CHAPTER 18

Advanced Time Series Topics

In this chapter, we cover some more advanced topics in time series econometrics. In Chapters 10, 11, and 12, we emphasized in several places that using time series data in regression analysis requires some care due to the trending, persistent nature of many economic time series. In addition to studying topics such as infinite distributed lag models and forecasting, we also discuss some recent advances in analyzing time series processes with unit roots.

In Section 18.1, we describe infinite distributed lag models, which allow a change in an explanatory variable to affect all future values of the dependent variable. Conceptually, these models are straightforward extensions of the finite distributed lag models in Chapter 10, but estimating these models poses some interesting challenges.

In Section 18.2, we show how to formally test for unit roots in a time series process. Recall from Chapter 11 that we excluded unit root processes to apply the usual asymptotic theory. Because the presence of a unit root implies that a shock today has a long-lasting impact, determining whether a process has a unit root is of interest in its own right.

We cover the notion of spurious regression between two time series processes, each of which has a unit root, in Section 18.3. The main result is that even if two unit root series are *independent*, it is quite likely that the regression of one on the other will yield a statistically significant t statistic. This emphasizes the potentially serious consequences of using standard inference when the dependent and independent variables are integrated processes.

The notion of cointegration applies when two series are I(1), but a linear combination of them is I(0); in this case, the regression of one on the other is not spurious, but instead tells us something about the long-run relationship between them. Cointegration between two series also implies a particular kind of model, called an error correction model, for the short-term dynamics. We cover these models in Section 18.4.

In Section 18.5, we provide an overview of forecasting and bring together all of the tools in this and previous chapters to show how regression methods can be used to forecast future outcomes of a time series. The forecasting literature is vast, so we focus only on the most common regression-based methods. We also touch on the related topic of Granger causality.

18.1 Infinite Distributed Lag Models

Let $\{(y_t, z_t): t = \ldots, -2, -1, 0, 1, 2, \ldots\}$ be a bivariate time series process (which is only partially observed). An **infinite distributed lag (IDL) model** relating y_t to current and all past values of z is

$$y_t = \alpha + \delta_0 z_t + \delta_1 z_{t-1} + \delta_2 z_{t-2} + \ldots + u_t, \qquad \textbf{18.1}$$

where the sum on lagged z extends back to the indefinite past. This model is only an approximation to reality, as no economic process started infinitely far into the past. Compared with a finite distributed lag model, an IDL model does not require that we truncate the lag at a particular value.

In order for model (18.1) to make sense, the lag coefficients, δ_j, must tend to zero as $j \to \infty$. This is not to say that δ_2 is smaller in magnitude than δ_1; it only means that the impact of z_{t-j} on y_t must eventually become small as j gets large. In most applications, this makes economic sense as well: the distant past of z should be less important for explaining y than the recent past of z.

Even if we decide that (18.1) is a useful model, we clearly cannot estimate it without some restrictions. For one, we only observe a finite history of data. Equation (18.1) involves an infinite number of parameters, $\delta_0, \delta_1, \delta_2, \ldots$, which cannot be estimated without restrictions. Later, we place restrictions on the δ_j that allow us to estimate (18.1).

As with finite distributed lag (FDL) models, the impact propensity in (18.1) is simply δ_0 (see Chapter 10). Generally, the δ_h have the same interpretation as in an FDL. Suppose that $z_s = 0$ for all $s < 0$ and that $z_0 = 1$ and $z_s = 0$ for all $s > 1$; in other words, at time $t = 0$, z increases temporarily by one unit and then reverts to its initial level of zero. For any $h \geq 0$, we have $y_h = \alpha + \delta_h + u_h$ for all $h \geq 0$, and so

$$\mathrm{E}(y_h) = \alpha + \delta_h, \qquad \textbf{18.2}$$

where we use the standard assumption that u_h has zero mean. It follows that δ_h is the change in $\mathrm{E}(y_h)$, given a one-unit, temporary change in z at time zero. We just said that δ_h must be tending to zero as h gets large for the IDL to make sense. This means that a temporary change in z has *no long-run effect* on expected y: $\mathrm{E}(y_h) = \alpha + \delta_h \to \alpha$ as $h \to \infty$.

We assumed that the process z starts at $z_s = 0$ and that the one-unit increase occurred at $t = 0$. These were only for the purpose of illustration. More generally, if z temporarily increases by one unit (from any initial level) at time t, then δ_h measures the change in the expected value of y after h periods. The lag distribution, which is δ_h plotted as a function of h, shows the expected path that future y follow given the one-unit, temporary increase in z.

The long-run propensity in model (18.1) is the sum of all of the lag coefficients:

$$LRP = \delta_0 + \delta_1 + \delta_2 + \delta_3 + \ldots, \qquad \textbf{18.3}$$

where we assume that the infinite sum is well defined. Because the δ_j must converge to zero, the LRP can often be well approximated by a finite sum of the form $\delta_0 + \delta_1 + \ldots + \delta_p$ for sufficiently large p. To interpret the LRP, suppose that the process z_t is steady at $z_s = 0$ for $s < 0$. At $t = 0$, the process permanently increases by one unit. For example, if

z_t is the percentage change in the money supply and y_t is the inflation rate, then we are interested in the effects of a permanent increase of one percentage point in money supply growth. Then, by substituting $z_s = 0$ for $s < 0$ and $z_t = 1$ for $t \geq 0$, we have

$$y_h = \alpha + \delta_0 + \delta_1 + \ldots + \delta_h + u_h,$$

where $h \geq 0$ is any horizon. Because u_t has a zero mean for all t, we have

$$\mathrm{E}(y_h) = \alpha + \delta_0 + \delta_1 + \ldots + \delta_h. \qquad \textbf{18.4}$$

[It is useful to compare (18.4) and (18.2).] As the horizon increases, that is, as $h \to \infty$, the right-hand side of (18.4) is, by definition, the long-run propensity, plus α. Thus, the LRP measures the long-run change in the expected value of y given a one-unit, *permanent* increase in z.

Question 18.1

Suppose that $z_s = 0$ for $s < 0$ and that $z_0 = 1$, $z_1 = 1$, and $z_s = 0$ for $s > 1$. Find $\mathrm{E}(y_{-1})$, $\mathrm{E}(y_0)$, and $\mathrm{E}(y_h)$ for $h \geq 1$. What happens as $h \to \infty$?

The previous derivation of the LRP and the interpretation of δ_j used the fact that the errors have a zero mean; as usual, this is not much of an assumption, provided an intercept is included in the model. A closer examination of our reasoning shows that we assumed that the change in z during any time period had no effect on the expected value of u_t. This is the infinite distributed lag version of the *strict exogeneity* assumption that we introduced in Chapter 10 (in particular, Assumption TS.3). Formally,

$$\mathrm{E}(u_t|\ldots, z_{t-2}, z_{t-1}, z_t, z_{t+1}, \ldots) = 0, \qquad \textbf{18.5}$$

so that the expected value of u_t does not depend on the z in *any* time period. Although (18.5) is natural for some applications, it rules out other important possibilities. In effect, (18.5) does not allow feedback from y_t to future z because z_{t+h} must be uncorrelated with u_t for $h > 0$. In the inflation/money supply growth example, where y_t is inflation and z_t is money supply growth, (18.5) rules out future changes in money supply growth that are tied to changes in today's inflation rate. Given that money supply policy often attempts to keep interest rates and inflation at certain levels, this might be unrealistic.

One approach to estimating the δ_j, which we cover in the next subsection, requires a strict exogeneity assumption in order to produce consistent estimators of the δ_j. A weaker assumption is

$$\mathrm{E}(u_t|z_t, z_{t-1}, \ldots) = 0. \qquad \textbf{18.6}$$

Under (18.6), the error is uncorrelated with current and *past* z, but it may be correlated with future z; this allows z_t to be a variable that follows policy rules that depend on past y. Sometimes, (18.6) is sufficient to estimate the δ_j; we explain this in the next subsection.

One thing to remember is that neither (18.5) nor (18.6) says anything about the serial correlation properties of $\{u_t\}$. (This is just as in finite distributed lag models.) If anything, we might expect the $\{u_t\}$ to be serially correlated because (18.1) is not generally

dynamically complete in the sense discussed in Section 11.4. We will study the serial correlation problem later.

How do we interpret the lag coefficients and the LRP if (18.6) holds but (18.5) does not? The answer is: the same way as before. We can still do the previous thought (or counterfactual) experiment, even though the data we observe are generated by some feedback between y_t and future z. For example, we can certainly ask about the long-run effect of a permanent increase in money supply growth on inflation, even though the data on money supply growth cannot be characterized as strictly exogenous.

The Geometric (or Koyck) Distributed Lag

Because there are generally an infinite number of δ_j, we cannot consistently estimate them without some restrictions. The simplest version of (18.1), which still makes the model depend on an infinite number of lags, is the **geometric (or Koyck) distributed lag**. In this model, the δ_j depend on only two parameters:

$$\delta_j = \gamma\rho^j, |\rho| < 1, \quad j = 0, 1, 2, \ldots. \tag{18.7}$$

The parameters γ and ρ may be positive or negative, but ρ must be less than one in absolute value. This ensures that $\delta_j \to 0$ as $j \to \infty$. In fact, this convergence happens at a very fast rate. (For example, with $\rho = .5$ and $j = 10$, $\rho^j = 1/1024 < .001$.)

The impact propensity (IP) in the GDL is simply $\delta_0 = \gamma$, so the sign of the IP is determined by the sign of γ. If $\gamma > 0$, say, and $\rho > 0$, then all lag coefficients are positive. If $\rho < 0$, the lag coefficients alternate in sign (ρ^j is negative for odd j). The long-run propensity is more difficult to obtain, but we can use a standard result on the sum of a geometric series: for $|\rho| < 1$, $1 + \rho + \rho^2 + \ldots + \rho^j + \ldots = 1/(1 - \rho)$, and so

$$LRP = \gamma/(1 - \rho).$$

The LRP has the same sign as γ.

If we plug (18.7) into (18.1), we still have a model that depends on the z back to the indefinite past. Nevertheless, a simple subtraction yields an estimable model. Write the IDL at times t and $t - 1$ as:

$$y_t = \alpha + \gamma z_t + \gamma\rho z_{t-1} + \gamma\rho^2 z_{t-2} + \ldots + u_t \tag{18.8}$$

and

$$y_{t-1} = \alpha + \gamma z_{t-1} + \gamma\rho z_{t-2} + \gamma\rho^2 z_{t-3} + \ldots + u_{t-1}. \tag{18.9}$$

If we multiply the second equation by ρ and subtract it from the first, all but a few of the terms cancel:

$$y_t - \rho y_{t-1} = (1 - \rho)\alpha + \gamma z_t + u_t - \rho u_{t-1},$$

which we can write as

$$y_t = \alpha_0 + \gamma z_t + \rho y_{t-1} + u_t - \rho u_{t-1}, \tag{18.10}$$

where $\alpha_0 = (1 - \rho)\alpha$. This equation looks like a standard model with a lagged dependent variable, where z_t appears contemporaneously. Because γ is the coefficient on z_t and ρ is the coefficient on y_{t-1}, it appears that we can estimate these parameters. [If, for some reason, we are interested in α, we can always obtain $\hat{\alpha} = \hat{\alpha}_0/(1 - \hat{\rho})$ after estimating ρ and α_0.]

The simplicity of (18.10) is somewhat misleading. The error term in this equation, $u_t - \rho u_{t-1}$, is generally correlated with y_{t-1}. From (18.9), it is pretty clear that u_{t-1} and y_{t-1} are correlated. Therefore, if we write (18.10) as

$$y_t = \alpha_0 + \gamma z_t + \rho y_{t-1} + v_t, \qquad \textbf{18.11}$$

where $v_t \equiv u_t - \rho u_{t-1}$, then we generally have correlation between v_t and y_{t-1}. Without further assumptions, OLS estimation of (18.11) produces inconsistent estimates of γ and ρ.

One case where v_t *must* be correlated with y_{t-1} occurs when u_t is independent of z_t and *all* past values of z and y. Then, (18.8) is dynamically complete, so u_t is uncorrelated with y_{t-1}. From (18.9), the covariance between v_t and y_{t-1} is $-\rho\text{Var}(u_{t-1}) = -\rho\sigma_u^2$, which is zero only if $\rho = 0$. We can easily see that v_t is serially correlated: because $\{u_t\}$ is serially uncorrelated, $\text{E}(v_t v_{t-1}) = \text{E}(u_t u_{t-1}) - \rho\text{E}(u_{t-1}^2) - \rho\text{E}(u_t u_{t-2}) + \rho^2\text{E}(u_{t-1}u_{t-2}) = -\rho\sigma_u^2$. For $j > 1$, $\text{E}(v_t v_{t-j}) = 0$. Thus, $\{v_t\}$ is a moving average process of order one (see Section 11.1). This, and equation (18.11), gives an example of a model—which is derived from the original model of interest—that has a lagged dependent variable *and* a particular kind of serial correlation.

If we make the strict exogeneity assumption (18.5), then z_t is uncorrelated with u_t and u_{t-1}, and therefore with v_t. Thus, if we can find a suitable instrumental variable for y_{t-1}, then we can estimate (18.11) by IV. What is a good IV candidate for y_{t-1}? By assumption, u_t and u_{t-1} are both uncorrelated with z_{t-1}, so v_t is uncorrelated with z_{t-1}. If $\gamma \neq 0$, z_{t-1} and y_{t-1} are correlated, even after partialling out z_t. Therefore, we can use instruments (z_t, z_{t-1}) to estimate (18.11). Generally, the standard errors need to be adjusted for serial correlation in the $\{v_t\}$, as we discussed in Section 15.7.

An alternative to IV estimation exploits the fact that $\{u_t\}$ may contain a specific kind of serial correlation. In particular, in addition to (18.6), suppose that $\{u_t\}$ follows the AR(1) model

$$u_t = \rho u_{t-1} + e_t \qquad \textbf{18.12}$$

$$\text{E}(e_t|z_t, y_{t-1}, z_{t-1}, \ldots) = 0. \qquad \textbf{18.13}$$

It is important to notice that the ρ appearing in (18.12) is the same parameter multiplying y_{t-1} in (18.11). If (18.12) and (18.13) hold, we can write equation (18.10) as

$$y_t = \alpha_0 + \gamma z_t + \rho y_{t-1} + e_t, \qquad \textbf{18.14}$$

which is a dynamically complete model under (18.13). From Chapter 11, we can obtain consistent, asymptotically normal estimators of the parameters by OLS. This is very convenient, as there is no need to deal with serial correlation in the errors. If e_t satisfies the homoskedasticity assumption $\text{Var}(e_t|z_t, y_{t-1}) = \sigma_e^2$, the usual inference applies. Once we have estimated γ and ρ, we can easily estimate the LRP: $\widehat{LRP} = \hat{\gamma}/(1 - \hat{\rho})$.

The simplicity of this procedure relies on the potentially strong assumption that $\{u_t\}$ follows an AR(1) process with the *same* ρ appearing in (18.7). This is usually no worse than assuming the $\{u_t\}$ are serially uncorrelated. Nevertheless, because consistency of the estimators relies heavily on this assumption, it is a good idea to test it. A simple test begins by specifying $\{u_t\}$ as an AR(1) process with a *different* parameter, say, $u_t = \lambda u_{t-1} + e_t$. McClain and Wooldridge (1995) devised a simple Lagrange multiplier test of H_0: $\lambda = \rho$ that can be computed after OLS estimation of (18.14).

The geometric distributed lag model extends to multiple explanatory variables—so that we have an infinite DL in each explanatory variable—but then we must be able to write the coefficient on $z_{t-j,h}$ as $\gamma_h \rho^j$. In other words, though γ_h is different for each explanatory variable, ρ is the same. Thus, we can write

$$y_t = \alpha_0 + \gamma_1 z_{t1} + \ldots + \gamma_k z_{tk} + \rho y_{t-1} + v_t. \qquad \textbf{18.15}$$

The same issues that arose in the case with one z arise in the case with many z. Under the natural extension of (18.12) and (18.13)—just replace z_t with $\mathbf{z}_t = (z_{t1}, \ldots, z_{tk})$—OLS is consistent and asymptotically normal. Or, an IV method can be used.

Rational Distributed Lag Models

The geometric DL implies a fairly restrictive lag distribution. When $\gamma > 0$ and $\rho > 0$, the δ_j are positive and monotonically declining to zero. It is possible to have more general infinite distributed lag models. The GDL is a special case of what is generally called a **rational distributed lag (RDL) model**. A general treatment is beyond our scope—Harvey (1990) is a good reference—but we can cover one simple, useful extension.

