Chapter Eighteen

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 issue 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

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

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 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

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

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:

$$LRP = \delta_0 + \delta_1 + \delta_2 + \delta_3 + \ldots, \quad \textbf{(18.3)}$$

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. 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.2). 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. While (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:

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

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 $j \to 0$. 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 in the GDL is simply $\delta_0 = \gamma$, and 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

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

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, and 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 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}, and 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

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

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

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

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

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: $\hat{LRP} = \hat{\gamma}/(1 - \hat{\rho})$.

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) devise 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, while 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):

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

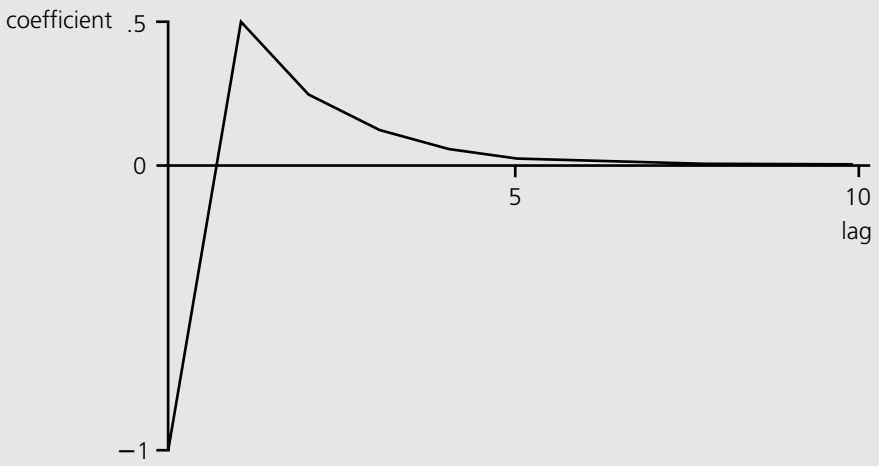
where $v_t = u_t - \rho u_{t-1}$, as before. By repeated substitution, it can be shown that (18.16) is equivalent to the infinite distributed lag model

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

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

Figure 18.1

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



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 $LRP = \Delta y^*/\Delta z^*$:

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

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.

Table 18.1

Distributed Lag Models for Housing Investment

Dependent Variable: log(*invpc*), detrended		
Independent Variables	**Geometric DL**	**Rational DL**
gprice	3.108 (0.933)	3.256 (0.970)
y_{-1}	.340 (.132)	.547 (.152)
$gprice_{-1}$	———	−2.936 (0.973)
constant	−.001 (.018)	−.578 (.307)
Long Run Propensity	4.688	.706
Sample Size Adjusted *R*-Squared	41 .375	40 .504

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.

18.2 TESTING FOR UNIT ROOTS

We now turn to the important problem of testing for **unit roots**. 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:

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

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:

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

[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 $\mathrm{E}(y_t)$ 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:

$$\mathrm{H}_0\colon \rho = 1. \tag{18.19}$$

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

$$\mathrm{H}_1\colon \rho < 1. \tag{18.20}$$

(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 $\mathrm{H}_1\colon \rho > 1$ 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 $\mathrm{Corr}(y_t,y_{t+h}) = \rho^h \to 0$ 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). [The reason we do not take the null to be I(0) in this setup is that $\{y_t\}$ 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 $\mathrm{H}_0\colon \theta = 0$ against $\mathrm{H}_1\colon \theta < 0$. 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 distribu-

tion 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).

While 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

$$\begin{gathered}\Delta\hat{r}3_t = .625 - .091\, r3_{t-1} \\ (.261)\quad(.037) \\ n = 123,\ R^2 = .048,\end{gathered} \tag{18.22}$$

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% level.

As with other hypotheses 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 provides *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, while $\hat{\rho} \approx .91$ implies a fair amount of persistence in $\{r3_t\}$, the correlation between observations which 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 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 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 twelve 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). The series spans the years from 1948 through 1996. Allowing for one lag of Δinf_t in the augmented Dickey-Fuller regression gives

$$\widehat{\Delta inf}_t = \underset{}{1.36} \underset{(.261)}{- .310\ inf_{t-1}} + \underset{(.103)}{.138\ \Delta inf_{t-1}}$$
$$\underset{(.126)}{}$$
$$n = 47,\ R^2 = .172.$$

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, and so a more appropriate null hypothesis is prob-

ably H_0: $\theta = 0$, $\delta = 0$. While 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 is 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. This does not change how we carry out the test.

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:

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

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 can be 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 only applicable 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

$$x_t = x_{t-1} + a_t \qquad \textbf{(18.27)}$$

and

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

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 (18.27) as

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

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 course. [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. 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, 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

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.

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 $y_t - \beta x_t$.]

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 while $r6_t$ 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. Since 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

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

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 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 is 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.

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, $E(y_t)$ and $E(x_t)$ 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 which 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 $R^2 = .032$. While 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 rhos 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

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

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.2—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 $\{u_t\}$ and, except for requiring that u_t is I(0), 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,

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

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

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

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 $\hat{\beta}$.

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 (1991).] 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 one is a relevant consideration.) There is little evidence of serial correlation in the residuals, and so we can use this t statistic as having an approximate normal distribution. [For comparison, the OLS estimate of β without the $\Delta r3$ terms—and using four 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 which 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:

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

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:

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

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 towards the equilibrium. Similarly, if $y_{t-1} < \beta x_{t-1}$, the error correction term induces a positive change in y back towards 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 Exercise 18.14). Under this assumption, an error correction model is

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

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

$$\begin{gathered}\widehat{\Delta hy6}_t = \underset{(.043)}{.090} + \underset{(.264)}{1.218}\,\Delta hy3_{t-1} - \underset{(.244)}{.840}\,(hy6_{t-1} - hy3_{t-2}) \\ n = 122,\ R^2 = .790.\end{gathered} \tag{18.39}$$

QUESTION 18.4

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

The error correction coefficient is negative and very significant. For example, if the holding yield on six-month bills is above that for three-month 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 is very convenient. 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 (1998) 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:

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

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

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

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:

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

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 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

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

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

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

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

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

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, which is 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*:

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

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

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

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, and so both equations over-predict 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 ± 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. While 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 Exercise 18.15).

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 equations 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

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

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, one or two lags are typically used; with quarterly data, usually four or eight; and with monthly data, perhaps six, 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)**:

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

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:

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

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 better in our sample than (18.48) did, and, at least for forecasting 1997, the model with lagged inflation worked better. Now, we estimate both models using data through 1989, saving 1990 through 1996 for out-of-sample comparisons. This leaves seven out-of-sample observations ($n = 41$ and $m = 7$, to be precise). For the AR(1) model, RMSE = .632, and MAE = .515. For the model that adds lagged inflation, RMSE = .550, and MAE = .362. Thus, by either measure, the model that includes inf_{t-1} produces better out-of-sample forecasts for the 1990s. In this case, the in-sample and out-of-sample criteria both 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, and so the forecast error variance increases with h. When $\rho^2 < 1$, 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 less precise.

Equation (18.54) shows that using the AR(1) model for multi-step 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

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

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 up 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 still ignores the estimation error in $\hat{\alpha}_0$ and $\hat{\gamma}_1$.

We can also compute multi-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, $y_{n+2} = \alpha + \rho_1 y_{n+1} + \rho_2 y_n + u_{n+2}$, and 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

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

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 multi-step-ahead forecasts for VAR models. To illustrate, suppose we have

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

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 $\mathrm{E}(u_{n+2}|I_n) = 0$], and 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 multi-step-ahead forecasts of y by using the recursive formula

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

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. You can find the actual civilian unemployment rate for 1998 in a recent *Economic Report of the President*. You will see that the one-step-ahead forecast turns out to be much closer 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 multi-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.

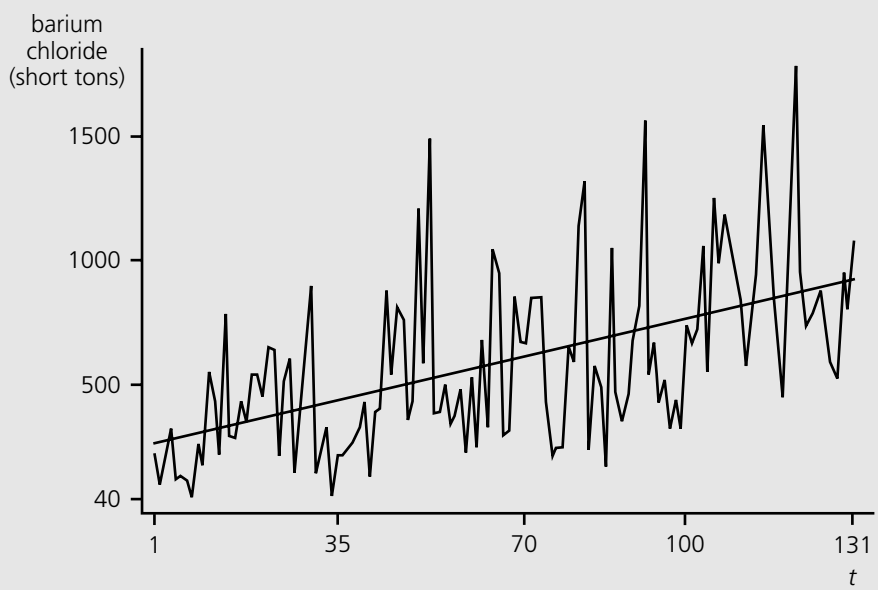
In the simplest case, suppose that $\{y_t\}$ has a linear trend but is unpredictable around that trend. Then, we can write

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

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.

Figure 18.2

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



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

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?

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

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 over-predicted. 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 under-prediction.

While 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 $\sigma^2(t + h)$, as 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. (Problem 18.17 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 eleven 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 up 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 multi-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

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

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 multi-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,

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

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 it directly.

As an example, let us estimate an AR(2) model for the general fertility rate in FERTIL3.RAW, using the observations up through 1979. (In Exercise 18.17 you are asked to use this model for forecasting, which is why we save some observations at the end of the sample.)

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

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 $\Delta gfr_t = \alpha + \theta gfr_{t-1} + \delta \Delta gfr_{t-1} + u_t$). 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 multi-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 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 which 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.

Multi-step-ahead forecasts present new challenges and are subject to large forecast error variances. Nevertheless, for models such as autoregressions and vector autore-

gressions, 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 Two-Step Procedure
Error Correction Model
Exponential Smoothing
Forecast Error
Forecast Interval
Geometric (or Koyck) Distributed Lag
Granger Causality
In-Sample Criteria
Infinite Distributed Lag (IDL) Model
Information Set
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 $t - 1$:

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

A natural assumption on $\{u_t\}$ is that $\mathrm{E}(u_t|I_{t-1}) = 0$, where I_{t-1} denotes all information on y and x observed at time $t - 1$; this means that $\mathrm{E}(y_t|I_{t-1}) = \alpha_0 + \alpha_1 x_t^*$. 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 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 & Poors 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

18.10 Use the data in WAGEPRC.RAW for this exercise. Problem 11.5 gives 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).

18.11 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 $\Delta\log(incpc_t)$. 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*)?

18.12 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 $\Delta i3$?

18.13 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.

18.14 Use INTQRT.RAW for this exercise.

(i) 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?

18.15 Use the data in PHILLIPS.RAW, adding the 1997 values for *unem* and *inf*: 4.9 and 2.3, respectively.

(i) Estimate the models in (18.48) and (18.49) using the data up 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. Use the *Economic Report of the President* (1999 or later) to obtain $unem_{1998}$. 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)?

18.16 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 twelve 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 twelve 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 twelve months in 1988. (You should obtain twelve 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.)

18.17 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 up 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?

18.18 Use CONSUMP.RAW for this exercise.

(i) Let y_t be real per capita disposable income. Use the data up 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?

18.19 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.

Chapter Nineteen

Carrying out an Empirical Project

In this chapter, we discuss the ingredients of a successful empirical analysis, with emphasis on completing a term project. In addition to reminding you of the important issues that have arisen throughout the text, we emphasize recurring themes that are important for applied research. We also provide suggestions for topics as a way of stimulating your imagination. Several sources of economic research and data are given as references.

19.1 POSING A QUESTION

The importance of posing a very specific question cannot be overstated. Without being explicit about the goal of your analysis, you cannot know where to even begin. The widespread availability of rich data sets makes it tempting to launch into data collection based on half-baked ideas, but this is often counterproductive. It is likely that, without carefully formulating your hypotheses and the kind of model you will need to estimate, you will forget to collect information on important variables, obtain a sample from the wrong population, or collect data for the wrong time period.

This does not mean that you should pose your question in a vacuum. Especially for a one-term project, you cannot be too ambitious. Therefore, when choosing a topic, you should be reasonably sure that data sources exist that will allow you to answer your question in the allotted time.

You need to decide what areas of economics or other social sciences interest you when selecting a topic. For example, if you have taken a course in labor economics, you have probably seen theories that can be tested empirically or relationships that have some policy relevance. Labor economists are constantly coming up with new variables that can explain wage differentials. Examples include quality of high school [Card and Krueger (1992) and Betts (1995)], amount of math and science taken in high school [Levine and Zimmerman (1995)], and physical appearance [Hamermesh and Biddle (1994), Averett and Korenman (1996), and Biddle and Hamermesh (1998)]. Researchers in state and local public finance study how local economic activity depends on economic policy variables, such as property taxes, sales taxes, level and quality of services (such as schools, fire, and police), and so on. [See, for example, White (1986), Papke (1987), Bartik (1991), and Netzer (1992).]

Economists that study education issues are interested in how spending affects performance [Hanushek (1986)], whether attending certain kinds of schools improves performance [for example, Evans and Schwab (1995)], and in determining factors that affect where private schools choose to locate [Downes and Greenstein (1996)].

Macroeconomists are interested in relationships between various aggregate time series, such as the link between growth in gross domestic product and growth in fixed investment or machinery [see De Long and Summers (1991)] or the effect of taxes on interest rates [for example, Peek (1982)].

There are certainly reasons for estimating models that are mostly descriptive. For example, property tax assessors use models (called *hedonic price models*—see Example 4.8) to estimate housing values for homes that have not been sold recently. This involves a regression model relating the price of a house to its characteristics (size, number of bedrooms, number of bathrooms, and so on). As a topic for a term paper, this is not very exciting: we are unlikely to learn much that is surprising, and such an analysis has no obvious policy implications. Adding the crime rate in the neighborhood as an explanatory variable would allow us to determine how important a factor crime is on housing prices, something that would be useful in estimating the costs of crime.

Several relationships have been estimated using macroeconomic data that are mostly descriptive. For example, an aggregate saving function can be used to estimate the aggregate marginal propensity to save, as well as the response of saving to asset returns (such as interest rates). Such an analysis could be made more interesting by using time series data on a country that has a history of political upheavals and determining whether savings rates decline during times of political uncertainty.

Once you decide on an area of research, there are a variety of ways to locate specific papers on the topic. The *Journal of Economic Literature* (JEL) has a detailed classification system so that each paper is given a set of identifying codes that places it within certain subfields of economics. The JEL also contains a list of articles published in a wide variety of journals, organized by topic, and it even contains short abstracts of some articles.

Especially convenient for finding published papers on various topics are **Internet** services, such as *EconLit*, which is subscribed to by many universities. *EconLit* allows users to do a comprehensive search of almost all economics journals by author, subject, words in the title, and so on. The *Social Science Citation Index* is useful for finding papers on a broad range of topics in the social sciences, including popular papers that have been cited often in other published works.

In thinking about a topic, there are some things to keep in mind. First, for a question to be interesting, it does not need to have broad-based policy implications; rather, it can be of local interest. For example, you might be interested in knowing whether living in a fraternity at your university causes students to have lower or higher grade point averages. This may or may not be of interest to people outside of your university, but it is probably of concern to at least some people within the university. On the other hand, you might study a problem that starts out being of local interest but turns out to have widespread interest, such as determining which factors affect, and which university policies can stem, alcohol abuse on college campuses.

Second, it is very difficult, especially for a quarter or semester project, to do truly original research using the standard macroeconomic aggregates on the U.S. economy.

For example, the question of whether money growth, government spending growth, and so on, affect economic growth has been and continues to be studied by professional macroeconomists. The question of whether stock or other asset returns can be systematically predicted using known information has, for obvious reasons, been studied pretty carefully. This does not mean that you should avoid estimating macroeconomic or empirical finance models, as even just using more recent data can add constructively to a debate. In addition, you can sometimes find a new variable that has an important effect on economic aggregates or financial returns; such a discovery can be exciting.

The point is that exercises such as using a few additional years to estimate a standard Phillips curve or an aggregate consumption function for the U.S. economy, or some other large economy, are unlikely to yield additional insights, although they can be instructive for the student. Instead, you might use data on a smaller country to estimate a static or dynamic Phillips curve, or to test the efficient markets hypothesis, and so on.

At the nonmacroeconomic level, there are also plenty of questions that have been studied extensively. For example, labor economists have published many papers on estimating the return to education. This question is still studied because it is very important, and new data sets, as well as new econometric approaches, continue to be developed. For example, as we saw in Chapter 9, certain data sets have better proxy variables for unobserved ability than other data sets. (Compare WAGE1.RAW and WAGE2.RAW.) In other cases, we can obtain panel data or data from a natural experiment—see Chapter 13—which allow us to approach an old question from a different perspective.

As another example, criminologists are interested in studying the effects of various laws on crimes. The question of whether capital punishment has a deterrent effect has long been debated. Similarly, economists have been interested in whether taxes on cigarettes and alcohol reduce consumption (as always, in a ceteris paribus sense). As more years of data at the state level become available, a richer panel data set can be created, and this can help us better answer major policy questions. Plus, there are fairly recent crime-fighting innovations—such as the advent of community policing—whose effectiveness can be evaluated empiricially.

While you are formulating your question, it is helpful to discuss your ideas with your classmates, instructor, and friends. You should be able to convince people that the answer to your question is of some interest. (Of course, whether you can persuasively answer your question is another issue, but you need to begin with an interesting question.) If someone asks you about your paper and you respond with "I'm doing my paper on crime" or "I'm doing my paper on interest rates," chances are you have only decided on a general area without formulating a true question. You should be able to say something like "I'm studying the effects of community policing on city crime rates in the United States" or "I'm looking at how inflation volatility affects short-term interest rates in Brazil."

19.2 LITERATURE REVIEW

All papers, even if they are relatively short, should contain a review of relevant literature. It is rare that one attempts an empirical project where there is not some published

precedent. If you search through journals or use **on-line search services** such as *EconLit* to come up with a topic, you are already well on your way to a literature review. If you select a topic on your own—such as studying the effects of drug usage on college performance at your university—then you will probably have to work a little harder. But on-line search services make that work a lot easier, as you can search by keywords, by words in the title, by author, and so on. You can then read abstracts of papers to see how relevant they are to your own work.

When doing your literature search, you should think of related topics that might not show up in a search using a handful of key words. For example, if you are studying the effects of drug usage on wages or grade point average, you should probably look at the literature on how alcohol usage affects such factors. Knowing how to do a thorough literature search is an acquired skill, but you can get a long way by thinking before searching.

Researchers differ on how a literature review should be incorporated into a paper. Some like to have a separate section called "literature review," while others like to include the literature review as part of the introduction. This is largely a matter of taste, although an extensive literature review probably deserves its own section. If the term paper is the focus of the course—say, in a senior seminar or an advanced econometrics course—your literature review probably will be lengthy. Term papers at the end of a first course are typically shorter, and the literature reviews are briefer.

19.3 DATA COLLECTION

Deciding on the Appropriate Data Set

Collecting data for a term paper can be educational, exciting, and sometimes even frustrating. You must first decide on the kind of data needed to answer your posed question. As we discussed in the introduction and have covered throughout this text, data sets come in a variety of forms. The most common kinds are cross-sectional, time series, pooled cross sections, and panel data sets.

Many questions can be addressed using any of the data structures we have described. For example, to study whether more law enforcement lowers crime, we could use a cross section of cities, a time series for a given city, or a panel data set of cities—which consists of data on the same cities over two or more years.

Deciding on which kind of data to collect often depends on the nature of the analysis. To answer questions at the individual or family level, we often only have access to a single cross section; typically, these are obtained via surveys. Then, we must ask whether we can obtain a rich enough data set to do a convincing ceteris paribus analysis. For example, suppose we want to know whether families who save through individual retirement accounts (IRAs)—which have certain tax advantages—have less non-IRA savings. In other words, does IRA saving simply crowd out other forms of saving? There are data sets, such as the Survey of Consumer Finances, which contain information on various kinds of saving for a different sample of families each year. There are several issues that arise in using such a data set. Perhaps the most important is whether there are enough controls—including income, demographics, and proxies for saving tastes—to do a reasonable ceteris paribus analysis. If these are the only kinds of data available, we must do what we can with them.

The same issues arise with cross-sectional data on firms, cities, states, and so on. In most cases, it is not obvious that we will be able to do a ceteris paribus analysis with a single cross section. For example, any study of the the effects of law enforcement on crime must recognize the endogeneity of law enforcement expenditures. When using standard regression methods, it may be very hard to complete a convincing ceteris paribus analysis, no matter how many controls we have. (See Section 19.4 for more discussion.)

If you have read the advanced chapters on panel data methods, you know that having the same cross-sectional units at two or more different points in time can allow us to control for time-constant unobserved effects that would normally confound regression on a single cross section. Panel data sets are relatively hard to obtain for individuals or families—although some important ones exist, such as the Panel Study of Income Dynamics—but they can be used in very convincing ways. Panel data sets on firms also exist. For example, CompuStat and the Center for Research on Securities Prices (CRSP) manage very large panel data sets of financial information on firms. Easier to obtain are panel data sets on larger units, such as schools, cities, counties, and states, as these tend not to disappear over time, and government agencies are responsible for collecting information on the same variables each year. For example, the Federal Bureau of Investigation collects and reports detailed information on crime rates at the city level. Sources of data are listed in the chapter appendix.

Data come in a variety of forms. Some data sets, especially historical ones, are available only in printed form. For small data sets, entering the data yourself from the printed source is manageable and convenient. Sometimes, articles are published with small data sets—especially time series applications. These can be used in an empirical study, perhaps by supplementing the data with more recent years.

Many data sets are available on computer diskettes or magnetic tapes. The former are especially easy to work with. Currently, very large data sets can be put on small diskettes. Various government agencies sell data diskettes, as do private firms. Authors of papers are often willing to provide their data sets in diskette form.

More and more data sets are available on the worldwide web. The web is a vast resource of **on-line data bases**. Numerous web sites containing economic and related data sets have recently been created. Several other web sites contain links to data sets that are of interest to economists; some of these are listed in the chapter appendix. Generally, searching the Internet for data sources is fairly easy and will become even more convenient in the future.

Entering and Storing Your Data

Once you have decided on a data type and have located a data source, you must put the data into usable form. If the data came on diskette, they are already in some form, hopefully one in widespread use. The most flexible way to obtain data in diskette form is as a standard **text (ASCII) file**. All statistics and econometrics software packages allow raw data to be stored this way. Typically, it is straightforward to read a text file directly into an econometrics package, provided the file is properly structured. The data files we have used throughout the text provide several examples of how cross-sectional, time series, pooled cross sections, and panel data sets are usually stored. As a general rule,

the data should have a tabular form, with each observation representing a different row; the columns in the data set represent different variables. Occasionally, you might encounter a data set stored with each column representing an observation and each row a different variable. This is not ideal, but most software packages allow data to be read in this form, and then reshaped. Naturally, it is crucial to know how the data are organized before reading them into your econometrics package.

For time series data sets, there is only one sensible way to enter and store the data: namely, chronologically, with the earliest time period listed as the first observation and the most recent time period as the last observation. It is often useful to include variables indicating year and, if relevant, quarter or month. This facilitates estimation of a variety of models later on, including allowing for seasonality and breaks at different time periods. For cross sections pooled over time, it is usually best to have the cross section for the earliest year fill the first block of observations, followed by the cross section for the second year, and so on. (See FERTIL1.RAW as an example.) This arrangement is not crucial, but it is very important to have a variable stating the year attached to each observation.

