.000	0.155	0.111
Panel B: Estimates and hypothesis tests for $S_{t+3} - f_{t,3} = a + b(s_t - f_{t,3}) + v_t$			
Estimate of a	0.00	−0.052	−0.89
Estimate of b	0.095	4.18	2.93
Joint test $\chi^2(2)$	31.923	22.06	5.39
p-value for joint test	0.000	0.000	0.07

Source: Ito (1988). Reprinted with permission from MIT Press Journals.

tests is

$$s_{t+3} - f_{t,3} = a + b(s_t - f_{t,3}) + v_t \tag{5.135}$$

where v_t is an error term and the hypothesis of interest in this case is H_0: $a = 0$ and $b = 0$.

Equation (5.134) tests whether past forecast errors have information useful for predicting the difference between the actual exchange rate at time $t + 3$, and the value of it that was predicted by the forward rate. Equation (5.135) tests whether the forward premium has any predictive power for the difference between the actual exchange rate at time $t + 3$, and the value of it that was predicted by the forward rate. The results for the three sample periods are presented in Ito's table 3, and are adapted and reported here in table 5.1.

The main conclusion is that UIP clearly failed to hold throughout the period of strictest controls, but there is less and less evidence against UIP as controls were relaxed.

5.10 Exponential smoothing

Exponential smoothing is another modelling technique (not based on the ARIMA approach) that uses only a linear combination of the previous values of a series for modelling it and for generating forecasts of its future

values. Given that only previous values of the series of interest are used, the only question remaining is how much weight should be attached to each of the previous observations. Recent observations would be expected to have the most power in helping to forecast future values of a series. If this is accepted, a model that places more weight on recent observations than those further in the past would be desirable. On the other hand, observations a long way in the past may still contain some information useful for forecasting future values of a series, which would not be the case under a centred moving average. An exponential smoothing model will achieve this, by imposing a geometrically declining weighting scheme on the lagged values of a series. The equation for the model is

$$S_t = \alpha y_t + (1-\alpha)S_{t-1} \tag{5.136}$$

where α is the smoothing constant, with $0 < \alpha < 1$, y_t is the current realised value, S_t is the current smoothed value.

Since $\alpha + (1-\alpha) = 1$, S_t is modelled as a weighted average of the current observation y_t and the previous smoothed value. The model above can be rewritten to express the exponential weighting scheme more clearly. By lagging (5.136) by one period, the following expression is obtained

$$S_{t-1} = \alpha y_{t-1} + (1-\alpha)S_{t-2} \tag{5.137}$$

and lagging again

$$S_{t-2} = \alpha y_{t-2} + (1-\alpha)S_{t-3} \tag{5.138}$$

Substituting into (5.136) for S_{t-1} from (5.137)

$$S_t = \alpha y_t + (1-\alpha)(\alpha y_{t-1} + (1-\alpha)S_{t-2}) \tag{5.139}$$

$$S_t = \alpha y_t + (1-\alpha)\alpha y_{t-1} + (1-\alpha)^2 S_{t-2} \tag{5.140}$$

Substituting into (5.140) for S_{t-2} from (5.138)

$$S_t = \alpha y_t + (1-\alpha)\alpha y_{t-1} + (1-\alpha)^2(\alpha y_{t-2} + (1-\alpha)S_{t-3}) \tag{5.141}$$

$$S_t = \alpha y_t + (1-\alpha)\alpha y_{t-1} + (1-\alpha)^2\alpha y_{t-2} + (1-\alpha)^3 S_{t-3} \tag{5.142}$$

T successive substitutions of this kind would lead to

$$S_t = \left(\sum_{i=0}^{T} \alpha(1-\alpha)^i y_{t-i}\right) + (1-\alpha)^{T+1} S_{t-1-T} \tag{5.143}$$

Since α 0, the effect of each observation declines geometrically as the variable moves another observation forward in time. In the limit as $T \rightarrow \infty$, $(1-\alpha)^T S_0 \rightarrow 0$, so that the current smoothed value is a geometrically weighted infinite sum of the previous realisations.

The forecasts from an exponential smoothing model are simply set to the current smoothed value, for any number of steps ahead, s

$$f_{t,s} = S_t, s = 1, 2, 3, \ldots \tag{5.144}$$

The exponential smoothing model can be seen as a special case of a Box–Jenkins model, an ARIMA(0,1,1), with MA coefficient $(1 - \alpha)$ – see Granger and Newbold (1986, p. 174).

The technique above is known as single or simple exponential smoothing, and it can be modified to allow for trends (Holt's method) or to allow for seasonality (Winter's method) in the underlying variable. These augmented models are not pursued further in this text since there is a much better way to model the trends (using a unit root process – see chapter 7) and the seasonalities (see chapters 1 and 9) of the form that are typically present in financial data.

Exponential smoothing has several advantages over the slightly more complex ARMA class of models discussed above. First, exponential smoothing is obviously very simple to use. There is no decision to be made on how many parameters to estimate (assuming only single exponential smoothing is considered). Thus it is easy to update the model if a new realisation becomes available.

Among the disadvantages of exponential smoothing is the fact that it is overly simplistic and inflexible. Exponential smoothing models can be viewed as but one model from the ARIMA family, which may not necessarily be optimal for capturing any linear dependence in the data. Also, the forecasts from an exponential smoothing model do not converge on the long-term mean of the variable as the horizon increases. The upshot is that long-term forecasts are overly affected by recent events in the history of the series under investigation and will therefore be sub-optimal.

A discussion of how exponential smoothing models can be estimated using EViews will be given after the following section on forecasting in econometrics.

5.11 Forecasting in econometrics

Although the words 'forecasting' and 'prediction' are sometimes given different meanings in some studies, in this text the words will be used synonymously. In this context, prediction or forecasting simply means an attempt to *determine the values that a series is likely to take*. Of course, forecasts might also usefully be made in a cross-sectional environment. Although the discussion below refers to time series data, some of the arguments will carry over to the cross-sectional context.

Determining the forecasting accuracy of a model is an important test of its adequacy. Some econometricians would go as far as to suggest that the statistical adequacy of a model in terms of whether it violates the CLRM assumptions or whether it contains insignificant parameters, is largely irrelevant if the model produces accurate forecasts. The following subsections of the book discuss why forecasts are made, how they are made from several important classes of models, how to evaluate the forecasts, and so on.

5.11.1 Why forecast?

Forecasts are made essentially because they are useful! Financial decisions often involve a long-term commitment of resources, the returns to which will depend upon what happens in the future. In this context, the decisions made today will reflect forecasts of the future state of the world, and the more accurate those forecasts are, the more utility (or money!) is likely to be gained from acting on them.