Such an RDL model is most easily described by adding a lag of z to equation (18.11):

$$y_t = \alpha_0 + \gamma_0 z_t + \rho y_{t-1} + \gamma_1 z_{t-1} + v_t, \qquad \textbf{18.16}$$

where $v_t = u_t - \rho u_{t-1}$, as before. By repeated substitution, it can be shown that (18.16) is equivalent to the infinite distributed lag model

$$\begin{aligned} y_t &= \alpha + \gamma_0(z_t + \rho z_{t-1} + \rho^2 z_{t-2} + \ldots) \\ &\quad + \gamma_1(z_{t-1} + \rho z_{t-2} + \rho^2 z_{t-3} + \ldots) + u_t \\ &= \alpha + \gamma_0 z_t + (\rho\gamma_0 + \gamma_1) z_{t-1} + \rho(\rho\gamma_0 + \gamma_1) z_{t-2} \\ &\quad + \rho^2(\rho\gamma_0 + \gamma_1) z_{t-3} + \ldots + u_t, \end{aligned}$$

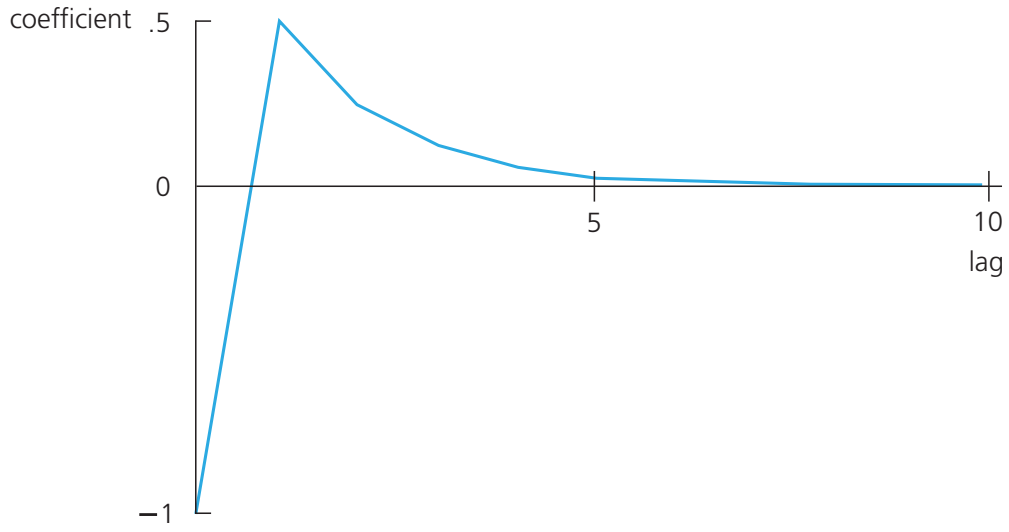
where we again need the assumption $|\rho| < 1$. From this last equation, we can read off the lag distribution. In particular, the impact propensity is γ_0, while the coefficient on z_{t-h} is $\rho^{h-1}(\rho\gamma_0 + \gamma_1)$ for $h \geq 1$. Therefore, this model allows the impact propensity to differ in sign from the other lag coefficients, even if $\rho > 0$. However, if $\rho > 0$, the δ_h have the same sign as $(\rho\gamma_0 + \gamma_1)$ for all $h \geq 1$. The lag distribution is plotted in Figure 18.1 for $\rho = .5$, $\gamma_0 = -1$, and $\gamma_1 = 1$.

The easiest way to compute the long-run propensity is to set y and z at their long-run values for all t, say, y^* and z^*, and then find the change in y^* with respect to z^* (see also Problem 10.3). We have $y^* = \alpha_0 + \gamma_0 z^* + \rho y^* + \gamma_1 z^*$, and solving gives $y^* = \alpha_0/(1 - \rho) + (\gamma_0 + \gamma_1)/(1 - \rho)z^*$. Now, we use the fact that $LRP = \Delta y^*/\Delta z^*$:

$$LRP = (\gamma_0 + \gamma_1)/(1 - \rho).$$

FIGURE 18.1

Lag distribution for the rational distributed lag (18.16) with $\rho = .5$, $\gamma_0 = -1$, and $\gamma_1 = 1$.



Because $|\rho| < 1$, the LRP has the same sign as $\gamma_0 + \gamma_1$, and the LRP is zero if, and only if, $\gamma_0 + \gamma_1 = 0$, as in Figure 18.1.

Example 18.1

[Housing Investment and Residential Price Inflation]

We estimate both the basic geometric and the rational distributed lag models by applying OLS to (18.14) and (18.16), respectively. The dependent variable is log(*invpc*) *after* a linear time trend has been removed [that is, we linearly detrend log(*invpc*)]. For z_t, we use the growth in the price index. This allows us to estimate how residential price inflation affects movements in housing investment around its trend. The results of the estimation, using the data in HSEINV.RAW, are given in Table 18.1.

The geometric distributed lag model is clearly rejected by the data, as $gprice_{-1}$ is very significant. The adjusted *R*-squareds also show that the RDL model fits much better.

The two models give very different estimates of the long-run propensity. If we incorrectly use the GDL, the estimated LRP is almost five: a permanent one percentage point increase in residential price inflation increases long-term housing investment by 4.7% (above its trend value). Economically, this seems implausible. The LRP estimated from the rational distributed lag model is below one. In fact, we cannot reject the null hypothesis H_0: $\gamma_0 + \gamma_1 = 0$ at any reasonable significance level (*p*-value = .83), so there is no evidence that the LRP is different from zero. This is a good example of how misspecifying the dynamics of a model by omitting relevant lags can lead to erroneous conclusions.

TABLE 18.1

Distributed Lag Models for Housing Investment

Dependent Variable: log(*invpc*), detrended		
Independent Variables	**Geometric DL**	**Rational DL**
gprice	3.095 (.933)	3.256 (.970)
y_{-1}	.340 (.132)	.547 (.152)
$gprice_{-1}$	—	−2.936 (.973)
constant	−.010 (.018)	.006 (.017)
Long-run propensity	4.689	.706
Sample size Adjusted *R*-squared	41 .375	40 .504

18.2 Testing for Unit Roots

We now turn to the important problem of testing whether a time series follows a **unit root process**. In Chapter 11, we gave some vague, necessarily informal guidelines to decide whether a series is I(1) or not. In many cases, it is useful to have a formal test for a unit root. As we will see, such tests must be applied with caution.

The simplest approach to testing for a unit root begins with an AR(1) model:

$$y_t = \alpha + \rho y_{t-1} + e_t, \; t = 1, 2, \ldots, \qquad \textbf{18.17}$$

where y_0 is the observed initial value. Throughout this section, we let $\{e_t\}$ denote a process that has zero mean, given past observed y:

$$\mathrm{E}(e_t | y_{t-1}, y_{t-2}, \ldots, y_0) = 0. \qquad \textbf{18.18}$$

[Under (18.18), $\{e_t\}$ is said to be a **martingale difference sequence** with respect to $\{y_{t-1}, y_{t-2}, \ldots\}$. If $\{e_t\}$ is assumed to be i.i.d. with zero mean and is independent of y_0, then it also satisfies (18.18).]

If $\{y_t\}$ follows (18.17), it has a unit root if, and only if, $\rho = 1$. If $\alpha = 0$ and $\rho = 1$, $\{y_t\}$ follows a random walk without drift [with the innovations e_t satisfying (18.18)]. If $\alpha \neq 0$ and $\rho = 1$, $\{y_t\}$ is a random walk with drift, which means that $E(y_t)$ is a linear function of t. A unit root process with drift behaves very differently from one without drift. Nevertheless, it is common to leave α unspecified under the null hypothesis, and this is the approach we take. Therefore, the null hypothesis is that $\{y_t\}$ has a unit root:

$$H_0: \rho = 1. \tag{18.19}$$

In almost all cases, we are interested in the one-sided alternative

$$H_1: \rho < 1. \tag{18.20}$$

(In practice, this means $0 < \rho < 1$, as $\rho < 0$ for a series that we suspect has a unit root would be very rare.) The alternative $H_1: \rho > 1$ is not usually considered, since it implies that y_t is explosive. In fact, if $\alpha > 0$, y_t has an exponential trend in its mean when $\rho > 1$.

When $|\rho| < 1$, $\{y_t\}$ is a stable AR(1) process, which means it is weakly dependent or asymptotically uncorrelated. Recall from Chapter 11 that $\text{Corr}(y_t, y_{t+h}) = \rho^h \to 0$ when $|\rho| < 1$. Therefore, testing (18.19) in model (18.17), with the alternative given by (18.20), is really a test of whether $\{y_t\}$ is I(1) against the alternative that $\{y_t\}$ is I(0). [We do not take the null to be I(0) in this setup because $\{y_t\}$ is I(0) for any value of ρ strictly between -1 and 1, something that classical hypothesis testing does not handle easily. There are tests where the null hypothesis is I(0) against the alternative of I(1), but these take a different approach. See, for example, Kwiatkowski, Phillips, Schmidt, and Shin (1992).]

A convenient equation for carrying out the unit root test is to subtract y_{t-1} from both sides of (18.17) and to define $\theta = \rho - 1$:

$$\Delta y_t = \alpha + \theta y_{t-1} + e_t. \tag{18.21}$$

Under (18.18), this is a dynamically complete model, and so it seems straightforward to test $H_0: \theta = 0$ against $H_1: \theta < 0$. The problem is that, under H_0, y_{t-1} is I(1), and so the usual central limit theorem that underlies the asymptotic standard normal distribution for the t statistic does not apply: the t statistic does not have an approximate standard normal distribution even in large sample sizes. The asymptotic distribution of the t statistic under H_0 has come to be known as the **Dickey-Fuller distribution** after Dickey and Fuller (1979).

Although we cannot use the usual critical values, we *can* use the usual t statistic for $\hat{\theta}$ in (18.21), at least once the appropriate critical values have been tabulated. The resulting test is known as the **Dickey-Fuller** (**DF**) **test** for a unit root. The theory used to obtain the asymptotic critical values is rather complicated and is covered in advanced texts on time series econometrics. [See, for example, Banerjee, Dolado, Galbraith, and Hendry (1993), or BDGH for short.] By contrast, using these results is very easy. The critical values for the t statistic have been tabulated by several authors, beginning with the original work by Dickey and Fuller (1979). Table 18.2 contains the large sample critical values for various significance levels, taken from BDGH (1993, Table 4.2). (Critical values adjusted for small sample sizes are available in BDGH.)

TABLE 18.2

Asymptotic Critical Values for Unit Root *t* Test: No Time Trend

Significance level	1%	2.5%	5%	10%
Critical value	−3.43	−3.12	−2.86	−2.57

We reject the null hypothesis H_0: $\theta = 0$ against H_1: $\theta < 0$ if $t_{\hat{\theta}} < c$, where c is one of the negative values in Table 18.2. For example, to carry out the test at the 5% significance level, we reject if $t_{\hat{\theta}} < -2.86$. This requires a t statistic with a much larger magnitude than if we used the standard normal critical value, which would be −1.65. If we use the standard normal critical value to test for a unit root, we would reject H_0 much more often than 5% of the time when H_0 is true.

Example 18.2

[Unit Root Test for Three-Month T-Bill Rates]

We use the quarterly data in INTQRT.RAW to test for a unit root in three-month T-bill rates. When we estimate (18.20), we obtain

$$\widehat{\Delta r3}_t = .625 - .091\, r3_{t-1} \quad (.261)\ \ (.037) \qquad n = 123,\ R^2 = .048, \tag{18.22}$$

where we keep with our convention of reporting standard errors in parentheses below the estimates. We must remember that these standard errors cannot be used to construct usual confidence intervals or to carry out traditional t tests because these do not behave in the usual ways when there is a unit root. The coefficient on $r3_{t-1}$ shows that the estimate of ρ is $\hat{\rho} = 1 + \hat{\theta} = .909$. While this is less than unity, we do not know whether it is *statistically* less than one. The t statistic on $r3_{t-1}$ is $-.091/.037 = -2.46$. From Table 18.2, the 10% critical value is −2.57; therefore, we fail to reject H_0: $\rho = 1$ against H_1: $\rho < 1$ at the 10% significance level.

As with other hypothesis tests, when we fail to reject H_0, we do *not* say that we accept H_0. Why? Suppose we test H_0: $\rho = .9$ in the previous example using a standard t test—which is asymptotically valid, because y_t is I(0) under H_0. Then, we obtain $t = .001/.037$, which is very small and provides no evidence against $\rho = .9$. Yet, it makes no sense to accept $\rho = 1$ and $\rho = .9$.

When we fail to reject a unit root, as in the previous example, we should only conclude that the data do not provide strong evidence against H_0. In this example, the test does provide *some* evidence against H_0 because the t statistic is close to the 10% critical value. (Ideally, we would compute a p-value, but this requires special software because of the nonnormal distribution.) In addition, though $\hat{\rho} \approx .91$ implies a fair amount of persistence

in $\{r3_t\}$, the correlation between observations that are 10 periods apart for an AR(1) model with $\rho = .9$ is about .35, rather than almost one if $\rho = 1$.

What happens if we now want to use $r3_t$ as an explanatory variable in a regression analysis? The outcome of the unit root test implies that we should be extremely cautious: if $r3_t$ does have a unit root, the usual asymptotic approximations need not hold (as we discussed in Chapter 11). One solution is to use the first difference of $r3_t$ in any analysis. As we will see in Section 18.4, that is not the only possibility.

We also need to test for unit roots in models with more complicated dynamics. If $\{y_t\}$ follows (18.17) with $\rho = 1$, then Δy_t is serially uncorrelated. We can easily allow $\{\Delta y_t\}$ to follow an AR model by augmenting equation (18.21) with additional lags. For example,

$$\Delta y_t = \alpha + \theta y_{t-1} + \gamma_1 \Delta y_{t-1} + e_t, \qquad \textbf{18.23}$$

where $|\gamma_1| < 1$. This ensures that, under H_0: $\theta = 0$, $\{\Delta y_t\}$ follows a stable AR(1) model. Under the alternative H_1: $\theta < 0$, it can be shown that $\{y_t\}$ follows a stable AR(2) model.

More generally, we can add p lags of Δy_t to the equation to account for the dynamics in the process. The way we test the null hypothesis of a unit root is very similar: we run the regression of

$$\Delta y_t \text{ on } y_{t-1}, \Delta y_{t-1}, \ldots, \Delta y_{t-p} \qquad \textbf{18.24}$$

and carry out the t test on $\hat{\theta}$, the coefficient on y_{t-1}, just as before. This extended version of the Dickey-Fuller test is usually called the **augmented Dickey-Fuller test** because the regression has been augmented with the lagged changes, Δy_{t-h}. The critical values and rejection rule are the same as before. The inclusion of the lagged changes in (18.24) is intended to clean up any serial correlation in Δy_t. The more lags we include in (18.24), the more initial observations we lose. If we include too many lags, the small sample power of the test generally suffers. But if we include too few lags, the size of the test will be incorrect, even asymptotically, because the validity of the critical values in Table 18.2 relies on the dynamics being completely modeled. Often, the lag length is dictated by the frequency of the data (as well as the sample size). For annual data, one or two lags usually suffice. For monthly data, we might include 12 lags. But there are no hard rules to follow in any case.

Interestingly, the t statistics on the lagged changes have approximate t distributions. The F statistics for joint significance of any group of terms Δy_{t-h} are also asymptotically valid. (These maintain the homoskedasticity assumption discussed in Section 11.5.) Therefore, we can use standard tests to determine whether we have enough lagged changes in (18.24).

Example 18.3

[Unit Root Test for Annual U.S. Inflation]

We use annual data on U.S. inflation, based on the CPI, to test for a unit root in inflation (see PHILLIPS.RAW), restricting ourselves to the years from 1948 through 1996. Allowing for one lag of Δinf_t in the augmented Dickey-Fuller regression gives

$$\widehat{\Delta inf_t} = 1.36 - .310\, inf_{t-1} + .138\, \Delta inf_{t-1}$$
$$(.517)\ (.103) \qquad (.126)$$
$$n = 47,\ R^2 = .172.$$

The t statistic for the unit root test is $-.310/.103 = -3.01$. Because the 5% critical value is -2.86, we reject the unit root hypothesis at the 5% level. The estimate of ρ is about .690. Together, this is reasonably strong evidence against a unit root in inflation. The lag Δinf_{t-1} has a t statistic of about 1.10, so we do not need to include it, but we could not know this ahead of time. If we drop Δinf_{t-1}, the evidence against a unit root is slightly stronger: $\hat{\theta} = -.335$ ($\hat{\rho} = .665$), and $t_{\hat{\theta}} = -3.13$.

For series that have clear time trends, we need to modify the test for unit roots. A trend-stationary process—which has a linear trend in its mean but is I(0) about its trend—can be mistaken for a unit root process if we do not control for a time trend in the Dickey-Fuller regression. In other words, if we carry out the usual DF or augmented DF test on a trending but I(0) series, we will probably have little power for rejecting a unit root.

To allow for series with time trends, we change the basic equation to

$$\Delta y_t = \alpha + \delta t + \theta y_{t-1} + e_t, \qquad \textbf{18.25}$$

where again the null hypothesis is H_0: $\theta = 0$, and the alternative is H_1: $\theta < 0$. Under the alternative, $\{y_t\}$ is a trend-stationary process. If y_t has a unit root, then $\Delta y_t = \alpha + \delta t + e_t$, and so the *change* in y_t has a mean linear in t unless $\delta = 0$. [It can be shown that $E(y_t)$ is actually a *quadratic* in t.] It is unusual for the first difference of an economic series to have a linear trend, so a more appropriate null hypothesis is probably H_0: $\theta = 0$, $\delta = 0$. Although it is possible to test this joint hypothesis using an F test—but with modified critical values—it is common to only test H_0: $\theta = 0$ using a t test. We follow that approach here. [See BDGH (1993, Section 4.4) for more details on the joint test.]