For panel data, as we discussed in Section 13.5, it is best if all the years for each cross-sectional observation are adjacent and in chronological order. With this ordering we can use all of the panel data methods from Chapters 13 and 14. With panel data, it is important to include a unique identifier for each cross-sectional unit, along with a year variable.

If you obtain your data in printed form, you have several options for entering it into a computer. First, you can create a text file using a standard **text editor**. (This is how several of the raw data sets included with the text were initially created.) Typically, it is required that each row starts a new observation, that each row contains the same ordering of the variables—in particular, each row should have the same number of entries—and that the values are separated by at least one space. Sometimes, a different separator, such as a comma, is better, but this depends on the software you are using. If you have missing observations on some variables, you must decide on how to denote that; simply leaving a blank does not generally work. Many regression packages accept a period as the missing value symbol. Some people prefer to use a number—presumably an impossible value for the variable of interest—to denote missing values. If you are not careful, this can be dangerous; we discuss this further later.

If you have nonnumerical data—for example, you want to include the names in a sample of colleges or the names of cities—then you should check the econometrics package you will use to see the best way to enter such variables (often called *strings*). Typically, strings are put between double or single quotations. Or, the text file can follow a rigid formatting, which usually requires a small program to read in the text file. But you need to check your econometrics package for details.

Another generally available option is to use a **spreadsheet** to enter your data, such as Excel. This has a couple of advantages over a text file. First, because each observation on each variable is a cell, it is less likely that numbers will be run together (as would happen if you forget to enter a space in a text file). Secondly, spreadsheets allow manipulation of data, such as sorting, computing averages, and so on. This second benefit is less important if you use a software package that allows for sophisticated data management; many software packages, including Eviews and Stata, fall into this cate-

gory. If you use a spreadsheet for initial data entry, then you must often export the data in a form that can be read by your econometrics package. This is usually straightforward, as spreadsheets export to text files using a variety of formats.

A third alternative is to enter the data directly into your econometrics package. While this obviates the need for a text editor or a spreadsheet, it is more awkward because you cannot freely move across different observations to make corrections or additions.

Data downloaded from the Internet may come in a variety of forms. Often data come as text files, but different conventions are used for separating variables; for panel data sets, the conventions on how to order the data may differ. Some Internet data sets come as spreadsheet files, in which case you must use an appropriate spreadsheet to read them.

Inspecting, Cleaning, and Summarizing Your Data

It is extremely important to become familiar with any data set you will use in an empirical analysis. If you enter the data yourself, you will be forced to know everything about it. But if you obtain data from an outside source, you should still spend some time understanding its structure and conventions. Even data sets that are widely used and heavily documented can contain glitches. If you are using a data set obtained from the author of a paper, you must be aware that methods of data set construction can be forgotten.

Earlier, we reviewed the standard ways that various data sets are stored. You also need to know how missing values are coded. Preferably, missing values are indicated with a nonnumeric character, such as a period. If a number is used as a missing value code, such as "999" or "-1", you must be very careful when using these observations in computing any statistics. Your econometrics package will probably not know that a certain number really represents a missing value: it is likely that such observations will be used as if they are valid, and this can produce rather misleading results. The best approach is to set any numerical codes for missing values to some other character (such as a period) that cannot be mistaken for real data.

You must also know the nature of the variables in the data set. Which are binary variables? Which are ordinal variables (such as a credit rating)? What are the units of measurement of the variables? For example, are monetary values expressed in dollars, thousands of dollars, millions of dollars, or so on? Are variables representing a rate—such as school dropout rates, inflation rates, unionization rates, or interest rates—measured as a percent or a proportion?

Especially for time series data, it is crucial to know if monetary values are in nominal (current) or real (constant) dollars. If the values are in real terms, what is the base year or period?

If you receive a data set from an author, some variables may already be transformed in certain ways. For example, sometimes only the log of a variable (such as wage or salary) is reported in the data set.

Detecting mistakes in a data set is necessary for preserving the integrity of any data analysis. It is always useful to find minimums, maximums, means, and standard deviations of all, or at least the most significant, variables in the analysis. For example, if you

find that the minimum value of education in your sample is −99, you know that at least one entry on education needs to be set to a missing value. If, upon further inspection, you find that several observations have −99 as the level of education, you can be confident that you have stumbled onto the missing value code for education. As another example, if you find that an average murder conviction rate across a sample of cities is .632, you know that conviction rate is measured as a proportion, not a percent. Then, if the maximum value is above one, this is likely a typographical error. (It is not uncommon to find data sets where most of the entries on a rate variable were entered as a percent, but where some were entered as a proportion, and vice versa. Such data coding errors can be difficult to detect, but it is important to try.)

We must also be careful in using time series data. If we are using monthly or quarterly data, we must know which variables, if any, have been seasonally adjusted. Transforming data also requires great care. Suppose we have a monthly data set and we want to create the change in a variable from one month to the next. To do this, we must be sure that the data are ordered chronologically, from earliest period to latest. If for some reason this is not the case, the differencing will result in garbage. To be sure the data are properly ordered, it is useful to have a time period indicator. With annual data, it is sufficient to know the year, but then we should know whether the year is entered as four digits or two digits (for example, 1998 versus 98). With monthly or quarterly data, it is also useful to have a variable or variables indicating month or quarter. With monthly data, we may have a set of dummy variables (11 or 12) or one variable indicating the month (1 through 12 or a string variable, such as *jan*, *feb*, and so on).

With or without yearly, monthly, or quarterly indicators, we can easily construct time trends in all econometrics software packages. Creating seasonal dummy variables is easy if the month or quarter is indicated; at a minimum, we need to know the month or quarter of the first observation.

Manipulating panel data can be even more challenging. In Chapter 13, we discussed pooled OLS on the differenced data as one general approach to controlling for unobserved effects. In constructing the differenced data, we must be careful not to create phantom observations. Suppose we have a balanced panel on cities from 1992 through 1997. Even if the data are ordered chronologically within each cross-sectional unit—something that should be done before proceeding—a mindless differencing will create an observation for 1992 for all cities except the first in the sample. This observation will be the 1992 value for city i, minus the 1997 value for city $i - 1$; this is clearly nonsense. Thus, we must make sure that 1992 is missing for all differenced variables.

With an unbalanced panel, things become much trickier because no single command works for all cross-sectional units. It is usually easier to use fixed effects estimation on unbalanced panels.

19.4 ECONOMETRIC ANALYSIS

This text has focused on econometric analysis, and we are not about to provide a review of econometric methods in this section. Nevertheless, we can give some general guidelines about the sorts of issues that need to be considered in an empirical analysis.

As we discussed earlier, after deciding on a topic, we must collect an appropriate data set. Assuming that this has also been done, we must next decide on the appropriate econometric methods.

If your course has focused on ordinary least squares estimation of a multiple linear regression model, using either cross-sectional or time series data, the econometric approach has pretty much been decided for you. This is not necessarily a weakness, as OLS is still the most widely used econometric method. Of course, you still have to decide whether any of the variants of OLS—such as weighted least squares or correcting for serial correlation in a time series regression—are required.

In order to justify OLS, you must also make a convincing case that the key OLS assumptions are satisfied for your model. As we have discussed at some length, the first issue is whether the error term is uncorrelated with the explanatory variables. Ideally, you have been able to control for enough other factors to assume that those that are left in the error are unrelated to the regressors. Especially when dealing with individual, family, or firm-level cross-sectional data, the self-selection problem—which we discussed in Chapters 7 and 15—is often relevant. For instance, in the IRA example from Section 19.3, it may be that families with unobserved taste for saving are also the ones that open IRAs. You should also be able to argue that the other potential sources of endogeneity—namely, measurement error and simultaneity—are not a serious problem.

When specifying your model you must also make functional form decisions. Should some variables appear in logarithmic form? (In econometric applications, the answer is often yes.) Should some variables be included in levels and squares, to possibly capture a diminishing effect? How should qualitative factors appear? Is it enough to just include binary variables for different attributes or groups? Or, do these need to be interacted with quantitative variables? (See Chapter 7 for details.)

For cross-sectional analysis, a secondary, but nevertheless important issue, is whether there is heteroskedasticity. In Chapter 8, we explained how this can be dealt with. The simplest way is to compute heteroskedasticity-robust statistics.

As we emphasized in Chapters 10, 11, and 12, time series applications require additional care. Should the equation be estimated in levels? If levels are used, are time trends needed? Is differencing the data more appropriate? If the data are monthly or quarterly, does seasonality have to be accounted for? If you are allowing for dynamics—for example, distributed lag dynamics—how many lags should be included? You must start with some lags based on intuition or common sense, but eventually it is an empirical matter.

If your model has some potential misspecification, such as omitted variables, and you use OLS, you should attempt some sort of misspecification analysis of the kinds we discussed in Chapters 3 and 5. Can you determine, based on reasonable assumptions, the direction of any bias in the estimators?

If you have studied the method of instrumental variables, you know that it can be used to solve various forms of endogeneity, including omitted variables (Chapter 15), errors-in-variables (Chapter 15), and simultaneity (Chapter 16). Naturally, you need to think hard about whether the instrumental variables you are considering are likely to be valid.

Good papers in the empirical social sciences contain **sensitivity analysis**. Broadly, this means you estimate your original model and modify it in ways that seem reason-

able. Hopefully, the important conclusions do not change. For example, if you use as an explanatory variable a measure of alcohol consumption (say, in a grade point average equation), do you get qualitatively similar results if you replace the quantitative measure with a dummy variable indicating alcohol usage? If the binary usage variable is significant but the alcohol quantity variable is not, it could be that usage reflects some unobserved attribute that affects GPA and is also correlated with alcohol usage. But this needs to be considered on a case-by-case basis.

If some observations are much different from the bulk of the sample—say, you have a few firms in a sample that are much larger than the other firms—do your results change much when those observations are excluded from the estimation? If so, you may have to alter functional forms to allow for these observations or argue that they follow a completely different model. The issue of outliers was discussed in Chapter 9.

Using panel data raises some additional econometric issues. Suppose you have collected two periods. There are at least four ways to use two periods of panel data without resorting to instrumental variables. You can pool the two years in a standard OLS analysis, as discussed in Chapter 13. While this might increase the sample size relative to a single cross section, it does not control for time-constant unobservables. In addition, the errors in such an equation are almost always serially correlated because of an unobserved effect. Random effects estimation corrects the serial correlation problem and produces asymptotically efficient estimators, provided the unobserved effect has zero mean given values of the explanatory variables in all time periods.

Another possibility is to include a lagged dependent variable in the equation for the second year. In Chapter 9, we presented this as a way to at least mitigate the omitted variables problem, as we are in any event holding fixed the initial outcome of the dependent variable. This often leads to similar results as differencing the data, as we covered in Chapter 13.

With more years of panel data, we have the same options, plus an additional choice. We can use the fixed effects transformation to eliminate the unobserved effect. (With two years of data, this is the same as differencing.) In Chapter 15, we showed how instrumental variables techniques can be combined with panel data transformations to relax exogeneity assumptions even more. As a general rule, it is a good idea to apply several reasonable econometric methods and compare the results. This often allows us to determine which of our assumptions are likely to be false.

Even if you are very careful in devising your topic, postulating your model, collecting your data, and carrying out the econometrics, it is quite possible that you will obtain puzzling results—at least some of the time. When that happens, the natural inclination is to try different models, different estimation techniques, or perhaps different subsets of data until the results correspond more closely to what was expected. Virtually all applied researchers search over various models before finding the "best" model. Unfortunately, this practice of **data mining** violates the assumptions we have made in our econometric analysis. The results on unbiasedness of OLS and other estimators, as well as the t and F distributions we derived for hypothesis testing, assume that we observe a sample following the population model and we estimate that model once. Estimating models that are variants of our original model violates that assumption because we are using the same set of data in a *specification search*. In effect, we use the

outcome of tests by using the data to respecify our model. The estimates and tests from different model specifications are not independent of one another.

Some specification searches have been programmed into standard software packages. A popular one is known as *stepwise regression*, where different combinations of explanatory variables are used in multiple regression analysis in an attempt to come up with the best model. There are various ways that stepwise regression can be used, and we have no intention of reviewing them here. The general idea is to either start with a large model and keep variables whose p-values are below a certain significance level or to start with a simple model and add variables that have significant p-values. Sometimes, groups of variables are tested with an F test. Unfortunately, the final model often depends on the order in which variables were dropped or added. [For more on stepwise regression, see Draper and Smith (1981).] In addition, this is a severe form of data mining, and it is difficult to interpret t and F statistics in the final model. One might argue that stepwise regression simply automates what researchers do anyway in searching over various models. However, in most applications, one or two explanatory variables are of primary interest, and then the goal is to see how robust the coefficients on those variables are to either adding or dropping other variables, or to changing functional form.

In principle, it is possible to incorporate the effects of data mining into our statistical inference; in practice, this is very difficult and is rarely done, especially in sophisticated empirical work. [See Leamer (1983) for an engaging discussion of this problem.] But we can try to minimize data mining by not searching over numerous models or estimation methods until a significant result is found and then reporting only that result. If a variable is statistically significant in only a small fraction of the models estimated, it is quite likely that the variable has no effect in the population.

19.5 WRITING AN EMPIRICAL PAPER

Writing a paper that uses econometric analysis is very challenging, but it can also be rewarding. A successful paper combines a careful, convincing data analysis with good explanations and exposition. Therefore, you must have a good grasp of your topic, good understanding of econometric methods, and solid writing skills. Do not be discouraged if you find writing an empirical paper difficult; most professional researchers have spent many years learning how to craft an empirical analysis and to write the results in a convincing form.

While writing styles vary, many papers follow the same general outline. The following paragraphs include ideas for section headings and explanations about what each section should contain. These are only suggestions and hardly need to be strictly followed. In the final paper, each section would be given a number, usually starting with one for the introduction.

Introduction

The introduction states the basic objectives of the study and explains why it is important. It generally entails a review of the literature, indicating what has been done and how previous work can be improved upon. (As discussed in Section 19.2, an extensive

literature review can be put in a separate section.) Presenting simple statistics or graphs that reveal a seemingly paradoxical relationship is a useful way to introduce the paper's topic. For example, suppose that you are writing a paper about factors affecting fertility in a developing country, with the focus on education levels of women. An appealing way to introduce the topic would be to produce a table or a graph showing that fertility has been falling (say) over time and a brief explanation of how you hope to examine the factors contributing to the decline. At this point, you may already know that, ceteris paribus, more highly educated women have fewer children and that average education levels have risen over time.

Most researchers like to summarize the findings of their paper in the introduction. This can be a useful device for grabbing the reader's attention. For example, you might state that your best estimate of the effect of missing 10 hours of lecture during a thirty-hour term is about one-half of a grade point. But the summary should not be too involved because neither the methods nor the data used to obtain the estimates have yet been introduced.

Conceptual (or Theoretical) Framework

This is the section where you describe the general approach to answering the question you have posed. It can be formal economic theory, but in many cases, it is an intuitive discussion about what conceptual problems arise in answering your question.

As an example, suppose you are studying the effects of economic opportunities and severity of punishment on criminal behavior. One approach to explaining participation in crime is to specify a utility maximization problem where the individual chooses the amount of time spent in legal and illegal activities, given wage rates in both kinds of activities, as well as variable measuring probability and severity of punishment for criminal activity. The usefulness of such an exercise is that it suggests which variables should be included in the empirical analysis; it gives guidance (but rarely specifics) as to how the variables should appear in the econometric model.

Often there is no need to write down an economic theory. For econometric policy analysis, common sense usually suffices for specifying a model. For example, suppose you are interested in estimating the effects of participation in Aid for Families with Dependent Children (AFDC) on the effects of child performance in school. AFDC provides supplemental income, but participation also makes it easier to receive Medicaid and other benefits. The hard part of such an analysis is deciding on the set of variables that should be controlled for. In this example, we could control for family income (including AFDC and any other welfare income), mother's education, whether the family lives in an urban area, and other variables. Then, the inclusion of an AFDC participation indicator (hopefully) measures the nonincome benefits of AFDC participation. A discussion of which factors should be controlled for and the mechanisms through which AFDC participation might improve school performance substitute for formal economic theory.

Econometric Models and Estimation Methods

It is very useful to have a section that contains a few equations of the sort you estimate and present in the results section of the paper. This allows you to fix ideas about what

the key explanatory variable is and what other factors you will control for. Writing equations containing error terms allows you to discuss whether a method such as OLS will be appropriate.

The distinction between a *model* and an estimation method should be made in this section. A model represents a population relationship (broadly defined to allow for time series equations). For example, we should write

$$colGPA = \beta_0 + \beta_1 alcohol + \beta_2 hsGPA + \beta_3 SAT + \beta_4 female + u \quad \textbf{(19.1)}$$

to describe the relationship between college GPA and alcohol consumption, with some other controls in the equation. Presumably, this equation represents a population, such as all undergraduates at a university. There are no "hats" (ˆ) on the β_j or on *colGPA* because this is a model, not an estimated equation. We do not put in numbers for the β_j because we do not know (and never will know) these numbers. Later, we will estimate them. In this section, do not anticipate the presentation of your empirical results. In other words, do not start with a general model and then say that you omitted certain variables because they turned out to be insignificant. Such discussions should be left for the results section.

A time series model to relate city-level car thefts to the unemployment rate (and other controls) could look like

$$thefts_t = \beta_0 + \beta_1 unem_t + \beta_2 unem_{t-1} + \beta_3 cars_t + \beta_4 convrate_t + \beta_5 convrate_{t-1} + u_t, \quad \textbf{(19.2)}$$

where the t subscript is useful for emphasizing any dynamics in the equation (in this case, allowing for unemployment and the automobile theft conviction rate to have lagged effects).

After specifying a model or models, it is appropriate to discuss estimation methods. In most cases, this will be OLS, but, for example, in a time series equation, you might use feasible GLS to do a serial correlation correction (as in Chapter 12). However, the method for estimating a model is quite distinct from the model itself. It is not meaningful, for instance, to talk about "an OLS model." Ordinary least squares is a method of estimation, and so are weighted least squares, Cochrane-Orcutt, and so on. There are usually many ways to estimate any model. You should explain why the method you are choosing is warranted.

Any assumptions that are used in obtaining an estimable econometric model from an underlying economic model should be clearly discussed. For example, in the quality of high school example mentioned in Section 19.1, the issue of how to measure school quality is central to the analysis. Should it be based on average SAT scores, percentage of graduates attending college, student-teacher ratios, average education level of teachers, some combination of these, or possibly other measures?

We always have to make assumptions about functional form whether or not a theoretical model has been presented. As you know, constant elasticity and constant semi-elasticity models are attractive because the coefficients are easy to interpret (as percentages). There are no hard rules on how to choose functional form, but the guidelines discussed in Section 6.2 seem to work well in practice. You do not need an exten-

sive discussion of functional form, but it is useful to mention whether you will be estimating elasticities or a semi-elasticity. For example, if you are estimating the effect of some variable on wage or salary, the dependent variable will almost surely be in logarithmic form, and you might as well include this in any equations from the beginning. You do not have to present every, or even most, of the functional form variations that you will report later in the results section.

Often the data used in empirical economics are at the city or county level. For example, suppose that for the population of small to mid-size cities, you wish to test the hypothesis that having a minor league baseball team causes a city to have a lower divorce rate. In this case, you must account for the fact that larger cities will have more divorces. One way to account for the size of the city is to scale divorces by the city or adult population. Thus, a reasonable model is

$$\log(div/pop) = \beta_0 + \beta_1 mlb + \beta_2 perCath + \beta_3 \log(inc/pop) + other\ factors, \quad \textbf{(19.3)}$$

where *mlb* is a dummy variable equal to one if the city has a minor league baseball team, *perCath* is the percentage of the population which is Catholic (so it is a number such as 34.6 to mean 34.6%). Note that *div*/*pop* is a divorce rate, which is generally easier to interpret than the absolute number of divorces.

Another way to control for population is to estimate the model

$$\log(div) = \gamma_0 + \gamma_1 mlb + \gamma_2 perCath + \gamma_3 \log(inc) + \gamma_4 \log(pop) + other\ factors. \quad \textbf{(19.4)}$$

The parameter of interest, γ_1, when multiplied by 100, gives the percentage difference between divorce rates, holding population, percent Catholic, income, and whatever else is in "other factors" constant. In equation (19.3), β_1 measures the percentage effect of minor league baseball on *div*/*pop*, which can change either because the number of divorces or the population changes. Using the fact that $\log(div/pop) = \log(div) - \log(pop)$ and $\log(inc/pop) = \log(inc) - \log(pop)$, we can rewrite (19.3) as

$$\log(div) = \beta_0 + \beta_1 mlb + \beta_2 perccath + \beta_3 \log(inc) + (1 - \beta_3)\log(pop) + other\ factors,$$

which shows that (19.3) is a special case of (19.4) with $\gamma_4 = (1 - \beta_3)$ and $\gamma_j = \beta_j$, $j = 0,1,2$, and 3. Alternatively, (19.4) is equivalent to adding $\log(pop)$ as an additional explanatory variable to (19.3). This makes it easy to test for a separate population effect on the divorce rate.

If you are using a more advanced estimation method, such as two stage least squares, you need to provide some reasons for why you are doing so. If you use 2SLS, you must provide a careful discussion on why your IV choices for the endogenous explanatory variable (or variables) are valid. As we mentioned in Chapter 15, there are two requirements for a variable to be considered a good IV. First, it must be omitted from and exogenous to the equation of interest (structural equation). This is something we must assume. Second, it must have some partial correlation with the endogenous explanatory variable. This we can test. For example, in equation (19.1), you might use

a binary variable for whether a student lives in a dormitory (*dorm*) as an IV for alcohol consumption. This requires that living situation has no direct impact on *colGPA*—so that it is omitted from (19.1)—and that it is uncorrelated with unobserved factors in u that have an effect on *colGPA*. We would also have to verify that *dorm* is partially correlated with *alcohol* by regressing *alcohol* on *dorm*, *hsGPA*, *SAT*, and *female*. (See Chapter 15 for details.)

You might account for the omitted variable problem (or omitted heterogeneity) by using panel data. Again, this is easily described by writing an equation or two. In fact, it is useful to show how to difference the equations over time to remove time-constant unobservables; this gives an equation that can be estimated by OLS. Or, if you are using fixed effects estimation instead, you simply state so.

As a simple example, suppose you are testing whether higher county tax rates reduce economic activity, as measured by per capita manufacturing output. Suppose that for the years 1982, 1987, and 1992, the model is

$$\log(manuf_{it}) = \beta_0 + \delta_1 d87_t + \delta_2 d92_t + \beta_1 tax_{it} + \ldots + a_i + u_{it},$$

where $d87_t$ and $d92_t$ are year dummy variables, and tax_{it} is the tax rate for county i at time t (in percent form). We would have other variables that change over time in the equation, including measures for costs of doing business (such as average wages), measures of worker productivity (as measured by average education), and so on. The term a_i is the fixed effect, containing all factors that do not vary over time, and u_{it} is the idiosyncratic error term. To remove a_i, we can either difference across the years or use time-demeaning (the fixed effects transformation).

The Data

You should always have a section that carefully describes the data used in the empirical estimation. This is particularly important if your data are nonstandard or have not been widely used by other researchers. Enough information should be presented so that a reader could, in principle, obtain the data and redo your analysis. In particular, all applicable public data sources should be included in the references, and short data sets can be listed in an appendix. If you used your own survey to collect the data, a copy of the questionaire should be presented in an appendix.

Along with a discussion of the data sources, be sure to discuss the units of each of the variables (for example, is income measured in hundreds or thousands of dollars?). Including a table of variable definitions is very useful to the reader. The names in the table should correspond to the names used in describing the econometric results in the following section.

It is also very informative to present a table of summary statistics, such as minimum and maximum values, means, and standard deviations for each variable. Having such a table makes it easier to interpret the coefficient estimates in the next section, and it emphasizes the units of measurement of the variables. For binary variables, the only necessary summary statistic is the fraction of ones in the sample (which is the same as the sample mean). For trending variables, things like means are less interesting. It is often useful to compute the average growth rate in a variable over the years in your sample.

You should always clearly state how many observations you have. For time series data sets, identify the years that you are using in the analysis, including a description of any special periods in history (such as World War II). If you use a pooled cross section or a panel data set, be sure to report how many cross-sectional units (people, cities, and so on) you have for each year.