Some examples in finance of where forecasts from econometric models might be useful include:

- Forecasting tomorrow's return on a particular *share*
- Forecasting the *price of a house* given its characteristics
- Forecasting the *riskiness of a portfolio* over the next year
- Forecasting the *volatility of bond returns*
- Forecasting the *correlation between US and UK stock market movements* tomorrow
- Forecasting the likely number of *defaults* on a portfolio of home loans.

Again, it is evident that forecasting can apply either in a cross-sectional or a time series context. It is useful to distinguish between two approaches to forecasting:

- *Econometric* (*structural*) *forecasting* – relates a dependent variable to one or more independent variables. Such models often work well in the long run, since a long-run relationship between variables often arises from no-arbitrage or market efficiency conditions. Examples of such forecasts would include return predictions derived from arbitrage pricing models, or long-term exchange rate prediction based on purchasing power parity or uncovered interest parity theory.
- *Time series forecasting* – involves trying to forecast the future values of a series given its previous values and/or previous values of an error term.

The distinction between the two types is somewhat blurred – for example, it is not clear where vector autoregressive models (see chapter 6 for an extensive overview) fit into this classification.

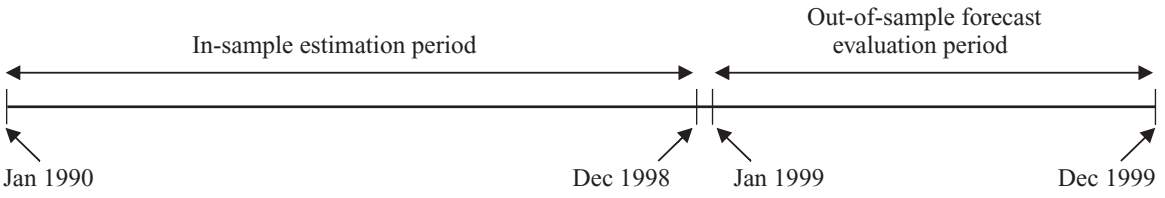


Figure 5.9 Use of an in-sample and an out-of-sample period for analysis

It is also worth distinguishing between point and interval forecasts. *Point* forecasts predict a single value for the variable of interest, while *interval* forecasts provide a range of values in which the future value of the variable is expected to lie with a given level of confidence.

5.11.2 *The difference between in-sample and out-of-sample forecasts*

In-sample forecasts are those generated for the same set of data that was used to estimate the model's parameters. One would expect the 'forecasts' of a model to be relatively good in-sample, for this reason. Therefore, a sensible approach to model evaluation through an examination of forecast accuracy is not to use all of the observations in estimating the model parameters, but rather to hold some observations back. The latter sample, sometimes known as a *holdout sample*, would be used to construct out-of-sample forecasts.

To give an illustration of this distinction, suppose that some monthly FTSE returns for 120 months (January 1990–December 1999) are available. It would be possible to use all of them to build the model (and generate only in-sample forecasts), or some observations could be kept back, as shown in figure 5.9.

What would be done in this case would be to use data from 1990*M*1 until 1998*M*12 to estimate the model parameters, and then the observations for 1999 would be forecasted from the estimated parameters. Of course, where each of the in-sample and out-of-sample periods should start and finish is somewhat arbitrary and at the discretion of the researcher. One could then compare how close the forecasts for the 1999 months were relative to their actual values that are in the holdout sample. This procedure would represent a better test of the model than an examination of the in-sample fit of the model since the information from 1999*M*1 onwards has not been used when estimating the model parameters.

5.11.3 *Some more terminology: one-step-ahead versus multi-step-ahead forecasts and rolling versus recursive samples*

A *one-step-ahead forecast* is a forecast generated for the next observation only, whereas *multi-step-ahead forecasts* are those generated for 1, 2, 3, . . . , s steps

ahead, so that the forecasting horizon is for the next s periods. Whether one-step- or multi-step-ahead forecasts are of interest will be determined by the forecasting horizon of interest to the researcher.

Suppose that the monthly FTSE data are used as described in the example above. If the in-sample estimation period stops in December 1998, then up to 12-step-ahead forecasts could be produced, giving 12 predictions that can be compared with the actual values of the series. Comparing the actual and forecast values in this way is not ideal, for the forecasting horizon is varying from 1 to 12 steps ahead. It might be the case, for example, that the model produces very good forecasts for short horizons (say, one or two steps), but that it produces inaccurate forecasts further ahead. It would not be possible to evaluate whether this was in fact the case or not since only a single one-step-ahead forecast, a single 2-step-ahead forecast, and so on, are available. An evaluation of the forecasts would require a considerably larger holdout sample.

A useful way around this problem is to use a *recursive or rolling window*, which generates a series of forecasts for a given number of steps ahead. A recursive forecasting model would be one where the initial estimation date is fixed, but additional observations are added one at a time to the estimation period. A rolling window, on the other hand, is one where the length of the in-sample period used to estimate the model is fixed, so that the start date and end date successively increase by one observation. Suppose now that only one-, two-, and three-step-ahead forecasts are of interest. They could be produced using the following recursive and rolling window approaches:

Objective: to produce	*Data used to estimate model parameters*	
1-, 2-, 3-step-ahead forecasts for:	*Rolling window*	*Recursive window*
1999*M*1, *M*2, *M*3	1990*M*1–1998*M*12	1990*M*1–1998*M*12
1999*M*2, *M*3, *M*4	1990*M*2–1999*M*1	1990*M*1–1999*M*1
1999*M*3, *M*4, *M*5	1990*M*3–1999*M*2	1990*M*1–1999*M*2
1999*M*4, *M*5, *M*6	1990*M*4–1999*M*3	1990*M*1–1999*M*3
1999*M*5, *M*6, *M*7	1990*M*5–1999*M*4	1990*M*1–1999*M*4
1999*M*6, *M*7, *M*8	1990*M*6–1999*M*5	1990*M*1–1999*M*5
1999*M*7, *M*8, *M*9	1990*M*7–1999*M*6	1990*M*1–1999*M*6
1999*M*8, *M*9, *M*10	1990*M*8–1999*M*7	1990*M*1–1999*M*7
1999*M*9, *M*10, *M*11	1990*M*9–1999*M*8	1990*M*1–1999*M*8
1999*M*10, *M*11, *M*12	1990*M*10–1999*M*9	1990*M*1–1999*M*9

The sample length for the rolling windows above is always set at 108 observations, while the number of observations used to estimate the

parameters in the recursive case increases as we move down the table and through the sample.

5.11.4 *Forecasting with time series versus structural models*

To understand how to construct forecasts, the idea of *conditional expectations* is required. A conditional expectation would be expressed as

$$E(y_{t+1} \mid \Omega_t)$$

This expression states that the expected value of y is taken for time $t+1$, conditional upon, or given, (|) all information available up to and including time t (Ω_t). Contrast this with the unconditional expectation of y, which is the expected value of y without any reference to time, i.e. the unconditional mean of y. The conditional expectations operator is used to generate forecasts of the series.