When we include a time trend in the regression, the critical values of the test change. Intuitively, this occurs because detrending a unit root process tends to make it look more like an I(0) process. Therefore, we require a larger magnitude for the t statistic in order to reject H_0. The Dickey-Fuller critical values for the t test that includes a time trend are given in Table 18.3; they are taken from BDGH (1993, Table 4.2).

TABLE 18.3

Asymptotic Critical Values for Unit Root *t* Test: Linear Time Trend

Significance level	1%	2.5%	5%	10%
Critical value	−3.96	−3.66	−3.41	−3.12

For example, to reject a unit root at the 5% level, we need the t statistic on $\hat{\theta}$ to be less than -3.41, as compared with -2.86 without a time trend.

We can augment equation (18.25) with lags of Δy_t to account for serial correlation, just as in the case without a trend.

Example 18.4

[Unit Root in the Log of U.S. Real Gross Domestic Product]

We can apply the unit root test with a time trend to the U.S. GDP data in INVEN.RAW. These annual data cover the years from 1959 through 1995. We test whether $\log(GDP_t)$ has a unit root. This series has a pronounced trend that looks roughly linear. We include a single lag of $\Delta\log(GDP_t)$, which is simply the growth in GDP (in decimal form), to account for dynamics:

$$\widehat{gGDP}_t = \underset{(.67)}{1.65} + \underset{(.0027)}{.0059}\, t - \underset{(.087)}{.210}\log(GDP_{t-1}) + \underset{(.165)}{.264}\, gGDP_{t-1} \qquad \textbf{18.26}$$
$$n = 35,\ R^2 = .268.$$

From this equation, we get $\hat{\rho} = 1 - .21 = .79$, which is clearly less than one. But we *cannot* reject a unit root in the log of GDP: the t statistic on $\log(GDP_{t-1})$ is $-.210/.087 = -2.41$, which is well above the 10% critical value of -3.12. The t statistic on $gGDP_{t-1}$ is 1.60, which is almost significant at the 10% level against a two-sided alternative.

What should we conclude about a unit root? Again, we cannot reject a unit root, but the point estimate of ρ is not especially close to one. When we have a small sample size—and $n = 35$ is considered to be pretty small—it is very difficult to reject the null hypothesis of a unit root if the process has something close to a unit root. Using more data over longer time periods, many researchers have concluded that there is little evidence against the unit root hypothesis for $\log(GDP)$. This has led most of them to assume that the *growth* in GDP is I(0), which means that $\log(GDP)$ is I(1). Unfortunately, given currently available sample sizes, we cannot have much confidence in this conclusion.

If we omit the time trend, there is much less evidence against H_0, as $\hat{\theta} = -.023$ and $t_{\hat{\theta}} = -1.92$. Here, the estimate of ρ is much closer to one, but this is misleading due to the omitted time trend.

It is tempting to compare the t statistic on the time trend in (18.26), with the critical value from a standard normal or t distribution, to see whether the time trend is significant. Unfortunately, the t statistic on the trend does not have an asymptotic standard normal distribution (unless $|\rho| < 1$). The asymptotic distribution of this t statistic is known, but it is rarely used. Typically, we rely on intuition (or plots of the time series) to decide whether to include a trend in the DF test.

There are many other variants on unit root tests. In one version that is applicable only to series that are clearly not trending, the intercept is omitted from the regression; that is, α is set to zero in (18.21). This variant of the Dickey-Fuller test is rarely used because of biases induced if $\alpha \neq 0$. Also, we can allow for more complicated time trends, such as quadratic. Again, this is seldom used.

Another class of tests attempts to account for serial correlation in Δy_t in a different manner than by including lags in (18.21) or (18.25). The approach is related to the serial correlation-robust standard errors for the OLS estimators that we discussed in Section 12.5. The idea is to be as agnostic as possible about serial correlation in Δy_t. In practice, the (augmented) Dickey-Fuller test has held up pretty well. [See BDGH (1993, Section 4.3) for a discussion on other tests.]

18.3 Spurious Regression

In a cross-sectional environment, we use the phrase "spurious correlation" to describe a situation where two variables are related through their correlation with a third variable. In particular, if we regress y on x, we find a significant relationship. But when we control for another variable, say, z, the partial effect of x on y becomes zero. Naturally, this can also happen in time series contexts with I(0) variables.

As we discussed in Section 10.5, it is possible to find a spurious relationship between time series that have increasing or decreasing trends. Provided the series are weakly dependent about their time trends, the problem is effectively solved by including a time trend in the regression model.

When we are dealing with processes that are integrated of order one, there is an additional complication. Even if the two series have means that are not trending, a simple regression involving two *independent* I(1) series will often result in a significant t statistic.

To be more precise, let $\{x_t\}$ and $\{y_t\}$ be random walks generated by

$$x_t = x_{t-1} + a_t, \quad t = 1, 2, \ldots, \tag{18.27}$$

and

$$y_t = y_{t-1} + e_t, \quad t = 1, 2, \ldots, \tag{18.28}$$

where $\{a_t\}$ and $\{e_t\}$ are independent, identically distributed innovations, with mean zero and variances σ_a^2 and σ_e^2, respectively. For concreteness, take the initial values to be $x_0 = y_0 = 0$. Assume further that $\{a_t\}$ and $\{e_t\}$ are independent processes. This implies that $\{x_t\}$ and $\{y_t\}$ are also independent. But what if we run the simple regression

$$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x_t \tag{18.29}$$

and obtain the usual t statistic for $\hat{\beta}_1$ and the usual R-squared? Because y_t and x_t are independent, we would hope that plim $\hat{\beta}_1 = 0$. Even more importantly, if we test H_0: $\beta_1 = 0$ against H_1: $\beta_1 \neq 0$ at the 5% level, we hope that the t statistic for $\hat{\beta}_1$ is insignificant 95% of the time. Through a simulation, Granger and Newbold (1974) showed that this is *not* the case: even though y_t and x_t are *independent*, the regression of y_t on x_t yields a statistically significant t statistic a large percentage of the time, much larger than the nominal significance level. Granger and Newbold called this the **spurious regression problem**: there is no sense in which y and x are related, but an OLS regression using the usual t statistics will often indicate a relationship.

Recent simulation results are given by Davidson and MacKinnon (1993, Table 19.1), where a_t and e_t are generated as independent, identically distributed normal random variables, and 10,000 different samples are generated. For a sample size of $n = 50$ at the 5% significance level, the standard t statistic for H_0: $\beta_1 = 0$ against the two-sided alternative rejects H_0 about 66.2% of the time under H_0, rather than 5% of the time. As the sample size increases, things get *worse*: with $n = 250$, the null is rejected 84.7% of the time!

Question 18.2

Under the preceding setup, where $\{x_t\}$ and $\{y_t\}$ are generated by (18.27) and (18.28) and $\{e_t\}$ and $\{a_t\}$ are i.i.d. sequences, what is the plim of the slope coefficient, say, $\hat{\gamma}_1$, from the regression of Δy_t on Δx_t? Describe the behavior of the t statistic of $\hat{\gamma}_1$.

Here is one way to see what is happening when we regress the level of y on the level of x. Write the model underlying (18.29) as

$$y_t = \beta_0 + \beta_1 x_t + u_t. \qquad \textbf{18.30}$$

For the t statistic of $\hat{\beta}_1$ to have an approximate standard normal distribution in large samples, at a minimum, $\{u_t\}$ should be a mean zero, serially uncorrelated process. But under H_0: $\beta_1 = 0$, $y_t = \beta_0 + u_t$, and, because $\{y_t\}$ is a random walk starting at $y_0 = 0$, equation (18.30) holds under H_0 only if $\beta_0 = 0$ and, more importantly, if $u_t = y_t = \sum_{j=1}^{t} e_j$. In other words, $\{u_t\}$ is a random walk under H_0. This clearly violates even the asymptotic version of the Gauss-Markov assumptions from Chapter 11.

Including a time trend does not really change the conclusion. If y_t or x_t is a random walk with drift and a time trend is not included, the spurious regression problem is even worse. The same qualitative conclusions hold if $\{a_t\}$ and $\{e_t\}$ are general I(0) processes, rather than i.i.d. sequences.

In addition to the usual t statistic not having a limiting standard normal distribution—in fact, it increases to infinity as $n \to \infty$—the behavior of R-squared is nonstandard. In cross-sectional contexts or in regressions with I(0) time series variables, the R-squared converges in probability to the population R-squared: $1 - \sigma_u^2/\sigma_y^2$. This is not the case in spurious regressions with I(1) processes. Rather than the R-squared having a well-defined plim, it actually converges to a random variable. Formalizing this notion is well beyond the scope of this text. [A discussion of the asymptotic properties of the t statistic and the R-squared can be found in BDGH (Section 3.1).] The implication is that the R-squared is large with high probability, even though $\{y_t\}$ and $\{x_t\}$ are independent time series processes.

The same considerations arise with multiple independent variables, each of which may be I(1) or some of which may be I(0). If $\{y_t\}$ is I(1) and at least some of the explanatory variables are I(1), the regression results may be spurious.

The possibility of spurious regression with I(1) variables is quite important and has led economists to reexamine many aggregate time series regressions whose t statistics were very significant and whose R-squareds were extremely high. In the next section, we show that regressing an I(1) dependent variable on an I(1) independent variable *can* be informative, but only if these variables are related in a precise sense.

18.4 Cointegration and Error Correction Models

The discussion of spurious regression in the previous section certainly makes one wary of using the levels of I(1) variables in regression analysis. In earlier chapters, we suggested that I(1) variables should be differenced before they are used in linear regression models, whether they are estimated by OLS or instrumental variables. This is certainly a safe course to follow, and it is the approach used in many time series regressions after Granger and Newbold's original paper on the spurious regression problem. Unfortunately, always differencing I(1) variables limits the scope of the questions that we can answer.

Cointegration

The notion of **cointegration**, which was given a formal treatment in Engle and Granger (1987), makes regressions involving I(1) variables potentially meaningful. A full treatment

of cointegration is mathematically involved, but we can describe the basic issues and methods that are used in many applications.

If $\{y_t: t = 0, 1, \ldots\}$ and $\{x_t: t = 0, 1, \ldots\}$ are two I(1) processes, then, in general, $y_t - \beta x_t$ is an I(1) process for any number β. Nevertheless, it is *possible* that for some $\beta \neq 0$, $y_t - \beta x_t$ is an I(0) process, which means it has constant mean, constant variance, and autocorrelations that depend only on the time distance between any two variables in the series, and it is asymptotically uncorrelated. If such a β exists, we say that y and x are *cointegrated*, and we call β the cointegration parameter. [Alternatively, we could look at $x_t - \gamma y_t$ for $\gamma \neq 0$: if $y_t - \beta x_t$ is I(0), then $x_t - (1/\beta)y_t$ is I(0). Therefore, the linear combination of y_t and x_t is not unique, but if we fix the coefficient on y_t at unity, then β is unique. See Problem 18.3. For concreteness, we consider linear combinations of the form $y_t - \beta x_t$.]

Question 18.3

Let $\{(y_t, x_t): t = 1, 2, \ldots\}$ be a bivariate time series where each series is I(1) without drift. Explain why, if y_t and x_t are cointegrated, y_t and x_{t-1} are also cointegrated.

For the sake of illustration, take $\beta = 1$, suppose that $y_0 = x_0 = 0$, and write $y_t = y_{t-1} + r_t$, $x_t = x_{t-1} + v_t$, where $\{r_t\}$ and $\{v_t\}$ are two I(0) processes with zero means. Then, y_t and x_t have a tendency to wander around and not return to the initial value of zero with any regularity. By contrast, if $y_t - x_t$ is I(0), it has zero mean and does return to zero with some regularity.

As a specific example, let $r6_t$ be the annualized interest rate for six-month T-bills (at the end of quarter t) and let $r3_t$ be the annualized interest rate for three-month T-bills. (These are typically called bond equivalent yields, and they are reported in the financial pages.) In Example 18.2, using the data in INTQRT.RAW, we found little evidence against the hypothesis that $r3_t$ has a unit root; the same is true of $r6_t$. Define the spread between six- and three-month T-bill rates as $spr_t = r6_t - r3_t$. Then, using equation (18.21), the Dickey-Fuller t statistic for spr_t is -7.71 (with $\hat{\theta} = -.67$ or $\hat{\rho} = .33$). Therefore, we strongly reject a unit root for spr_t in favor of I(0). The upshot of this is that though $r6_t$ and $r3_t$ each appear to be unit root processes, the difference between them is an I(0) process. In other words, *r6* and *r3* are cointegrated.

Cointegration in this example, as in many examples, has an economic interpretation. If *r6* and *r3* were not cointegrated, the difference between interest rates could become very large, with no tendency for them to come back together. Based on a simple arbitrage argument, this seems unlikely. Suppose that the spread spr_t continues to grow for several time periods, making six-month T-bills a much more desirable investment. Then, investors would shift away from three-month and toward six-month T-bills, driving up the price of six-month T-bills, while lowering the price of three-month T-bills. Because interest rates are inversely related to price, this would lower *r6* and increase *r3*, until the spread is reduced. Therefore, large deviations between *r6* and *r3* are not expected to continue: the spread has a tendency to return to its mean value. (The spread actually has a slightly positive mean because long-term investors are more rewarded relative to short-term investors.)

There is another way to characterize the fact that spr_t will not deviate for long periods from its average value: *r6* and *r3* have a *long-run* relationship. To describe what we mean by this, let $\mu = \mathrm{E}(spr_t)$ denote the expected value of the spread. Then, we can write

$$r6_t = r3_t + \mu + e_t,$$

where $\{e_t\}$ is a zero mean, I(0) process. The equilibrium or long-run relationship occurs when $e_t = 0$, or $r6^* = r3^* + \mu$. At any time period, there can be deviations from equilibrium, but they will be temporary: there are economic forces that drive *r6* and *r3* back toward the equilibrium relationship.

In the interest rate example, we used economic reasoning to tell us the value of β if y_t and x_t are cointegrated. If we have a hypothesized value of β, then *testing* whether two series are cointegrated is easy: we simply define a new variable, $s_t = y_t - \beta x_t$, and apply either the usual DF or augmented DF test to $\{s_t\}$. If we *reject* a unit root in $\{s_t\}$ in favor of the I(0) alternative, then we find that y_t and x_t *are cointegrated*. In other words, the null hypothesis is that y_t and x_t are *not* cointegrated.

Testing for cointegration is more difficult when the (potential) cointegration parameter β is unknown. Rather than test for a unit root in $\{s_t\}$, we must first estimate β. *If* y_t and x_t are cointegrated, it turns out that the OLS estimator $\hat{\beta}$ from the regression

$$y_t = \hat{\alpha} + \hat{\beta} x_t \qquad \textbf{18.31}$$

is consistent for β. The problem is that the null hypothesis states that the two series are *not* cointegrated, which means that, under H_0, we are running a spurious regression. Fortunately, it is possible to tabulate critical values even when β is estimated, where we apply the Dickey-Fuller or augmented Dickey-Fuller test to the residuals, say, $\hat{u}_t = y_t - \hat{\alpha} - \hat{\beta} x_t$, from (18.31). The only difference is that the critical values account for estimation of β. The resulting test is called the **Engle-Granger test**, and the asymptotic critical values are given in Table 18.4. These are taken from Davidson and MacKinnon (1993, Table 20.2).

TABLE 18.4

Asymptotic Critical Values for Cointegration Test: No Time Trend

Significance level	1%	2.5%	5%	10%
Critical value	−3.90	−3.59	−3.34	−3.04

In the basic test, we run the regression of $\Delta\hat{u}_t$ on $\hat{u}_{t-1}$ and compare the t statistic on $\hat{u}_{t-1}$ to the desired critical value in Table 18.4. If the t statistic is below the critical value, we have evidence that $y_t - \beta x_t$ is I(0) for some β; that is, y_t and x_t are cointegrated. We can add lags of $\Delta\hat{u}_t$ to account for serial correlation. If we compare the critical values in Table 18.4 with those in Table 18.2, we must get a t statistic much larger in magnitude to find cointegration than if we used the usual DF critical values. This happens because OLS, which minimizes the sum of squared residuals, tends to produce residuals that look like an I(0) sequence even if y_t and x_t are *not* cointegrated.

As with the usual Dickey-Fuller test, we can augment the Engle-Granger test by including lags of $\Delta\hat{u}_t$ as additional regressors.

If y_t and x_t are not cointegrated, a regression of y_t on x_t is spurious and tells us nothing meaningful: there is no long-run relationship between y and x. We can still run a regression involving the first differences, Δy_t and Δx_t, including lags. But we should interpret these regressions for what they are: they explain the difference in y in terms of the difference in x and have nothing necessarily to do with a relationship in levels.

If y_t and x_t *are* cointegrated, we can use this to specify more general dynamic models, as we will see in the next subsection.