Results

The results section should include your estimates of any models formulated in the models section. You might start with a very simple analysis. For example, suppose that percent of students attending college from the graduating class (*percoll*) is used as a measure of the quality of the high school a person attended. Then, an equation to estimate is

$$\log(wage) = \beta_0 + \beta_1 percoll + u.$$

Of course, this does not control for several other factors that may determine wages and that may be correlated with *percoll*. But a simple analysis can draw the reader into the more sophisticated analysis and reveal the importance of controlling for other factors.

If only a few equations are estimated, you can present the results in equation form with standard errors in parentheses below estimated coefficients. If your model has several explanatory variables and you are presenting several variations on the general model, it is better to report the results in tabular rather than equation form. Most of you should have at least one table, which should always include at least the *R*-squared and the number of observations for each equation. Other statistics, such as the adjusted *R*-squared, can also be listed.

The most important thing is to discuss the interpretation and strength of your empirical results. Do the coefficients have the expected signs? Are they statistically significant? If a coefficient is statistically significant but has a counterintuitive sign, why might this be true? It might be revealing a problem with the data or the econometric method (for example, OLS may be inappropriate due to omitted variables problems).

Be sure to describe the *magnitudes* of the coefficients on the major explanatory variables. Often there are one or two policy variables that are central to the study. Their signs, magnitudes, and statistical significance should be treated in detail. Remember to distinguish between economic and statistical significance. If a *t* statistic is small, is it because the coefficient is practically small or because its standard error is large?

In addition to discussing estimates from the most general model, you can provide interesting special cases, especially those needed to test certain multiple hypotheses. For example, in a study to determine wage differentials across industries, you might present the equation without the industry dummies; this allows the reader to easily test whether the industry differentials are statistically significant (using the *R*-squared form of the *F* test). Do not worry too much about dropping various variables to find the "best" combination of explanatory variables. As we mentioned earlier, this is a difficult and not even very well-defined task. Only if eliminating a set of variables substantially alters the magnitudes and/or significance of the coefficients of interest is this important. Dropping a group of variables to simplify the model—such as quadratics or interactions—can be justified via an *F* test.

If you have used at least two different methods—such as OLS and 2SLS, or levels and differencing for a time series, or pooled OLS versus differencing with a panel data

set—then you should comment on any critical differences. In particular, if OLS gives counterintuitive results, did using 2SLS or panel data methods improve the estimates?

Conclusions

This can be a short section that summarizes what you have learned. For example, you might want to present the magnitude of a coefficient that was of particular interest. The conclusion should also discuss caveats to the conclusions drawn, and it might even suggest directions for further research. It is useful to imagine readers turning first to the conclusion in order to decide whether to read the rest of the paper.

Style Hints

You should give your paper a title that reflects its topic. Papers should be typed and double-spaced. All equations should begin on a new line, and they should be centered and numbered consecutively, that is, (1), (2), (3), and so on. Large graphs and tables may be included after the main body. In the text, refer to papers by author and date, for example, White (1980). The reference section at the end of the paper should be done in standard format. Several examples are given in the references at the back of the text.

When you introduce an equation in the "Econometric Models" section, you should describe the important variables: the dependent variable and the key independent variable or variables. To focus on a single independent variable, you can write an equation, such as

$$GPA = \beta_0 + \beta_1 alcohol + \boldsymbol{x\delta} + u$$

or

$$\log(wage) = \beta_0 + \beta_1 educ + \boldsymbol{x\delta} + u,$$

where the notation $\boldsymbol{x\delta}$ is shorthand for several other explanatory variables. At this point, you need only describe them generally; they can be described specifically in the data section in a table. For example, in a study of the factors affecting chief executive officer salaries, you might include the following table in the data section:

Table 1: Variable Descriptions

salary:	annual salary (including bonuses) in 1990 (in thousands)
sales:	firm sales in 1990 (in millions)
roe:	average return on equity from 1988–1990 (in percent)
pcsal:	percentage change in salary from 1988–1990
pcroe:	percentage change in roe from 1988–1990

continued

Table 1: (*concluded*)

indust:	= 1 if an industrial company, 0 otherwise
finance:	= 1 if a financial company, 0 otherwise
consprod:	= 1 if a consumer products company, 0 otherwise
util:	= 1 if a utility company, 0 otherwise
ceoten:	number of years as CEO of the company

A table of summary statistics using the data set 401K.RAW, which we used for studying the factors that affect participation in 401(k) pension plans, might be set up as follows:

Table 2: Summary Statistics

Variable	Mean	Standard Deviation	Minimum	Maximum
prate	.869	.167	.023	1
mrate	.746	.844	.011	5
employ	4,621.01	16,299.64	53	443,040
age	13.14	9.63	4	76
sole	.415	.493	0	1
Number of Observations = 3,784				

In the results section, you can either write the estimates in equation form, as we often have done, or in a table. Especially when several models have been estimated with different sets of explanatory variables, tables are very useful. If you write out the estimates as an equation, for example,

$$\widehat{\log(salary)} = \underset{(0.93)}{2.45} + \underset{(.115)}{.236}\log(sales) + \underset{(.003)}{.008}\,roe + \underset{(.028)}{.061}\,ceoten$$
$$n = 204,\ R^2 = .351,$$

be sure to state near the first equation that standard errors are in parentheses. It is acceptable to report the t statistics for testing H_0: $\beta_j = 0$, or their absolute values, but it is most important to state what you are doing.

If you report your results in tabular form, make sure the dependent and independent variables are clearly indicated. Again, state whether standard errors or *t* statistics are below the coefficients (with the former preferred). Some authors like to use asterisks to indicate statistical significance at different significance levels (for example, one star means significant at 5%, two stars mean significant at 10% but not 5% and so on). This is not necessary if you carefully discuss the significance of the explanatory variables in the text.

A sample table of results follows:

Table 3: OLS Results

Dependent Variable: Participation Rate

Independent Variables			
mrate	.156 (.012)	.239 (.042)	.218 (.342)
$mrate^2$	———	−.087 (.043)	−.096 (.073)
log(*emp*)	−.112 (.014)	−.112 (.014)	−.098 (.111)
$\log(emp)^2$	.0057 (.0009)	.0057 (.0009)	.0052 (.0007)
age	.0060 (.0010)	.0059 (.0010)	.0050 (.0021)
age^2	−.00007 (.00002)	−.00007 (.00002)	−.00006 (.00002)
sole	−.0001 (.0058)	.0008 (.0058)	.0006 (.0061)
constant	1.213 (0.051)	.198 (.052)	.085 (.041)
industry dummies?	no	no	yes
Observations: *R*-Squared:	3,784 .143	3,784 .152	3,784 .152

Note: The quantities in parentheses below the estimates are the standard errors.

Your results will be easier to read and interpret if you choose the units of both your dependent and independent variables so that coefficients are not too large or too small. You should never report numbers such as 1.051e-007 or 3.524e+006 for your coefficients or standard errors, and you should not use scientific notation. If coefficients are either extremely small or large, rescale the dependent or independent variables, as we discussed in Chapter 6. You should limit the number of digits reported after the decimal point. For example, if your regression package estimates a coefficient to be .54821059, you should report this as .548, or even .55, in the paper.

As a general rule, the commands that your particular econometrics package uses to produce results should not appear in the paper; only the results are important. If some special command was used to carry out a certain estimation method, this can be given in an appendix. An appendix is also a good place to include extra results that support your analysis but are not central to it.

SUMMARY

In this chapter, we have discussed the ingredients of a successful empirical study and have provided hints that can improve the quality of an analysis. Ultimately, the success of any study depends crucially on the care and effort put into it.

KEY TERMS

Data Mining
Internet
On-Line Data Bases
On-Line Search Services
Sensitivity Analysis
Spreadsheet
Text Editor
Text (ASCII) File

SAMPLE EMPIRICAL PROJECTS

Throughout the text, we have seen examples of econometric analysis that either came from or were motivated by published works. Hopefully, these have given you a good idea about the scope of empirical analysis. We include the following list as additional examples of questions that others have found or are likely to find interesting. These are intended to stimulate your imagination; no attempt is made to fill in all of the details of specific models, data requirements, or alternative estimation methods. It should be possible to complete these projects in one term.

1. Do your own campus survey to answer a question of interest at your university. For example: What is the effect of working, on college GPA? You can ask students about high school GPA, college GPA, ACT or SAT scores, hours worked per week, participation in athletics, major, gender, race, and so on. Then, use these variables to create a model that explains GPA. How much of an effect, if any, does another hour worked per week have on GPA? One issue of concern is that hours worked might be endogenous: it might be correlated with unobserved factors that affect college GPA, or lower GPAs might cause students to work more.

A better approach would be to collect cumulative GPA prior to the semester and then to obtain GPA for the most recent semester, along with amount worked during that semester, and the other variables. Now, cumulative GPA could be used as a control (explanatory variable) in the equation.

2. There are many variants on the preceding topic. You can study the effects of drug or alcohol usage, or of living in a fraternity, on grade point average. You would want to control for many family background variables, as well as previous performance variables.
3. Do gun control laws at the city level reduce violent crimes? Such questions can be difficult to answer with a single cross section because city and state laws are often endogenous. [See Kleck and Patterson (1993) for an example. They used cross-sectional data and instrumental variables methods, but their IVs are questionable.] Panel data can be very useful for inferring causality in these contexts. At a minimum, you could control for a previous year's violent crime rate.
4. Low and McPheters (1983) used city cross-sectional data on wage rates and estimates of risk of death for police officers, along with other controls. The idea is to determine whether police officers are compensated for working in cities with a higher risk of on-the-job injury or death.
5. Do parental consent laws increase the teenage birth rate? You can use state-level data for this: either a time series for a given state or, even better, a panel data set of states. Do the same laws reduce abortion rates among teenagers? The *Statistical Abstract of the United States* contains all kinds of state-level data. Levine, Trainor, and Zimmerman (1996) studied the effects of abortion funding restrictions on similar outcomes. Other factors, such as access to abortions, may affect teen birth and abortion rates.
6. Do changes in traffic laws affect traffic fatalities? McCarthy (1994) contains an analysis of monthly time series data for the state of California. A set of dummy variables can be used to indicate the months in which certain laws were in effect. The file TRAFFIC2.RAW contains the data used by McCarthy. An alternative is to obtain a panel data set on states in the United States, where you can exploit variation in laws across states, as well as across time. (See the file TRAFFIC1.RAW.)

   Mullahy and Sindelar (1994) used individual-level data matched with state laws and taxes on alcohol to estimate the effects of laws and taxes on the probability of driving drunk.
7. Are blacks discriminated against in the lending market? Hunter and Walker (1996) looked at this question; in fact, we used their data in Exercises 7.16 and 17.9.
8. Is there a marriage premium for professional athletes? Korenman and Neumark (1991) found a significant wage premium for married men after using a variety of econometric methods. Professional athletes—such as National Basketball Association players, major league baseball players, and professional golfers—provide an interesting group in which to study the marriage premium because we can observe several productivity measures. With players in individual sports, such as golf or tennis, earnings directly reflect productivity. In team sports, salary may not entirely reflect productivity—for example, years in the

league might matter. So we can include a marriage indicator in an equation with something like scoring as the dependent variable, as well as in a regression where log(*salary*) is the dependent variable and several productivity controls are among the independent variables.

9. Answer the question: Are cigarette smokers less productive? A variant on this is: Do workers who smoke take more sick days (everything else being equal)? Mullahy and Portney (1990) use individual-level data to evaluate this question. You could use data at, say, the metropolitan level. Something like average productivity in manufacturing can be related to percent of manufacturing workers who smoke. Other variables, such as average worker education, capital per worker, and size of the city (you can think of more) should be controlled for.
10. Do minimum wages alleviate poverty? You can use state or county data to answer this question. The idea is that the minimum wage varies across state because some states have higher minimums than the federal minimum. Further, there are changes over time in the nominal minimum within a state, some due to changes at the federal level and some because of changes at the state level. Neumark and Wascher (1995) used a panel data set on states to estimate the effects of the minimum wage on the employment rates of young workers, as well as on school enrollment rates.
11. What factors affect student performance at public schools? It is fairly easy to get school-level or at least district-level data in most states. Does spending per student matter? Do student-teacher ratios have any effects? It is difficult to estimate ceteris paribus effects because spending is related to other factors, such as family incomes or poverty rates. The data set MEAP93.RAW, for Michigan high schools, contains a measure of the poverty rates. Another possibility is to use panel data, or to at least control for a previous year's performance measure (such as average test score or percentage of students passing an exam).

    You can look at less obvious factors that affect student performance. For example, after controlling for income, does family structure matter? Perhaps families with two parents, but only one working for a wage, have a positive effect on performance. (There could be at least two channels: parents spend more time with the children, and they might also volunteer at school.) What about the effect of single-parent households, controlling for income and other factors? You can merge census data for one or two years with school district data.

    Do public schools with more private schools nearby better educate their students because of competition? There is a tricky simultaneity issue here because private schools are probably located in areas where the public schools are already poor. Hoxby (1994) used an instrumental variables approach, where population proportions of various religions were IVs for the number of private schools.

    Rouse (1998) studied a different question: Did students who were able to attend a private school due to the Milwaukee voucher program perform better than those who did not? She used panel data and was able to control for an unobserved student effect.

12. Can excess returns on a stock, or a stock index, be predicted by the lagged price/dividend ratio? Or, by lagged interest rates or weekly monetary policy? It would be interesting to pick a foreign stock index, or one of the less well-known U.S. indexes. Cochrane (1997) contains a nice survey of recent theories and empirical results for explaining excess stock returns.
13. Is there racial discrimination in the market for baseball cards? This involves relating the prices of baseball cards to factors that should affect their prices, such as career statistics, whether the player is in the Hall of Fame, and so on. Holding other factors fixed, do cards of black or Hispanic players sell at a discount?
14. You can test whether the market for gambling on sports is efficient. For example, does the spread on football or basketball games contain all usable information for picking against the spread? The data set PNTSPRD.RAW contains information on men's college basketball games. The outcome variable is binary. Was the spread covered or not? Then, you can try to find information that was known prior to each game's being played in order to predict whether the spread is covered. (Good luck!)
15. What effect, if any, does success in college athletics have on other aspects of the university (applications, quality of students, quality of nonathletic departments)? McCormick and Tinsley (1987) looked at the effects of athletic success at major colleges on changes in SAT scores of entering freshman. Timing is important here: presumably, it is recent past success that affects current applications and student quality. One must control for many other factors—such as tuition and measures of school quality—to make the analysis convincing because, without controlling for other factors, there is a negative correlation between academics and athletic performance.

    A variant is to match up natural rivals in football or men's basketball and to look at differences across school as a function of which school won the football game or one or more basketball games. ATHLET1.RAW and ATHLET2.RAW are small data sets that could be expanded and updated.
16. Collect murder rates for a sample of cities or counties (say, from the FBI uniform crime reports) for two years. Make the latter year such that economic and demographic variables are easy to obtain from the County and City Data Book. From the Statistical Abstract of the United States, you can obtain the total number of people on death row, plus executions for intervening years at the state level. If the years are 1990 and 1985, you might estimate

$$mrdrte_{90} = \beta_0 + \beta_1 mrdrte_{85} + \beta_2 executions + other\ factors,$$

    where interest is in the coefficient on *executions*. The lagged murder rate and other factors serve as controls.

    Other factors may also act as a deterrent to crime. For example, Cloninger (1991) presented a cross-sectional analysis of the effects of lethal police response on crime rates.

    As a different twist, what factors affect crime rates on college campuses? Does the fraction of students living in fraternities or sororities have an effect? Does the size of the police force matter, or the kind of policing used? (Be care-

ful about inferring causality here.) Does having an escort program help reduce crime? What about crime rates in nearby communities? Recently, colleges and universities have been required to report crime statistics; in previous years, reporting was voluntary.

17. What factors affect manufacturing productivity at the state level? In addition to levels of capital and worker education, you could look at degree of unionization. A panel data analysis would be most convincing here, using two census years (say 1980 and 1990). Clark (1984) provides an analysis of how unionization affects firm performance and productivity. What other variables might explain productivity?

    Firm-level data can be obtained from *Compustat*. For example, other factors being fixed, do changes in unionization affect stock price of a firm?
18. Use state- or county-level data or, if possible, school district-level data to look at the factors that affect education spending per pupil. An interesting question is: Other things being equal (such as income and education levels of residents), do districts with a larger percentage of elderly people spend less on schools? Census data can be matched with school district spending data to obtain a very large cross section. The U.S. Department of Education compiles such data.
19. What are the effects of state regulations, such as motorcycle helmet laws, on motorcycle fatalities? Or, do differences in boating laws—such as minimum operating age—help to explain boating accident rates? The U.S. Department of Transportation compiles such information. This can be merged with data from the Statistical Abstract of the United States. A panel data analysis seems to be warranted here.
20. What factors affect output growth? Two factors of interest are inflation and investment [for example, Blomström, Lipsey, and Zejan (1996)]. You might use time series data on a country you find interesting. Or, you could use a cross section of countries, as in De Long and Summers (1991). Friedman and Kuttner (1992) found evidence that, at least in the 1980s, the spread between the commercial paper rate and the treasury bill rate affects real output.
21. What is the behavior of mergers in the U.S. economy (or some other economy)? Shughart and Tollison (1984) characterize (the log of) annual mergers in the U.S. economy as a random walk by showing that the difference in logs—roughly, the growth rate—is unpredictable given past growth rates. Does this still hold? Does it hold across various industries? What past measures of economic activity can be used to forecast mergers?
22. What factors might explain racial and gender differences in employment and wages? For example, Holzer (1991) reviewed the evidence on the "spatial mismatch hypothesis" to explain differences in employment rates between blacks and whites. Korenman and Neumark (1992) examined the effects of childbearing on women's wages, while Hersch and Stratton (1997) looked at the effects of household responsibilities on men's and women's wages.
23. Obtain monthly or quarterly data on teenage employment rates, the minimum wage, and factors that affect teen employment, to estimate the effects of the minimum wage on teen employment. Solon (1985) used quarterly U.S. data, while Castillo-Freeman and Freeman (1992) used annual data on Puerto Rico.

It might be informative to analyze time series data on a low-wage state in the United States—where changes in the minimum wage are likely to have the largest effect.

24. At the city level, estimate a time series model for crime. An example is Cloninger and Sartorius (1979). As a recent twist, you might estimate the effects of community policing or midnight basketball programs, relatively new innovations in fighting crime. Inferring causality is tricky. Including a lagged dependent variable might be helpful. Because you are using time series data, you should be aware of the spurious regression problem.

    Grogger (1990) used data on daily homicide counts to estimate the deterrent effects of capital punishment. Might there be other factors—such as news on lethal response by police—that have an effect on daily crime counts?
25. Are there aggregate productivity effects of computer usage? You would need to obtain time series data, perhaps at the national level, on productivity, percentage of employees using computers, and other factors. What about spending (probably as a fraction of total sales) on research and development? What sociological factors might affect productivity? alcohol usage? divorce rates?
26. What factors affect chief executive officer salaries? The files CEOSAL1.RAW and CEOSAL2.RAW are data sets that have various firm performance measures, as well as information such as tenure and education. You can certainly update these data files and look for other interesting factors. Rose and Shepard (1997) considered firm diversification as one important determinant of CEO compensation.
27. Do differences in tax codes across states affect the amount of foreign direct investment? Hines (1996) studied the effects of state corporate taxes, along with the ability to apply foreign tax credits, on investment from outside the United States.
28. What factors affect election outcomes? Does spending matter? Do votes on specific issues matter? Does the state of the local economy matter? See, for example, Levitt (1994) and the data sets VOTE1.RAW and VOTE2.RAW. Fair (1996) performed a time series analysis of U.S. presidential elections.

LIST OF JOURNALS

The following is a partial list of popular journals containing research in empirical business, economics, and other social sciences. A complete set of journals can be found on the Internet.

American Economic Review
American Journal of Agricultural Economics
American Political Science Review
Applied Economics
Brookings Papers on Economic Activity
Canadian Journal of Economics
Demography
Economic Inquiry

Economica
Economics Letters
Empirical Economics
Federal Reserve Bulletin
International Economic Review
Journal of Applied Econometrics
Journal of Business and Economic Statistics
Journal of Development Economics
Journal of Economic Education
Journal of Empirical Finance
Journal of Environmental Economics and Management
Journal of Finance
Journal of Health Economics
Journal of Human Resources
Journal of Industrial Economics
Journal of International Economics
Journal of Labor Economics
Journal of Political Economy
Journal of Public Economics
Journal of Monetary Economics
Journal of Money, Credit, and Banking
Journal of Quantitative Criminology
Journal of Urban Economics
National Bureau of Economic Research Working Paper Series
National Tax Journal
Public Finance Quarterly
Quarterly Journal of Economics
Regional Science & Urban Economics
Review of Economic Studies
Review of Economics and Statistics

DATA SOURCES

There are numerous data sources available throughout the world. Governments of most countries compile a wealth of data; some general and easily accessible data sources for the United States, such as the *Economic Report of the President*, the *Statistical Abstract of the United States*, and the *County and City Data Book*, have already been mentioned. International financial data on many countries are published annually in *International Financial Statistics*. Various magazines, like *Business Week* and *U.S. News and World Report*, often publish statistics—such as CEO salaries and firm performance, or ranking of academic programs—that are novel and can be used in an econometric analysis.

Rather than attempting to provide a list here, we instead give some Internet addresses that are comprehensive sources for economists. A very useful site for economists, called *Resources for Economists on the Internet*, is maintained by Bill Goffe at the University of Southern Mississippi. The address is

http://econwpa.wustl.edu/EconFAQ/EconFAQ.html.

This site provides links to journals, data sources, and lists of professional and academic economists. It is quite simple to use.

The Business and Economic Statistics section of the American Statistical Association contains an extremely detailed list of data sources and provides links to them. The address is

http://www.econ-datalinks.org.

In addition, the *Journal of Applied Econometrics* and the *Journal of Business and Economics Statistics* have data archives that contain data sets used in most papers published in the journals over the past several years. If you find a data set that interests you, this is a good way to go, as much of the cleaning and formatting of the data have already been done. The downside is that some of these data sets are used in econometric analyses that are more advanced than we have learned about in this text. On the other hand, it is often useful to estimate simpler models using standard econometric methods for comparison.

Many universities, such as the University of California, Berkeley, the University of Michigan, and the University of Maryland, maintain very extensive data sets as well as links to a variety of data sets. Your own library possibly contains an extensive set of links to data bases in business, economics, and the other social sciences. The regional federal reserve banks, such as the one in St. Louis, manage a variety of data. The National Bureau of Economic Research posts data sets used by some of its researchers. Naturally, state and federal governments now publish a wealth of data that can be accessed via the Internet. Census data are publicly available from the Department of Census. (Two useful publications are the *Census of Manufacturing*, published in years ending with two and seven, and the *Census of the Population*, published at the beginning of each decade.) Other agencies, such as the Department of Justice, also make data available to the public.

Appendix A

Basic Mathematical Tools

This appendix covers some basic mathematics that are used in econometric analysis. We summarize various properties of the summation operator, study properties of linear and certain nonlinear equations, and review proportions and percents. We also present some special functions that often arise in applied econometrics, including quadratic functions and the natural logarithm. The first four sections require only basic algebra skills. Section A.5 contains a brief review of differential calculus; while a knowledge of calculus is not necessary to understand most of the text, it is used in some end-of-chapter appendices and in several of the more advanced chapters in Part III.