How this conditional expectation is evaluated will of course depend on the model under consideration. Several families of models for forecasting will be developed in this and subsequent chapters.

A first point to note is that by definition the optimal forecast for a zero mean white noise process is zero

$$\mathrm{E}(u_{t+s}|\Omega_t) = 0 \,\forall\, s > 0 \tag{5.145}$$

The two simplest forecasting 'methods' that can be employed in almost every situation are shown in box 5.3.

Box 5.3 Naive forecasting methods

(1) Assume no change so that the forecast, f, of the value of y, s steps into the future is the current value of y

$$E(y_{t+s}|\Omega_t) = y_t \tag{5.146}$$

Such a forecast would be optimal if y_t followed a random walk process.

(2) In the absence of a full model, forecasts can be generated using the long-term average of the series. Forecasts using the unconditional mean would be more useful than 'no change' forecasts for any series that is 'mean-reverting' (i.e. stationary).

Time series models are generally better suited to the production of time series forecasts than structural models. For an illustration of this, consider the following linear regression model

$$y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \cdots + \beta_k x_{kt} + u_t \tag{5.147}$$

To forecast y, the conditional expectation of its future value is required. Taking expectations of both sides of (5.147) yields

$$\mathrm{E}(y_t \,|\Omega_{t-1}) = \mathrm{E}(\beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \cdots + \beta_k x_{kt} + u_t) \tag{5.148}$$

The parameters can be taken through the expectations operator, since this is a population regression function and therefore they are assumed known. The following expression would be obtained

$$\mathrm{E}(y_t \,|\Omega_{t-1}) = \beta_1 + \beta_2 \mathrm{E}(x_{2t}) + \beta_3 \mathrm{E}(x_{3t}) + \cdots + \beta_k \mathrm{E}(x_{kt}) \tag{5.149}$$

But there is a problem: what are $\mathrm{E}(x_{2t})$, etc.? Remembering that information is available only until time $t-1$, the values of these variables are unknown. It may be possible to forecast them, but this would require another set of forecasting models for every explanatory variable. To the extent that forecasting the explanatory variables may be as difficult, or even more difficult, than forecasting the explained variable, this equation has achieved nothing! In the absence of a set of forecasts for the explanatory variables, one might think of using $\bar{x}_2$, etc., i.e. the mean values of the explanatory variables, giving

$$\mathrm{E}(y_t) = \beta_1 + \beta_2 \bar{x}_2 + \beta_3 \bar{x}_3 + \cdots + \beta_k \bar{x}_k = \bar{y}\,! \tag{5.150}$$

Thus, if the mean values of the explanatory variables are used as inputs to the model, all that will be obtained as a forecast is the average value of y. Forecasting using pure time series models is relatively common, since it avoids this problem.

5.11.5 *Forecasting with ARMA models*

Forecasting using ARMA models is a fairly simple exercise in calculating conditional expectations. Although any consistent and logical notation could be used, the following conventions will be adopted in this book. Let $f_{t,s}$ denote a forecast made using an ARMA(p,q) model at time t for s steps into the future for some series y. The forecasts are generated by what is known as a forecast function, typically of the form

$$f_{t,s} = \sum_{i=1}^{p} a_i f_{t,s-i} + \sum_{j=1}^{q} b_j u_{t+s-j} \tag{5.151}$$

where $f_{t,s} = y_{t+s},\ s \leq 0; \quad u_{t+s} = 0,\ s > 0$
$= u_{t+s},\ s \leq 0$

and a_i and b_i are the autoregressive and moving average coefficients, respectively.

A demonstration of how one generates forecasts for separate AR and MA processes, leading to the general equation (5.151) above, will now be given.

5.11.6 *Forecasting the future value of an MA(q) process*

A moving average process has a memory only of length q, and this limits the sensible forecasting horizon. For example, suppose that an MA(3) model has been estimated

$$y_t = \mu + \theta_1 u_{t-1} + \theta_2 u_{t-2} + \theta_3 u_{t-3} + u_t \tag{5.152}$$

Since parameter constancy over time is assumed, if this relationship holds for the series y at time t, it is also assumed to hold for y at time $t+1, t+2, \ldots$, so 1 can be added to each of the time subscripts in (5.152), and 2 added to each of the time subscripts, and then 3, and so on, to arrive at the following

$$y_{t+1} = \mu + \theta_1 u_t + \theta_2 u_{t-1} + \theta_3 u_{t-2} + u_{t+1} \tag{5.153}$$

$$y_{t+2} = \mu + \theta_1 u_{t+1} + \theta_2 u_t + \theta_3 u_{t-1} + u_{t+2} \tag{5.154}$$

$$y_{t+3} = \mu + \theta_1 u_{t+2} + \theta_2 u_{t+1} + \theta_3 u_t + u_{t+3} \tag{5.155}$$

Suppose that all information up to and including that at time t is available and that forecasts for $1, 2, \ldots, s$ steps ahead – i.e. forecasts for y at times $t+1, t+2, \ldots, t+s$ are wanted. y_t, $y_{t-1}, \ldots$, and u_t, u_{t-1}, are known, so producing the forecasts is just a matter of taking the conditional expectation of (5.153)

$$f_{t,1} = E(y_{t+1|t}) = E(\mu + \theta_1 u_t + \theta_2 u_{t-1} + \theta_3 u_{t-2} + u_{t+1} | \Omega_t) \tag{5.156}$$

where $\mathrm{E}(y_{t+1|t})$ is a short-hand notation for $\mathrm{E}(y_{t+1}|\Omega_t)$

$$f_{t,1} = E(y_{t+1|t}) = \mu + \theta_1 u_t + \theta_2 u_{t-1} + \theta_3 u_{t-2} \tag{5.157}$$

Thus the forecast for y, 1 step ahead, made at time t, is given by this linear combination of the disturbance terms. Note that it would not be appropriate to set the values of these disturbance terms to their unconditional mean of zero. This arises because it is the *conditional expectation* of their values that is of interest. Given that all information is known up to and including that at time t is available, the values of the error terms up to time t are known. But u_{t+1} is not known at time t and therefore $\mathrm{E}(u_{t+1|t}) = 0$, and so on.