The previous discussion assumes that neither y_t nor x_t has a drift. This is reasonable for interest rates but not for other time series. If y_t and x_t contain drift terms, $\mathrm{E}(y_t)$ and $\mathrm{E}(x_t)$ are linear (usually increasing) functions of time. The strict definition of cointegration requires $y_t - \beta x_t$ to be I(0) *without* a trend. To see what this entails, write $y_t = \delta t + g_t$ and $x_t = \lambda t + h_t$, where $\{g_t\}$ and $\{h_t\}$ are I(1) processes, δ is the drift in y_t [$\delta = \mathrm{E}(\Delta y_t)$], and λ is the drift in x_t [$\lambda = \mathrm{E}(\Delta x_t)$]. Now, if y_t and x_t are cointegrated, there must exist β such that $g_t - \beta h_t$ is I(0). But then

$$y_t - \beta x_t = (\delta - \beta\lambda)t + (g_t - \beta h_t),$$

which is generally a *trend-stationary* process. The strict form of cointegration requires that there not be a trend, which means $\delta = \beta\lambda$. For I(1) processes with drift, it is possible that the stochastic parts—that is, g_t and h_t—are cointegrated, but that the parameter β that causes $g_t - \beta h_t$ to be I(0) does not eliminate the linear time trend.

We can test for cointegration between g_t and h_t, without taking a stand on the trend part, by running the regression

$$\hat{y}_t = \hat{\alpha} + \hat{\eta}t + \hat{\beta}x_t \qquad \textbf{18.32}$$

and applying the usual DF or augmented DF test to the residuals $\hat{u}_t$. The asymptotic critical values are given in Table 18.5 [from Davidson and MacKinnon (1993, Table 20.2)].

TABLE 18.5

Asymptotic Critical Values for Cointegration Test: Linear Time Trend

Significance level	1%	2.5%	5%	10%
Critical value	−4.32	−4.03	−3.78	−3.50

A finding of cointegration in this case leaves open the possibility that $y_t - \beta x_t$ has a linear trend. But at least it is not I(1).

Example 18.5

[Cointegration between Fertility and Personal Exemption]

In Chapters 10 and 11, we studied various models to estimate the relationship between the general fertility rate (*gfr*) and the real value of the personal tax exemption (*pe*) in the United States. The static regression results in levels and first differences are notably different. The regression in levels, with a time trend included, gives an OLS coefficient on *pe* equal to .187 (se = .035) and $R^2 = .500$. In first differences (without a trend), the coefficient on Δpe is $-.043$ (se = .028), and $R^2 = .032$. Although there are other reasons for these differences—such as misspecified distributed lag dynamics—the discrepancy between the levels and changes regressions suggests that we should test for cointegration. Of course, this presumes that *gfr* and *pe* are I(1) processes. This appears to be the case: the augmented DF tests, with a single lagged change and a linear time trend, each yield *t* statistics of about -1.47, and the estimated AR(1) coefficients are close to one.

When we obtain the residuals from the regression of *gfr* on *t* and *pe* and apply the augmented DF test with one lag, we obtain a *t* statistic on $\hat{u}_{t-1}$ of -2.43, which is nowhere near the 10% critical value, -3.50. Therefore, we must conclude that there is little evidence of cointegration between *gfr* and *pe*, even allowing for separate trends. It is very likely that the earlier regression results we obtained in levels suffer from the spurious regression problem.

The good news is that, when we used first differences and allowed for two lags—see equation (11.27)—we found an overall positive and significant long-run effect of Δpe on Δgfr.

If we think two series are cointegrated, we often want to test hypotheses about the cointegrating parameter. For example, a theory may state that the cointegrating parameter is one. Ideally, we could use a *t* statistic to test this hypothesis.

We explicitly cover the case without time trends, although the extension to the linear trend case is immediate. When y_t and x_t are I(1) and cointegrated, we can write

$$y_t = \alpha + \beta x_t + u_t, \qquad \textbf{18.33}$$

where u_t is a zero mean, I(0) process. Generally, $\{u_t\}$ contains serial correlation, but we know from Chapter 11 that this does not affect consistency of OLS. As mentioned earlier, OLS applied to (18.33) consistently estimates β (and α). Unfortunately, because x_t is I(1), the usual inference procedures do not necessarily apply: OLS is not asymptotically normally distributed, and the *t* statistic for $\hat{\beta}$ does not necessarily have an approximate *t* distribution. We do know from Chapter 10 that, if $\{x_t\}$ is strictly exogenous—see Assumption TS.3—and the errors are homoskedastic, serially uncorrelated, and normally distributed, the OLS estimator is also normally distributed (conditional on the explanatory variables) *and* the *t* statistic has an exact *t* distribution. Unfortunately, these assumptions are too strong to apply to most situations. The notion of cointegration implies nothing about the relationship between $\{x_t\}$ and $\{u_t\}$—indeed, they can be arbitrarily correlated. Further, except for requiring that $\{u_t\}$ is I(0), cointegration between y_t and x_t does not restrict the serial dependence in $\{u_t\}$.

Fortunately, the feature of (18.33) that makes inference the most difficult—the lack of strict exogeneity of $\{x_t\}$—can be fixed. Because x_t is I(1), the proper notion of strict exogeneity is that u_t is uncorrelated with Δx_s, for all *t* and *s*. We can always arrange this

for a *new* set of errors, at least approximately, by writing u_t as a function of the Δx_s for all s close to t. For example,

$$u_t = \eta + \phi_0 \Delta x_t + \phi_1 \Delta x_{t-1} + \phi_2 \Delta x_{t-2} + \gamma_1 \Delta x_{t+1} + \gamma_2 \Delta x_{t+2} + e_t, \qquad \textbf{18.34}$$

where, by construction, e_t is uncorrelated with each Δx_s appearing in the equation. The hope is that e_t is uncorrelated with further lags and leads of Δx_s. We know that, as $|s - t|$ gets large, the correlation between e_t and Δx_s approaches zero, because these are I(0) processes. Now, if we plug (18.34) into (18.33), we obtain

$$y_t = \alpha_0 + \beta x_t + \phi_0 \Delta x_t + \phi_1 \Delta x_{t-1} + \phi_2 \Delta x_{t-2} + \gamma_1 \Delta x_{t+1} + \gamma_2 \Delta x_{t+2} + e_t. \qquad \textbf{18.35}$$

This equation looks a bit strange because future Δx_s appear with both current and lagged Δx_t. The key is that the coefficient on x_t is still β, and, by construction, x_t is now strictly exogenous in this equation. The strict exogeneity assumption is the important condition needed to obtain an approximately normal t statistic for $\hat{\beta}$. If u_t is uncorrelated with all Δx_s, $s \neq t$, then we can drop the leads and lags of the changes and simply include the contemporaneous change, Δx_t. Then, the equation we estimate looks more standard but still includes the first difference of x_t along with its level: $y_t = \alpha_0 + \beta x_t + \phi_0 \Delta x_t + e_t$. In effect, adding Δx_t solves any contemporaneous endogeneity between x_t and u_t. (Remember, any endogeneity does not cause inconsistency. But we are trying to obtain an asymptotically normal t statistic.) Whether we need to include leads and lags of the changes, and how many, is really an empirical issue. Each time we add an additional lead or lag, we lose one observation, and this can be costly unless we have a large data set.

The OLS estimator of β from (18.35) is called the **leads and lags estimator** of β because of the way it employs Δx. [See, for example, Stock and Watson (1993).] The only issue we must worry about in (18.35) is the possibility of serial correlation in $\{e_t\}$. This can be dealt with by computing a serial correlation-robust standard error for $\hat{\beta}$ (as described in Section 12.5) or by using a standard AR(1) correction (such as Cochrane-Orcutt).

Example 18.6

[Cointegrating Parameter for Interest Rates]

Earlier, we tested for cointegration between *r6* and *r3*—six- and three-month T-bill rates—by assuming that the cointegrating parameter was equal to one. This led us to find cointegration and, naturally, to conclude that the cointegrating parameter is equal to unity. Nevertheless, let us estimate the cointegrating parameter directly and test H_0: $\beta = 1$. We apply the leads and lags estimator with two leads and two lags of $\Delta r3$, as well as the contemporaneous change. The estimate of β is $\hat{\beta} = 1.038$, and the usual OLS standard error is .0081. Therefore, the t statistic for H_0: $\beta = 1$ is $(1.038 - 1)/.0081 \approx 4.69$, which is a strong statistical rejection of H_0. (Of course, whether 1.038 is economically different from 1 is a relevant consideration.) There is little evidence of serial correlation in the residuals, so we can use this t statistic as having an approximate normal distribution. [For comparison, the OLS estimate of β without the leads, lags, or contemporaneous $\Delta r3$ terms—and using five more observations—is 1.026 (se = .0077). But the t statistic from (18.33) is not necessarily valid.]

There are many other estimators of cointegrating parameters, and this continues to be a very active area of research. The notion of cointegration applies to more than two processes, but the interpretation, testing, and estimation are much more complicated. One issue is that, even after we normalize a coefficient to be one, there can be many cointegrating relationships. BDGH provide some discussion and several references.

Error Correction Models

In addition to learning about a potential long-run relationship between two series, the concept of cointegration enriches the kinds of dynamic models at our disposal. If y_t and x_t are I(1) processes and are *not* cointegrated, we might estimate a dynamic model in first differences. As an example, consider the equation

$$\Delta y_t = \alpha_0 + \alpha_1 \Delta y_{t-1} + \gamma_0 \Delta x_t + \gamma_1 \Delta x_{t-1} + u_t, \qquad \textbf{18.36}$$

where u_t has zero mean given Δx_t, Δy_{t-1}, Δx_{t-1}, and further lags. This is essentially equation (18.16), but in first differences rather than in levels. If we view this as a rational distributed lag model, we can find the impact propensity, long-run propensity, and lag distribution for Δy as a distributed lag in Δx.

If y_t and x_t are cointegrated with parameter β, then we have additional I(0) variables that we can include in (18.36). Let $s_t = y_t - \beta x_t$, so that s_t is I(0), and assume for the sake of simplicity that s_t has zero mean. Now, we can include lags of s_t in the equation. In the simplest case, we include one lag of s_t:

$$\begin{aligned} \Delta y_t &= \alpha_0 + \alpha_1 \Delta y_{t-1} + \gamma_0 \Delta x_t + \gamma_1 \Delta x_{t-1} + \delta s_{t-1} + u_t \\ &= \alpha_0 + \alpha_1 \Delta y_{t-1} + \gamma_0 \Delta x_t + \gamma_1 \Delta x_{t-1} + \delta(y_{t-1} - \beta x_{t-1}) + u_t, \end{aligned} \qquad \textbf{18.37}$$

where $\mathrm{E}(u_t | I_{t-1}) = 0$, and I_{t-1} contains information on Δx_t and all past values of x and y. The term $\delta(y_{t-1} - \beta x_{t-1})$ is called the *error correction term*, and (18.37) is an example of an **error correction model**. (In some error correction models, the contemporaneous change in x, Δx_t, is omitted. Whether it is included or not depends partly on the purpose of the equation. In forecasting, Δx_t is rarely included, for reasons we will see in Section 18.5.)

An error correction model allows us to study the short-run dynamics in the relationship between y and x. For simplicity, consider the model without lags of Δy_t and Δx_t:

$$\Delta y_t = \alpha_0 + \gamma_0 \Delta x_t + \delta(y_{t-1} - \beta x_{t-1}) + u_t, \qquad \textbf{18.38}$$

where $\delta < 0$. If $y_{t-1} > \beta x_{t-1}$, then y in the previous period has overshot the equilibrium; because $\delta < 0$, the error correction term works to push y back toward the equilibrium. Similarly, if $y_{t-1} < \beta x_{t-1}$, the error correction term induces a positive change in y back toward the equilibrium.

How do we estimate the parameters of an error correction model? If we know β, this is easy. For example, in (18.38), we simply regress Δy_t on Δx_t and s_{t-1}, where $s_{t-1} = (y_{t-1} - \beta x_{t-1})$.

Example 18.7

[Error Correction Model for Holding Yields]

In Problem 11.6, we regressed $hy6_t$, the three-month holding yield (in percent) from buying a six-month T-bill at time $t-1$ and selling it at time t as a three-month T-bill, on $hy3_{t-1}$, the three-month holding yield from buying a three-month T-bill at time $t-1$. The expectations hypothesis implies that the slope coefficient should not be statistically different from one. It turns out that there is evidence of a unit root in $\{hy3_t\}$, which calls into question the standard regression analysis. We will assume that both holding yields are I(1) processes. The expectations hypothesis implies, at a minimum, that $hy6_t$ and $hy3_{t-1}$ are cointegrated with β equal to one, which appears to be the case (see Computer Exercise C18.5). Under this assumption, an error correction model is

Question 18.4

How would you test H_0: $\gamma_0 = 1, \delta = -1$ in the holding yield error correction model?

$$\Delta hy6_t = \alpha_0 + \gamma_0 \Delta hy3_{t-1} + \delta(hy6_{t-1} - hy3_{t-2}) + u_t,$$

where u_t has zero mean, given all *hy3* and *hy6* dated at time $t-1$ and earlier. The lags on the variables in the error correction model are dictated by the expectations hypothesis.

Using the data in INTQRT.RAW gives

$$\widehat{\Delta hy6_t} = .090 + 1.218\ \Delta hy3_{t-1} - .840\ (hy6_{t-1} - hy3_{t-2}) \quad (18.39)$$
$$(.043) \quad (.264) \quad (.244)$$
$$n = 122,\ R^2 = .790.$$

The error correction coefficient is negative and very significant. For example, if the holding yield on six-month T-bills is above that for three-month T-bills by one point, *hy6* falls by .84 points on average in the next quarter. Interestingly, $\hat{\delta} = -.84$ is not statistically different from -1, as is easily seen by computing the 95% confidence interval.

In many other examples, the cointegrating parameter must be estimated. Then, we replace s_{t-1} with $\hat{s}_{t-1} = y_{t-1} - \hat{\beta}x_{t-1}$, where $\hat{\beta}$ can be various estimators of β. We have covered the standard OLS estimator as well as the leads and lags estimator. This raises the issue about how sampling variation in $\hat{\beta}$ affects inference on the other parameters in the error correction model. Fortunately, as shown by Engle and Granger (1987), we can ignore the preliminary estimation of β (asymptotically). This property is very convenient and implies that the asymptotic efficiency of the estimators of the parameters in the error correction model is unaffected by whether we use the OLS estimator or the leads and lags estimator for $\hat{\beta}$. Of course, the choice of $\hat{\beta}$ will generally have an effect on the estimated error correction parameters in any particular sample, but we have no systematic way of deciding which preliminary estimator of β to use. The procedure of replacing β with $\hat{\beta}$ is called the **Engle-Granger two-step procedure**.

18.5 Forecasting

Forecasting economic time series is very important in some branches of economics, and it is an area that continues to be actively studied. In this section, we focus on regression-based forecasting methods. Diebold (2001) provides a comprehensive introduction to forecasting, including recent developments.

We assume in this section that the primary focus is on forecasting future values of a time series process and not necessarily on estimating causal or structural economic models.

It is useful to first cover some fundamentals of forecasting that do not depend on a specific model. Suppose that at time t we want to forecast the outcome of y at time $t + 1$, or y_{t+1}. The time period could correspond to a year, a quarter, a month, a week, or even a day. Let I_t denote information that we can observe at time t. This **information set** includes y_t, earlier values of y, and often other variables dated at time t or earlier. We can combine this information in innumerable ways to forecast y_{t+1}. Is there one best way?

The answer is yes, provided we specify the *loss* associated with forecast error. Let f_t denote the forecast of y_{t+1} made at time t. We call f_t a **one-step-ahead forecast**. The **forecast error** is $e_{t+1} = y_{t+1} - f_t$, which we observe once the outcome on y_{t+1} is observed. The most common measure of loss is the same one that leads to ordinary least squares estimation of a multiple linear regression model: the squared error, e_{t+1}^2. The squared forecast error treats positive and negative prediction errors symmetrically, and larger forecast errors receive relatively more weight. For example, errors of $+2$ and -2 yield the same loss, and the loss is four times as great as forecast errors of $+1$ or -1. The squared forecast error is an example of a **loss function**. Another popular loss function is the absolute value of the prediction error, $|e_{t+1}|$. For reasons to be seen shortly, we focus now on squared error loss.

Given the squared error loss function, we can determine how to best use the information at time t to forecast y_{t+1}. But we must recognize that at time t, we do not know e_{t+1}: it is a random variable, because y_{t+1} is a random variable. Therefore, any useful criterion for choosing f_t must be based on what we know at time t. It is natural to choose the forecast to minimize the *expected* squared forecast error, given I_t:

$$\mathrm{E}(e_{t+1}^2|I_t) = \mathrm{E}[(y_{t+1} - f_t)^2|I_t]. \qquad \textbf{18.40}$$

A basic fact from probability (see Property CE.6 in Appendix B) is that the conditional expectation, $\mathrm{E}(y_{t+1}|I_t)$, minimizes (18.40). In other words, if we wish to minimize the expected squared forecast error given information at time t, our forecast should be the expected value of y_{t+1} given variables we know at time t.

For many popular time series processes, the conditional expectation is easy to obtain. Suppose that $\{y_t: t = 0, 1, \ldots\}$ is a martingale difference sequence (MDS) and take I_t to be $\{y_t, y_{t-1}, \ldots, y_0\}$, the observed past of y. By definition, $\mathrm{E}(y_{t+1}|I_t) = 0$ for all t; the best prediction of y_{t+1} at time t is always zero! Recall from Section 18.2 that an i.i.d. sequence with zero mean is a martingale difference sequence.