A.1 THE SUMMATION OPERATOR AND DESCRIPTIVE STATISTICS

The **summation operator** is a useful shorthand for manipulating expressions involving the sums of many numbers, and it plays a key role in statistics and econometric analysis. If $\{x_i: i = 1, \ldots, n\}$ denotes a sequence of n numbers, then we write the sum of these numbers as

$$\sum_{i=1}^{n} x_i \equiv x_1 + x_2 + \ldots + x_n. \tag{A.1}$$

With this definition, the summation operator is easily shown to have the following properties:

PROPERTY SUM. 1: For any constant c,

$$\sum_{i=1}^{n} c = nc. \tag{A.2}$$

PROPERTY SUM. 2: For any constant c,

$$\sum_{i=1}^{n} cx_i = c\sum_{i=1}^{n} x_i. \tag{A.3}$$

PROPERTY SUM. 3: If $\{(x_i,y_i): i = 1,2,\ldots,n\}$ is a set of n pairs of numbers, and a and b are constants, then

$$\sum_{i=1}^{n} (ax_i + by_i) = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} y_i. \qquad \textbf{(A.4)}$$

It is also important to be aware of some things that *cannot* be done with the summation operator. Let $\{(x_i,y_i): i = 1,2,\ldots,n\}$ again be a set of n pairs of numbers with $y_i \neq 0$ for each i. Then,

$$\sum_{i=1}^{n} (x_i/y_i) \neq \left(\sum_{i=1}^{n} x_i\right)\Big/\left(\sum_{i=1}^{n} y_i\right).$$

In other words, the sum of the ratios is not the ratio of the sums. In the $n = 2$ case, the application of familiar elementary algebra also reveals this lack of equality: $x_1/y_1 + x_2/y_2 \neq (x_1 + x_2)/(y_1 + y_2)$. Similarly, the sum of the squares is not the square of the sum: $\sum_{i=1}^{n} x_i^2 \neq \left(\sum_{i=1}^{n} x_i\right)^2$, except in special cases. That these two quantities are not generally equal is easiest to see when $n = 2$: $x_1^2 + x_2^2 \neq (x_1 + x_2)^2 = x_1^2 + 2x_1x_2 + x_2^2$.

Given n numbers $\{x_i: i = 1,\ldots,n\}$, we compute their **average** or *mean* by adding them up and dividing by n:

$$\bar{x} = (1/n)\sum_{i=1}^{n} x_i. \qquad \textbf{(A.5)}$$

When the x_i are a sample of data on a particular variable (such as years of education), we often call this the *sample average* (or *sample mean*) to emphasize that it is computed from a particular set of data. The sample average is an example of a **descriptive statistic**; in this case, the statistic describes the central tendency of the set of points x_i.

There are some basic properties about averages that are important to understand. First, suppose we take each observation on x and subtract off the average: $d_i \equiv x_i - \bar{x}$ (the "d" here stands for *deviation* from the average). Then the sum of these deviations is always zero:

$$\sum_{i=1}^{n} d_i = \sum_{i=1}^{n} (x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} = \sum_{i=1}^{n} x_i - n\bar{x} = n\bar{x} - n\bar{x} = 0.$$

We summarize this as

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0. \qquad \textbf{(A.6)}$$

A simple numerical example shows how this works. Suppose $n = 5$ and $x_1 = 6$, $x_2 = 1$, $x_3 = -2$, $x_4 = 0$, and $x_5 = 5$. Then $\bar{x} = 2$, and the demeaned sample is $\{4,-1,-4,-2,3\}$. Adding these up gives zero, which is just what equation (A.6) says.

In our treatment of regression analysis in Chapter 2, we need to know some additional algebraic facts involving deviations from sample averages. An important one is

that the sum of squared deviations is the sum of the squared x_i minus n times the square of $\bar{x}$:

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n(\bar{x})^2. \tag{A.7}$$

This can be shown using basic properties of the summation operator:

$$\begin{aligned}\sum_{i=1}^{n} (x_i - \bar{x})^2 &= \sum_{i=1}^{n} (x_i^2 - 2x_i\bar{x} + \bar{x}^2) \\ &= \sum_{i=1}^{n} x_i^2 - 2\bar{x} \sum_{i=1}^{n} x_i + n(\bar{x})^2 \\ &= \sum_{i=1}^{n} x_i^2 - 2n(\bar{x})^2 + n(\bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n(\bar{x})^2.\end{aligned}$$

Given a data set on two variables, $\{(x_i,y_i): i = 1,2,\ldots,n\}$, it can also be shown that

$$\begin{aligned}\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) &= \sum_{i=1}^{n} x_i(y_i - \bar{y}) \\ = \sum_{i=1}^{n} (x_i - \bar{x})y_i &= \sum_{i=1}^{n} x_i y_i - n(\bar{x}\cdot\bar{y});\end{aligned} \tag{A.8}$$

this is a generalization of equation (A.7) (there, $y_i = x_i$ for all i).

The average is the measure of central tendency that we will focus on in most of this text. However, it is sometimes informative to use the **median** (or *sample median*) to describe the central value. To obtain the median of the n numbers $\{x_1,\ldots,x_n\}$, we first order the values of the x_i from smallest to largest. Then, if n is odd, the sample median is the middle number of the ordered observations. For example, given the numbers $\{-4,8,2,0,21,-10,18\}$, the median value is 2 (since the ordered sequence is $\{-10,-4,0,2,8,18,21\}$). If we change the largest number in this list, 21, to twice its value, 42, the median is still 2. By contrast, the sample average would increase from 5 to 8, a sizable change. Generally, the median is less sensitive than the average to changes in the extreme values (large or small) in a list of numbers. This is why "median incomes" or "median housing values" are often reported, rather than averages, when summarizing income or housing values in a city or county.

If n is even, there is no unique way to define the median because there are two numbers at the center. Usually the median is defined to be the average of the two middle values (again, after ordering the numbers from smallest to largest). Using this rule, the median for the set of numbers $\{4,12,2,6\}$ would be $(4 + 6)/2 = 5$.

A.2 PROPERTIES OF LINEAR FUNCTIONS

Linear functions play an important role in econometrics because they are simple to interpret and manipulate. If x and y are two variables related by

$$y = \beta_0 + \beta_1 x, \tag{A.9}$$

then we say that y is a **linear function** of x, and β_0 and β_1 are two parameters (numbers) describing this relationship. The **intercept** is β_0, and the **slope** is β_1.

The defining feature of a linear function is that the change in y is always β_1 times the change in x:

$$\Delta y = \beta_1 \Delta x, \tag{A.10}$$

where Δ denotes "change." In other words, the **marginal effect** of x on y is constant and equal to β_1.

EXAMPLE A.1
(Linear Housing Expenditure Function)

Suppose that the relationship between monthly housing expenditure and monthly income is

$$housing = 164 + .27\, income. \tag{A.11}$$

Then, for each additional dollar of income, 27 cents is spent on housing. If family income increases by \$200, then housing expenditure increases by (.27)200 = \$54. This function is graphed in Figure A.1.

According to equation (A.11), a family with no income spends \$164 on housing, which of course cannot be literally true. For low levels of income, this linear function would not describe the relationship between *housing* and *income* very well, which is why we will eventually have to use other types of functions to describe such relationships.

In (A.11), the *marginal propensity to consume* (MPC) housing out of income is .27. This is different from the *average propensity to consume* (APC), which is

$$\frac{housing}{income} = 164/income + .27.$$

The APC is not constant, it is always larger than the MPC, and it gets closer to the MPC as income increases.

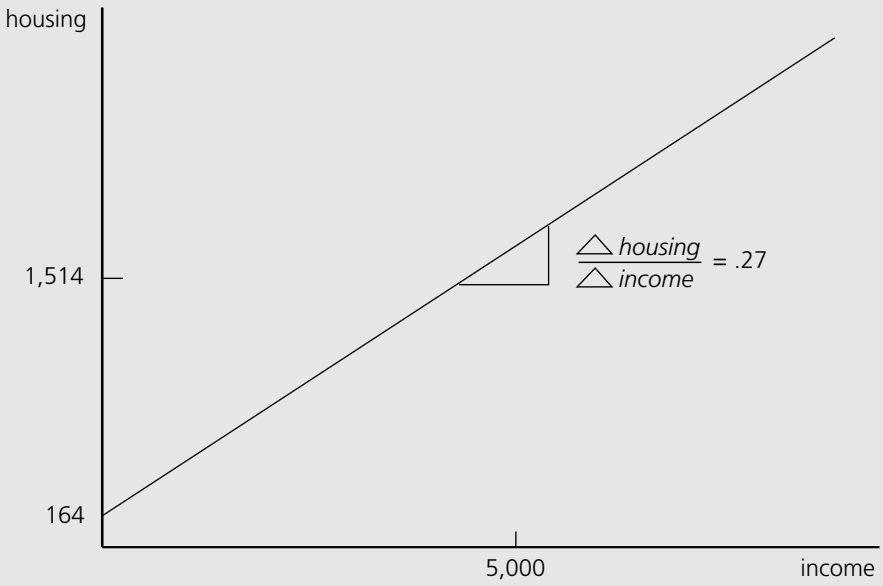
Linear functions are easily defined for more than two variables. Suppose that y is related to two variables, x_1 and x_2, in the general form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2. \tag{A.12}$$

It is rather difficult to envision this function because its graph is three-dimensional. Nevertheless, β_0 is still the intercept (the value of y when $x_1 = 0$ and $x_2 = 0$), and β_1 and β_2 measure particular slopes. From (A.12), the change in y, for given changes in x_1 and x_2, is

Figure A.1

Graph of *housing* = 164 + .27 *income*.



$$\Delta y = \beta_1 \Delta x_1 + \beta_2 \Delta x_2. \qquad \textbf{(A.13)}$$

If x_2 does not change, that is, $\Delta x_2 = 0$, then we have

$$\Delta y = \beta_1 \Delta x_1 \text{ if } \Delta x_2 = 0,$$

so that β_1 is the slope of the relationship in the direction of x_1:

$$\beta_1 = \frac{\Delta y}{\Delta x_1} \text{ if } \Delta x_2 = 0.$$

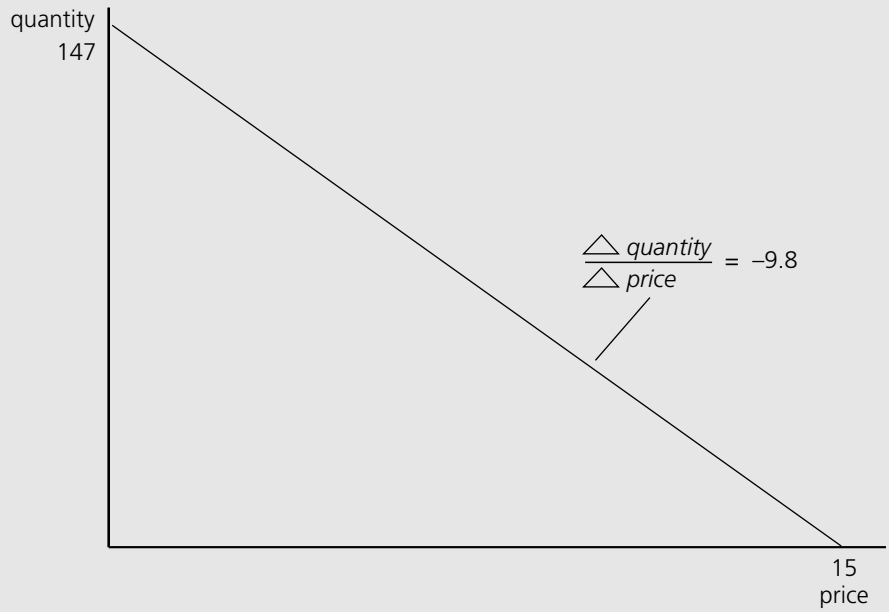
Because it measures how y changes with x_1, holding x_2 fixed, β_1 is often called the **partial effect** of x_1 on y. Since the partial effect involves holding other factors fixed, it is closely linked to the notion of **ceteris paribus**. The parameter β_2 has a similar interpretation: $\beta_2 = \Delta y/\Delta x_2$ if $\Delta x_1 = 0$, so that β_2 is the partial effect of x_2 on y.

EXAMPLE A.2
(Demand for Compact Discs)

For college students, suppose that the monthly quantity demanded of compact discs is related to the price of compact discs and monthly discretionary income by

Figure A.2

Graph of *quantity* = 120 − 9.8 *price* + .03 *income*, with *income* fixed at $900.



$$quantity = 120 - 9.8\, price + .03\, income,$$

where *price* is dollars per disk and *income* is measured in dollars. The *demand curve* is the relationship between *quantity* and *price*, holding *income* (and other factors) fixed. This is graphed in two dimensions in Figure A.2 at an income level of $900. The slope of the demand curve, −9.8, is the *partial* effect of price on quantity: holding income fixed, if the price of compact discs increases by one dollar, then the quantity demanded falls by 9.8. (We abstract from the fact that CDs can only be purchased in discrete units.) An increase in income simply shifts the demand curve up (changes the intercept), but the slope remains the same.

A.3 PROPORTIONS AND PERCENTAGES

Proportions and percentages play such an important role in applied economics that it is necessary to become very comfortable in working with them. Many quantities reported in the popular press are in the form of percentages; a few examples include interest rates, unemployment rates, and high school graduation rates.

An important skill is being able to convert between proportions and percentages. A percentage is easily obtained by multiplying a proportion by 100. For example, if the proportion of adults in a county with a high school degree is .82, then we say that 82% (82 percent) of adults have a high school degree. Another way to think of percents and proportions is that a proportion is the decimal form of a percent. For example, if the marginal tax rate for a family earning $30,000 per year is reported as 28%, then the proportion of the next dollar of income that is paid in income taxes is .28 (or 28 cents).

When using percentages, we often need to convert them to decimal form. For example, if a state sales tax is 6% and $200 is spent on a taxable item, then the sales tax paid is 200(.06) = 12 dollars. If the annual return on a certificate of deposit (CD) is 7.6% and we invest $3,000 in such a CD at the beginning of the year, then our interest income is 3,000(.076) = 228 dollars. As much as we would like it, the interest income is not obtained by multiplying 3,000 by 7.6.

We must be wary of proportions that are sometimes incorrectly reported as percentages in the popular media. If we read, "The percentage of high school students who drink alcohol is .57," we know that this really means 57% (not just over one-half of a percent, as the statement literally implies). College volleyball fans are probably familiar with press clips containing statements such as "Her hitting percentage was .372." This really means that her hitting percentage was 37.2%.

In econometrics, we are often interested in measuring the *changes* in various quantities. Let x denote some variable, such as an individual's income, the number of crimes committed in a community, or the profits of a firm. Let x_0 and x_1 denote two values for x: x_0 is the initial value, and x_1 is the subsequent value. For example, x_0 could be the annual income of an individual in 1994 and x_1 the income of the same individual in 1995. The **proportionate change** in x in moving from x_0 to x_1 is simply

$$(x_1 - x_0)/x_0 = \Delta x/x_0, \tag{A.14}$$

assuming, of course, that $x_0 \neq 0$. In other words, to get the proportionate change, we simply divide the change in x by its initial value. This is a way of standardizing the change so that it is free of units. For example, if an individual's income goes from $30,000 per year to $36,000 per year, then the proportionate change is 6,000/30,000 = .20.

It is more common to state changes in terms of percentages. The **percentage change** in x in going from x_0 to x_1 is simply 100 times the proportionate change:

$$\%\Delta x = 100(\Delta x/x_0); \tag{A.15}$$

the notation "$\%\Delta x$" is read as "the percentage change in x." For example, when income goes from $30,000 to $33,750, income has increased by 12.5%; to get this, we simply multiply the proportionate change, .125, by 100.

Again, we must be on guard for proportionate changes that are reported as percentage changes. In the previous example, for instance, reporting the percentage change in income as .125 is incorrect and could lead to confusion.

When we look at changes in things like dollar amounts or population, there is no ambiguity about what is meant by a percentage change. By contrast, interpreting per-

centage change calculations can be tricky when the variable of interest is itself a percentage, something that happens often in economics and other social sciences. To illustrate, let x denote the percentage of adults in a particular city having a college education. Suppose the initial value is $x_0 = 24$ (24% have a college education), and the new value is $x_1 = 30$. There are two quantities we can compute to describe how the percentage of college-educated people has changed. The first is the change in x, Δx. In this case, $\Delta x = x_1 - x_0 = 6$: the percentage of people with a college education has increased by six *percentage points*. On the other hand, we can compute the percentage change in x using equation (A.15): $\%\Delta x = 100[(30 - 24)/24] = 25$.

In this example, the percentage point change and the percentage change are very different. The **percentage point change** is just the change in the percentages. The percentage change is the change relative to the initial value. Generally, we must pay close attention to which number is being computed. The careful researcher makes this distinction perfectly clear; unfortunately, in the popular press as well as in academic research, the type of reported change is often unclear.

EXAMPLE A.3
(Michigan Sales Tax Increase)

In March 1994, Michigan voters approved a sales tax increase from 4% to 6%. In political advertisements, supporters of the measure referred to this as a two percentage point increase, or an increase of two cents on the dollar. Opponents to the tax increase called it a 50% increase in the sales tax rate. Both claims are correct; they are simply different ways of measuring the increase in the sales tax. Naturally, each group reported the measure that made their position most favorable.

For a variable such as salary, it makes no sense to talk of a "percentage point change in salary" because salary is not measured as a percentage. We can describe a change in salary either in dollar or percentage terms.

A.4 SOME SPECIAL FUNCTIONS AND THEIR PROPERTIES

In Section A.2, we reviewed the basic properties of linear functions. We already indicated one important feature of functions like $y = \beta_0 + \beta_1 x$: a one-unit change in x results in the *same* change in y, regardless of the initial value of x. As we noted earlier, this is the same as saying the marginal effect of x on y is constant, something that is not realistic for many economic relationships. For example, the important economic notion of *diminishing marginal returns* is not consistent with a linear relationship.

In order to model a variety of economic phenomena, we need to study several nonlinear functions. A **nonlinear function** is characterized by the fact that the change in y for a given change in x depends on the starting value of x. Certain nonlinear functions appear frequently in empirical economics, so it is important to know how to interpret them. A complete understanding of nonlinear functions takes us into the realm of cal-

culus. Here, we simply summarize the most significant aspects of the functions, leaving the details of some derivations for Section A.5.

Quadratic Functions

One simple way to capture diminishing returns is to add a quadratic term to a linear relationship. Consider the equation

$$y = \beta_0 + \beta_1 x + \beta_2 x^2, \tag{A.16}$$

where β_0, β_1, and β_2 are parameters. When $\beta_1 > 0$ and $\beta_2 < 0$, the relationship between y and x has the parabolic shape given in Figure A.3, where $\beta_0 = 6$, $\beta_1 = 8$, and $\beta_2 = -2$.

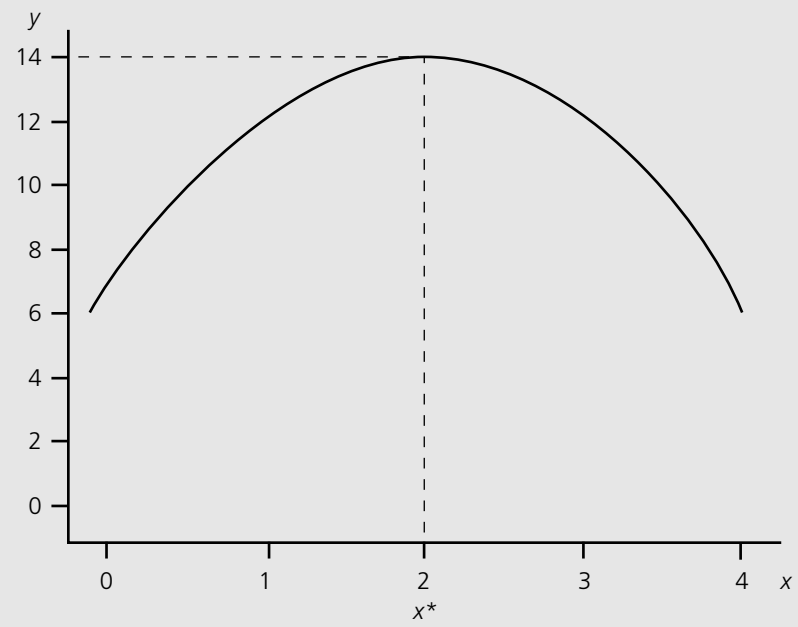
When $\beta_1 > 0$ and $\beta_2 < 0$, it can be shown (using calculus in the next section) that the *maximum* of the function occurs at the point

$$x^* = \beta_1/(-2\beta_2). \tag{A.17}$$

For example, if $y = 6 + 8x - 2x^2$ (so $\beta_1 = 8$, $\beta_2 = -2$), then the largest value of y occurs at $x^* = 8/4 = 2$, and this value is $6 + 8(2) - 2(2)^2 = 14$ (see Figure A.3).

Figure A.3

Graph of $y = 6 + 8x - 2x^2$.



The fact that equation (A.16) implies a **diminishing marginal effect** of x on y is easily seen from its graph. Suppose we start at a low value of x and then increase x by some amount, say c. This has a larger effect on y than if we start at a higher value of x and increase x by the same amount c. In fact, once $x > x^*$, an increase in x actually decreases y.

The statement that x has a diminishing marginal effect on y is the same as saying that the slope of the function in Figure A.3 decreases as x increases. While this is clear from looking at the graph, we usually want to quantify how quickly the slope is changing. An application of calculus gives the approximate slope of the quadratic function as

$$slope = \frac{\Delta y}{\Delta x} \approx \beta_1 + 2\beta_2 x, \tag{A.18}$$

for "small" changes in x. [The right-hand side of equation (A.18) is the **derivative** of the function in equation (A.16) with respect to x.] Another way to write this is

$$\Delta y \approx (\beta_1 + 2\beta_2 x)\Delta x \text{ for "small" } \Delta x. \tag{A.19}$$

To see how well this approximation works, consider again the function $y = 6 + 8x - 2x^2$. Then, according to equation (A.19), $\Delta y \approx (8 - 4x)\Delta x$. Now, suppose we start at $x = 1$ and change x by $\Delta x = .1$. Using (A.19), $\Delta y \approx (8 - 4)(.1) = .4$. Of course, we can compute the change exactly by finding the values of y when $x = 1$ and $x = 1.1$: $y_0 = 6 + 8(1) - 2(1)^2 = 12$ and $y_1 = 6 + 8(1.1) - 2(1.1)^2 = 12.38$, and so the exact change in y is .38. The approximation is pretty close in this case.

Now, suppose we start at $x = 1$ but change x by a larger amount: $\Delta x = .5$. Then, the approximation gives $\Delta y \approx 4(.5) = 2$. The exact change is determined by finding the difference in y when $x = 1$ and $x = 1.5$. The former value of y was 12, and the latter value is $6 + 8(1.5) - 2(1.5)^2 = 13.5$, so the actual change is 1.5 (not 2). The approximation is worse in this case because the change in x is larger.

For many applications, equation (A.19) can be used to compute the approximate marginal effect of x on y for any initial value of x and small changes. And, we can always compute the exact change if necessary.

EXAMPLE A.4
(A Quadratic Wage Function)

Suppose the relationship between hourly wages and years in the work force (*exper*) is given by

$$wage = 5.25 + .48\, exper - .008\, exper^2. \tag{A.20}$$

This function has the same general shape as the one in Figure A.3. Using equation (A.17), *exper* has a positive effect on wage up to the turning point, $exper^* = .48/[2(.008)] = 30$. The first year of experience is worth approximately .48, or 48 cents [see (A.19) with $x = 0$, $\Delta x = 1$]. Each additional year of experience increases wage by less than the previous

year—reflecting a diminishing marginal return to experience. At 30 years, an additional year of experience would actually lower the wage. This is not very realistic, but it is one of the consequences of using a quadratic function to capture a diminishing marginal effect: at some point, the function must reach a maximum and curve downward. For practical purposes, the point at which this happens is often large enough to be inconsequential, but not always.

The graph of the quadratic function in (A.16) has a U-shape if $\beta_1 < 0$ and $\beta_2 > 0$, in which case there is an increasing marginal return. The minimum of the function is at the point $-\beta_1/(2\beta_2)$.

The Natural Logarithm

The nonlinear function that plays the most important role in econometric analysis is the **natural logarithm**. In this text, we denote the natural logarithm, which we often refer to simply as the **log function**, as

$$y = \log(x). \tag{A.21}$$

You might remember learning different symbols for the natural log; $\ln(x)$ or $\log_e(x)$ are the most common. These different notations are useful when logarithms with several different bases are being used. For our purposes, only the natural logarithm is important, and so $\log(x)$ denotes the natural logarithm throughout this text. This corresponds to the notation usage in many statistical packages, although some use $\ln(x)$ [and most calculators use $\ln(x)$]. Economists use both $\log(x)$ and $\ln(x)$, which is useful to know when you are reading papers in applied economics.