The forecast for 2 steps ahead is formed by taking the conditional expectation of (5.154)

$$f_{t,2} = E(y_{t+2|t}) = E(\mu + \theta_1 u_{t+1} + \theta_2 u_t + \theta_3 u_{t-1} + u_{t+2} \mid \Omega_t) \tag{5.158}$$

$$f_{t,2} = E(y_{t+2|t}) = \mu + \theta_2 u_t + \theta_3 u_{t-1} \tag{5.159}$$

In the case above, u_{t+2} is not known since information is available only to time t, so E(u_{t+2}) is set to zero. Continuing and applying the same rules to generate 3-, 4-, ..., s-step-ahead forecasts

$$f_{t,3} = E(y_{t+3|t}) = E(\mu + \theta_1 u_{t+2} + \theta_2 u_{t+1} + \theta_3 u_t + u_{t+3} \mid \Omega_t) \tag{5.160}$$

$$f_{t,3} = E(y_{t+3|t}) = \mu + \theta_3 u_t \tag{5.161}$$

$$f_{t,4} = E(y_{t+4|t}) = \mu \tag{5.162}$$

$$f_{t,s} = E(y_{t+s|t}) = \mu \ \forall \ s \geq 4 \tag{5.163}$$

As the MA(3) process has a memory of only three periods, all forecasts four or more steps ahead collapse to the intercept. Obviously, if there had been no constant term in the model, the forecasts four or more steps ahead for an MA(3) would be zero.

5.11.7 *Forecasting the future value of an AR(p) process*

Unlike a moving average process, an autoregressive process has infinite memory. To illustrate, suppose that an AR(2) model has been estimated

$$y_t = \mu + \phi_1 y_{t-1} + \phi_2 y_{t-2} + u_t \tag{5.164}$$

Again, by appealing to the assumption of parameter stability, this equation will hold for times $t+1$, $t+2$, and so on

$$y_{t+1} = \mu + \phi_1 y_t + \phi_2 y_{t-1} + u_{t+1} \tag{5.165}$$

$$y_{t+2} = \mu + \phi_1 y_{t+1} + \phi_2 y_t + u_{t+2} \tag{5.166}$$

$$y_{t+3} = \mu + \phi_1 y_{t+2} + \phi_2 y_{t+1} + u_{t+3} \tag{5.167}$$

Producing the one-step-ahead forecast is easy, since all of the information required is known at time t. Applying the expectations operator to (5.165), and setting E(u_{t+1}) to zero would lead to

$$f_{t,1} = E(y_{t+1|t}) = E(\mu + \phi_1 y_t + \phi_2 y_{t-1} + u_{t+1} \mid \Omega_t) \tag{5.168}$$

$$f_{t,1} = E(y_{t+1|t}) = \mu + \phi_1 E(y_t \mid t) + \phi_2 E(y_{t-1} \mid t) \tag{5.169}$$

$$f_{t,1} = E(y_{t+1|t}) = \mu + \phi_1 y_t + \phi_2 y_{t-1} \tag{5.170}$$

Applying the same procedure in order to generate a two-step-ahead forecast

$$f_{t,2} = E(y_{t+2|t}) = E(\mu + \phi_1 y_{t+1} + \phi_2 y_t + u_{t+2} \mid \Omega_t) \tag{5.171}$$

$$f_{t,2} = E(y_{t+2|t}) = \mu + \phi_1 E(y_{t+1} \mid t) + \phi_2 E(y_t \mid t) \tag{5.172}$$

The case above is now slightly more tricky, since E(y_{t+1}) is not known, although this in fact is the one-step-ahead forecast, so that (5.172) becomes

$$f_{t,2} = E(y_{t+2|t}) = \mu + \phi_1 f_{t,1} + \phi_2 y_t \tag{5.173}$$

Similarly, for three, four, ... and s steps ahead, the forecasts will be, respectively, given by

$$f_{t,3} = E(y_{t+3|t}) = E(\mu + \phi_1 y_{t+2} + \phi_2 y_{t+1} + u_{t+3} \mid \Omega_t) \tag{5.174}$$

$$f_{t,3} = E(y_{t+3|t}) = \mu + \phi_1 E(y_{t+2} \mid t) + \phi_2 E(y_{t+1} \mid t) \tag{5.175}$$

$$f_{t,3} = E(y_{t+3|t}) = \mu + \phi_1 f_{t,2} + \phi_2 f_{t,1} \tag{5.176}$$

$$f_{t,4} = \mu + \phi_1 f_{t,3} + \phi_2 f_{t,2} \tag{5.177}$$

etc. so

$$f_{t,s} = \mu + \phi_1 f_{t,s-1} + \phi_2 f_{t,s-2} \tag{5.178}$$

Thus the s-step-ahead forecast for an AR(2) process is given by the intercept + the coefficient on the one-period lag multiplied by the time $s-1$ forecast + the coefficient on the two-period lag multiplied by the $s-2$ forecast.

ARMA(p,q) forecasts can easily be generated in the same way by applying the rules for their component parts, and using the general formula given by (5.151).

5.11.8 *Determining whether a forecast is accurate or not*

For example, suppose that tomorrow's return on the FTSE is predicted to be 0.2, and that the outcome is actually −0.4. Is this an accurate forecast? Clearly, one cannot determine whether a forecasting model is good or not based upon only one forecast and one realisation. Thus in practice, forecasts would usually be produced for the whole of the out-of-sample period, which would then be compared with the actual values, and the difference between them aggregated in some way. The forecast error for observation i is defined as the difference between the actual value for observation i and the forecast made for it. The forecast error, defined in this way, will be positive (negative) if the forecast was too low (high). Therefore, it is not possible simply to sum the forecast errors, since the

Table 5.2 Forecast error aggregation

Steps ahead	Forecast	Actual	Squared error	Absolute error
1	0.20	−0.40	$(0.20 - -0.40)^2 = 0.360$	$\|0.20 - -0.40\| = 0.600$
2	0.15	0.20	$(0.15-0.20)^2 = 0.002$	$\|0.15-0.20\| = 0.050$
3	0.10	0.10	$(0.10-0.10)^2 = 0.000$	$\|0.10-0.10\| = 0.000$
4	0.06	−0.10	$(0.06 - -0.10)^2 = 0.026$	$\|0.06 - -0.10\| = 0.160$
5	0.04	−0.05	$(0.04 - -0.05)^2 = 0.008$	$\|0.04 - -0.05\| = 0.090$

positive and negative errors will cancel one another out. Thus, before the forecast errors are aggregated, they are usually squared or the absolute value taken, which renders them all positive. To see how the aggregation works, consider the example in table 5.2, where forecasts are made for a series up to 5 steps ahead, and are then compared with the actual realisations (with all calculations rounded to 3 decimal places).

The mean squared error, *MSE*, and mean absolute error, *MAE*, are now calculated by taking the average of the fourth and fifth columns, respectively

$$MSE = (0.360 + 0.002 + 0.000 + 0.026 + 0.008)/5 = 0.079 \quad (5.179)$$

$$MAE = (0.600 + 0.050 + 0.000 + 0.160 + 0.090)/5 = 0.180 \quad (5.180)$$

Taken individually, little can be gleaned from considering the size of the *MSE* or *MAE*, for the statistic is unbounded from above (like the residual sum of squares or *RSS*). Instead, the *MSE* or *MAE* from one model would be compared with those of other models for the same data and forecast period, and the model(s) with the lowest value of the error measure would be argued to be the most accurate.