A martingale difference sequence is one in which the past is not useful for predicting the future. Stock returns are widely thought to be well approximated as an MDS or, perhaps, with a positive mean. The key is that $\mathrm{E}(y_{t+1}|y_t, y_{t-1}, \ldots) = \mathrm{E}(y_{t+1})$: the conditional mean is equal to the unconditional mean, in which case past y do not help to predict future y.

A process $\{y_t\}$ is a **martingale** if $\mathrm{E}(y_{t+1}|y_t, y_{t-1}, \ldots, y_0) = y_t$ for all $t \geq 0$. [If $\{y_t\}$ is a martingale, then $\{\Delta y_t\}$ is a martingale difference sequence, which is where the latter name

comes from.] The predicted value of y for the next period is always the value of y for this period.

A more complicated example is

$$\mathrm{E}(y_{t+1}|I_t) = \alpha y_t + \alpha(1-\alpha)y_{t-1} + \ldots + \alpha(1-\alpha)^t y_0, \qquad \textbf{18.41}$$

where $0 < \alpha < 1$ is a parameter that we must choose. This method of forecasting is called **exponential smoothing** because the weights on the lagged y decline to zero exponentially.

The reason for writing the expectation as in (18.41) is that it leads to a very simple recurrence relation. Set $f_0 = y_0$. Then, for $t \geq 1$, the forecasts can be obtained as

$$f_t = \alpha y_t + (1-\alpha)f_{t-1}.$$

In other words, the forecast of y_{t+1} is a weighted average of y_t and the forecast of y_t made at time $t-1$. Exponential smoothing is suitable only for very specific time series and requires choosing α. Regression methods, which we turn to next, are more flexible.

The previous discussion has focused on forecasting y only one period ahead. The general issues that arise in forecasting y_{t+h} at time t, where h is any positive integer, are similar. In particular, if we use expected squared forecast error as our measure of loss, the best predictor is $\mathrm{E}(y_{t+h}|I_t)$. When dealing with a **multiple-step-ahead forecast**, we use the notation $f_{t,h}$ to indicate the forecast of y_{t+h} made at time t.

Types of Regression Models Used for Forecasting

There are many different regression models that we can use to forecast future values of a time series. The first regression model for time series data from Chapter 10 was the static model. To see how we can forecast with this model, assume that we have a single explanatory variable:

$$y_t = \beta_0 + \beta_1 z_t + u_t. \qquad \textbf{18.42}$$

Suppose, for the moment, that the parameters β_0 and β_1 are known. Write this equation at time $t+1$ as $y_{t+1} = \beta_0 + \beta_1 z_{t+1} + u_{t+1}$. Now, *if* z_{t+1} is known at time t, so that it is an element of I_t and $\mathrm{E}(u_{t+1}|I_t) = 0$, then

$$\mathrm{E}(y_{t+1}|I_t) = \beta_0 + \beta_1 z_{t+1},$$

where I_t contains $z_{t+1}, y_t, z_t, \ldots, y_1, z_1$. The right-hand side of this equation is the forecast of y_{t+1} at time t. This kind of forecast is usually called a **conditional forecast** because it is conditional on knowing the value of z at time $t+1$.

Unfortunately, at any time, we rarely know the value of the explanatory variables in future time periods. Exceptions include time trends and seasonal dummy variables, which we cover explicitly below, but otherwise knowledge of z_{t+1} at time t is rare. Sometimes, we wish to generate conditional forecasts for several values of z_{t+1}.

Another problem with (18.42) as a model for forecasting is that $\mathrm{E}(u_{t+1}|I_t) = 0$ means that $\{u_t\}$ cannot contain serial correlation, something we have seen to be false in most

static regression models. [Problem 18.8 asks you to derive the forecast in a simple distributed lag model with AR(1) errors.]

If z_{t+1} is not known at time t, we cannot include it in I_t. Then, we have

$$\mathrm{E}(y_{t+1}|I_t) = \beta_0 + \beta_1 \mathrm{E}(z_{t+1}|I_t).$$

This means that in order to forecast y_{t+1}, we must first forecast z_{t+1}, based on the same information set. This is usually called an **unconditional forecast** because we do not assume knowledge of z_{t+1} at time t. Unfortunately, this is somewhat of a misnomer, as our forecast is still conditional on the information in I_t. But the name is entrenched in the forecasting literature.

For forecasting, unless we are wedded to the static model in (18.42) for other reasons, it makes more sense to specify a model that depends only on lagged values of y and z. This saves us the extra step of having to forecast a right-hand side variable before forecasting y. The kind of model we have in mind is

$$\begin{aligned} y_t &= \delta_0 + \alpha_1 y_{t-1} + \gamma_1 z_{t-1} + u_t \\ \mathrm{E}(u_t|I_{t-1}) &= 0, \end{aligned} \qquad \textbf{18.43}$$

where I_{t-1} contains y and z dated at time $t-1$ and earlier. Now, the forecast of y_{t+1} at time t is $\delta_0 + \alpha_1 y_t + \gamma_1 z_t$; if we know the parameters, we can just plug in the values of y_t and z_t.

If we only want to use past y to predict future y, then we can drop z_{t-1} from (18.43). Naturally, we can add more lags of y or z and lags of other variables. Especially for forecasting one step ahead, such models can be very useful.

One-Step-Ahead Forecasting

Obtaining a forecast one period after the sample ends is relatively straightforward using models such as (18.43). As usual, let n be the sample size. The forecast of y_{n+1} is

$$\hat{f}_n = \hat{\delta}_0 + \hat{\alpha}_1 y_n + \hat{\gamma}_1 z_n, \qquad \textbf{18.44}$$

where we assume that the parameters have been estimated by OLS. We use a hat on f_n to emphasize that we have estimated the parameters in the regression model. (If we knew the parameters, there would be no estimation error in the forecast.) The forecast error—which we will not know until time $n+1$—is

$$\hat{e}_{n+1} = y_{n+1} - \hat{f}_n. \qquad \textbf{18.45}$$

If we add more lags of y or z to the forecasting equation, we simply lose more observations at the beginning of the sample.

The forecast $\hat{f}_n$ of y_{n+1} is usually called a **point forecast**. We can also obtain a **forecast interval**. A forecast interval is essentially the same as a prediction interval, which we studied in Section 6.4. There we showed how, under the classical linear model assumptions, to obtain an exact 95% prediction interval. A forecast interval is obtained in *exactly* the same way. If the model does not satisfy the classical linear model assumptions—for example, if it contains lagged dependent variables, as in (18.44)—the forecast interval is

still approximately valid, provided u_t given I_{t-1} is normally distributed with zero mean and constant variance. (This ensures that the OLS estimators are approximately normally distributed with the usual OLS variances and that u_{n+1} is independent of the OLS estimators with mean zero and variance σ^2.) Let $\text{se}(\hat{f}_n)$ be the standard error of the forecast and let $\hat{\sigma}$ be the standard error of the regression. [From Section 6.4, we can obtain $\hat{f}_n$ and $\text{se}(\hat{f}_n)$ as the intercept and its standard error from the regression of y_t on $(y_{t-1} - y_n)$ and $(z_{t-1} - z_n)$, $t = 1, 2, \ldots, n$; that is, we subtract the time n value of y from each lagged y, and similarly for z, before doing the regression.] Then,

$$\text{se}(\hat{e}_{n+1}) = \{[\text{se}(\hat{f}_n)]^2 + \hat{\sigma}^2\}^{1/2}, \tag{18.46}$$

and the (approximate) 95% forecast interval is

$$\hat{f}_n \pm 1.96 \cdot \text{se}(\hat{e}_{n+1}). \tag{18.47}$$

Because $\text{se}(\hat{f}_n)$ is roughly proportional to $1/\sqrt{n}$, $\text{se}(\hat{f}_n)$ is usually small relative to the uncertainty in the error u_{n+1}, as measured by $\hat{\sigma}$. [Some econometrics packages compute forecast intervals routinely, but others require some simple manipulations to obtain (18.47).]

Example 18.8

[Forecasting the U.S. Unemployment Rate]

We use the data in PHILLIPS.RAW, but only for the years 1948 through 1996, to forecast the U.S. civilian unemployment rate for 1997. We use two models. The first is a simple AR(1) model for *unem*:

$$\begin{aligned} \widehat{unem}_t &= 1.572 + .732\, unem_{t-1} \\ &\quad (.577)\ \ (.097) \\ n &= 48, \bar{R}^2 = .544, \hat{\sigma} = 1.049. \end{aligned} \tag{18.48}$$

In a second model, we add inflation with a lag of one year:

$$\begin{aligned} \widehat{unem}_t &= 1.304 + .647\, unem_{t-1} + .184\, inf_{t-1} \\ &\quad (.490)\ \ (.084)\qquad\quad (.041) \\ n &= 48, \bar{R}^2 = .677, \hat{\sigma} = .883. \end{aligned} \tag{18.49}$$

The lagged inflation rate is very significant in (18.49) ($t \approx 4.5$), and the adjusted R-squared from the second equation is much higher than that from the first. Nevertheless, this does *not* necessarily mean that the second equation will produce a better forecast for 1997. All we can say so far is that, using the data up through 1996, a lag of inflation helps to explain variation in the unemployment rate.

To obtain the forecasts for 1997, we need to know *unem* and *inf* in 1996. These are 5.4 and 3.0, respectively. Therefore, the forecast of $unem_{1997}$ from equation (18.48) is 1.572 + .732(5.4), or about 5.52. The forecast from equation (18.49) is 1.304 + .647(5.4) + .184(3.0), or about 5.35. The actual civilian unemployment rate for 1997 was 4.9, so both equations overpredict the actual rate. The second equation does provide a somewhat better forecast.

We can easily obtain a 95% forecast interval. When we regress $unem_t$ on $(unem_{t-1} - 5.4)$ and $(inf_{t-1} - 3.0)$, we obtain 5.35 as the intercept—which we already computed as the forecast—and $se(\hat{f}_n) = .137$. Therefore, because $\hat{\sigma} = .883$, we have $se(\hat{e}_{n+1}) = [(.137)^2 + (.883)^2]^{1/2} \approx .894$. The 95% forecast interval from (18.47) is $5.35 \pm 1.96(.894)$, or about [3.6, 7.1]. This is a wide interval, and the realized 1997 value, 4.9, is well within the interval. As expected, the standard error of u_{n+1}, which is .883, is a very large fraction of $se(\hat{e}_{n+1})$.

A professional forecaster must usually produce a forecast for every time period. For example, at time n, she or he produces a forecast of y_{n+1}. Then, when y_{n+1} and z_{n+1} become available, he or she must forecast y_{n+2}. Even if the forecaster has settled on model (18.43), there are two choices for forecasting y_{n+2}. The first is to use $\hat{\delta}_0 + \hat{\alpha}_1 y_{n+1} + \hat{\gamma}_1 z_{n+1}$, where the parameters are estimated using the first n observations. The second possibility is to *reestimate* the parameters using all $n + 1$ observations and then to use the same formula to forecast y_{n+2}. To forecast in subsequent time periods, we can generally use the parameter estimates obtained from the initial n observations, or we can update the regression parameters each time we obtain a new data point. Although the latter approach requires more computation, the extra burden is relatively minor, and it can (although it need not) work better because the regression coefficients adjust at least somewhat to the new data points.

As a specific example, suppose we wish to forecast the unemployment rate for 1998, using the model with a single lag of *unem* and *inf*. The first possibility is to just plug the 1997 values of unemployment and inflation into the right-hand side of (18.49). With $unem = 4.9$ and $inf = 2.3$ in 1997, we have a forecast for $unem_{1998}$ of about 4.9. (It is just a coincidence that this is the same as the 1997 unemployment rate.) The second possibility is to reestimate the equation by adding the 1997 observation and then using this new equation (see Computer Exercise C18.6).

The model in equation (18.43) is one equation in what is known as a **vector autoregressive (VAR) model**. We know what an autoregressive model is from Chapter 11: we model a single series, $\{y_t\}$, in terms of its own past. In vector autoregressive models, we model several series—which, if you are familiar with linear algebra, is where the word "vector" comes from—in terms of their own past. If we have two series, y_t and z_t, a vector autoregression consists of equations that look like

$$y_t = \delta_0 + \alpha_1 y_{t-1} + \gamma_1 z_{t-1} + \alpha_2 y_{t-2} + \gamma_2 z_{t-2} + \ldots \qquad \textbf{18.50}$$

and

$$z_t = \eta_0 + \beta_1 y_{t-1} + \rho_1 z_{t-1} + \beta_2 y_{t-2} + \rho_2 z_{t-2} + \ldots,$$

where each equation contains an error that has zero expected value given past information on y and z. In equation (18.43)—and in the example estimated in (18.49)—we assumed that one lag of each variable captured all of the dynamics. (An F test for joint significance of $unem_{t-2}$ and inf_{t-2} confirms that only one lag of each is needed.)

As Example 18.8 illustrates, VAR models can be useful for forecasting. In many cases, we are interested in forecasting only one variable, y, in which case we only need to estimate and analyze the equation for y. Nothing prevents us from adding other lagged variables, say, w_{t-1}, w_{t-2}, …, to equation (18.50). Such equations are efficiently estimated by OLS,

provided we have included enough lags of all variables and the equation satisfies the homoskedasticity assumption for time series regressions.

Equations such as (18.50) allow us to test whether, *after controlling for past y*, past *z* help to forecast y_t. Generally, we say that *z Granger causes y* if

$$\mathrm{E}(y_t|I_{t-1}) \neq \mathrm{E}(y_t|J_{t-1}), \tag{18.51}$$

where I_{t-1} contains past information on *y and z*, and J_{t-1} contains only information on past *y*. When (18.51) holds, past *z* is useful, *in addition to past y*, for predicting y_t. The term "causes" in "Granger causes" should be interpreted with caution. The only sense in which *z* "causes" *y* is given in (18.51). In particular, it has nothing to say about *contemporaneous* causality between *y* and *z*, so it does not allow us to determine whether z_t is an exogenous or endogenous variable in an equation relating y_t to z_t. (This is also why the notion of **Granger causality** does not apply in pure cross-sectional contexts.)

Once we assume a linear model and decide how many lags of *y* should be included in $\mathrm{E}(y_t|y_{t-1}, y_{t-2}, \ldots)$, we can easily test the null hypothesis that *z* does *not* Granger cause *y*. To be more specific, suppose that $\mathrm{E}(y_t|y_{t-1}, y_{t-2}, \ldots)$ depends on only three lags:

$$y_t = \delta_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \alpha_3 y_{t-3} + u_t$$
$$\mathrm{E}(u_t|y_{t-1}, y_{t-2}, \ldots) = 0.$$

Now, under the null hypothesis that *z* does not Granger cause *y*, *any* lags of *z* that we add to the equation should have zero population coefficients. If we add z_{t-1}, then we can simply do a *t* test on z_{t-1}. If we add two lags of *z*, then we can do an *F* test for joint significance of z_{t-1} and z_{t-2} in the equation

$$y_t = \delta_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \alpha_3 y_{t-3} + \gamma_1 z_{t-1} + \gamma_2 z_{t-2} + u_t.$$

(If there is heteroskedasticity, we can use a robust form of the test. There cannot be serial correlation under H_0 because the model is dynamically complete.)

As a practical matter, how do we decide on which lags of *y* and *z* to include? First, we start by estimating an autoregressive model for *y* and performing *t* and *F* tests to determine how many lags of *y* should appear. With annual data, the number of lags is typically small, say, one or two. With quarterly or monthly data, there are usually many more lags. Once an autoregressive model for *y* has been chosen, we can test for lags of *z*. The choice of lags of *z* is less important because, when *z* does not Granger cause *y*, no set of lagged *z*'s should be significant. With annual data, 1 or 2 lags are typically used; with quarterly data, usually 4 or 8; and with monthly data, perhaps 6, 12, or maybe even 24, given enough data.

We have already done one example of testing for Granger causality in equation (18.49). The autoregressive model that best fits unemployment is an AR(1). In equation (18.49), we added a single lag of inflation, and it was very significant. Therefore, inflation Granger causes unemployment.

There is an extended definition of Granger causality that is often useful. Let $\{w_t\}$ be a third series (or, it could represent several additional series). Then, *z Granger causes y conditional on w* if (18.51) holds, but now I_{t-1} contains past information on *y*, *z*, and *w*, while J_{t-1} contains past information on *y* and *w*. It is certainly possible that *z* Granger causes *y*, but *z* does not Granger cause *y* conditional on *w*. A test of the null that *z* does *not* Granger cause *y* conditional on *w* is obtained by testing for significance of lagged *z* in a model for

y that also depends on lagged y and lagged w. For example, to test whether growth in the money supply Granger causes growth in real GDP, conditional on the change in interest rates, we would regress $gGDP_t$ on lags of $gGDP$, Δint, and gM and do significance tests on the lags of gM. [See, for example, Stock and Watson (1989).]

Comparing One-Step-Ahead Forecasts

In almost any forecasting problem, there are several competing methods for forecasting. Even when we restrict attention to regression models, there are many possibilities. Which variables should be included, and with how many lags? Should we use logs, levels of variables, or first differences?