The function $y = \log(x)$ is defined only for $x > 0$, and it is plotted in Figure A.4. It is not very important to know how the values of $\log(x)$ are obtained. For our purposes, the function can be thought of as a black box: we can plug in any $x > 0$ and obtain $\log(x)$ from a calculator or a computer.

Several things are apparent from Figure A.4. First, when $y = \log(x)$, the relationship between y and x displays diminishing marginal returns. One important difference between the log and the quadratic function in Figure A.3 is that when $y = \log(x)$, the effect of x on y never becomes negative: the slope of the function gets closer and closer to zero as x gets large, but the slope never quite reaches zero and certainly never becomes negative.

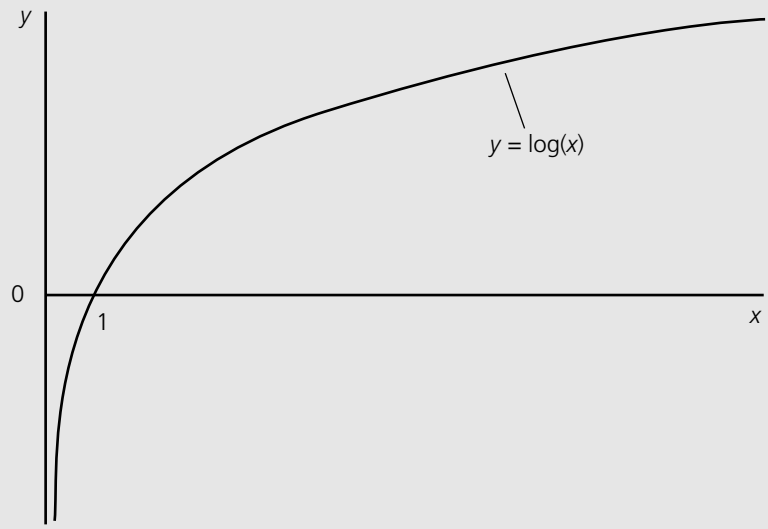
The following are also apparent from Figure A.4:

$$\log(x) < 0 \text{ for } 0 < x < 1$$
$$\log(1) = 0$$
$$\log(x) > 0 \text{ for } x > 1.$$

In particular, $\log(x)$ can be positive or negative. Some useful algebraic facts about the log function are

Figure A.4

Graph of $y = \log(x)$.



$$\log(x_1 \cdot x_2) = \log(x_1) + \log(x_2),\ x_1, x_2 > 0$$
$$\log(x_1/x_2) = \log(x_1) - \log(x_2),\ x_1, x_2 > 0$$
$$\log(x^c) = c\log(x),\ x > 0,\ c \text{ any number.}$$

Occasionally, we will need to rely on these properties.

The logarithm can be used for various approximations that arise in econometric applications. First, $\log(1 + x) \approx x$ for $x \approx 0$. You can try this with $x = .02$, .1, and .5 to see how the quality of the approximation deteriorates as x gets larger. Even more useful is the fact that the difference in logs can be used to approximate proportionate changes. Let x_0 and x_1 be positive values. Then, it can be shown (using calculus) that

$$\log(x_1) - \log(x_0) \approx (x_1 - x_0)/x_0 = \Delta x/x_0 \qquad \textbf{(A.22)}$$

for small changes in x. If we multiply equation (A.22) by 100 and write $\Delta\log(x) = \log(x_1) - \log(x_0)$, then

$$100 \cdot \Delta\log(x) \approx \%\Delta x \qquad \textbf{(A.23)}$$

for small changes in x. The meaning of small depends on the context, and we will encounter several examples throughout this text.

Why should we approximate the percentage change using (A.23) when the exact percentage change is so easy to compute? Momentarily, we will see why the approxi-

mation in (A.23) is useful in econometrics. First, let us see how good the approximation is in two examples.

First, suppose $x_0 = 40$ and $x_1 = 41$. Then, the percentage change in x in moving from x_0 to x_1 is 2.5%, using $100(x_1 - x_0)/x_0$. Now, $\log(41) - \log(40) = .0247$ to four digits, which when multiplied by 100 is very close to 2.5. The approximation works pretty well. Now, consider a much bigger change: $x_0 = 40$ and $x_1 = 60$. The exact percentage change is 50%. However, $\log(60) - \log(40) \approx .4055$, so the approximation gives 40.55%, which is much farther off.

Why is the approximation in (A.23) useful if it is only satisfactory for small changes? To build up to the answer, we first define the **elasticity** of y with respect to x as

$$\frac{\Delta y}{\Delta x} \cdot \frac{x}{y} = \frac{\%\Delta y}{\%\Delta x}. \tag{A.24}$$

In other words, the elasticity of y with respect to x is the percentage change in y, when x increases by 1%. This notion should be familiar from introductory economics.

If y is a linear function of x, $y = \beta_0 + \beta_1 x$, then the elasticity is

$$\frac{\Delta y}{\Delta x} \cdot \frac{x}{y} = \beta_1 \cdot \frac{x}{y} = \beta_1 \cdot \frac{x}{\beta_0 + \beta_1 x}, \tag{A.25}$$

which clearly depends on the value of x. (This is a generalization of the well-known result from basic demand theory: the elasticity is not constant along a straight-line demand curve.)

Elasticities are of critical importance in many areas of applied economics—not just in demand theory. It is convenient in many situations to have *constant* elasticity models, and the log function allows us to specify such models. If we use the approximation (A.23) for both x and y, then the elasticity is approximately equal to $\Delta\log(y)/\Delta\log(x)$. Thus, a constant elasticity model is approximated by the equation

$$\log(y) = \beta_0 + \beta_1 \log(x), \tag{A.26}$$

and β_1 is the elasticity of y with respect to x (assuming that $x, y > 0$).

EXAMPLE A.5
(Constant Elasticity Demand Function)

If q is quantity demanded and p is price, and these variables are related by

$$\log(q) = 4.7 - 1.25 \log(p),$$

then the price elasticity of demand is -1.25. Roughly, a 1% increase in price leads to a 1.25% fall in the quantity demanded.

For our purposes, the fact that β_1 in (A.26) is only close to the elasticity is not important. In fact, when the elasticity is defined using calculus—as in Section A.5—the definition is exact. For the purposes of econometric analysis, (A.26) defines a **constant elasticity model**. Such models play a large role in empirical economics.

There are other possibilities for using the log function that often arise in empirical work. Suppose that $y > 0$, and

$$\log(y) = \beta_0 + \beta_1 x. \tag{A.27}$$

Then $\Delta\log(y) = \beta_1 \Delta x$, so $100 \cdot \Delta\log(y) = (100 \cdot \beta_1)\Delta x$. It follows that, when y and x are related by equation (A.27),

$$\%\Delta y \approx (100 \cdot \beta_1)\Delta x. \tag{A.28}$$

EXAMPLE A.6
(Logarithmic Wage Equation)

Suppose that hourly wage and years of education are related by

$$\log(wage) = 2.78 + .094\, educ.$$

Then, using equation (A.28),

$$\%\Delta wage \approx 100(.094)\, \Delta educ = 9.4\, \Delta educ.$$

It follows that one more year of education increases hourly wage by about 9.4%.

Generally, the quantity $\%\Delta y/\Delta x$ is called the **semi-elasticity** of y with respect to x. The semi-elasticity is the percentage change in y when x increases by one *unit*. What we have just shown is that, in model (A.27), the semi-elasticity is constant and equal to $100 \cdot \beta_1$. In Example A.6, we can conveniently summarize the relationship between wages and education by saying that one more year of education—starting from any amount of education—increases the wage by about 9.4%. This is why such models play an important role in economics.

Another relationship of some interest in applied economics is:

$$y = \beta_0 + \beta_1 \log(x), \tag{A.29}$$

where $x > 0$. How can we interpret this equation? If we take the change in y, we get $\Delta y = \beta_1 \Delta\log(x)$, which can be rewritten as $\Delta y = (\beta_1/100)[100 \cdot \Delta\log(x)]$. Thus, using the approximation in (A.23), we have

$$\Delta y \approx (\beta_1/100)(\%\Delta x). \tag{A.30}$$

In other words, $\beta_1/100$ is the unit change in y when x increases by 1%.

EXAMPLE A.7
(Labor Supply Function)

Assume that the labor supply of a worker can be described by

$$hours = 33 + 45.1 \log(wage),$$

where *wage* is hourly wage and *hours* is hours worked per week. Then, from (A.30),

$$\Delta hours \approx (45.1/100)(\%\Delta wage) = .451\ \%\Delta wage.$$

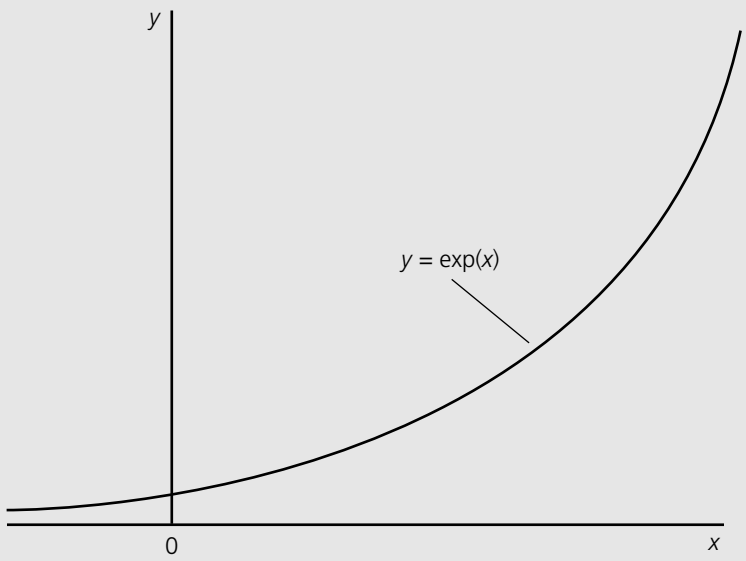
In other words, a 1% increase in *wage* increases the weekly hours worked by about .45, or slightly less than one-half of an hour. If the wage increases by 10%, then $\Delta hours = .451(10) = 4.51$, or about four and one-half hours. We would not want to use this approximation for much larger percentage changes in wages.

The Exponential Function

Before leaving this section, we need to discuss one more special function, one that is related to the log. As motivation, consider equation (A.27). There, $\log(y)$ is a linear function of x. But how do we find y itself as a function of x? The answer is given by the **exponential function**.

We will write the exponential function as $y = \exp(x)$, which is graphed in Figure A.5.

Figure A.5
Graph of $y = \exp(x)$.



From Figure A.5, we see that $\exp(x)$ is defined for any value of x and is always greater than zero. Sometimes the exponential function is written as $y = e^x$, but we will not use this notation. Two important values of the exponential function are $\exp(0) = 1$ and $\exp(1) = 2.7183$ (to four decimals).

The exponential function is the inverse of the log function in the following sense: $\log[\exp(x)] = x$ for all x, and $\exp[\log(x)] = x$ for $x > 0$. In other words, the log "undoes" the exponential, and vice versa. (This is why the exponential function is sometimes called the *anti-log* function.) In particular, note that $\log(y) = \beta_0 + \beta_1 x$ is equivalent to

$$y = \exp(\beta_0 + \beta_1 x).$$

If $\beta_1 > 0$, the relationship between x and y has the same shape as in Figure A.5. Thus, if $\log(y) = \beta_0 + \beta_1 x$ with $\beta_1 > 0$, then x has an *increasing* marginal effect on y. In Example A.6, this means that another year of education leads to a larger change in wage than the previous year of education.

Two useful facts about the exponential function are $\exp(x_1 + x_2) = \exp(x_1)\exp(x_2)$ and $\exp[c \cdot \log(x)] = x^c$.

A.5 DIFFERENTIAL CALCULUS

In the previous section, we asserted several approximations that have foundations in calculus. Let $y = f(x)$ for some function f. Then, for small changes in x,

$$\Delta y \approx \frac{df}{dx} \cdot \Delta x, \tag{A.31}$$

where df/dx is the derivative of the function f, evaluated at the initial point x_0. We also write the derivative as dy/dx.

For example, if $y = \log(x)$, then $dy/dx = 1/x$. Using (A.31), with dy/dx evaluated at x_0, we have $\Delta y \approx (1/x_0)\Delta x$, or $\Delta \log(x) \approx \Delta x/x_0$, which is the approximation given in (A.22).

In applying econometrics, it helps to recall the derivatives of a handful of functions because we use the derivative to define the slope of a function at a given point. We can then use (A.31) to find the approximate change in y for small changes in x. In the linear case, the derivative is simply the slope of the line, as we would hope: if $y = \beta_0 + \beta_1 x$, then $dy/dx = \beta_1$.

If $y = x^c$, then $dy/dx = cx^{c-1}$. The derivative of a sum of two functions is the sum of the derivatives: $d[f(x) + g(x)]/dx = df(x)/dx + dg(x)/dx$. The derivative of a constant times any function is that same constant times the derivative of the function: $d[cf(x)]/dx = c[df(x)/dx]$. These simple rules allow us to find derivatives of more complicated functions. Other rules, such as the product, quotient, and chain rules will be familiar to those who have taken calculus, but we will not review those here.

Some functions that are often used in economics, along with their derivatives, are

$$y = \beta_0 + \beta_1 x + \beta_2 x^2;\ dy/dx = \beta_1 + 2\beta_2 x$$
$$y = \beta_0 + \beta_1/x;\ dy/dx = -\beta_1/(x^2)$$
$$y = \beta_0 + \beta_1\sqrt{x};\ dy/dx = (1/2)x^{-1/2}$$
$$y = \beta_0 + \beta_1 \log(x);\ dy/dx = \beta_1/x$$
$$y = \exp(\beta_0 + \beta_1 x);\ dy/dx = \beta_1 \exp(\beta_0 + \beta_1 x).$$

If $\beta_0 = 0$ and $\beta_1 = 1$ in this last expression, we get $dy/dx = \exp(x)$, when $y = \exp(x)$.

In Section A.4, we noted that equation (A.26) defines a constant elasticity model when calculus is used. The calculus definition of elasticity is $\frac{dy}{dx}\cdot\frac{x}{y}$. It can be shown using properties of logs and exponentials that, when (A.26) holds, $\frac{dy}{dx}\cdot\frac{x}{y} = \beta_1$.

When y is a function of multiple variables, the notion of a **partial derivative** becomes important. Suppose that

$$y = f(x_1, x_2). \tag{A.32}$$

Then, there are two partial derivatives, one with respect to x_1 and one with respect to x_2. The partial derivative of y with respect to x_1, denoted here by $\frac{\partial y}{\partial x_1}$, is just the usual derivative of (A.32) with respect to x_1, where x_2 is treated as a *constant*. Similarly, $\frac{\partial y}{\partial x_2}$ is just the derivative of (A.32) with respect to x_2, holding x_1 fixed.

Partial derivatives are useful for much the same reason as ordinary derivatives. We can approximate the change in y as

$$\Delta y \approx \frac{\partial y}{\partial x_1}\cdot\Delta x_1, \text{ holding } x_2 \text{ fixed.} \tag{A.33}$$

Thus, calculus allows us to define partial effects in nonlinear models just as we could in linear models. In fact, if

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

then

$$\frac{\partial y}{\partial x_1} = \beta_1,\ \frac{\partial y}{\partial x_2} = \beta_2.$$

These can be recognized as the partial effects defined in Section A.2.

A more complicated example is

$$y = 5 + 4x_1 + x_1^2 - 3x_2 + 7x_1\cdot x_2. \tag{A.34}$$

Now, the derivative of (A.34), with respect to x_1 (treating x_2 as a constant), is simply

$$\frac{\partial y}{\partial x_1} = 4 + 2x_1 + 7x_2;$$

note how this depends on x_1 and x_2. The derivative of (A.34), with respect to x_2, is $\frac{\partial y}{\partial x_2} = -3 + 7x_1$, so this depends only on x_1.

EXAMPLE A.8
(Wage Function with Interaction)

A function relating wages to years of education and experience is

$$wage = 3.10 + .41\ educ + .19\ exper - .004\ exper^2 + .007\ educ \cdot exper. \quad \textbf{(A.35)}$$

The partial effect of *exper* on *wage* is the partial derivative of (A.35):

$$\frac{\partial wage}{\partial exper} = .19 - .008\ exper + .007\ educ.$$

This is the approximate change in wage due to increasing experience by one year. Notice that this partial effect depends on the initial level of *exper* and *educ*. For example, for a worker who is starting with *educ* = 12 and *exper* = 5, the next year of experience increases wage by about .19 − .008(5) + .007(12) = .234, or 23.4 cents per hour. The exact change can be calculated by computing (A.35) at *exper* = 5, *educ* = 12 and at *exper* = 6, *educ* = 12, and then taking the difference. This turns out to be .23, which is very close to the approximation.

Differential calculus plays an important role in minimizing and maximizing functions of one or more variables. If $f(x_1,x_2,\ldots,x_k)$ is a differentiable function of k variables, then a necessary condition for $x_1^*, x_2^*, \ldots, x_k^*$ to either minimize or maximize f over all possible values of x_j is

$$\frac{\partial f}{\partial x_j}(x_1^*,x_2^*,\ldots,x_k^*) = 0, j = 1,2,\ldots,k. \quad \textbf{(A.36)}$$

In other words, all of the partial derivatives of f must be zero when they are evaluated at the x_h^*. These are called the *first order conditions* for minimizing or maximizing a function. Practically, we hope to solve equation (A.36) for the x_h^*. Then, we can use other criteria to determine whether we have minimized or maximized the function. We will not need those here. [See Sydsaeter and Hammond (1995) for a discussion of multivariable calculus and its use in optimizing functions.]

SUMMARY

The math tools reviewed here are crucial for understanding regression analysis and the probability and statistics that are covered in Appendices B and C. The material on nonlinear functions—especially quadratic, logarithmic, and exponential functions—is critical for understanding modern applied economic research. The level of comprehension

required of these functions does not include a deep knowledge of calculus, although calculus is needed for certain derivations.

KEY TERMS

Average
Ceteris Paribus
Constant Elasticity Model
Derivative
Descriptive Statistic
Diminishing Marginal Effect
Elasticity
Exponential Function
Intercept
Linear Function
Log Function
Marginal Effect
Median
Natural Logarithm
Nonlinear Function
Partial Derivative
Partial Effect
Percentage Change
Percentage Point Change
Proportionate Change
Semi-Elasticity
Slope
Summation Operator

PROBLEMS

A.1 The following table contains monthly housing expenditures for 10 families.

Family	Monthly Housing Expenditures (Dollars)
1	300
2	440
3	350
4	1,100
5	640
6	480
7	450
8	700
9	670
10	530

(i) Find the average monthly housing expenditure.
(ii) Find the median monthly housing expenditure.
(iii) If monthly housing expenditures were measured in hundreds of dollars, rather than in dollars, what would be the average and median expenditures?
(iv) Suppose that family number 8 increases its monthly housing expenditure to $900 dollars, but the expenditures of all other families remain the same. Compute the average and median housing expenditures.

A.2 Suppose the following equation describes the relationship between the average number of classes missed during a semester (*missed*) and the distance from school (*distance*, measured in miles):

$$missed = 3 + 0.2\ distance.$$

(i) Sketch this line, being sure to label the axes. How do you interpret the intercept in this equation?
(ii) What is the average number of classes missed for someone who lives five miles away?
(iii) What is the difference in the average number of classes missed for someone who lives 10 miles away and someone who lives 20 miles away?

A.3 In Example A.2, quantity of compact disks was related to price and income by $quantity = 120 - 9.8\ price + .03\ income$. What is the demand for CDs if $price = 15$ and $income = 200$? What does this suggest about using linear functions to describe demand curves?

A.4 Suppose the unemployment rate in the United States goes from 6.4% in one year to 5.6% in the next.
(i) What is the percentage point decrease in the unemployment rate?
(ii) By what percent has the unemployment rate fallen?

A.5 Suppose that the return from holding a particular firm's stock goes from 15% in one year to 18% in the following year. The majority shareholder claims that "the stock return only increased by 3%," while the chief executive officer claims that "the return on the firm's stock has increased by 20%." Reconcile their disagreement.

A.6 Suppose that Person A earns $35,000 per year and Person B earns $42,000.
(i) Find the exact percent by which Person B's salary exceeds Person A's.
(ii) Now use the difference in natural logs to find the approximate percentage difference.

A.7 Suppose the following model describes the relationship between annual salary (*salary*) and the number of previous years of labor market experience (*exper*):

$$\log(salary) = 10.6 + .027\ exper.$$

(i) What is *salary* when $exper = 0$? when $exper = 5$? (*Hint*: You will need to exponentiate.)

(ii) Use equation (A.28) to approximate the percentage increase in *salary* when *exper* increases by five years.

(iii) Use the results of part (i) to compute the exact percentage difference in salary when $exper = 5$ and $exper = 0$. Comment on how this compares with the approximation in part (ii).

A.8 Let *grthemp* denote the proportionate growth in employment, at the county level, from 1990 to 1995, and let *salestax* denote the county sales tax rate, stated as a proportion. Interpret the intercept and slope in the equation

$$grthemp = .043 - .78\ salestax.$$

A.9 Suppose the yield of a certain crop (in bushels per acre) is related to fertilizer amount (in pounds per acre) as

$$yield = 120 + .19\sqrt{fertilizer}.$$

(i) Graph this relationship by plugging in several values for *fertilizer*.

(ii) Describe how the shape of this relationship compares with a linear function between *yield* and *fertilizer*.

Appendix B

Fundamentals of Probability

This appendix covers key concepts from basic probability. Appendices B and C are primarily for review; they are not intended to replace a course in probability and statistics. Nevertheless, all of the probability and statistics concepts that we use in the text are covered in these appendices.

Probability is of interest in its own right for students in business, economics, and other social sciences. For example, consider the problem of an airline trying to decide how many reservations to accept for a flight that has 100 available seats. If fewer than 100 people want reservations, then these should all be accepted. But what if more than 100 people request reservations? A safe solution is to accept at most 100 reservations. However, since some people book reservations and then do not show up for the flight, there is some chance that the plane will not be full even if 100 reservations are booked. This results in lost revenue to the airline. A different strategy is to book more than 100 reservations and to hope that some people do not show up, and so the final number of passengers is as close to 100 as possible. This policy runs the risk of the airline having to compensate people who are necessarily bumped from an overbooked flight.

A natural question in this context is: Can we decide on the optimal (or best) number of reservations the airline should make? This is a nontrivial problem. Nevertheless, given certain information (on airline costs and how frequently people show up for reservations), we can use basic probability to arrive at a solution.

B.1 RANDOM VARIABLES AND THEIR PROBABILITY DISTRIBUTIONS

Suppose that we flip a coin 10 times and count the number of times the coin turns up heads. This is an example of an **experiment**. Generally, an experiment is any procedure that can, at least in theory, be infinitely repeated, and has a well-defined set of outcomes. We could, in principle, carry out the coin-flipping procedure again and again. Before we flip the coin, we know that the number of heads appearing is an integer from 0 to 10, so the outcomes of the experiment are well-defined.

A **random variable** is one that takes on numerical values and has an outcome that is determined by an experiment. In the coin-flipping example, the number of heads appearing in 10 flips of a coin is an example of a random variable. Before we flip the

coin 10 times, we do not know how many times the coin will come up heads. Once we flip the coin 10 times and count the number of heads, we obtain the outcome of the random variable for this particular trial of the experiment. Another trial can produce a different outcome.

In the airline reservation example mentioned earlier, the number of people showing up for their flight is a random variable: before any particular flight, we do not know how many people will show up.

To analyze data collected in business and the social sciences, it is important to have a basic understanding of random variables and their properties. Following the usual conventions in probability and statistics throughout Appendices B and C, we denote random variables by upper case letters, usually W, X, Y, and Z; particular outcomes of random variables are denoted by the corresponding lower case letters, w, x, y, and z. For example, in the coin-flipping experiment, let X denote the number of heads appearing in 10 flips of a coin. Then, X is not associated with any particular value, but we know X will take on a value in the set $\{0,1,2,\ldots,10\}$. A particular outcome is, say, $x = 6$.