MSE provides a quadratic loss function, and so may be particularly useful in situations where large forecast errors are disproportionately more serious than smaller errors. This may, however, also be viewed as a disadvantage if large errors are not disproportionately more serious, although the same critique could also, of course, be applied to the whole least squares methodology. Indeed Dielman (1986) goes as far as to say that when there are outliers present, least absolute values should be used to determine model parameters rather than least squares. Makridakis (1993, p. 528) argues that mean absolute percentage error (*MAPE*) is 'a relative measure that incorporates the best characteristics among the various accuracy criteria'. Once again, denoting s-step-ahead forecasts of a variable made at time t as $f_{t,s}$ and the actual value of the variable at time t as y_t,

then the mean square error can be defined as

$$MSE = \frac{1}{T-(T_1-1)}\sum_{t=T_1}^{T}(y_{t+s}-f_{t,s})^2 \tag{5.181}$$

where T is the total sample size (in-sample + out-of-sample), and T_1 is the first out-of-sample forecast observation. Thus in-sample model estimation initially runs from observation 1 to (T_1-1), and observations T_1 to T are available for out-of-sample estimation, i.e. a total holdout sample of $T-(T_1-1)$.

Mean absolute error (*MAE)* measures the average absolute forecast error, and is given by

$$MAE = \frac{1}{T-(T_1-1)}\sum_{t=T_1}^{T}|y_{t+s}-f_{t,s}| \tag{5.182}$$

Adjusted MAPE (*AMAPE*) or symmetric *MAPE* corrects for the problem of asymmetry between the actual and forecast values

$$AMAPE = \frac{100}{T-(T_1-1)}\sum_{t=T_1}^{T}\left|\frac{y_{t+s}-f_{t,s}}{y_{t+s}+f_{t,s}}\right| \tag{5.183}$$

The symmetry in (5.183) arises since the forecast error is divided by twice the average of the actual and forecast values. So, for example, *AMAPE* will be the same whether the forecast is 0.5 and the actual value is 0.3, or the actual value is 0.5 and the forecast is 0.3. The same is not true of the standard *MAPE* formula, where the denominator is simply y_{t+s}, so that whether y_t or $f_{t,s}$ is larger will affect the result

$$MAPE = \frac{100}{T-(T_1-1)}\sum_{t=T_1}^{T}\left|\frac{y_{t+s}-f_{t,s}}{y_{t+s}}\right| \tag{5.184}$$

MAPE also has the attractive additional property compared to *MSE* that it can be interpreted as a percentage error, and furthermore, its value is bounded from below by 0.

Unfortunately, it is not possible to use the adjustment if the series and the forecasts can take on opposite signs (as they could in the context of returns forecasts, for example). This is due to the fact that the prediction and the actual value may, purely by coincidence, take on values that are almost equal and opposite, thus almost cancelling each other out in the denominator. This leads to extremely large and erratic values of *AMAPE*. In such an instance, it is not possible to use *MAPE* as a criterion either. Consider the following example: say we forecast a value of $f_{t,s} = 3$, but the out-turn is that $y_{t+s} = 0.0001$. The addition to total *MSE* from this one

observation is given by

$$\frac{1}{391} \times (0.0001 - 3)^2 = 0.0230 \tag{5.185}$$

This value for the forecast is large, but perfectly feasible since in many cases it will be well within the range of the data. But the addition to total *MAPE* from just this single observation is given by

$$\frac{100}{391} \left| \frac{0.0001 - 3}{0.0001} \right| = 7670 \tag{5.186}$$

MAPE has the advantage that for a random walk in the log levels (i.e. a zero forecast), the criterion will take the value one (or 100 if we multiply the formula by 100 to get a percentage, as was the case for the equation above. So if a forecasting model gives a *MAPE* smaller than one (or 100), it is superior to the random walk model. In fact the criterion is also not reliable if the series can take on absolute values less than one. This point may seem somewhat obvious, but it is clearly important for the choice of forecast evaluation criteria.

Another criterion which is popular is Theil's U-statistic (1966). The metric is defined as follows

$$U = \frac{\sqrt{\sum_{t=T_1}^{T} \left(\frac{y_{t+s} - f_{t,s}}{y_{t+s}} \right)^2}}{\sqrt{\sum_{t=T_1}^{T} \left(\frac{y_{t+s} - fb_{t,s}}{y_{t+s}} \right)^2}} \tag{5.187}$$

where $fb_{t,s}$ is the forecast obtained from a benchmark model (typically a simple model such as a naive or random walk). A U-statistic of one implies that the model under consideration and the benchmark model are equally (in)accurate, while a value of less than one implies that the model is superior to the benchmark, and vice versa for $U > 1$. Although the measure is clearly useful, as Makridakis and Hibon (1995) argue, it is not without problems since if $fb_{t,s}$ is the same as y_{t+s}, Uwill be infinite since the denominator will be zero. The value of U will also be influenced by outliers in a similar vein to *MSE* and has little intuitive meaning.[3]

5.11.9 Statistical versus financial or economic loss functions

Many econometric forecasting studies evaluate the models' success using statistical loss functions such as those described above. However, it is not

[3] Note that the Theil's U-formula reported by EViews is slightly different.

necessarily the case that models classed as accurate because they have small mean squared forecast errors are useful in practical situations. To give one specific illustration, it has recently been shown (Gerlow, Irwin and Liu, 1993) that the accuracy of forecasts according to traditional statistical criteria may give little guide to the potential profitability of employing those forecasts in a market trading strategy. So models that perform poorly on statistical grounds may still yield a profit if used for trading, and vice versa.

On the other hand, models that can accurately forecast the sign of future returns, or can predict turning points in a series have been found to be more profitable (Leitch and Tanner, 1991). Two possible indicators of the ability of a model to predict direction changes irrespective of their magnitude are those suggested by Pesaran and Timmerman (1992) and by Refenes (1995). The relevant formulae to compute these measures are, respectively

$$\%\text{ correct sign predictions} = \frac{1}{T-(T_1-1)}\sum_{t=T_1}^{T} z_{t+s} \tag{5.188}$$

where $z_{t+s} = 1$ if $(y_{t+s} f_{t,s}) > 0$
$z_{t+s} = 0$ *otherwise*

and

$$\%\text{ correct direction change predictions} = \frac{1}{T-(T_1-1)}\sum_{t=T_1}^{T} z_{t+s} \tag{5.189}$$

where $z_{t+s} = 1$ if $(y_{t+s} - y_t)(f_{t,s} - y_t) > 0$
$z_{t+s} = 0$ *otherwise*

Thus, in each case, the criteria give the proportion of correctly predicted signs and directional changes for some given lead time s, respectively.