In order to decide on a forecasting method, we need a way to choose which one is most suitable. Broadly, we can distinguish between **in-sample criteria** and **out-of-sample criteria**. In a regression context, in-sample criteria include R-squared and especially adjusted R-squared. There are many other *model selection statistics*, but we will not cover those here [see, for example, Ramanathan (1995, Chapter 4)].

For forecasting, it is better to use out-of-sample criteria, as forecasting is essentially an out-of-sample problem. A model might provide a good fit to y in the sample used to estimate the parameters. But this need not translate to good forecasting performance. An out-of-sample comparison involves using the first part of a sample to estimate the parameters of the model and saving the latter part of the sample to gauge its forecasting capabilities. This mimics what we would have to do in practice if we did not yet know the future values of the variables.

Suppose that we have $n + m$ observations, where we use the first n observations to estimate the parameters in our model and save the last m observations for forecasting. Let $\hat{f}_{n+h}$ be the one-step-ahead forecast of y_{n+h+1} for $h = 0, 1, \ldots, m - 1$. The m forecast errors are $\hat{e}_{n+h+1} = y_{n+h+1} - \hat{f}_{n+h}$. How should we measure how well our model forecasts y when it is out of sample? Two measures are most common. The first is the **root mean squared error** (**RMSE**):

$$RMSE = \left(m^{-1} \sum_{h=0}^{m-1} \hat{e}_{n+h+1}^2 \right)^{1/2}. \qquad \textbf{18.52}$$

This is essentially the sample standard deviation of the forecast errors (without any degrees of freedom adjustment). If we compute RMSE for two or more forecasting methods, then we prefer the method with the smallest out-of-sample RMSE.

A second common measure is the **mean absolute error** (**MAE**), which is the average of the absolute forecast errors:

$$MAE = m^{-1} \sum_{h=0}^{m-1} |\hat{e}_{n+h+1}|. \qquad \textbf{18.53}$$

Again, we prefer a smaller MAE. Other possible criteria include minimizing the largest of the absolute values of the forecast errors.

Example 18.9

[Out-of-Sample Comparisons of Unemployment Forecasts]

In Example 18.8, we found that equation (18.49) fit notably better over the years 1948 through 1996 than did equation (18.48), and, at least for forecasting unemployment in 1997, the model that included lagged inflation worked better. Now, we use the two models, still estimated using the data

only through 1996, to compare one-step-ahead forecasts for 1997 through 2003. This leaves seven out-of-sample observations ($n = 48$ and $m = 7$) to use in equations (18.52) and (18.53). For the AR(1) model, RMSE = .962 and MAE = .778. For the model that adds lagged inflation (a VAR model of order one), RMSE = .673 and MAE = .628. Thus, by either measure, the model that includes inf_{t-1} produces better out-of-sample forecasts for 1997 through 2003. In this case, the in-sample and out-of-sample criteria choose the same model.

Rather than using only the first n observations to estimate the parameters of the model, we can reestimate the models each time we add a new observation and use the new model to forecast the next time period.

Multiple-Step-Ahead Forecasts

Forecasting more than one period ahead is generally more difficult than forecasting one period ahead. We can formalize this as follows. Suppose we consider forecasting y_{t+1} at time t and at an earlier time period s (so that $s < t$). Then $\text{Var}[y_{t+1} - \text{E}(y_{t+1}|I_t)] \leq \text{Var}[y_{t+1} - \text{E}(y_{t+1}|I_s)]$, where the inequality is usually strict. We will not prove this result generally, but, intuitively, it makes sense: the forecast error variance in predicting y_{t+1} is larger when we make that forecast based on less information.

If $\{y_t\}$ follows an AR(1) model (which includes a random walk, possibly with drift), we can easily show that the error variance increases with the forecast horizon. The model is

$$y_t = \alpha + \rho y_{t-1} + u_t$$
$$\text{E}(u_t|I_{t-1}) = 0, I_{t-1} = \{y_{t-1}, y_{t-2}, \ldots\},$$

and $\{u_t\}$ has constant variance σ^2 conditional on I_{t-1}. At time $t + h - 1$, our forecast of y_{t+h} is $\alpha + \rho y_{t+h-1}$, and the forecast error is simply u_{t+h}. Therefore, the one-step-ahead forecast variance is simply σ^2. To find multiple-step-ahead forecasts, we have, by repeated substitution,

$$y_{t+h} = (1 + \rho + \ldots + \rho^{h-1})\alpha + \rho^h y_t + \rho^{h-1}u_{t+1} + \rho^{h-2}u_{t+2} + \ldots + u_{t+h}.$$

At time t, the expected value of u_{t+j}, for all $j \geq 1$, is zero. So

$$\text{E}(y_{t+h}|I_t) = (1 + \rho + \ldots + \rho^{h-1})\alpha + \rho^h y_t, \qquad \textbf{18.54}$$

and the forecast error is $e_{t,h} = \rho^{h-1}u_{t+1} + \rho^{h-2}u_{t+2} + \ldots + u_{t+h}$. This is a sum of uncorrelated random variables, and so the variance of the sum is the sum of the variances: $\text{Var}(e_{t,h}) = \sigma^2[\rho^{2(h-1)} + \rho^{2(h-2)} + \ldots + \rho^2 + 1]$. Because $\rho^2 > 0$, each term multiplying σ^2 is positive, so the forecast error variance increases with h. When $\rho^2 < 1$, as h gets large the forecast variance converges to $\sigma^2/(1 - \rho^2)$, which is just the unconditional variance of y_t. In the case of a random walk ($\rho = 1$), $f_{t,h} = \alpha h + y_t$, and $\text{Var}(e_{t,h}) = \sigma^2 h$: the forecast variance grows without bound as the horizon h increases. This demonstrates that it is very difficult to forecast a random walk, with or without drift, far out into the future. For example, forecasts of interest rates farther into the future become dramatically less precise.

Equation (18.54) shows that using the AR(1) model for multistep forecasting is easy, once we have estimated ρ by OLS. The forecast of y_{n+h} at time n is

$$\hat{f}_{n,h} = (1 + \hat{\rho} + \ldots + \hat{\rho}^{h-1})\hat{\alpha} + \hat{\rho}^h y_n. \qquad \textbf{18.55}$$

Obtaining forecast intervals is harder, unless $h = 1$, because obtaining the standard error of $\hat{f}_{n,h}$ is difficult. Nevertheless, the standard error of $\hat{f}_{n,h}$ is usually small compared with the standard deviation of the error term, and the latter can be estimated as $\hat{\sigma}[\hat{\rho}^{2(h-1)} + \hat{\rho}^{2(h-2)} + \ldots + \hat{\rho}^2 + 1]^{1/2}$, where $\hat{\sigma}$ is the standard error of the regression from the AR(1) estimation. We can use this to obtain an approximate confidence interval. For example, when $h = 2$, an approximate 95% confidence interval (for large n) is

$$\hat{f}_{n,2} \pm 1.96\hat{\sigma}(1 + \hat{\rho}^2)^{1/2}. \qquad \textbf{18.56}$$

Because we are underestimating the standard deviation of y_{n+h}, this interval is too narrow, but perhaps not by much, especially if n is large.

A less traditional, but useful, approach is to estimate a different model for each forecast horizon. For example, suppose we wish to forecast y two periods ahead. If I_t depends only on y through time t, we might assume that $\mathrm{E}(y_{t+2}|I_t) = \alpha_0 + \gamma_1 y_t$ [which, as we saw earlier, holds if $\{y_t\}$ follows an AR(1) model]. We can estimate α_0 and γ_1 by regressing y_t on an intercept and on y_{t-2}. Even though the errors in this equation contain serial correlation—errors in adjacent periods are correlated—we can obtain consistent and approximately normal estimators of α_0 and γ_1. The forecast of y_{n+2} at time n is simply $\hat{f}_{n,2} = \hat{\alpha}_0 + \hat{\gamma}_1 y_n$. Further, and very importantly, the standard error of the regression is just what we need for computing a confidence interval for the forecast. Unfortunately, to get the standard error of $\hat{f}_{n,2}$, using the trick for a one-step-ahead forecast requires us to obtain a serial correlation-robust standard error of the kind described in Section 12.5. This standard error goes to zero as n gets large while the variance of the error is constant. Therefore, we can get an approximate interval by using (18.56) and by putting the SER from the regression of y_t on y_{t-2} in place of $\hat{\sigma}(1 + \hat{\rho}^2)^{1/2}$. But we should remember that this ignores the estimation error in $\hat{\alpha}_0$ and $\hat{\gamma}_1$.

We can also compute multiple-step-ahead forecasts with more complicated autoregressive models. For example, suppose $\{y_t\}$ follows an AR(2) model and that at time n, we wish to forecast y_{n+2}. Now, $y_{n+2} = \alpha + \rho_1 y_{n+1} + \rho_2 y_n + u_{n+2}$, so

$$\mathrm{E}(y_{n+2}|I_n) = \alpha + \rho_1 \mathrm{E}(y_{n+1}|I_n) + \rho_2 y_n.$$

We can write this as

$$f_{n,2} = \alpha + \rho_1 f_{n,1} + \rho_2 y_n,$$

so that the two-step-ahead forecast at time n can be obtained once we get the one-step-ahead forecast. If the parameters of the AR(2) model have been estimated by OLS, then we operationalize this as

$$\hat{f}_{n,2} = \hat{\alpha} + \hat{\rho}_1 \hat{f}_{n,1} + \hat{\rho}_2 y_n. \qquad \textbf{18.57}$$

Now, $\hat{f}_{n,1} = \hat{\alpha} + \hat{\rho}_1 y_n + \hat{\rho}_2 y_{n-1}$, which we can compute at time n. Then, we plug this into (18.57), along with y_n, to obtain $\hat{f}_{n,2}$. For any $h > 2$, obtaining any h-step-ahead forecast for an AR(2) model is easy to find in a recursive manner: $\hat{f}_{n,h} = \hat{\alpha} + \hat{\rho}_1 \hat{f}_{n,h-1} + \hat{\rho}_2 \hat{f}_{n,h-2}$.

Similar reasoning can be used to obtain multiple-step-ahead forecasts for VAR models. To illustrate, suppose we have

$$y_t = \delta_0 + \alpha_1 y_{t-1} + \gamma_1 z_{t-1} + u_t \qquad \textbf{18.58}$$

and

$$z_t = \eta_0 + \beta_1 y_{t-1} + \rho_1 z_{t-1} + v_t.$$

Now, if we wish to forecast y_{n+1} at time n, we simply use $\hat{f}_{n,1} = \hat{\delta}_0 + \hat{\alpha}_1 y_n + \hat{\gamma}_1 z_n$. Likewise, the forecast of z_{n+1} at time n is (say) $\hat{g}_{n,1} = \hat{\eta}_0 + \hat{\beta}_1 y_n + \hat{\rho}_1 z_n$. Now, suppose we wish to obtain a two-step-ahead forecast of y at time n. From (18.58), we have

$$\mathrm{E}(y_{n+2}|I_n) = \delta_0 + \alpha_1 \mathrm{E}(y_{n+1}|I_n) + \gamma_1 \mathrm{E}(z_{n+1}|I_n)$$

[because $\mathrm{E}(u_{n+2}|I_n) = 0$], so we can write the forecast as

$$\hat{f}_{n,2} = \hat{\delta}_0 + \hat{\alpha}_1 \hat{f}_{n,1} + \hat{\gamma}_1 \hat{g}_{n,1}. \qquad \textbf{18.59}$$

This equation shows that the two-step-ahead forecast for y depends on the one-step-ahead forecasts for y *and* z. Generally, we can build up multiple-step-ahead forecasts of y by using the recursive formula

$$\hat{f}_{n,h} = \hat{\delta}_0 + \hat{\alpha}_1 \hat{f}_{n,h-1} + \hat{\gamma}_1 \hat{g}_{n,h-1}, \quad h \geq 2.$$

Example 18.10

[Two-Year-Ahead Forecast for the Unemployment Rate]

To use equation (18.49) to forecast unemployment two years out—say, the 1998 rate using the data through 1996—we need a model for inflation. The best model for *inf* in terms of lagged *unem* and *inf* appears to be a simple AR(1) model ($unem_{-1}$ is not significant when added to the regression):

$$\widehat{inf}_t = \underset{(.558)}{1.277} + \underset{(.107)}{.665}\, inf_{t-1}$$
$$n = 48, R^2 = .457, \bar{R}^2 = .445.$$

If we plug the 1996 value of *inf* into this equation, we get the forecast of *inf* for 1997: $\widehat{inf}_{1997} = 3.27$. Now, we can plug this, along with $\widehat{unem}_{1997} = 5.35$ (which we obtained earlier), into (18.59) to forecast $unem_{1998}$:

$$\widehat{unem}_{1998} = 1.304 + .647(5.35) + .184(3.27) \approx 5.37.$$

Remember, this forecast uses information only through 1996. The one-step-ahead forecast of $unem_{1998}$, obtained by plugging the 1997 values of *unem* and *inf* into (18.48), was about 4.90. The actual unemployment rate in 1998 was 4.5%, which means that, in this case, the one-step-ahead forecast does quite a bit better than the two-step-ahead forecast.

Just as with one-step-ahead forecasting, an out-of-sample root mean squared error or a mean absolute error can be used to choose among multiple-step-ahead forecasting methods.

Forecasting Trending, Seasonal, and Integrated Processes

We now turn to forecasting series that either exhibit trends, have seasonality, or have unit roots. Recall from Chapters 10 and 11 that one approach to handling trending dependent or independent variables in regression models is to include time trends, the most popular being a linear trend. Trends can be included in forecasting equations as well, although they must be used with caution.

In the simplest case, suppose that $\{y_t\}$ has a linear trend but is unpredictable around that trend. Then, we can write

$$y_t = \alpha + \beta t + u_t, \mathrm{E}(u_t|I_{t-1}) = 0, t = 1, 2, \ldots, \qquad \textbf{18.60}$$

where, as usual, I_{t-1} contains information observed through time $t - 1$ (which includes at least past y). How do we forecast y_{n+h} at time n for any $h \geq 1$? This is simple because $\mathrm{E}(y_{n+h}|I_n) = \alpha + \beta(n + h)$. The forecast error variance is simply $\sigma^2 = \mathrm{Var}(u_t)$ (assuming a constant variance over time). If we estimate α and β by OLS using the first n observations, then our forecast for y_{n+h} at time n is $\hat{f}_{n,h} = \hat{\alpha} + \hat{\beta}(n + h)$. In other words, we simply plug the time period corresponding to y into the estimated trend function. For example, if we use the $n = 131$ observations in BARIUM.RAW to forecast monthly Chinese imports of barium chloride to the United States, we obtain $\hat{\alpha} = 249.56$ and $\hat{\beta} = 5.15$. The sample period ends in December 1988, so the forecast of Chinese imports six months later is $249.56 + 5.15(137) = 955.11$, measured as short tons. For comparison, the December 1988 value is 1,087.81, so it is greater than the forecasted value six months later. The series and its estimated trend line are shown in Figure 18.2.

As we discussed in Chapter 10, most economic time series are better characterized as having, at least approximately, a constant growth rate, which suggests that $\log(y_t)$ follows a linear time trend. Suppose we use n observations to obtain the equation

$$\widehat{\log(y_t)} = \hat{\alpha} + \hat{\beta}t, t = 1, 2, \ldots, n. \qquad \textbf{18.61}$$

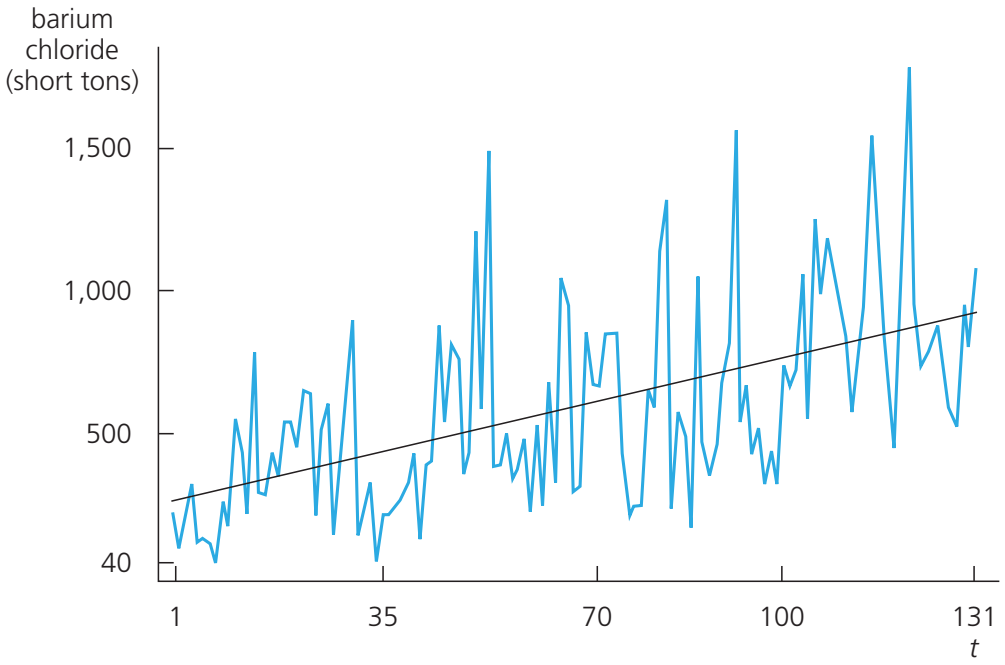
Then, to forecast $\log(y)$ at any future time period $n + h$, we just plug $n + h$ into the trend equation, as before. But this does not allow us to forecast y, which is usually what we want. It is tempting to simply exponentiate $\hat{\alpha} + \hat{\beta}(n + h)$ to obtain the forecast for y_{n+h}, but this is not quite right, for the same reasons we gave in Section 6.4. We must properly account for the error implicit in (18.61). The simplest way to do this is to use the n

Question 18.5

Suppose you model $\{y_t: t = 1, 2, \ldots, 46\}$ as a linear time trend, where data are annual starting in 1950 and ending in 1995. Define the variable $year_t$ as ranging from 50 when $t = 1$ to 95 when $t = 46$. If you estimate the equation $\hat{y}_t = \hat{\gamma} + \hat{\delta} year_t$, how do $\hat{\gamma}$ and $\hat{\delta}$ compare with $\hat{\alpha}$ and $\hat{\beta}$ in $\hat{y}_t = \hat{\alpha} + \hat{\beta}t$? How will forecasts from the two equations compare?