We indicate large collections of random variables by using subscripts. For example, if we record last year's income of 20 randomly chosen households in the United States, we might denote these random variables by $X_1, X_2, \ldots, X_{20}$; the particular outcomes would be denoted $x_1, x_2, \ldots, x_{20}$.

As stated in the definition, random variables are always defined to take on numerical values, even when they describe qualitative events. For example, consider tossing a single coin, where the two outcomes are heads and tails. We can define a random variable as follows: $X = 1$ if the coin turns up heads, and $X = 0$ if the coin turns up tails.

A random variable that can only take on the values zero and one is called a **Bernoulli (or binary) random variable**. In basic probability, it is traditional to call the event $X = 1$ a "success" and the event $X = 0$ a "failure." For a particular application, the success-failure nomenclature might not correspond to our notion of a success or failure, but it is a useful terminology that we will adopt.

Discrete Random Variables

A **discrete random variable** is one that takes on only a finite or countably infinite number of values. The notion of "countably infinite" means that even though an infinite number of values can be taken on by a random variable, those values can be put in a one-to-one correspondence with the positive integers. Because the distinction between "countably infinite" and "uncountably infinite" is somewhat subtle, we will concentrate on discrete random variables that take on only a finite number of values. Larsen and Marx (1986, Chapter 3) contains a detailed treatment.

A Bernoulli random variable is the simplest example of a discrete random variable. The only thing we need to completely describe the behavior of a Bernoulli random variable is the probability that it takes on the value one. In the coin-flipping example, if the coin is "fair," then $P(X = 1) = 1/2$ (read as "the probability that X equals one is one-half"). Because probabilities must sum to one, $P(X = 0) = 1/2$, also.

Social scientists are interested in more than flipping coins, so we must allow for more general situations. Again, consider the example where the airline must decide how many people to book for a flight with 100 available seats. This problem can be analyzed

in the context of several Bernoulli random variables as follows: for a randomly selected customer, define a Bernoulli random variable as $X = 1$ if the person shows up for the reservation, and $X = 0$ if not.

There is no reason to think that the probability of any particular customer showing up is 1/2; in principle, the probability can be any number between zero and one. Call this number θ, so that

$$P(X = 1) = \theta \tag{B.1}$$

$$P(X = 0) = 1 - \theta. \tag{B.2}$$

For example, if $\theta = .75$, then there is a 75% chance that a customer shows up after making a reservation, and a 25% chance that the customer does not show up. Intuitively, the value of θ is crucial in determining the airline's strategy for booking reservations. Methods for *estimating* θ, given historical data on airline reservations, is a subject of mathematical statistics, something we turn to in Appendix C.

More generally, any discrete random variable is completely described by listing its possible values and the associated probability that it takes on each value. If X takes on the k possible values $\{x_1, \ldots, x_k\}$, then the probabilities $p_1, p_2, \ldots, p_k$ are defined by

$$p_j = P(X = x_j), j = 1,2, \ldots, k, \tag{B.3}$$

where each p_j is between 0 and 1, and

$$p_1 + p_2 + \ldots + p_k = 1. \tag{B.4}$$

Equation (B.3) is read as: "The probability that X takes on the value x_j is equal to p_j."

Equations (B.1) and (B.2) show that the probabilities of success and failure for a Bernoulli random variable are determined entirely by the value of θ. Because Bernoulli random variables are so prevalent, we have a special notation for them: $X \sim \text{Bernoulli}(\theta)$ is read as "X has a Bernoulli distribution with probability of success equal to θ."

The **probability density function (pdf)** of X summarizes the information concerning the possible outcomes of X and the corresponding probabilities:

$$f(x_j) = p_j, j = 1,2,\ldots,k, \tag{B.5}$$

with $f(x) = 0$ for any x not equal to x_j for some j. In other words, for any real number x, $f(x)$ is the probability that the random variable X takes on the particular value x. When dealing with more than one random variable, it is sometimes useful to subscript the pdf in question: f_X is the pdf of X, f_Y is the pdf of Y, and so on.

Given the pdf of any discrete random variable, it is simple to compute the probability of any event involving that random variable. For example, suppose that X is the number of free throws made by a basketball player out of two attempts, so that X can take on the three values $\{0,1,2\}$. Assume that the pdf of X is given by

$$f(0) = .20, f(1) = .44, \text{ and } f(2) = .36.$$

The three probabilities sum to one, as they must. Using this pdf, we can calculate the probability that the player makes *at least* one free throw: $P(X \geq 1) = P(X = 1) + P(X = 2) = .44 + .36 = .80$. The pdf of X is shown in Figure B.1.

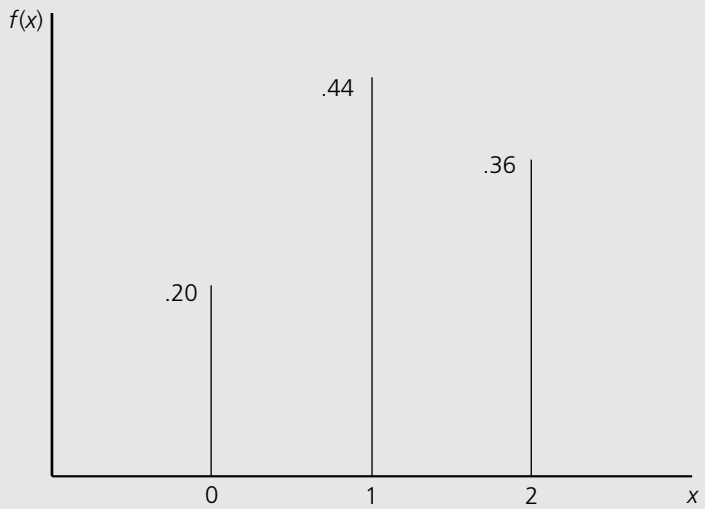
Continuous Random Variables

A variable X is a **continuous random variable** if it takes on any real value with *zero* probability. This definition is somewhat counterintuitive, since in any application, we eventually observe some outcome for a random variable. The idea is that a continuous random variable X can take on so many possible values that we cannot count them or match them up with the positive integers, so logical consistency dictates that X can take on each value with probability zero. While measurements are always discrete in practice, random variables that take on numerous values are best treated as continuous. For example, the most refined measure of the price of a good is in terms of cents. We can imagine listing all possible values of price in order (even though the list may continue indefinitely), which technically makes price a discrete random variable. However, there are so many possible values of price that using the mechanics of discrete random variables is not feasible.

We can define a probability density function for continuous random variables, and, as with discrete random variables, the pdf provides information on the likely outcomes of the random variable. However, because it makes no sense to discuss the probability that a continuous random variable takes on a particular value, we use the pdf of a con-

Figure B.1

The pdf of the number of free throws made out of two attempts.



tinuous rv only to compute events involving a range of values. For example, if a and b are constants where $a < b$, the probability that X lies between the numbers a and b, $P(a \leq X \leq b)$, is the *area* under the pdf between points a and b, as shown in Figure B.2. If you are familiar with calculus, you recognize this as the *integral* of the function f between the points a and b. The entire area under the pdf must always equal one.

When computing probabilities for continuous random variables, it is easiest to work with the **cumulative distribution function (cdf)**. If X is any random variable, then its cdf is defined for any real number x by

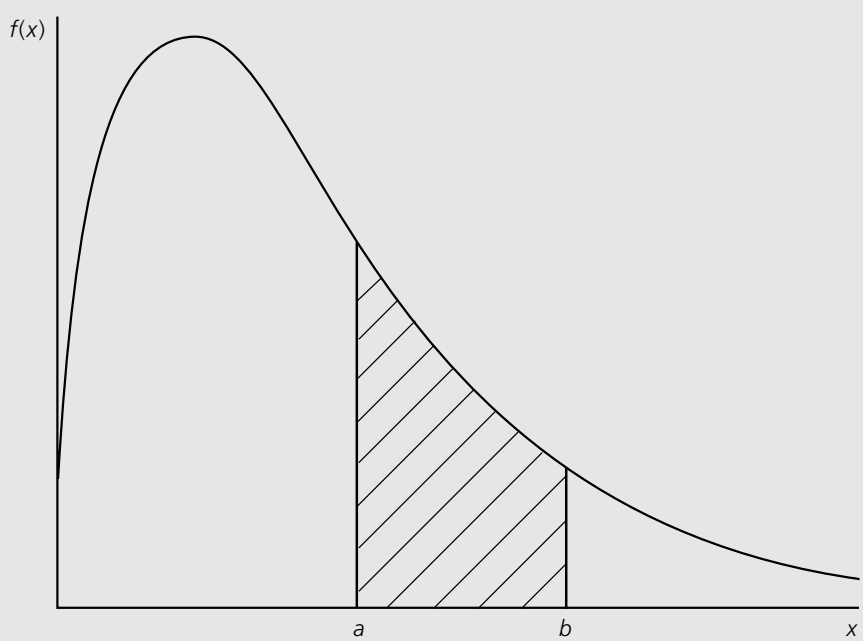
$$F(x) \equiv P(X \leq x). \tag{B.6}$$

For discrete random variables, (B.6) is obtained by summing the pdf over all values x_j such that $x_j \leq x$. For a continuous random variable, $F(x)$ is the area under the pdf, f, to the left of the point x. Since $F(x)$ is simply a probability, it is always between 0 and 1. Further, if $x_1 < x_2$, then $P(X \leq x_1) \leq P(X \leq x_2)$, that is, $F(x_1) \leq F(x_2)$. This means that a cdf is an increasing (or at least nondecreasing) function of x.

Two important properties of cdfs that are useful for computing probabilities are the following:

Figure B.2

The probability that X lies between the points a and b.



For any number c, $P(X > c) = 1 - F(c)$. **(B.7)**

For any numbers $a < b$, $P(a < X \leq b) = F(b) - F(a)$. **(B.8)**

In our study of econometrics, we will use cdfs to compute probabilities only for continuous random variables, in which case it does not matter whether inequalities in probability statements are strict or not. That is, for a continuous random variable X,

$$P(X \geq c) = P(X > c), \tag{B.9}$$

and

$$P(a < X < b) = P(a \leq X \leq b) = P(a \leq X < b) = P(a < X \leq b). \tag{B.10}$$

Combined with (B.7) and (B.8), equations (B.9) and (B.10) greatly expand the probability calculations that can be done using continuous cdfs.

Cumulative distribution functions have been tabulated for all of the important continuous distributions in probability and statistics. The most well-known of these is the normal distribution, which we cover along with some related distributions in Section B.5.

B.2 JOINT DISTRIBUTIONS, CONDITIONAL DISTRIBUTIONS, AND INDEPENDENCE

In economics, we are usually interested in the occurrence of events involving more than one random variable. For example, in the airline reservation example referred to earlier, the airline might be interested in the probability that a person who makes a reservation shows up *and* is a business traveler; this is an example of a *joint probability*. Or, the airline might be interested in the following *conditional probability*: conditional on the person being a business traveler, what is the probability of he or she showing up? In the next two subsections, we formalize the notions of joint and conditional distributions and the important notion of *independence* of random variables.

Joint Distributions and Independence

Let X and Y be discrete random variables. Then, (X,Y) have a **joint distribution**, which is fully described by the *joint probability density function* of (X,Y):

$$f_{X,Y}(x,y) = P(X = x, Y = y), \tag{B.11}$$

where the right-hand side is the probability that $X = x$ and $Y = y$. When X and Y are continuous, a joint pdf can also be defined, but we will not cover such details because joint pdfs for continuous random variables are not used explicitly in this text.

In one case, it is easy to obtain the joint pdf if we are given the pdfs of X and Y. In particular, random variables X and Y are said to be independent if and only if

$$f_{X,Y}(x,y) = f_X(x)f_Y(y) \qquad \textbf{(B.12)}$$

for all x and y, where f_X is the pdf of X, and f_Y is the pdf of Y. In the context of more than one random variable, the pdfs f_X and f_Y are often called *marginal probability density functions* to distinguish them from the joint pdf $f_{X,Y}$. This definition of independence is valid for discrete and continuous random variables.

To understand the meaning of (B.12), it is easiest to deal with the discrete case. If X and Y are discrete, then (B.12) is the same as

$$\mathrm{P}(X = x, Y = y) = \mathrm{P}(X = x)\mathrm{P}(Y = y); \qquad \textbf{(B.13)}$$

in other words, the probability that $X = x$ and $Y = y$ is the product of the two probabilities $\mathrm{P}(X = x)$ and $\mathrm{P}(Y = y)$. One implication of (B.13) is that joint probabilities are fairly easy to compute, since they only require knowledge of $\mathrm{P}(X = x)$ and $\mathrm{P}(Y = y)$.

If random variables are not independent, then they are said to be *dependent.*

EXAMPLE B.1
(Free Throw Shooting)

Consider a basketball player shooting two free throws. Let X be the Bernoulli random variable equal to one if she or he makes the first free throw, and zero otherwise. Let Y be a Bernoulli random variable equal to one if he or she makes the second free throw. Suppose that she or he is an 80% free-throw shooter, so that $\mathrm{P}(X = 1) = \mathrm{P}(Y = 1) = .8$. What is the probability of the player making both free throws?

If X and Y are independent, we can easily answer this question: $\mathrm{P}(X = 1, Y = 1) = \mathrm{P}(X = 1)\mathrm{P}(Y = 1) = (.8)(.8) = .64$. Thus, there is a 64% chance of making both free throws. If the chance of making the second free throw depends on whether the first was made—that is, X and Y are not independent—then this simple calculation is not valid.

Independence of random variables is a very important concept. In the next subsection, we will show that if X and Y are independent, then knowing the outcome of X does not change the probabilities of the possible outcomes of Y, and vice versa. One useful fact about independence is that if X and Y are independent and we define new random variables $g(X)$ and $h(Y)$ for any functions g and h, then these new random variables are also independent.

There is no need to stop at two random variables. If $X_1, X_2, \ldots, X_n$ are discrete random variables, then their joint pdf is $f(x_1, x_2, \ldots, x_n) = \mathrm{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$. The random variables $X_1, X_2, \ldots, X_n$ are **independent random variables** if and only if their joint pdf is the product of the individual pdfs for any $(x_1, x_2, \ldots, x_n)$. This definition of independence also holds for continuous random variables.

The notion of independence plays an important role in obtaining some of the classic distributions in probability and statistics. Earlier we defined a Bernoulli random variable as a zero-one random variable indicating whether or not some event occurs. Often, we are interested in the number of successes in a sequence of *independent*

Bernoulli trials. A standard example of independent Bernoulli trials is flipping a coin again and again. Since the outcome on any particular flip has nothing to do with the outcomes on other flips, independence is an appropriate assumption.

Independence is often a reasonable approximation in more complicated situations. In the airline reservation example, suppose that the airline accepts n reservations for a particular flight. For each $i = 1,2,\ldots,n$, let Y_i denote the Bernolli random variable indicating whether customer i shows up: $Y_i = 1$ if customer i appears, and $Y_i = 0$ otherwise. Letting θ again denote the probability of success (using reservation), each Y_i has a Bernoulli(θ) distribution. As an approximation, we might assume that the Y_i are independent of one another, although this is not exactly true in reality: some people travel in groups, which means that whether or not a person shows up is not truly independent of whether all others show up. Modeling this kind of dependence is complex, however, so we might be willing to use independence as an approximation.

The variable of primary interest is the total number of customers showing up out of the n reservations; call this variable X. Since each Y_i is unity when a person shows up, we can write $X = Y_1 + Y_2 + \ldots + Y_n$. Now, assuming that each Y_i has probability of success θ and that the Y_i are independent, X can be shown to have a **binomial distribution**. That is, the probability density function of X is

$$f(x) = \binom{n}{x}\theta^x(1-\theta)^{n-x},\ x = 0,1,2,\ldots,n, \qquad \textbf{(B.14)}$$

where $\binom{n}{x} = \dfrac{n!}{x!(n-x)!}$, and for any integer n, $n!$ (read "n factorial") is defined as $n! = n\cdot(n-1)\cdot(n-2)\cdots 1$. By convention, $0! = 1$. When a random variable X has the pdf given in (B.14), we write $X \sim \text{Binomial}(n,\theta)$. Equation (B.14) can be used to compute $P(X = x)$ for any value of x from 0 to n.

If the flight has 100 available seats, the airline is interested in $P(X > 100)$. Suppose, initially, that $n = 120$, so that the airline accepts 120 reservations, and the probability that each person shows up is $\theta = .80$. Then, $P(X > 100) = P(X = 101) + P(X = 102) + \ldots + P(X = 120)$, and each of the probabilities in the sum can be found from equation (B.14) with $n = 120$, $\theta = .80$, and the appropriate value of x (101 to 120). This is a difficult hand calculation, but many statistical packages have commands for computing this kind of probability. In this case, the probability that more than 100 people will show up is about .659, which is probably more risk of overbooking than the airline wants to tolerate. If, instead, the number of reservations is 110, the probability of more than 100 passengers showing up is only about .024.

Conditional Distributions

In econometrics, we are usually interested in how one random variable, call it Y, is related to one or more other variables. For now, suppose that there is only variable whose effects we are interested in, call it X. The most we can know about how X affects Y is contained in the **conditional distribution** of Y given X. This information is summarized by the *conditional probability density function*, defined by

$$f_{Y|X}(y|x) = f_{X,Y}(x,y)/f_X(x) \quad \textbf{(B.15)}$$

for all values of x such that $f_X(x) > 0$. The interpretation of (B.15) is most easily seen when X and Y are discrete. Then,

$$f_{Y|X}(y|x) = \mathrm{P}(Y = y|X = x), \quad \textbf{(B.16)}$$

where the right-hand side is read as "the probability that $Y = y$ given that $X = x$." When Y is continuous, $f_{Y|X}(y|x)$ is not interpretable directly as a probability, for the reasons discussed earlier, but conditional probabilities are found by computing areas under the conditional pdf.

An important feature of conditional distributions is that, if X and Y are independent random variables, knowledge of the value taken on by X tells us nothing about the probability that Y takes on various values (and vice versa). That is, $f_{Y|X}(y|x) = f_Y(y)$, and $f_{X|Y}(x|y) = f_X(x)$.

EXAMPLE B.2
(Free Throw Shooting)

Consider again the basketball-shooting example, where two free throws are to be attempted. Assume that the conditional density is

$$f_{Y|X}(1|1) = .85, f_{Y|X}(0|1) = .15$$
$$f_{Y|X}(1|0) = .70, f_{Y|X}(0|0) = .30.$$

This means that the probability of the player making the second free throw depends on whether the first free throw was made: if the first free throw is made, the chance of making the second is .85; if the first free throw is missed, the chance of making the second is .70. This implies that X and Y are *not* independent; they are dependent.

We can still compute $\mathrm{P}(X = 1, Y = 1)$, provided we know $\mathrm{P}(X = 1)$. Assume that the probability of making the first free throw is .8, that is, $\mathrm{P}(X = 1) = .8$. Then, from (B.15), we have

$$\mathrm{P}(X = 1, Y = 1) = \mathrm{P}(Y = 1|X = 1)\cdot\mathrm{P}(X = 1) = (.85)(.8) = .68.$$

B.3 FEATURES OF PROBABILITY DISTRIBUTIONS

For many purposes, we will be interested in only a few aspects of the distributions of random variables. The features of interest can be put into three categories: measures of central tendency, measures of variability or spread, and measures of association between two random variables. We cover the last of these in Section B.4.

A Measure of Central Tendency: The Expected Value

The expected value is one of the most important probabilistic concepts that we will encounter in our study of econometrics. If X is a random variable, the **expected value**

(or expectation) of X, denoted $E(X)$ and sometimes μ_X or simply μ, is a weighted average of all possible values of X. The weights are determined by the probability density function. Sometimes, the expected value is called the *population mean*, especially when we want to emphasize that X represents some variable in a population.

The precise definition of expected value is simplest in the case that X is a discrete random variable taking on a finite number of values, say $\{x_1, \ldots, x_k\}$. Let $f(x)$ denote the probability density function of X. The expected value of X is the weighted average

$$E(X) = x_1 f(x_1) + x_2 f(x_2) + \ldots + x_k f(x_k) \equiv \sum_{j=1}^{k} x_j f(x_j). \quad \textbf{(B.17)}$$

This is easily computed given the values of the pdf at each possible outcome of X.

EXAMPLE B.3
(Computing an Expected Value)

Suppose that X takes on the values -1, 0, and 2 with probabilities 1/8, 1/2, and 3/8, respectively. Then,

$$E(X) = (-1)\cdot(1/8) + 0\cdot(1/2) + 2\cdot(3/8) = 5/8.$$

This example illustrates something curious about expected values: the expected value of X can be a number that is not even a possible outcome of X. We know that X takes on the value -1, 0, or 2, yet its expected value is 5/8. This makes the expected value deficient for summarizing the central tendency of certain discrete random variables, but calculations such as those just mentioned can be useful, as we will see later.

If X is a continuous random variable, then $E(X)$ is defined as an integral:

$$E(X) = \int_{-\infty}^{\infty} x f(x) \mathrm{d}x, \quad \textbf{(B.18)}$$

which we assume is well-defined. This can still be interpreted as a weighted average. Unlike in the discrete case, $E(X)$ is always a number that is a possible outcome of X. In this text, we will not need to compute expected values using integration, although we will draw on some well-known results from probability for expected values of special random variables.

Given a random variable X and a function $g(\cdot)$, we can create a new random variable $g(X)$. For example, if X is a random variable, then so is X^2 and $\log(X)$ (if $X > 0$). The expected value of $g(X)$ is, again, simply a weighted average:

$$E[g(X)] = \sum_{j=1}^{k} g(x_j) f_X(x_j) \quad \textbf{(B.19)}$$

or, for a continuous random variable,

$$\mathrm{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) \mathrm{d}x. \tag{B.20}$$

EXAMPLE B.4
(Expected Value of X^2)

For the random variable in Example B.3, let $g(X) = X^2$. Then,

$$\mathrm{E}(X^2) = (-1)^2(1/8) + (0)^2(1/2) + (2)^2(3/8) = 13/8.$$

In Example B.3, we computed $\mathrm{E}(X) = 5/8$, so that $[\mathrm{E}(X)]^2 = 25/64$. This shows that $\mathrm{E}(X^2)$ is *not* the same as $[\mathrm{E}(X)]^2$. In fact, for a nonlinear function $g(X)$, $\mathrm{E}[g(X)] \neq g[\mathrm{E}(X)]$ (except in very special cases).

If X and Y are random variables, then $g(X,Y)$ is a random variable for any function g, and so we can define its expectation. When X and Y are both discrete, taking on values $\{x_1,x_2,\ldots,x_k\}$ and $\{y_1,y_2,\ldots,y_m\}$, respectively, the expected value is

$$\mathrm{E}[g(X,Y)] = \sum_{h=1}^{k}\sum_{j=1}^{m} g(x_h,y_j) f_{X,Y}(x_h,y_j),$$

where $f_{X,Y}$ is the joint pdf of (X,Y). The definition is more complicated for continuous random variables since it involves integration; we do not need it here. The extension to more than two random variables is straightforward.

Properties of Expected Value

In econometrics, we are not so concerned with computing expected values from various distributions; the major calculations have been done many times, and we will largely take these on faith. We will need to manipulate some expected values using a few simple rules. These are so important that we give them labels:

PROPERTY E.1
For any constant c, $\mathrm{E}(c) = c$.

PROPERTY E.2
For any constants a and b, $\mathrm{E}(aX + b) = a\mathrm{E}(X) + b$.

One useful implication of E.2 is that, if $\mu = \mathrm{E}(X)$, and we define a new random variable as $Y = X - \mu$, then $\mathrm{E}(Y) = 0$; in E.2, take $a = 1$ and $b = -\mu$.

As an example of Property E.2, let X be the temperature measured in Celsius, at noon on a particular day at a given location; suppose the expected temperature is $\mathrm{E}(X) = 25$. If Y is the temperature measured in Fahrenheit, then $Y = 32 + (9/5)X$. From Property E.2, the expected temperature in Fahrenheit is $\mathrm{E}(Y) = 32 + (9/5)\cdot\mathrm{E}(X) = 32 + (9/5)\cdot 25 = 77$.