Considering how strongly each of the three criteria outlined above (*MSE*, *MAE* and proportion of correct sign predictions) penalises large errors relative to small ones, the criteria can be ordered as follows:

Penalises large errors least → penalises large errors most heavily
Sign prediction → *MAE* →*MSE*

MSE penalises large errors disproportionately more heavily than small errors, *MAE* penalises large errors proportionately equally as heavily as small errors, while the sign prediction criterion does not penalise large errors any more than small errors.

5.11.10 Finance theory and time series analysis

An example of ARIMA model identification, estimation and forecasting in the context of commodity prices is given by Chu (1978). He finds ARIMA models useful compared with structural models for short-term forecasting, but also finds that they are less accurate over longer horizons. It also observed that ARIMA models have limited capacity to forecast unusual movements in prices.

Chu (1978) argues that, although ARIMA models may appear to be completely lacking in theoretical motivation, and interpretation, this may not necessarily be the case. He cites several papers and offers an additional example to suggest that ARIMA specifications quite often arise naturally as reduced form equations (see chapter 6) corresponding to some underlying structural relationships. In such a case, not only would ARIMA models be convenient and easy to estimate, they could also be well grounded in financial or economic theory after all.

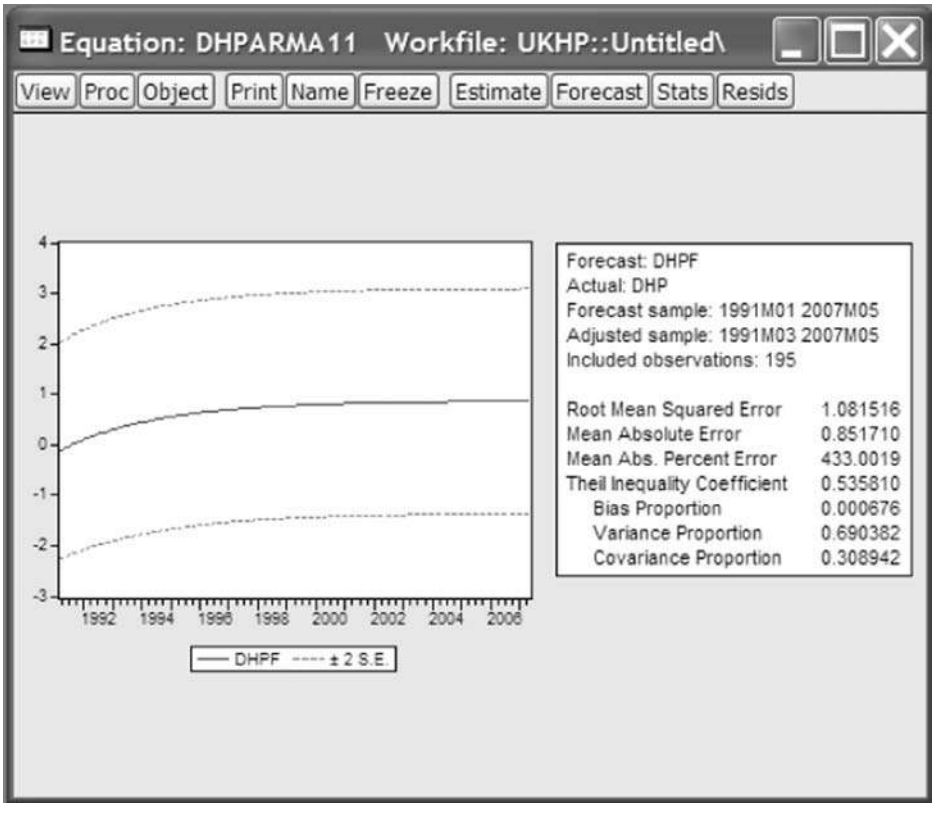
5.12 Forecasting using ARMA models in EViews

Once a specific model order has been chosen and the model estimated for a particular set of data, it may be of interest to use the model to forecast future values of the series. Suppose that the AR(2) model selected for the house price percentage changes series were estimated using observations February 1991–December 2004, leaving 29 remaining observations to construct forecasts for and to test forecast accuracy (for the period January 2005–May 2007).

Once the required model has been estimated and EViews has opened a window displaying the output, click on the **Forecast** icon. In this instance, the sample range to forecast would, of course, be 169–197 (which should be entered as 2005M01–2007M05). There are two methods available in EViews for constructing forecasts: dynamic and static. Select the option **Dynamic** to calculate multi-step forecasts starting from the first period in the forecast sample or **Static** to calculate a sequence of one-step-ahead forecasts, rolling the sample forwards one observation after each forecast to use actual rather than forecasted values for lagged dependent variables. The outputs for the dynamic and static forecasts are given in screenshots 5.2 and 5.3.

The forecasts are plotted using the continuous line, while a confidence interval is given by the two dotted lines in each case. For the dynamic forecasts, it is clearly evident that the forecasts quickly converge upon the long-term unconditional mean value as the horizon increases. Of course,

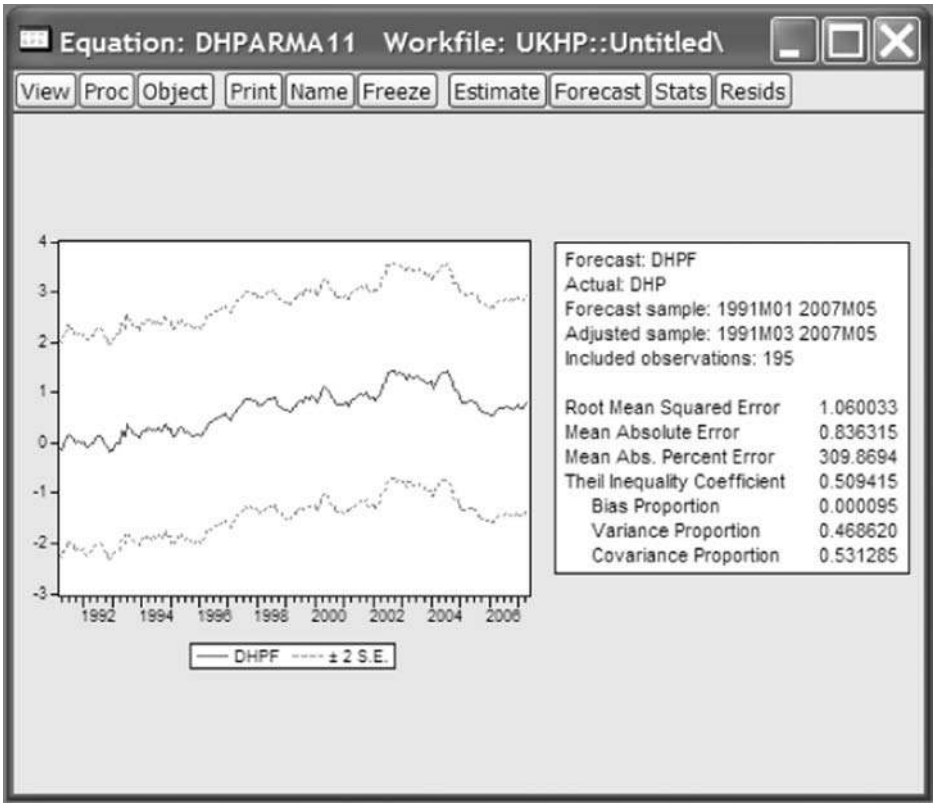
Screenshot 5.2
Plot and summary statistics for the dynamic forecasts for the percentage changes in house prices using an AR(2)