FIGURE 18.2

Chinese barium chloride imports into the United States (in short tons) and its estimated linear trend line, 249.56 + 5.15*t*.



observations to regress y_t on $\exp(\widehat{logy}_t)$ *without* an intercept. Let $\hat{\gamma}$ be the slope coefficient on $\exp(\widehat{logy}_t)$. Then, the forecast of y in period $n + h$ is simply

$$\hat{f}_{n,h} = \hat{\gamma}\exp[\hat{\alpha} + \hat{\beta}(n + h)]. \qquad \textbf{18.62}$$

As an example, if we use the first 687 weeks of data on the New York Stock Exchange index in NYSE.RAW, we obtain $\hat{\alpha} = 3.782$ and $\hat{\beta} = .0019$ [by regressing $\log(price_t)$ on a linear time trend]; this shows that the index grows about .2% per week, on average. When we regress *price* on the exponentiated fitted values, we obtain $\hat{\gamma} = 1.018$. Now, we forecast *price* four weeks out, which is the last week in the sample, using (18.62): $1.018 \cdot \exp[3.782 + .0019(691)] \approx 166.12$. The actual value turned out to be 164.25, so we have somewhat overpredicted. But this result is much better than if we estimate a linear time trend for the first 687 weeks: the forecasted value for week 691 is 152.23, which is a substantial underprediction.

Although trend models can be useful for prediction, they must be used with caution, especially for forecasting far into the future integrated series that have drift. The potential problem can be seen by considering a random walk with drift. At time $t + h$, we can write y_{t+h} as

$$y_{t+h} = \beta h + y_t + u_{t+1} + \ldots + u_{t+h},$$

where β is the drift term (usually $\beta > 0$), and each u_{t+j} has zero mean given I_t and constant variance σ^2. As we saw earlier, the forecast of y_{t+h} at time t is $\mathrm{E}(y_{t+h}|I_t) = \beta h + y_t$, and the forecast error variance is $\sigma^2 h$. What happens if we use a linear trend model? Let y_0 be the initial value of the process at time zero, which we take as nonrandom. Then, we can also write

$$\begin{aligned} y_{t+h} &= y_0 + \beta(t + h) + u_1 + u_2 + \ldots + u_{t+h} \\ &= y_0 + \beta(t + h) + v_{t+h}. \end{aligned}$$

This looks like a linear trend model with the intercept $\alpha = y_0$. But the error, v_{t+h}, while having mean zero, has variance $\sigma^2(t + h)$. Therefore, if we use the linear trend $y_0 + \beta(t + h)$ to forecast y_{t+h} at time t, the forecast error variance is $\sigma^2(t + h)$, compared with $\sigma^2 h$ when we use $\beta h + y_t$. The ratio of the forecast variances is $(t + h)/h$, which can be big for large t. The bottom line is that we should not use a linear trend to forecast a random walk with drift. (Computer Exercise C18.8 asks you to compare forecasts from a cubic trend line and those from the simple random walk model for the general fertility rate in the United States.)

Deterministic trends can also produce poor forecasts if the trend parameters are estimated using old data and the process has a subsequent shift in the trend line. Sometimes, exogenous shocks—such as the oil crises of the 1970s—can change the trajectory of trending variables. If an old trend line is used to forecast far into the future, the forecasts can be way off. This problem can be mitigated by using the most recent data available to obtain the trend line parameters.

Nothing prevents us from combining trends with other models for forecasting. For example, we can add a linear trend to an AR(1) model, which can work well for forecasting series with linear trends but which are also stable AR processes around the trend.

It is also straightforward to forecast processes with deterministic seasonality (monthly or quarterly series). For example, the file BARIUM.RAW contains the monthly production of gasoline in the United States from 1978 through 1988. This series has no obvious trend, but it does have a strong seasonal pattern. (Gasoline production is higher in the summer months and in December.) In the simplest model, we would regress *gas* (measured in gallons) on 11 month dummies, say, for February through December. Then, the forecast for any future month is simply the intercept plus the coefficient on the appropriate month dummy. (For January, the forecast is just the intercept in the regression.) We can also add lags of variables and time trends to allow for general series with seasonality.

Forecasting processes with unit roots also deserves special attention. Earlier, we obtained the expected value of a random walk conditional on information through time n. To forecast a random walk, with possible drift α, h periods into the future at time n, we use $\hat{f}_{n,h} = \hat{\alpha}h + y_n$, where $\hat{\alpha}$ is the sample average of the Δy_t up through $t = n$. (If there is no drift, we set $\hat{\alpha} = 0$.) This approach imposes the unit root. An alternative would be to estimate an AR(1) model for $\{y_t\}$ and to use the forecast formula (18.55). This approach does not impose a unit root, but if one is present, $\hat{\rho}$ converges in probability to one as n gets large. Nevertheless, $\hat{\rho}$ can be substantially different than one, especially if the sample size is not very large. The matter of which approach produces better out-of-sample forecasts is an empirical issue. If in the AR(1) model, ρ is less than one, even slightly, the AR(1) model will tend to produce better long-run forecasts.

Generally, there are two approaches to producing forecasts for I(1) processes. The first is to impose a unit root. For a one-step-ahead forecast, we obtain a model to forecast the change in y, Δy_{t+1}, given information through time t. Then, because $y_{t+1} = \Delta y_{t+1} + y_t$, $\mathrm{E}(y_{t+1}|I_t) = \mathrm{E}(\Delta y_{t+1}|I_t) + y_t$. Therefore, our forecast of y_{n+1} at time n is just

$$\hat{f}_n = \hat{g}_n + y_n,$$

where $\hat{g}_n$ is the forecast of Δy_{n+1} at time n. Typically, an AR model (which is necessarily stable) is used for Δy_t, or a vector autoregression.

This can be extended to multiple-step-ahead forecasts by writing y_{n+h} as

$$y_{n+h} = (y_{n+h} - y_{n+h-1}) + (y_{n+h-1} - y_{n+h-2}) + \ldots + (y_{n+1} - y_n) + y_n,$$

or

$$y_{n+h} = \Delta y_{n+h} + \Delta y_{n+h-1} + \ldots + \Delta y_{n+1} + y_n.$$

Therefore, the forecast of y_{n+h} at time n is

$$\hat{f}_{n,h} = \hat{g}_{n,h} + \hat{g}_{n,h-1} + \ldots + \hat{g}_{n,1} + y_n, \qquad \textbf{18.63}$$

where $\hat{g}_{n,j}$ is the forecast of Δy_{n+j} at time n. For example, we might model Δy_t as a stable AR(1), obtain the multiple-step-ahead forecasts from (18.55) (but with $\hat{\alpha}$ and $\hat{\rho}$ obtained from Δy_t on Δy_{t-1}, and y_n replaced with Δy_n), and then plug these into (18.63).

The second approach to forecasting I(1) variables is to use a general AR or VAR model for $\{y_t\}$. This does not impose the unit root. For example, if we use an AR(2) model,

$$y_t = \alpha + \rho_1 y_{t-1} + \rho_2 y_{t-2} + u_t, \qquad \textbf{18.64}$$

then $\rho_1 + \rho_2 = 1$. If we plug in $\rho_1 = 1 - \rho_2$ and rearrange, we obtain $\Delta y_t = \alpha - \rho_2 \Delta y_{t-1} + u_t$, which is a stable AR(1) model in the difference that takes us back to the first approach described earlier. Nothing prevents us from estimating (18.64) directly by OLS. One nice thing about this regression is that we *can* use the usual t statistic on $\hat{\rho}_2$ to determine if y_{t-2} is significant. (This assumes that the homoskedasticity assumption holds; if not, we can use the heteroskedasticity-robust form.) We will not show this formally, but, intuitively, it follows by rewriting the equation as $y_t = \alpha + \gamma y_{t-1} - \rho_2 \Delta y_{t-1} + u_t$, where $\gamma = \rho_1 + \rho_2$. Even if $\gamma = 1$, ρ_2 is minus the coefficient on a stationary, weakly dependent process $\{\Delta y_{t-1}\}$. Because the regression results will be identical to (18.64), we can use (18.64) directly.

As an example, let us estimate an AR(2) model for the general fertility rate in FERTIL3.RAW, using the observations through 1979. (In Computer Exercise C18.8, you are asked to use this model for forecasting, which is why we save some observations at the end of the sample.)

$$\begin{aligned} \widehat{gfr}_t &= 3.22 + 1.272\, gfr_{t-1} - .311\, gfr_{t-2} \\ &\quad (2.92) \quad (.120) \qquad (.121) \\ n &= 65,\ R^2 = .949,\ \bar{R}^2 = .947. \end{aligned} \qquad \textbf{18.65}$$

The t statistic on the second lag is about -2.57, which is statistically different from zero at about the 1% level. (The first lag also has a very significant t statistic, which has an approximate t distribution by the same reasoning used for $\hat{\rho}_2$.) The R-squared, adjusted or not, is not especially informative as a goodness-of-fit measure because *gfr* apparently contains a unit root, and it makes little sense to ask how much of the variance in *gfr* we are explaining.

The coefficients on the two lags in (18.65) add up to .961, which is close to and not statistically different from one (as can be verified by applying the augmented Dickey-Fuller test to the equation $\Delta gfr_t = \alpha + \theta gfr_{t-1} + \delta_1 \Delta gfr_{t-1} + u_t$). Even though we have not imposed the unit root restriction, we can still use (18.65) for forecasting, as we discussed earlier.

Before ending this section, we point out one potential improvement in forecasting in the context of vector autoregressive models with I(1) variables. Suppose $\{y_t\}$ and $\{z_t\}$ are each I(1) processes. One approach for obtaining forecasts of y is to estimate a bivariate autoregression in the variables Δy_t and Δz_t and then to use (18.63) to generate one- or multiple-step-ahead forecasts; this is essentially the first approach we described earlier. However, if y_t and z_t are *cointegrated*, we have more stationary, stable variables in the information set that can be used in forecasting Δy: namely, lags of $y_t - \beta z_t$, where β is the cointegrating parameter. A simple error correction model is

$$\Delta y_t = \alpha_0 + \alpha_1 \Delta y_{t-1} + \gamma_1 \Delta z_{t-1} + \delta_1 (y_{t-1} - \beta z_{t-1}) + e_t, \quad \mathrm{E}(e_t | I_{t-1}) = 0. \tag{18.66}$$

To forecast y_{n+1}, we use observations up through n to estimate the cointegrating parameter, β, and then estimate the parameters of the error correction model by OLS, as described in Section 18.4. Forecasting Δy_{n+1} is easy: we just plug Δy_n, Δz_n, and $y_n - \hat{\beta} z_n$ into the estimated equation. Having obtained the forecast of Δy_{n+1}, we add it to y_n.

By rearranging the error correction model, we can write

$$y_t = \alpha_0 + \rho_1 y_{t-1} + \rho_2 y_{t-2} + \delta_1 z_{t-1} + \delta_2 z_{t-2} + u_t, \tag{18.67}$$

where $\rho_1 = 1 + \alpha_1 + \delta$, $\rho_2 = -\alpha_1$, and so on, which is the first equation in a VAR model for y_t and z_t. Notice that this depends on five parameters, just as many as in the error correction model. The point is that, for the purposes of forecasting, the VAR model in the levels and the error correction model are essentially the same. This is not the case in more general error correction models. For example, suppose that $\alpha_1 = \gamma_1 = 0$ in (18.66), but we have a second error correction term, $\delta_2 (y_{t-2} - \beta z_{t-2})$. Then, the error correction model involves only four parameters, whereas (18.67)—which has the same order of lags for y and z—contains five parameters. Thus, error correction models can economize on parameters; that is, they are generally more *parsimonious* than VARs in levels.

If y_t and z_t are I(1) but not cointegrated, the appropriate model is (18.66) without the error correction term. This can be used to forecast Δy_{n+1}, and we can add this to y_n to forecast y_{n+1}.

SUMMARY

The time series topics covered in this chapter are used routinely in empirical macroeconomics, empirical finance, and a variety of other applied fields. We began by showing how infinite distributed lag models can be interpreted and estimated. These can provide flexible lag distributions with fewer parameters than a similar finite distributed lag model. The geometric distributed lag and, more generally, rational distributed lag models are the most popular. They can be estimated using standard econometric procedures on simple dynamic equations.

Testing for a unit root has become very common in time series econometrics. If a series has a unit root, then, in many cases, the usual large sample normal approximations are no longer valid. In addition, a unit root process has the property that an innovation has a long-lasting effect, which is of interest in its own right. While there are many tests for unit roots, the Dickey-Fuller t test—and its extension, the augmented Dickey-Fuller test—is probably the most popular and easiest to implement. We can allow for a linear trend when testing for unit roots by adding a trend to the Dickey-Fuller regression.

When an I(1) series, y_t, is regressed on another I(1) series, x_t, there is serious concern about spurious regression, even if the series do not contain obvious trends. This has been studied thoroughly in the case of a random walk: even if the two random walks are independent, the usual t test for significance of the slope coefficient, based on the usual critical values, will reject much more than the nominal size of the test. In addition, the R^2 tends to a random variable, rather than to zero (as would be the case if we regress the difference in y_t on the difference in x_t).

In one important case, a regression involving I(1) variables is not spurious, and that is when the series are cointegrated. This means that a linear function of the two I(1) variables is I(0). If y_t and x_t are I(1) but $y_t - x_t$ is I(0), y_t and x_t cannot drift arbitrarily far apart. There are simple tests of the null of no cointegration against the alternative of cointegration, one of which is based on applying a Dickey-Fuller unit root test to the residuals from a static regression. There are also simple estimators of the cointegrating parameter that yield t statistics with approximate standard normal distributions (and asymptotically valid confidence intervals). We covered the leads and lags estimator in Section 18.4.

Cointegration between y_t and x_t implies that error correction terms may appear in a model relating Δy_t to Δx_t; the error correction terms are lags in $y_t - \beta x_t$, where β is the cointegrating parameter. A simple two-step estimation procedure is available for estimating error correction models. First, β is estimated using a static regression (or the leads and lags regression). Then, OLS is used to estimate a simple dynamic model in first differences that includes the error correction terms.

Section 18.5 contained an introduction to forecasting, with emphasis on regression-based forecasting methods. Static models or, more generally, models that contain explanatory variables dated contemporaneously with the dependent variable, are limited because then the explanatory variables need to be forecasted. If we plug in hypothesized values of unknown future explanatory variables, we obtain a conditional forecast. Unconditional forecasts are similar to simply modeling y_t as a function of *past* information we have observed at the time the forecast is needed. Dynamic regression models, including autoregressions and vector autoregressions, are used routinely. In addition to obtaining one-step-ahead point forecasts, we also discussed the construction of forecast intervals, which are very similar to prediction intervals.

Various criteria are used for choosing among forecasting methods. The most common performance measures are the root mean squared error and the mean absolute error. Both estimate

the size of the average forecast error. It is most informative to compute these measures using out-of-sample forecasts.

Multiple-step-ahead forecasts present new challenges and are subject to large forecast error variances. Nevertheless, for models such as autoregressions and vector autoregressions, multi-step-ahead forecasts can be computed, and approximate forecast intervals can be obtained.

Forecasting trending and I(1) series requires special care. Processes with deterministic trends can be forecasted by including time trends in regression models, possibly with lags of variables. A potential drawback is that deterministic trends can provide poor forecasts for long-horizon forecasts: once it is estimated, a linear trend continues to increase or decrease. The typical approach to forecasting an I(1) process is to forecast the difference in the process and to add the level of the variable to that forecasted difference. Alternatively, vector autoregressive models can be used in the levels of the series. If the series are cointegrated, error correction models can be used instead.

KEY TERMS

Augmented Dickey-Fuller Test
Cointegration
Conditional Forecast
Dickey-Fuller Distribution
Dickey-Fuller (DF) Test
Engle-Granger Test
Engle-Granger Two-Step Procedure
Error Correction Model
Exponential Smoothing
Forecast Error
Forecast Interval
Geometric (or Koyck) Distributed Lag
Granger Causality
Infinite Distributed Lag (IDL) Model
Information Set
In-Sample Criteria
Leads and Lags Estimator
Loss Function
Martingale
Martingale Difference Sequence
Mean Absolute Error (MAE)
Multiple-Step-Ahead Forecast
One-Step-Ahead Forecast
Out-of-Sample Criteria
Point Forecast
Rational Distributed Lag (RDL) Model
Root Mean Squared Error (RMSE)
Spurious Regression Problem
Unconditional Forecast
Unit Roots
Vector Autoregressive (VAR) Model

PROBLEMS

18.1 Consider equation (18.15) with $k = 2$. Using the IV approach to estimating the γ_h and ρ, what would you use as instruments for y_{t-1}?