Generally, it is easy to compute the expected value of a linear function of many random variables.

PROPERTY E.3

If $\{a_1, a_2, \ldots, a_n\}$ are constants and $\{X_1, X_2, \ldots, X_n\}$ are random variables, then

$$\mathrm{E}(a_1X_1 + a_2X_2 + \ldots + a_nX_n) = a_1\mathrm{E}(X_1) + a_2\mathrm{E}(X_2) + \ldots + a_n\mathrm{E}(X_n).$$

Or, using summation notation,

$$\mathrm{E}\left(\sum_{i=1}^{n} a_iX_i\right) = \sum_{i=1}^{n} a_i\mathrm{E}(X_i). \tag{B.21}$$

As a special case of this, we have (with each $a_i = 1$)

$$\mathrm{E}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{E}(X_i), \tag{B.22}$$

so that the expected value of the sum is the sum of expected values. This property is used often for derivations in mathematical statistics.

EXAMPLE B.5
(Finding Expected Revenue)

Let X_1, X_2, and X_3 be the numbers of small, medium, and large pizzas, respectively, sold during the day at a pizza parlor. These are random variables with expected values $\mathrm{E}(X_1) = 25$, $\mathrm{E}(X_2) = 57$, and $\mathrm{E}(X_3) = 40$. The prices of small, medium, and large pizzas are \$5.50, \$7.60, and \$9.15. Therefore, the expected revenue from pizza sales on a given day is

$$\begin{aligned}\mathrm{E}(5.50\,X_1 + 7.60\,X_2 + 9.15\,X_3) &= 5.50\,\mathrm{E}(X_1) + 7.60\,\mathrm{E}(X_2) + 9.15\,\mathrm{E}(X_3)\\ &= 5.50(25) + 7.60(57) + 9.15(40) = 936.70,\end{aligned}$$

that is, \$936.70. The actual revenue on any particular day will generally differ from this value, but this is the *expected* revenue.

We can also use Property E.3 to show show that if $X \sim \text{Binomial}(n,\theta)$, then $\mathrm{E}(X) = n\theta$. That is, the expected number of successes in n Bernoulli trials is simply the number of trials times the probability of success on any particular trial. This is easily seen by writing X as $X = Y_1 + Y_2 + \ldots + Y_n$, where each $Y_i \sim \text{Bernoulli}(\theta)$. Then,

$$\mathrm{E}(X) = \sum_{i=1}^{n} \mathrm{E}(Y_i) = \sum_{i=1}^{n} \theta = n\theta.$$

We can apply this to the airline reservation example, where the airline makes $n = 120$ reservations, and the probability of showing up is $\theta = .85$. The *expected* number of people showing up is $120(.85) = 102$. Therefore, if there are 100 seats available, the expected number of people showing up is too large; this has some bearing on whether it is a good idea for the airline to make 120 reservations.

Actually, what the airline should do is define a profit function that accounts for the net revenue earned per seat sold and the cost per passenger bumped from the flight. This

profit function is random because the actual number of people showing up is random. Let r be the net revenue from each passenger. (You can think of this as the price of the ticket for simplicity.) Let c be the compensation owed to any passenger bumped from the flight. Neither r nor c is random; these are assumed to be known to the airline. Let Y denote profits for the flight. Then, with 100 seats available,

$$\begin{aligned} Y &= rX \text{ if } X \leq 100 \\ &= 100r - c(X - 100) \text{ if } X > 100. \end{aligned}$$

The first equation gives profit if no more than 100 people show up for the flight; the second equation is profit if more than 100 people show up. (In the latter case, the net revenue from ticket sales is $100r$, since all 100 seats are sold, and then $c(X - 100)$ is the cost of making more than 100 reservations.) Using the fact that X has a Binomial(n,.85) distribution, where n is the number of reservations made, expected profits, E(Y), can be found as a function of n (and r and c). Computing E(Y) directly would be quite difficult, but it can be found quickly using a computer. Once values for r and c are given, the value of n that maximizes expected profits can be found by searching over different values of n.

Another Measure of Central Tendency: The Median

The expected value is only one possibility for defining the central tendency of a random variable. Another measure of central tendency is the **median**. A general definition of median is too complicated for our purposes. If X is continuous, then the median of X, say m, is the value such that one-half of the area under pdf is to the left of m, and one-half of the area is to the right of m.

When X is discrete and takes on a finite number of odd values, the median is obtained by ordering the possible values of X and then selecting the value in the middle. For example, if X can take on the values $\{-4,0,2,8,10,13,17\}$, then the median value of X is 8. If X takes on an even number of values, there are really two median values; sometimes these are averaged to get a unique median value. Thus, if X takes on the values $\{-5,3,9,17\}$, then the median values are 3 and 9; if we average these, we get a median equal to 6.

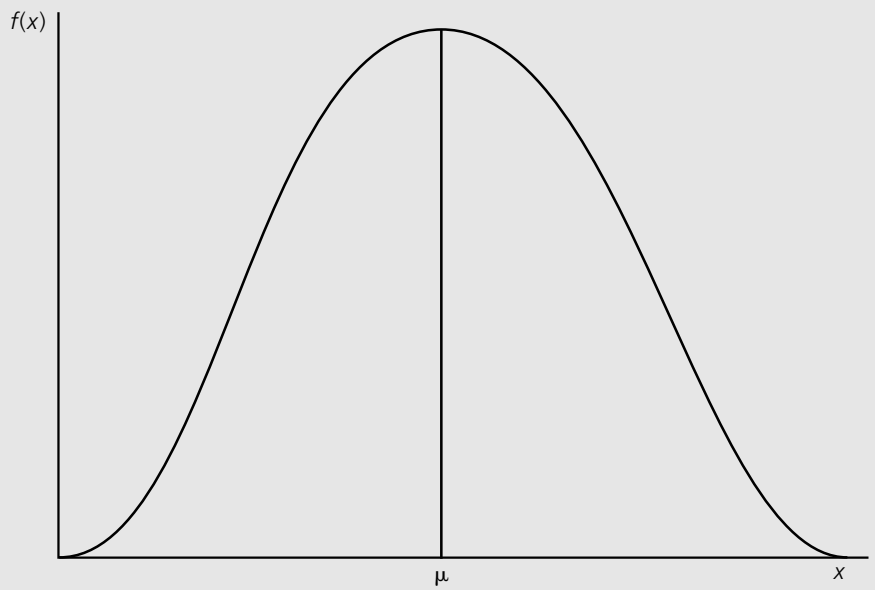
In general, the median, sometimes denoted Med(X), and the expected value, E(X), are different. Neither is "better" than the other as a measure of central tendency; they are both valid ways to measure the center of the distribution of X. In one special case, the median and expected value (or mean) are the same. If the probability distribution of X is *symmetrically distributed* about the value μ, then μ is both the expected value and the median. Mathematically, the condition is $f(\mu + x) = f(\mu - x)$ for all x. This case is illustrated in Figure B.3.

Measures of Variability: Variance and Standard Deviation

While the central tendency of a random variable is valuable, it does not tell us everything we want to know about the distribution of a random variable. Figure B.4 shows the pdfs of two random variables with the same mean. Clearly, the distribution of X is

Figure B.3

A symmetric probability distribution.



more tightly centered about its mean than is the distribution of Y. We would like to have a simple way of summarizing this.

Variance

For a random variable X, let $\mu = \mathrm{E}(X)$. There are various ways to measure how far X is from its expected value, but the simplest one to work with algebraically is the squared difference, $(X - \mu)^2$. (The squaring serves to eliminate the sign from the distance measure; the resulting positive value corresponds to our intuitive notion of distance.) This distance is itself a random variable since it can change with every outcome of X. Just as we needed a number to summarize the central tendency of X, we need a number that tells us how far X is from μ, *on average*. One such number is the **variance**, which tells us the expected distance from X to its mean:

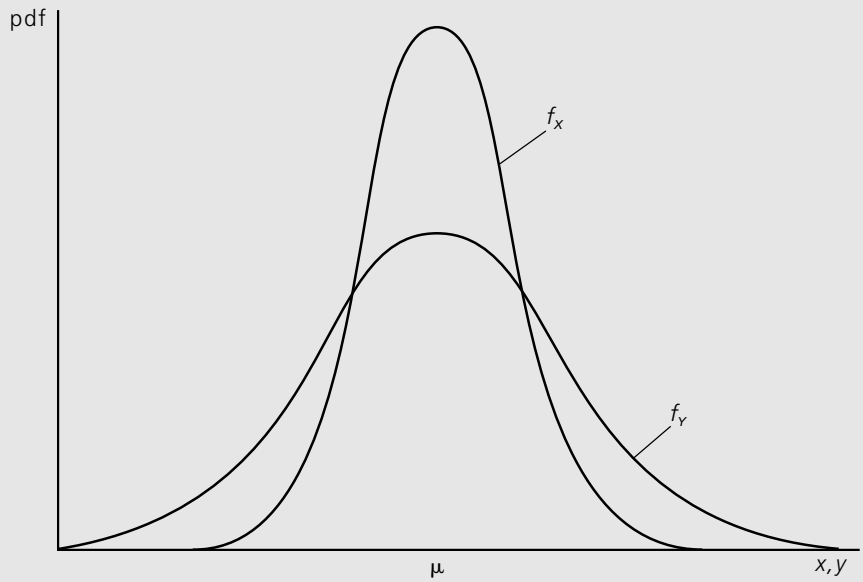
$$\mathrm{Var}(X) \equiv \mathrm{E}[(X - \mu)^2]. \qquad \textbf{(B.23)}$$

Variance is sometimes denoted σ_X^2, or simply σ^2, when the context is clear. From (B.23), it follows that the variance is always nonnegative.

As a computational device, it is useful to observe that

Figure B.4

Random variables with the same mean but different distributions.



$$\sigma^2 = \mathrm{E}(X^2 - 2X\mu + \mu^2) = \mathrm{E}(X^2) - 2\mu^2 + \mu^2 = \mathrm{E}(X^2) - \mu^2. \quad \textbf{(B.24)}$$

In using either (B.23) or (B.24), we need not distinguish between discrete and continuous random variables: the definition of variance is the same in either case. Most often, we first compute $\mathrm{E}(X)$, then $\mathrm{E}(X^2)$, and then we use the formula in (B.24). For example, if $X \sim$ Bernoulli(θ), then $\mathrm{E}(X) = \theta$, and, since $X^2 = X$, $\mathrm{E}(X^2) = \theta$. It follows from equation (B.24) that $\mathrm{Var}(X) = \mathrm{E}(X^2) - \mu^2 = \theta - \theta^2 = \theta(1 - \theta)$.

Two important properties of the variance follow.

PROPERTY VAR.1

$\mathrm{Var}(X) = 0$ if and only if there is a constant c, such that $\mathrm{P}(X = c) = 1$, in which case, $\mathrm{E}(X) = c$.

This first property says that the variance of any constant is zero and if a random variable has zero variance, then it is essentially constant.

PROPERTY VAR.2

For any constants a and b, $\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$.

This means that adding a constant to a random variable does not change the variance, but multiplying a random variable by a constant increases the variance by a factor equal to the *square* of that constant. For example, if X denotes temperature in Celsius and $Y = 32 + (9/5)X$ is temperature in Fahrenheit, then $\text{Var}(Y) = (9/5)^2\text{Var}(X) = (81/25)\text{Var}(X)$.

Standard Deviation

The **standard deviation** of a random variable, denoted $\text{sd}(X)$, is simply the positive square root of the variance: $\text{sd}(X) \equiv +\sqrt{\text{Var}(X)}$. The standard deviation is sometimes denoted σ_X, or simply σ, when the random variable is understood. Two standard deviation properties immediately follow from Properties VAR.1 and VAR.2.

PROPERTY SD.1

For any constant c, $\text{sd}(c) = 0$.

PROPERTY SD.2

For any constants a and b,

$$\text{sd}(aX + b) = |a|\text{sd}(X).$$

In particular, if $a > 0$, then $\text{sd}(aX) = a \cdot \text{sd}(X)$.

This last property makes the standard deviation more natural to work with than the variance. For example, suppose that X is a random variable measured in thousands of dollars, say income. If we define $Y = 1{,}000X$, then Y is income measured in dollars. Suppose that $\text{E}(X) = 20$, and $\text{sd}(X) = 6$. Then $\text{E}(Y) = 1{,}000\text{E}(X) = 20{,}000$, and $\text{sd}(Y) = 1{,}000 \cdot \text{sd}(X) = 6{,}000$, so that the expected value and standard deviation both increase by the same factor, 1,000. If we worked with variance, we would have $\text{Var}(Y) = (1{,}000)^2\text{Var}(X)$, so that the variance of Y is one million times larger than the variance of X.

Standardizing a Random Variable

As an application of the properties of variance and standard deviation—and a topic of practical interest in its own right—suppose that given a random variable X, we define a new random variable by subtracting off its mean μ and dividing by its standard deviation σ:

$$Z \equiv \frac{X - \mu}{\sigma}, \tag{B.25}$$

which we can write as $Z = aX + b$, where $a \equiv (1/\sigma)$, and $b \equiv -(\mu/\sigma)$. Then, from Property E.2,

$$\text{E}(Z) = a\text{E}(X) + b = (\mu/\sigma) - (\mu/\sigma) = 0.$$

From Property VAR.2,

$$\text{Var}(Z) = a^2\text{Var}(X) = (\sigma^2/\sigma^2) = 1.$$

Thus, the random variable Z has a mean of zero and a variance (and therefore a standard deviation) equal to one. This procedure is sometimes known as *standardizing* the random variable X, and Z is called a **standardized random variable**. (In introductory statistics courses, it is sometimes called the *z-transform* of X.) It is important to remember that the standard deviation, not the variance, appears in the denominator of (B.25). As we will see, this transformation is frequently used in statistical inference.

As a specific example, suppose that $\text{E}(X) = 2$, and $\text{Var}(X) = 9$. Then $Z = (X - 2)/3$ has expected value zero and variance one.

B.4 FEATURES OF JOINT AND CONDITIONAL DISTRIBUTIONS

Measures of Association: Covariance and Correlation

While the joint pdf of two random variables completely describes the relationship between them, it is useful to have summary measures of how, on average, two random variables vary with one another. As with the expected value and variance, this is similar to using a single number to summarize something about an entire distribution, which in this case is a joint distribution of two random variables.

Covariance

Let $\mu_X = \text{E}(X)$ and $\mu_Y = \text{E}(Y)$ and consider the random variable $(X - \mu_X)(Y - \mu_Y)$. Now, if X is above its mean and Y is above its mean, then $(X - \mu_X)(Y - \mu_Y) > 0$. This is also true if $X < \mu_X$ and $Y < \mu_Y$. On the other hand, if $X > \mu_X$ and $Y < \mu_Y$, or vice versa, then $(X - \mu_X)(Y - \mu_Y) < 0$. How, then, can this product tell us anything about the relationship between X and Y?

The **covariance** between two random variables X and Y, sometimes called the population covariance to emphasize that it concerns the relationship between two variables describing a population, is defined as the expected value of the product $(X - \mu_X)(Y - \mu_Y)$:

$$\text{Cov}(X,Y) \equiv \text{E}[(X - \mu_X)(Y - \mu_Y)], \tag{B.26}$$

which is sometimes denoted σ_{XY}. If $\sigma_{XY} > 0$, then, on average, when X is above its mean, Y is also above its mean. If $\sigma_{XY} < 0$, then, on average, when X is above its mean, Y is below its mean.

Several expressions useful for computing $\text{Cov}(X,Y)$ are as follows:

$$\begin{aligned}\text{Cov}(X,Y) &= \text{E}[(X - \mu_X)(Y - \mu_Y)] = \text{E}[(X - \mu_X)Y] \\ &= \text{E}[X(Y - \mu_Y)] = \text{E}(XY) - \mu_X\mu_Y.\end{aligned} \tag{B.27}$$

It follows from (B.27), that if $\text{E}(X) = 0$ or $\text{E}(Y) = 0$, then $\text{Cov}(X,Y) = \text{E}(XY)$.

Covariance measures the amount of *linear* dependence between two random variables. A positive covariance indicates that two random variables move in the same

direction, while a negative covariance indicates they move in opposite directions. Interpreting the *magnitude* of a covariance can be a little tricky, as we will see shortly.

Since covariance is a measure of how two random variables are related, it is natural to ask how covariance is related to the notion of independence. This is given by the following property.

PROPERTY COV.1
If X and Y are independent, then $\text{Cov}(X,Y) = 0$.

This property follows from equation (B.27) and the fact that $\text{E}(XY) = \text{E}(X)\text{E}(Y)$ when X and Y are independent. It is important to remember that the converse of COV.1 is *not* true: zero covariance between X and Y does not imply that X and Y are independent. In fact, there are random variables X such that, if $Y = X^2$, $\text{Cov}(X,Y) = 0$. (Any random variable with $\text{E}(X) = 0$ and $\text{E}(X^3) = 0$ has this property.) If $Y = X^2$, then X and Y are clearly not independent: once we know X, we know Y. It seems rather strange that X and X^2 could have zero covariance, and this reveals a weakness of covariance as a general measure of association between random variables. The covariance is useful in contexts when relationships are at least approximately linear.

The second major property of covariance involves covariances between linear functions.

PROPERTY COV.2
For any constants a_1, b_1, a_2, and b_2,

$$\text{Cov}(a_1X + b_1, a_2Y + b_2) = a_1a_2\text{Cov}(X,Y). \tag{B.28}$$

An important implication of COV.2 is that the covariance between two random variables can be altered simply by multiplying one or both of the random variables by a constant. This is important in economics since monetary variables, inflation rates, and so on, can be defined with different units of measurement without changing their meaning.

Finally, it is useful to know that the absolute value of the covariance between any two random variables is bounded by the product of their standard deviations; this is known as the *Cauchy-Schwartz inequality*.

PROPERTY COV.3
$|\text{Cov}(X,Y)| \leq \text{sd}(X)\text{sd}(Y)$.

Correlation Coefficient

Suppose we want to know the relationship between amount of education and annual earnings in the working population. We could let X denote education and Y denote earnings and then compute their covariance. But the answer we get will depend on how we choose to measure education and earnings. Property COV.2 implies that the covariance between education and earnings depends on whether earnings are measured in dollars or thousands of dollars, or whether education is measured in months or years. It is pretty

clear that how we measure these variables has no bearing on how strongly they are related. But the covariance between them does depend on the units of measurement.

The fact that the covariance depends on units of measurement is a deficiency that is overcome by the **correlation coefficient** between X and Y:

$$\text{Corr}(X,Y) \equiv \frac{\text{Cov}(X,Y)}{\text{sd}(X)\cdot\text{sd}(Y)} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}; \tag{B.29}$$

the correlation coefficient between X and Y is sometimes denoted ρ_{XY} (and is sometimes called the population correlation).

Because σ_X and σ_Y are positive, $\text{Cov}(X,Y)$ and $\text{Corr}(X,Y)$ always have the same sign, and $\text{Corr}(X,Y) = 0$ if and only if $\text{Cov}(X,Y) = 0$. Some of the properties of covariance carry over to correlation. If X and Y are independent, then $\text{Corr}(X,Y) = 0$, but zero correlation does not imply lack of independence. (The correlation coefficient is also a measure of linear dependence.) However, the magnitude of the correlation coefficient is easier to interpret than the size of the covariance due to the following property.

PROPERTY CORR.1
$-1 \leq \text{Corr}(X,Y) \leq 1$.

If $\text{Corr}(X,Y) = 0$, or equivalently $\text{Cov}(X,Y) = 0$, then there is no linear relationship between X and Y, and X and Y are said to be *uncorrelated*; otherwise, X and Y are *correlated.* $\text{Corr}(X,Y) = 1$ implies a perfect positive linear relationship, which means that we can write $Y = a + bX$, for some constant a and some constant $b > 0$. $\text{Corr}(X,Y) = -1$ implies a perfect negative relationship, so that $Y = a + bX$, for some $b < 0$. The extreme cases of positive or negative one rarely occur. Values of ρ_{XY} closer to 1 or -1 indicate stronger linear relationships.

As mentioned earlier, the correlation between X and Y is invariant to the units of measurement of either X or Y. This is stated more generally as follows.

PROPERTY CORR.2
For constants a_1, b_1, a_2, and b_2, with $a_1a_2 > 0$,

$$\text{Corr}(a_1X + b_1, a_2Y + b_2) = \text{Corr}(X,Y).$$

If $a_1a_2 < 0$, then

$$\text{Corr}(a_1X + b_1, a_2Y + b_2) = -\text{Corr}(X,Y).$$

As an example, suppose that the correlation between earnings and education in the working population is .15. This measure does not depend on whether earnings are measured in dollars, thousands of dollars, or any other unit; it also does not depend on whether education is measured in years, quarters, months, and so on.

Variance of Sums of Random Variables

Now that we have defined covariance and correlation, we can complete our list of major properties of the variance.

PROPERTY VAR.3

For constants a and b,

$$\text{Var}(aX + bY) = a^2\text{Var}(X) + b^2\text{Var}(Y) + 2ab\text{Cov}(X,Y).$$

It follows immediately that, if X and Y are uncorrelated—so that $\text{Cov}(X,Y) = 0$—then

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) \tag{B.30}$$

and

$$\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y). \tag{B.31}$$

In the latter case, note how the variance of the difference is the *sum*, not the difference, in the variances.

As an example of (B.30), let X denote profits earned by a restaurant during a Friday night and let Y be profits earned on the following Saturday night. Then, $Z = X + Y$ is profits for the two nights. Suppose X and Y each have an expected value of \$300 and a standard deviation of \$15 (so that the variance is 225). Expected profits for the two nights is $\text{E}(Z) = \text{E}(X) + \text{E}(Y) = 2\cdot(300) = 600$ dollars. If X and Y are independent, and therefore uncorrelated, then the variance of total profits is the sum of the variances: $\text{Var}(Z) = \text{Var}(X) + \text{Var}(Y) = 2\cdot(225) = 450$. It follows that the standard deviation of total profits is $\sqrt{450}$ or about \$21.21.

Expressions (B.30) and (B.31) extend to more than two random variables. To state this extension, we need a definition. The random variables $\{X_1, \ldots, X_n\}$ are **pairwise uncorrelated random variables** if each variable in the set is uncorrelated with every other variable in the set. That is, $\text{Cov}(X_i,X_j) = 0$, for all $i \neq j$.

PROPERTY VAR.4

If $\{X_1, \ldots, X_n\}$ are pairwise uncorrelated random variables and $\{a_i: i = 1, \ldots, n\}$ are constants, then

$$\text{Var}(a_1X_1 + \ldots + a_nX_n) = a_1^2\text{Var}(X_1) + \ldots + a_n^2\text{Var}(X_n).$$

In summation notation, we can write

$$\text{Var}\left(\sum_{i=1}^{n} a_iX_i\right) = \sum_{i=1}^{n} a_i^2\text{Var}(X_i). \tag{B.32}$$

A special case of Property VAR.4 occurs when we take $a_i = 1$ for all i. Then, for pairwise uncorrelated random variables, the variance of the sum is the sum of the variances:

$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{Var}(X_i). \tag{B.33}$$

Since independent random variables are uncorrelated (see Property COV.1), the variance of a sum of independent random variables is the sum of the variances.

If the X_i are not pairwise uncorrelated, then the expression for $\text{Var}(\sum_{i=1}^{n} a_i X_i)$ is much more complicated; it depends on each covariance, as well as on each variance. We will not need the more general formula for our purposes.

We can use (B.33) to derive the variance for a binomial random variable. Let $X \sim \text{Binomial}(n,\theta)$ and write $X = Y_1 + \ldots + Y_n$, where the Y_i are independent Bernoulli(θ) random variables. Then, by (B.33), $\text{Var}(X) = \text{Var}(Y_1) + \ldots + \text{Var}(Y_n) = n\theta(1 - \theta)$.

In the airline reservations example with $n = 120$ and $\theta = .85$, the variance of the number of passengers arriving for their reservations is $120(.85)(.15) = 15.3$, and so the standard deviation is about 3.9.