this does not occur with the series of one-step-ahead forecasts produced by the 'static' command. Several other useful measures concerning the forecast errors are displayed in the plot box, including the square root of the mean squared error (RMSE), the MAE, the MAPE and Theil's U-statistic. The MAPE for the dynamic and static forecasts for DHP are well over 100% in both cases, which can sometimes happen for the reasons outlined above. This indicates that the model forecasts are unable to account for much of the variability of the out-of-sample part of the data. This is to be expected as forecasting changes in house prices, along with the changes in the prices of any other assets, is difficult!

EViews provides another piece of useful information – a decomposition of the forecast errors. The mean squared forecast error can be decomposed into a bias proportion, a variance proportion and a covariance proportion. The *bias component* measures the extent to which the mean of the forecasts is different to the mean of the actual data (i.e. whether the forecasts are biased). Similarly, the *variance component* measures the difference between the variation of the forecasts and the variation of the actual data, while the *covariance component* captures any remaining unsystematic part of the

Screenshot 5.3
Plot and summary statistics for the static forecasts for the percentage changes in house prices using an AR(2)



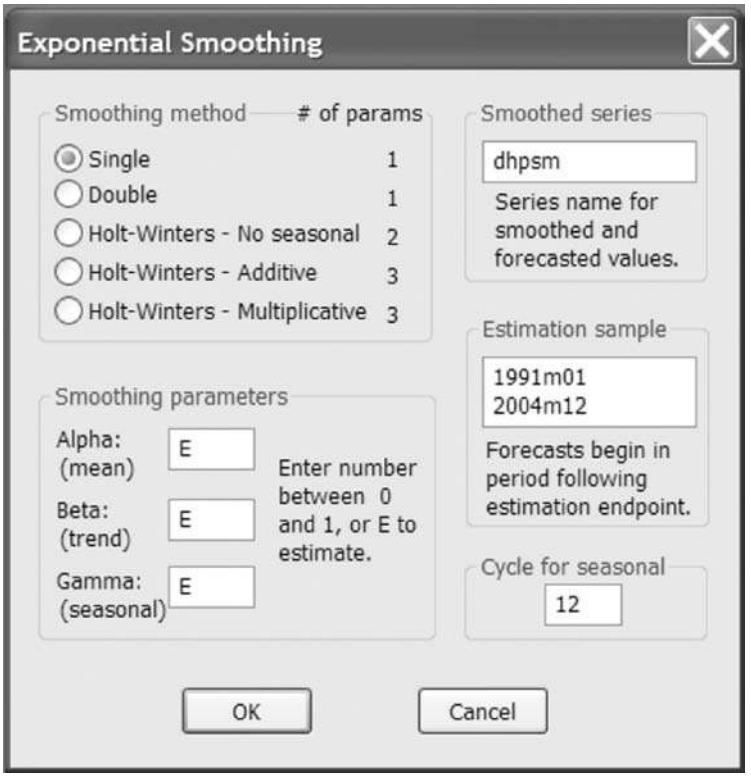
forecast errors. As one might have expected, the forecasts are not biased. Accurate forecasts would be unbiased and also have a small variance proportion, so that most of the forecast error should be attributable to the covariance (unsystematic or residual) component. For further details, see Granger and Newbold (1986).

A robust forecasting exercise would of course employ a longer out-of-sample period than the two years or so used here, would perhaps employ several competing models in parallel, and would also compare the accuracy of the predictions by examining the error measures given in the box after the forecast plots.

5.13 Estimating exponential smoothing models using EViews

This class of models can be easily estimated in EViews by double clicking on the desired variable in the workfile, so that the spreadsheet for that variable appears, and selecting **Proc** on the button bar for that variable and then **Exponential Smoothing**. . . . The screen with options will appear as in screenshot 5.4.

Screenshot 5.4
Estimating exponential smoothing models



There is a variety of smoothing methods available, including single and double, or various methods to allow for seasonality and trends in the data. Select **Single** (exponential smoothing), which is the only smoothing method that has been discussed in this book, and specify the estimation sample period as **1991M1 – 2004M12** to leave 29 observations for out-of-sample forecasting. Clicking **OK** will give the results in the following table.

Date: 09/02/07 Time: 14:46
Sample: 1991M02 2004M12
Included observations: 167
Method: Single Exponential
Original Series: DHP
Forecast Series: DHPSM

Parameters: Alpha		0.0760
Sum of Squared Residuals		208.5130
Root Mean Squared Error		1.117399
End of Period Levels:	Mean	0.994550

The output includes the value of the estimated smoothing coefficient (= 0.076 in this case), together with the RSS for the in-sample estimation period and the RMSE for the 29 forecasts. The final in-sample smoothed value will be the forecast for those 29 observations (which in this case would be 0.994550). EViews has automatically saved the smoothed values (i.e. the model fitted values) and the forecasts in a series called 'DHPSM'.

Key concepts

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

- ARIMA models
- invertible MA
- autocorrelation function
- Box-Jenkins methodology
- exponential smoothing
- rolling window
- multi-step forecast
- mean absolute percentage error
- Ljung–Box test
- Wold's decomposition theorem
- partial autocorrelation function
- information criteria
- recursive window
- out-of-sample
- mean squared error

Review questions

1. What are the differences between autoregressive and moving average models?
2. Why might ARMA models be considered particularly useful for financial time series? Explain, without using any equations or mathematical notation, the difference between AR, MA and ARMA processes.
3. Consider the following three models that a researcher suggests might be a reasonable model of stock market prices

$$y_t = y_{t-1} + u_t \tag{5.190}$$
$$y_t = 0.5y_{t-1} + u_t \tag{5.191}$$
$$y_t = 0.8u_{t-1} + u_t \tag{5.192}$$

   (a) What classes of models are these examples of?
   (b) What would the autocorrelation function for each of these processes look like? (You do not need to calculate the acf, simply consider what shape it might have given the class of model from which it is drawn.)
   (c) Which model is more likely to represent stock market prices from a theoretical perspective, and why? If any of the three models truly represented the way stock market prices move, which could

potentially be used to make money by forecasting future values of the series?