18.2 An interesting economic model that leads to an econometric model with a lagged dependent variable relates y_t to the *expected value* of x_t, say, x_t^*, where the expectation is based on all observed information at time $t - 1$:

$$y_t = \alpha_0 + \alpha_1 x_t^* + u_t. \qquad \textbf{18.68}$$

A natural assumption on $\{u_t\}$ is that $E(u_t|I_{t-1}) = 0$, where I_{t-1} denotes all information on y and x observed at time $t - 1$; this means that $E(y_t|I_{t-1}) = \alpha_0 + \alpha_1 x_t^*$. To complete this model, we need an assumption about how the expectation x_t^* is formed. We saw a simple

example of adaptive expectations in Section 11.2, where $x_t^* = x_{t-1}$. A more complicated adaptive expectations scheme is

$$x_t^* - x_{t-1}^* = \lambda(x_{t-1} - x_{t-1}^*), \tag{18.69}$$

where $0 < \lambda < 1$. This equation implies that the change in expectations reacts to whether last period's realized value was above or below its expectation. The assumption $0 < \lambda < 1$ implies that the change in expectations is a fraction of last period's error.

(i) Show that the two equations imply that

$$y_t = \lambda\alpha_0 + (1 - \lambda)y_{t-1} + \lambda\alpha_1 x_{t-1} + u_t - (1 - \lambda)u_{t-1}.$$

[*Hint:* Lag equation (18.68) one period, multiply it by $(1 - \lambda)$, and subtract this from (18.68). Then, use (18.69).]

(ii) Under $\mathrm{E}(u_t|I_{t-1}) = 0$, $\{u_t\}$ is serially uncorrelated. What does this imply about the new errors, $v_t = u_t - (1 - \lambda)u_{t-1}$?

(iii) If we write the equation from part (i) as

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 x_{t-1} + v_t,$$

how would you consistently estimate the β_j?

(iv) Given consistent estimators of the β_j, how would you consistently estimate λ and α_1?

18.3 Suppose that $\{y_t\}$ and $\{z_t\}$ are I(1) series, but $y_t - \beta z_t$ is I(0) for some $\beta \neq 0$. Show that for any $\delta \neq \beta$, $y_t - \delta z_t$ must be I(1).

18.4 Consider the error correction model in equation (18.37). Show that if you add another lag of the error correction term, $y_{t-2} - \beta x_{t-2}$, the equation suffers from perfect collinearity. (*Hint:* Show that $y_{t-2} - \beta x_{t-2}$ is a perfect linear function of $y_{t-1} - \beta x_{t-1}$, Δx_{t-1}, and Δy_{t-1}.)

18.5 Suppose the process $\{(x_t, y_t): t = 0, 1, 2, \ldots\}$ satisfies the equations

$$y_t = \beta x_t + u_t$$

and

$$\Delta x_t = \gamma \Delta x_{t-1} + v_t,$$

where $\mathrm{E}(u_t|I_{t-1}) = \mathrm{E}(v_t|I_{t-1}) = 0$, I_{t-1} contains information on x and y dated at time $t - 1$ and earlier, $\beta \neq 0$, and $|\gamma| < 1$ [so that x_t, and therefore y_t, is I(1)]. Show that these two equations imply an error correction model of the form

$$\Delta y_t = \gamma_1 \Delta x_{t-1} + \delta(y_{t-1} - \beta x_{t-1}) + e_t,$$

where $\gamma_1 = \beta\gamma$, $\delta = -1$, and $e_t = u_t + \beta v_t$. (*Hint:* First subtract y_{t-1} from both sides of the first equation. Then, add and subtract βx_{t-1} from the right-hand side and rearrange. Finally, use the second equation to get the error correction model that contains Δx_{t-1}.)

18.6 Using the monthly data in VOLAT.RAW, the following model was estimated:

$$\widehat{pcip} = \underset{(.56)}{1.54} + \underset{(.042)}{.344}\, pcip_{-1} + \underset{(.045)}{.074}\, pcip_{-2} + \underset{(.042)}{.073}\, pcip_{-3} + \underset{(.013)}{.031}\, pcsp_{-1}$$

$$n = 554,\ R^2 = .174,\ \bar{R}^2 = .168,$$

where *pcip* is the percentage change in monthly industrial production, at an annualized rate, and *pcsp* is the percentage change in the Standard & Poor's 500 Index, also at an annualized rate.

(i) If the past three months of *pcip* are zero and $pcsp_{-1} = 0$, what is the predicted growth in industrial production for this month? Is it statistically different from zero?

(ii) If the past three months of *pcip* are zero but $pcsp_{-1} = 10$, what is the predicted growth in industrial production?

(iii) What do you conclude about the effects of the stock market on real economic activity?

18.7 Let gM_t be the annual growth in the money supply and let $unem_t$ be the unemployment rate. Assuming that $unem_t$ follows a stable AR(1) process, explain in detail how you would test whether *gM* Granger causes *unem*.

18.8 Suppose that y_t follows the model

$$y_t = \alpha + \delta_1 z_{t-1} + u_t$$
$$u_t = \rho u_{t-1} + e_t$$
$$\mathrm{E}(e_t|I_{t-1}) = 0,$$

where I_{t-1} contains y and z dated at $t-1$ and earlier.

(i) Show that $\mathrm{E}(y_{t+1}|I_t) = (1-\rho)\alpha + \rho y_t + \delta_1 z_t - \rho\delta_1 z_{t-1}$. (*Hint*: Write $u_{t-1} = y_{t-1} - \alpha - \delta_1 z_{t-2}$ and plug this into the second equation; then, plug the result into the first equation and take the conditional expectation.)

(ii) Suppose that you use n observations to estimate α, δ_1, and ρ. Write the equation for forecasting y_{n+1}.

(iii) Explain why the model with one lag of z and AR(1) serial correlation is a special case of the model

$$y_t = \alpha_0 + \rho y_{t-1} + \gamma_1 z_{t-1} + \gamma_2 z_{t-2} + e_t.$$

(iv) What does part (iii) suggest about using models with AR(1) serial correlation for forecasting?

18.9 Let $\{y_t\}$ be an I(1) sequence. Suppose that $\hat{g}_n$ is the one-step-ahead forecast of Δy_{n+1} and let $\hat{f}_n = \hat{g}_n + y_n$ be the one-step-ahead forecast of y_{n+1}. Explain why the forecast errors for forecasting Δy_{n+1} and y_{n+1} are identical.

COMPUTER EXERCISES

C18.1 Use the data in WAGEPRC.RAW for this exercise. Problem 11.5 gave estimates of a finite distributed lag model of *gprice* on *gwage*, where 12 lags of *gwage* are used.

(i) Estimate a simple geometric DL model of *gprice* on *gwage*. In particular, estimate equation (18.11) by OLS. What are the estimated impact propensity and LRP? Sketch the estimated lag distribution.

(ii) Compare the estimated IP and LRP to those obtained in Problem 11.5. How do the estimated lag distributions compare?

(iii) Now, estimate the rational distributed lag model from (18.16). Sketch the lag distribution and compare the estimated IP and LRP to those obtained in part (ii).

C18.2 Use the data in HSEINV.RAW for this exercise.

(i) Test for a unit root in log(*invpc*), including a linear time trend and two lags of $\Delta\log(invpc_t)$. Use a 5% significance level.

(ii) Use the approach from part (i) to test for a unit root in log(*price*).

(iii) Given the outcomes in parts (i) and (ii), does it make sense to test for cointegration between log(*invpc*) and log(*price*)?

C18.3 Use the data in VOLAT.RAW for this exercise.

(i) Estimate an AR(3) model for *pcip*. Now, add a fourth lag and verify that it is very insignificant.

(ii) To the AR(3) model from part (i), add three lags of *pcsp* to test whether *pcsp* Granger causes *pcip*. Carefully, state your conclusion.

(iii) To the model in part (ii), add three lags of the change in *i3*, the three-month T-bill rate. Does *pcsp* Granger cause *pcip* conditional on past $\Delta i3$?

C18.4 In testing for cointegration between *gfr* and *pe* in Example 18.5, add t^2 to equation (18.32) to obtain the OLS residuals. Include one lag in the augmented DF test. The 5% critical value for the test is -4.15.

C18.5 Use INTQRT.RAW for this exercise.

(i) In Example 18.7, we estimated an error correction model for the holding yield on six-month T-bills, where one lag of the holding yield on three-month T-bills is the explanatory variable. We assumed that the cointegration parameter was one in the equation $hy6_t = \alpha + \beta hy3_{t-1} + u_t$. Now, add the lead change, $\Delta hy3_t$, the contemporaneous change, $\Delta hy3_{t-1}$, and the lagged change, $\Delta hy3_{t-2}$, of $hy3_{t-1}$. That is, estimate the equation

$$hy6_t = \alpha + \beta hy3_{t-1} + \phi_0 \Delta hy3_t + \phi_1 \Delta hy3_{t-1} + \rho_1 \Delta hy3_{t-2} + e_t$$

and report the results in equation form. Test H_0: $\beta = 1$ against a two-sided alternative. Assume that the lead and lag are sufficient so that $\{hy3_{t-1}\}$ is strictly exogenous in this equation and do not worry about serial correlation.

(ii) To the error correction model in (18.39), add $\Delta hy3_{t-2}$ and $(hy6_{t-2} - hy3_{t-3})$. Are these terms jointly significant? What do you conclude about the appropriate error correction model?

C18.6 Use the data in PHILLIPS.RAW to answer these questions.

(i) Estimate the models in (18.48) and (18.49) using the data through 1997. Do the parameter estimates change much compared with (18.48) and (18.49)?

(ii) Use the new equations to forecast $unem_{1998}$; round to two places after the decimal. Which equation produces a better forecast?

(iii) As we discussed in the text, the forecast for $unem_{1998}$ using (18.49) is 4.90. Compare this with the forecast obtained using the data through 1997. Does using the extra year of data to obtain the parameter estimates produce a better forecast?

(iv) Use the model estimated in (18.48) to obtain a two-step-ahead forecast of *unem*. That is, forecast $unem_{1998}$ using equation (18.55) with $\hat{\alpha} = 1.572$, $\hat{\rho} = .732$, and

$h = 2$. Is this better or worse than the one-step-ahead forecast obtained by plugging $unem_{1997} = 4.9$ into (18.48)?

C18.7 Use the data in BARIUM.RAW for this exercise.

(i) Estimate the linear trend model $chnimp_t = \alpha + \beta t + u_t$, using the first 119 observations (this excludes the last 12 months of observations for 1988). What is the standard error of the regression?

(ii) Now, estimate an AR(1) model for *chnimp*, again using all data but the last 12 months. Compare the standard error of the regression with that from part (i). Which model provides a better in-sample fit?

(iii) Use the models from parts (i) and (ii) to compute the one-step-ahead forecast errors for the 12 months in 1988. (You should obtain 12 forecast errors for each method.) Compute and compare the RMSEs and the MAEs for the two methods. Which forecasting method works better out-of-sample for one-step-ahead forecasts?

(iv) Add monthly dummy variables to the regression from part (i). Are these jointly significant? (Do not worry about the slight serial correlation in the errors from this regression when doing the joint test.)

C18.8 Use the data in FERTIL3.RAW for this exercise.

(i) Graph *gfr* against time. Does it contain a clear upward or downward trend over the entire sample period?

(ii) Using the data through 1979, estimate a cubic time trend model for *gfr* (that is, regress *gfr* on t, t^2, and t^3, along with an intercept). Comment on the R-squared of the regression.

(iii) Using the model in part (ii), compute the mean absolute error of the one-step-ahead forecast errors for the years 1980 through 1984.

(iv) Using the data through 1979, regress Δgfr_t on a constant only. Is the constant statistically different from zero? Does it make sense to assume that any drift term is zero, if we assume that gfr_t follows a random walk?

(v) Now, forecast *gfr* for 1980 through 1984, using a random walk model: the forecast of gfr_{n+1} is simply gfr_n. Find the MAE. How does it compare with the MAE from part (iii)? Which method of forecasting do you prefer?

(vi) Now, estimate an AR(2) model for *gfr*, again using the data only through 1979. Is the second lag significant?

(vii) Obtain the MAE for 1980 through 1984, using the AR(2) model. Does this more general model work better out-of-sample than the random walk model?

C18.9 Use CONSUMP.RAW for this exercise.

(i) Let y_t be real per capita disposable income. Use the data through 1989 to estimate the model

$$y_t = \alpha + \beta t + \rho y_{t-1} + u_t$$

and report the results in the usual form.

(ii) Use the estimated equation from part (i) to forecast y in 1990. What is the forecast error?

(iii) Compute the mean absolute error of the one-step-ahead forecasts for the 1990s, using the parameters estimated in part (i).

(iv) Now, compute the MAE over the same period, but drop y_{t-1} from the equation. Is it better to include y_{t-1} in the model or not?

C18.10 Use the data in INTQRT.RAW for this exercise.

(i) Using the data from all but the last four years (16 quarters), estimate an AR(1) model for $\Delta r6_t$. (We use the difference because it appears that $r6_t$ has a unit root.) Find the RMSE of the one-step-ahead forecasts for $\Delta r6$, using the last 16 quarters.

(ii) Now, add the error correction term $spr_{t-1} = r6_{t-1} - r3_{t-1}$ to the equation from part (i). (This assumes that the cointegrating parameter is one.) Compute the RMSE for the last 16 quarters. Does the error correction term help with out-of-sample forecasting in this case?

(iii) Now, estimate the cointegrating parameter, rather than setting it to one. Use the last 16 quarters again to produce the out-of-sample RMSE. How does this compare with the forecasts from parts (i) and (ii)?

(iv) Would your conclusions change if you wanted to predict $r6$ rather than $\Delta r6$? Explain.

C18.11 Use the data in VOLAT.RAW for this exercise.

(i) Confirm that $lsp500 = \log(sp500)$ and $lip = \log(ip)$ appear to contain unit roots. Use Dickey-Fuller tests with four lagged changes and do the tests with and without a linear time trend.

(ii) Run a simple regression of $lsp500$ on lip. Comment on the sizes of the t statistic and R-squared.

(iii) Use the residuals from part (ii) to test whether $lsp500$ and lip are cointegrated. Use the standard Dickey-Fuller test and the ADF test with two lags. What do you conclude?

(iv) Add a linear time trend to the regression from part (ii) and now test for cointegration using the same tests from part (iii).

(v) Does it appear that stock prices and real economic activity have a long-run equilibrium relationship?

C18.12 This exercise also uses the data from VOLAT.RAW. Computer Exercise 18.11 studies the long-run relationship between stock prices and industrial production. Here, you will study the question of Granger causality using the percentage changes.

(i) Estimate an AR(3) model for $pcip_t$, the percentage change in industrial production (reported at an annualized rate). Show that the second and third lags are jointly significant at the 2.5% level.

(ii) Add one lag of $pcsp_t$ to the equation estimated in part (i). Is the lag statistically significant? What does this tell you about Granger causality between the growth in industrial production and the growth in stock prices?

(iii) Redo part (ii) but obtain a heteroskedasticity-robust t statistic. Does the robust test change your conclusions from part (ii)?

C18.13 Use the data in TRAFFIC2.RAW for this exercise. These monthly data, on traffic accidents in California over the years 1981 to 1989, were used in Computer Exercise C10.11.

(i) Using the standard Dickey-Fuller regression, test whether $ltotacc_t$ has a unit root. Can you reject a unit root at the 2.5% level?

(ii) Now, add two lagged changes to the test from part (i) and compute the augmented Dickey-Fuller test. What do you conclude?

(iii) Add a linear time trend to the ADF regression from part (ii). Now what happens?

(iv) Given the findings from parts (i) through (iii), what would you say is the best characterization of $ltotacc_t$: an I(1) process or an I(0) process about a linear time trend?

(v) Test the percentage of fatalities, $prcfat_t$, for a unit root, using two lags in an ADF regression. In this case, does it matter whether you include a linear time trend?

C18.14 Use the data in MINWAGE.DTA for sector 232 to answer the following questions.

(i) Confirm that $lwage232_t$ and $lemp232_t$ are best characterized as I(1) processes. Use the augmented DF test with one lag of $gwage232$ and $gemp232$, respectively, and a linear time trend. Is there any doubt that these series should be assumed to have unit roots?

(ii) Regress $lemp232_t$ on $lwage232_t$ and test for cointegration, both with and without a time trend, allowing for two lags in the augmented Engle-Granger test. What do you conclude?

(iii) Now regress $lemp232_t$ on log of the real wage rate, $lrwage232_t = lwage232_t - lcpi_t$, and a time trend. Do you find cointegration? Are they "closer" to being cointegrated when you use real wages than nominal wages?

(iv) What are some factors that might be missing from the cointegrating regression in part (iii)?