(d) By making a series of successive substitutions or from your knowledge of the behaviour of these types of processes, consider the extent of persistence of shocks in the series in each case.

4. (a) Describe the steps that Box and Jenkins (1976) suggested should be involved in constructing an ARMA model.
   (b) What particular aspect of this methodology has been the subject of criticism and why?
   (c) Describe an alternative procedure that could be used for this aspect.
5. You obtain the following estimates for an AR(2) model of some returns data

$$y_t = 0.803y_{t-1} + 0.682y_{t-2} + u_t$$

where u_t is a white noise error process. By examining the characteristic equation, check the estimated model for stationarity.

6. A researcher is trying to determine the appropriate order of an ARMA model to describe some actual data, with 200 observations available. She has the following figures for the log of the estimated residual variance (i.e. log $(\hat{\sigma}^2)$) for various candidate models. She has assumed that an order greater than (3,3) should not be necessary to model the dynamics of the data. What is the 'optimal' model order?

ARMA(p,q) *model order*	$\log(\hat{\sigma}^2)$
(0,0)	0.932
(1,0)	0.864
(0,1)	0.902
(1,1)	0.836
(2,1)	0.801
(1,2)	0.821
(2,2)	0.789
(3,2)	0.773
(2,3)	0.782
(3,3)	0.764

7. How could you determine whether the order you suggested for question 6 was in fact appropriate?
8. 'Given that the objective of any econometric modelling exercise is to find the model that most closely 'fits' the data, then adding more lags

to an ARMA model will almost invariably lead to a better fit. Therefore a large model is best because it will fit the data more closely.'
Comment on the validity (or otherwise) of this statement.

9. (a) You obtain the following sample autocorrelations and partial autocorrelations for a sample of 100 observations from actual data:

Lag	1	2	3	4	5	6	7	8
acf	0.420	0.104	0.032	−0.206	−0.138	0.042	−0.018	0.074
pacf	0.632	0.381	0.268	0.199	0.205	0.101	0.096	0.082

Can you identify the most appropriate time series process for this data?

(b) Use the Ljung–Box Q* test to determine whether the first three autocorrelation coefficients taken together are jointly significantly different from zero.

10. You have estimated the following ARMA(1,1) model for some time series data

$$y_t = 0.036 + 0.69y_{t-1} + 0.42u_{t-1} + u_t$$

Suppose that you have data for time to $t-1$, i.e. you know that $y_{t-1} = 3.4$, and $\hat{u}_{t-1} = -1.3$

(a) Obtain forecasts for the series y for times t, $t+1$, and $t+2$ using the estimated ARMA model.

(b) If the actual values for the series turned out to be -0.032, 0.961, 0.203 for t, $t+1$, $t+2$, calculate the (out-of-sample) mean squared error.

(c) A colleague suggests that a simple exponential smoothing model might be more useful for forecasting the series. The estimated value of the smoothing constant is 0.15, with the most recently available smoothed value, S_{t-1} being 0.0305. Obtain forecasts for the series y for times t, $t+1$, and $t+2$ using this model.

(d) Given your answers to parts (a) to (c) of the question, determine whether Box–Jenkins or exponential smoothing models give the most accurate forecasts in this application.

11. (a) Explain what stylised shapes would be expected for the autocorrelation and partial autocorrelation functions for the following stochastic processes:

- white noise
- an AR(2)
- an MA(1)
- an ARMA (2,1).

(b) Consider the following ARMA process.

$$y_t = 0.21 + 1.32y_{t-1} + 0.58u_{t-1} + u_t$$

Determine whether the MA part of the process is invertible.

(c) Produce 1-,2-,3- and 4-step-ahead forecasts for the process given in part (b).

(d) Outline two criteria that are available for evaluating the forecasts produced in part (c), highlighting the differing characteristics of each.

(e) What procedure might be used to estimate the parameters of an ARMA model? Explain, briefly, how such a procedure operates, and why OLS is not appropriate.

12. (a) Briefly explain any difference you perceive between the characteristics of macroeconomic and financial data. Which of these features suggest the use of different econometric tools for each class of data?

(b) Consider the following autocorrelation and partial autocorrelation coefficients estimated using 500 observations for a weakly stationary series, y_t:

Lag	acf	pacf
1	0.307	0.307
2	−0.013	0.264
3	0.086	0.147
4	0.031	0.086
5	−0.197	0.049

Using a simple 'rule of thumb', determine which, if any, of the acf and pacf coefficients are significant at the 5% level. Use both the Box–Pierce and Ljung–Box statistics to test the joint null hypothesis that the first five autocorrelation coefficients are jointly zero.

(c) What process would you tentatively suggest could represent the most appropriate model for the series in part (b)? Explain your answer.

(d) Two researchers are asked to estimate an ARMA model for a daily USD/GBP exchange rate return series, denoted x_t. Researcher *A* uses Schwarz's criterion for determining the appropriate model order and arrives at an ARMA(0,1). Researcher *B* uses Akaike's information criterion which deems an ARMA(2,0) to be optimal. The

estimated models are

$$A : \hat{x}_t = 0.38 + 0.10u_{t-1}$$
$$B : \hat{x}_t = 0.63 + 0.17x_{t-1} - 0.09x_{t-2}$$

where u_t is an error term.
You are given the following data for time until day z (i.e. $t = z$)

$$x_z = 0.31,\ x_{z-1} = 0.02,\ x_{z-2} = -0.16$$
$$u_z = -0.02,\ u_{z-1} = 0.13,\ u_{z-2} = 0.19$$

Produce forecasts for the next 4 days (i.e. for times $z+1$, $z+2$, $z+3$, $z+4$) from both models.

(e) Outline two methods proposed by Box and Jenkins (1970) for determining the adequacy of the models proposed in part (d).

(f) Suppose that the actual values of the series x on days $z+1$, $z+2$, $z+3$, $z+4$ turned out to be 0.62, 0.19, −0.32, 0.72, respectively. Determine which researcher's model produced the most accurate forecasts.

13. Select two of the stock series from the 'CAPM.XLS' Excel file, construct a set of continuously compounded returns, and then perform a time-series analysis of these returns. The analysis should include

(a) An examination of the autocorrelation and partial autocorrelation functions.

(b) An estimation of the information criteria for each ARMA model order from (0,0) to (5,5).

(c) An estimation of the model that you feel most appropriate given the results that you found from the previous two parts of the question.

(d) The construction of a forecasting framework to compare the forecasting accuracy of

i. Your chosen ARMA model
ii. An arbitrary ARMA(1,1)
iii. An single exponential smoothing model
iv. A random walk with drift in the log price levels (hint: this is easiest achieved by treating the returns as an ARMA(0,0) - i.e. simply estimating a model including only a constant).

(e) Then compare the fitted ARMA model with the models that were estimated in chapter 4 based on exogenous variables. Which type of model do you prefer and why?