Conditional Expectation

Covariance and correlation measure the linear relationship between two random variables and treat them symmetrically. More often in the social sciences, we would like to explain one variable, called Y, in terms of another variable, say X. Further, if Y is related to X in a nonlinear fashion, we would like to know this. Call Y the explained variable and X the explanatory variable. For example, Y might be hourly wage, and X might be years of formal education.

We have already introduced the notion of the conditional probability density function of Y given X. Thus, we might want to see how the distribution of wages changes with education level. However, we usually want to have a simple way of summarizing this distribution. A single number will no longer suffice, since the distribution of Y, given $X = x$, generally depends on the value of x. Nevertheless, we can summarize the relationship between Y and X by looking at the **conditional expectation** of Y given X, sometimes called the *conditional mean.* The idea is this. Suppose we know that X has taken on a particular value, say x. Then, we can compute the expected value of Y, given that we know this outcome of X. We denote this expected value by $\text{E}(Y|X = x)$, or sometimes $\text{E}(Y|x)$ for shorthand. Generally, as x changes, so does $\text{E}(Y|x)$.

When Y is a discrete random variable taking on values $\{y_1, \ldots, y_m\}$, then

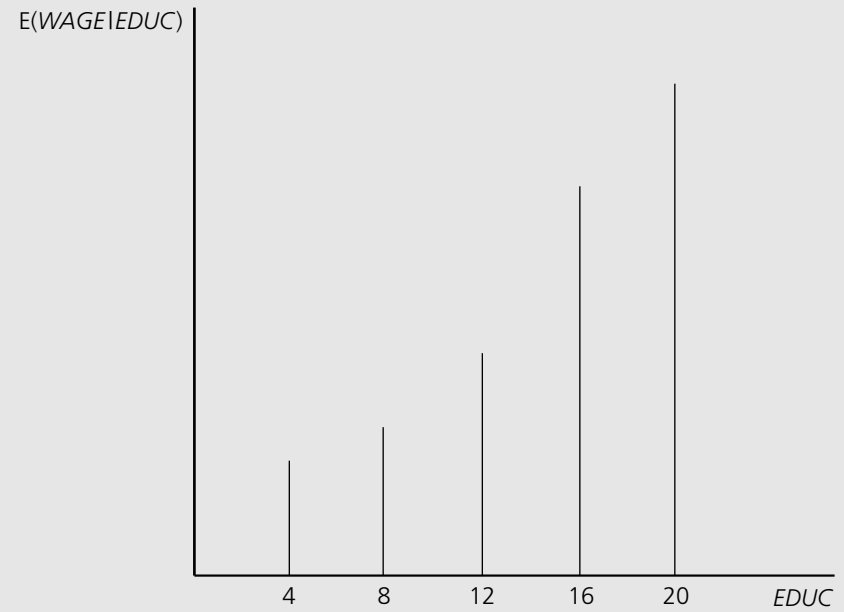
$$\text{E}(Y|x) = \sum_{j=1}^{m} y_j f_{Y|X}(y_j|x).$$

When Y is continuous, $\text{E}(Y|x)$ is defined by integrating $y f_{Y|X}(y|x)$ over all possible values of y. As with unconditional expectations, the conditional expectation is a weighted average of possible values of Y, but now the weights reflect the fact that X has taken on a specific value. Thus, $\text{E}(Y|x)$ is just some function of x, which tells us how the expected value of Y varies with x.

As an example, let (X,Y) represent the population of all working individuals, where X is years of education, and Y is hourly wage. Then, $\text{E}(Y|X = 12)$ is the average hourly wage for all people in the population with 12 years of education (roughly a high school education). $\text{E}(Y|X = 16)$ is the average hourly wage for all people with 16 years of education. Tracing out the expected value for various levels of education provides important information on how wages and education are related. See Figure B.5 for an illustration.

Figure B.5

The expected value of hourly wage given various levels of education.



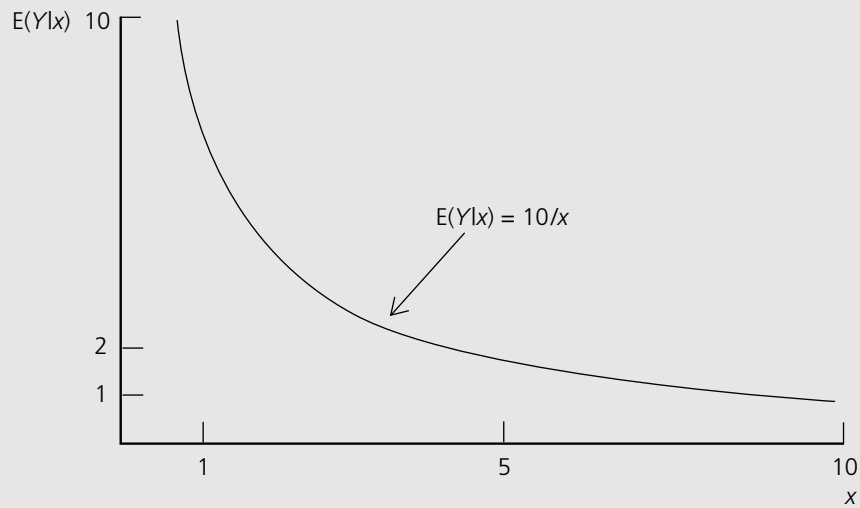
In principle, the expected value of hourly wage can be found at each level of education, and these expectations can be summarized in a table. Since education can vary widely—and can even be measured in fractions of a year—this is a cumbersome way to show the relationship between average wage and amount of education. In econometrics, we typically specify simple functions that capture this relationship. As an example, suppose that the expected value of *WAGE* given *EDUC* is the linear function

$$\mathrm{E}(WAGE|EDUC) = 1.05 + .45\ EDUC.$$

If this relationship holds in the population of working people, the average wage for people with eight years of education is $1.05 + .45(8) = 4.65$, or \$4.65. The average wage for people with 16 years of education is 8.25, or \$8.25. The coefficient on *EDUC* implies that each year of education increases the expected hourly wage by .45, or 45 cents.

Conditional expectations can also be nonlinear functions. For example, suppose that $\mathrm{E}(Y|x) = 10/x$, where X is a random variable that is always greater than zero. This function is graphed in Figure B.6. This could represent a demand function, where Y is quantity demanded, and X is price. If Y and X are related in this way, an analysis of linear association, such as correlation analysis, would be inadequate.

Figure B.6
Graph of $E(Y|x) = 10/x$.



Properties of Conditional Expectation

Several basic properties of conditional expectations are useful for derivations in econometric analysis.

PROPERTY CE.1:
$E[c(X)|X] = c(X)$, for any function $c(X)$.

This first property means that functions of X behave as constants when we compute expectations conditional on X. For example, $E(X^2|X) = X^2$. Intuitively, this simply means that if we know X, then we also know X^2.

PROPERTY CE.2
For functions $a(X)$ and $b(X)$,

$$E[a(X)Y + b(X)|X] = a(X)E(Y|X) + b(X).$$

For example, we can easily compute the conditional expectation of a function such as $XY + 2X^2$: $E(XY + 2X^2|X) = XE(Y|X) + 2X^2$.

The next property ties together the notions of independence and conditional expectations.

PROPERTY CE.3
If X and Y are independent, then $E(Y|X) = E(Y)$.

This property means that, if X and Y are independent, then the expected value of Y given X does not depend on X, in which case, $E(Y|X)$ always equals the (unconditional) expected value of Y. In the wage and education example, if wages were independent of education, then the average wages of high school and college graduates would be the same. Since this is almost certainly false, we cannot assume that wage and education are independent.

A special case of Property CE.3 is the following: if U and X are independent and $E(U) = 0$, then $E(U|X) = 0$.

There are also properties of the conditional expectation that have to do with the fact that $E(Y|X)$ is a function of X, say $E(Y|X) = \mu(X)$. Since X is a random variable, $\mu(X)$ is also a random variable. Furthermore, $\mu(X)$ has a probability distribution and therefore an expected value. Generally, the expected value of $\mu(X)$ could be very difficult to compute directly. The **law of iterated expectations** says that the expected value of $\mu(X)$ is simply equal to the expected value of Y. We write this as follows.

PROPERTY CE.4
$E[E(Y|X)] = E(Y)$.

This property is a little hard to grasp at first. It means that, if we first obtain $E(Y|X)$ as a function of X and take the expected value of this (with respect to the distribution of X, of course), then we end up with $E(Y)$. This is hardly obvious, but it can be derived using the definition of expected values.

Suppose $Y = WAGE$ and $X = EDUC$, where $WAGE$ is measured in hours, and $EDUC$ is measured in years. Suppose the expected value of $WAGE$ given $EDUC$ is $E(WAGE|EDUC) = 4 + .60\ EDUC$. Further, $E(EDUC) = 11.5$. Then, the law of iterated expectations implies that $E(WAGE) = E(4 + .60\ EDUC) = 4 + .60\ E(EDUC) = 4 + .60(11.5) = 10.90$, or \$10.90 an hour.

The next property states a more general version of the law of iterated expectations.

PROPERTY CE.4′
$E(Y|X) = E[E(Y|X,Z)|X]$.

In other words, we can find $E(Y|X)$ in two steps. First, find $E(Y|X,Z)$ for any other random variable Z. Then, find the expected value of $E(Y|X,Z)$, conditional on X.

PROPERTY CE.5
If $E(Y|X) = E(Y)$, then $Cov(X,Y) = 0$ (and so $Corr(X,Y) = 0$). In fact, *every* function of X is uncorrelated with Y.

This property means that, if knowledge of X does not change the expected value of Y, then X and Y *must* be uncorrelated, which implies that if X and Y are correlated, then $E(Y|X)$ must depend on X. The converse of Property CE.5 is not true: if X and Y are uncorrelated, $E(Y|X)$ *could* still depend on X. For example, suppose $Y = X^2$. Then, $E(Y|X) = X^2$, which is clearly a function of X. However, as we mentioned in our discussion of covariance and correlation, it is possible that X and X^2 are uncorrelated. The

conditional expectation captures the nonlinear relationship between X and Y that correlation analysis would miss entirely.

Properties CE.4 and CE.5 have two major implications: if U and X are random variables such that $E(U|X) = 0$, then $E(U) = 0$, and U and X are uncorrelated.

PROPERTY CE.6
If $E(Y^2) < \infty$ and $E[g(X)^2] < \infty$ for some function g, then $E\{[Y - \mu(X)]^2|X\} \leq E\{[Y - g(X)]^2|X\}$ and $E\{[Y - \mu(X)]^2\} \leq E\{[Y - g(X)]^2\}$.

This last property is very useful in predicting or forecasting contexts. The first inequality says that, if we measure prediction inaccuracy as the *expected* squared prediction error, conditional on X, then the conditional mean is better than any other function of X for predicting Y. The conditional mean also minimizes the unconditional expected squared prediction error.

Conditional Variance

Given random variables X and Y, the variance of Y, conditional on $X = x$, is simply the variance associated with the conditional distribution of Y, given $X = x$: $E\{[Y - E(Y|x)]^2|x\}$. The formula

$$\text{Var}(Y|X = x) = E(Y^2|x) - [E(Y|x)]^2$$

is often useful for calculations. Only occasionally will we have to compute a conditional variance. But we will have to make assumptions about and manipulate conditional variances for certain topics in regression analysis.

As an example, let $Y = SAVING$ and $X = INCOME$ (both of these measured annually for the population of all families). Suppose that $\text{Var}(SAVING|INCOME) = 400 + .25\ INCOME$. This says that, as income increases, the variance in saving levels also increases. It is important to see that the relationship between the variance of *SAVING* and *INCOME* is totally separate from that between the *expected value* of *SAVING* and *INCOME*.

We state one useful property about the conditional variance.

PROPERTY CV.1
If X and Y are independent, then $\text{Var}(Y|X) = \text{Var}(Y)$.

This property is pretty clear, since the distribution of Y given X does not depend on X, and $\text{Var}(Y|X)$ is just one feature of this distribution.

B.5 THE NORMAL AND RELATED DISTRIBUTIONS

The Normal Distribution

The normal distribution, and those derived from it, are the most widely used distributions in statistics and econometrics. Assuming that random variables defined over populations are normally distributed simplifies probability calculations. In addition, we will

rely heavily on the normal and related distributions to conduct inference in statistics and econometrics—even when the underlying population is not necessarily normal. We must postpone the details, but be assured that these distributions will arise many times throughout this text.

A normal random variable is a continuous random variable that can take on any value. Its probability density function has the familiar bell shape graphed in Figure B.7.

Mathematically, the pdf of X can be written as

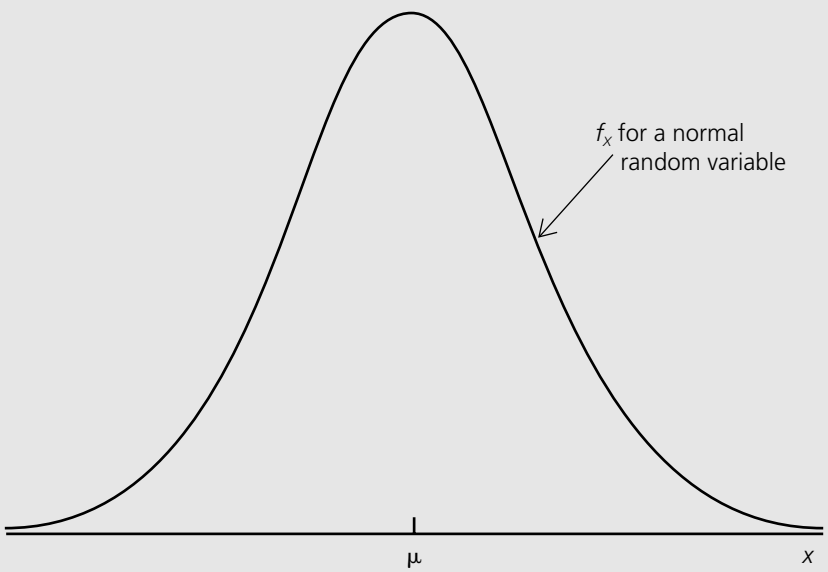
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-(x - \mu)^2/2\sigma^2], \quad -\infty < x < \infty, \tag{B.34}$$

where $\mu = \mathrm{E}(X)$, and $\sigma^2 = \mathrm{Var}(X)$. We say that X has a **normal distribution** with expected value μ and variance σ^2, written as $X \sim \mathrm{Normal}(\mu,\sigma^2)$. Because the normal distribution is symmetric about μ, μ is also the median of X. The normal distribution is sometimes called the *Gaussian distribution* after the famous statistician C. F. Gauss.

Certain random variables appear to roughly follow a normal distribution. Human heights and weights, test scores, and county unemployment rates have pdfs roughly the shape in Figure B.7. Other distributions, such as income distributions, do not appear to follow the normal probability function. In most countries, income is not symmetrically distributed about any value; the distribution is skewed towards the upper tail. In some

Figure B.7

The general shape of the normal probability density function.



cases, a variable can be transformed to achieve normality. A popular transformation is the natural log, which makes sense for positive random variables. If X is a positive random variable, such as income, and $Y = \log(X)$ has a normal distribution, then we say that X has a *lognormal distribution*. It turns out that the lognormal distribution fits income distribution pretty well in many countries. Other variables, such as prices of goods, appear to be well-described as lognormally distributed.

The Standard Normal Distribution

One special case of the normal distribution occurs when the mean is zero and the variance (and, therefore, the standard deviation) is unity. If a random variable Z has a Normal(0,1) distribution, then we say it has a **standard normal distribution**. The pdf of a standard normal random variable is denoted $\phi(z)$; from (B.34), with $\mu = 0$ and $\sigma^2 = 1$, it is given by

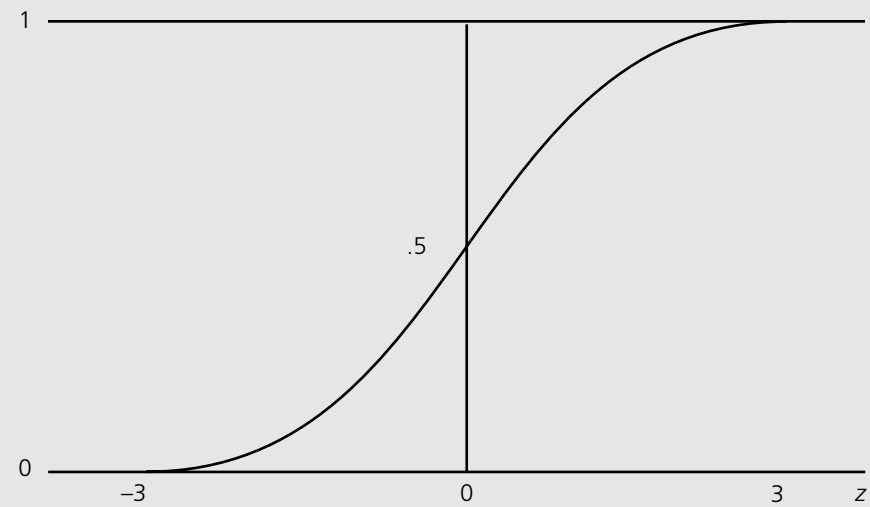
$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp(-z^2/2), \quad -\infty < z < \infty. \tag{B.35}$$

The standard normal cumulative distribution function is denoted $\Phi(z)$ and is obtained as the area under ϕ, to the left of z; see Figure B.8. Recall that $\Phi(z) = P(Z \leq z)$; since Z is continuous, $\Phi(z) = P(Z < z)$, as well.

There is no simple formula that can be used to obtain the values of $\Phi(z)$ [because $\Phi(z)$ is the integral of the function in (B.35), and this intregral has no closed form].

Figure B.8

The standard normal cumulative distribution function.



Nevertheless, the values for $\Phi(z)$ are easily tabulated; they are given for z between -3.1 and 3.1 in Table G.1. For $z \leq -3.1$, $\Phi(z)$ is less than .001, and for $z \geq 3.1$, $\Phi(z)$ is greater than .999. Most statistics and econometrics software packages include simple commands for computing values of the standard normal cdf, so we can often avoid printed tables entirely and obtain the probabilities for any value of z.

Using basic facts from probability—and, in particular, properties (B.7) and (B.8) concerning cdfs—we can use the standard normal cdf for computing the probability of any event involving a standard normal random variable. The most important formulas are

$$P(Z > z) = 1 - \Phi(z), \tag{B.36}$$

$$P(Z < -z) = P(Z > z), \tag{B.37}$$

and

$$P(a \leq Z \leq b) = \Phi(b) - \Phi(a). \tag{B.38}$$

Because Z is a continuous random variable, all three formulas hold whether or not the inequalities are strict. Some examples include $P(Z > .44) = 1 - .67 = .33$, $P(Z < -.92) = P(Z > .92) = 1 - .821 = .179$, and $P(-1 < Z \leq .5) = .692 - .159 = .533$.

Another useful expression is that, for any $c > 0$,

$$\begin{aligned} P(|Z| > c) &= P(Z > c) + P(Z < -c) \\ &= 2 \cdot P(Z > c) = 2[1 - \Phi(c)]. \end{aligned} \tag{B.39}$$

Thus, the probability that the absolute value of Z is bigger than some positive constant c is simply twice the probability $P(Z > c)$; this reflects the symmetry of the standard normal distribution.

In most applications, we start with a normally distributed random variable, $X \sim \text{Normal}(\mu,\sigma^2)$, where μ is different from zero, and $\sigma^2 \neq 1$. Any normal random variable can be turned into a standard normal using the following property.

PROPERTY NORMAL.1
If $X \sim \text{Normal}(\mu,\sigma^2)$, then $(X - \mu)/\sigma \sim \text{Normal}(0,1)$.

Property Normal.1 shows how to turn any normal random variable into a standard normal. Thus, suppose $X \sim \text{Normal}(3,4)$, and we would like to compute $P(X \leq 1)$. The steps always involve the normalization of X to a standard normal:

$$\begin{aligned} P(X \leq 1) &= P(X - 3 \leq 1 - 3) = P\left(\frac{X - 3}{2} \leq -1\right) \\ &= P(Z \leq -1) = \Phi(-1) = .159. \end{aligned}$$

EXAMPLE B.6
(Probabilities for a Normal Random Variable)

First, let us compute $P(2 < X \leq 6)$ when $X \sim \text{Normal}(4,9)$ (whether we use $<$ or $\leq$ is irrelevant because X is a continuous random variable). Now,

$$P(2 < X \leq 6) = P\left(\frac{2-4}{3} < \frac{X-4}{3} \leq \frac{6-4}{3}\right) = P(-2/3 < Z \leq 2/3)$$
$$= \Phi(.67) - \Phi(-.67) = .749 - .251 = .498.$$

Now, let us compute $P(|X| > 2)$:

$$P(|X| > 2) = P(X > 2) + P(X < -2) = 2 \cdot P(X > 2)$$
$$= 2 \cdot P\left(\frac{X-4}{3} > \frac{2-4}{3}\right) = 2 \cdot P(Z > -.67)$$
$$= 2[1 - \Phi(-.67)] = .772.$$

Additional Properties of the Normal Distribution

We end this subsection by collecting several other facts about normal distributions that we will later use.

PROPERTY NORMAL.2
If $X \sim \text{Normal}(\mu,\sigma^2)$, then $aX + b \sim \text{Normal}(a\mu + b, a^2\sigma^2)$.

Thus, if $X \sim \text{Normal}(1,9)$, then $Y = 2X + 3$ is distributed as normal with mean $2E(X) + 3 = 5$ and variance $2^2 \cdot 9 = 36$; $sd(Y) = 2sd(X) = 2 \cdot 3 = 6$.

Earlier we discussed how, in general, zero correlation and independence are not the same. In the case of normally distributed random variables, it turns out that zero correlation suffices for independence.

PROPERTY NORMAL.3
If X and Y are jointly normally distributed, then they are independent if and only if $\text{Cov}(X,Y) = 0$.

PROPERTY NORMAL.4
Any linear combination of independent, identically distributed normal random variables has a normal distribution.

For example, let X_i, $i = 1,2$, and 3, be independent random variables distributed as $\text{Normal}(\mu,\sigma^2)$. Define $W = X_1 + 2X_2 - 3X_3$. Then, W is normally distributed; we must simply find its mean and variance. Now,

$$E(W) = E(X_1) + 2E(X_2) - 3E(X_3) = \mu + 2\mu - 3\mu = 0.$$

Also,

$$\text{Var}(W) = \text{Var}(X_1) + 4\text{Var}(X_2) + 9\text{Var}(X_3) = 14\sigma^2.$$

Property Normal.4 also implies that the average of independent, normally distributed random variables has a normal distribution. If $Y_1, Y_2, \ldots, Y_n$ are independent random variables and each is distributed as Normal(μ,σ^2), then

$$\bar{Y} \sim \text{Normal}(\mu,\sigma^2/n). \tag{B.40}$$

This result is critical for statistical inference about the mean in a normal population.

The Chi-Square Distribution

The chi-square distribution is obtained directly from independent, standard normal random variables. Let Z_i, $i = 1,2, \ldots, n$, be independent random variables, each distributed as standard normal. Define a new random variable as the sum of the squares of the Z_i:

$$X = \sum_{i=1}^{n} Z_i^2. \tag{B.41}$$

Then, X has what is known as a **chi-square distribution** with n **degrees of freedom** (or *df* for short). We write this as $X \sim \chi_n^2$. The *df* in a chi-square distribution corresponds to the number of terms in the sum (B.41). The concept of degrees of freedom will play an important role in our statistical and econometric analyses.

The pdf for chi-square distributions with varying degrees of freedom is given in Figure B.9; we will not need the formula for this pdf, and so we do not reproduce it here. From equation (B.41), it is clear that a chi-square random variable is always nonnegative, and that, unlike the normal distribution, the chi-square distribution is not symmetric about any point. It can be shown that if $X \sim \chi_n^2$, then the expected value of X is n [the number of terms in (B.41)], and the variance of X is $2n$.

The *t* Distribution

The t distribution is the workhorse in classical statistics and multiple regression analysis. We obtain a t distribution from a standard normal and a chi-square random variable.

Let Z have a standard normal distribution and let X have a chi-square distribution with n degrees of freedom. Further, assume that Z and X are independent. Then, the random variable

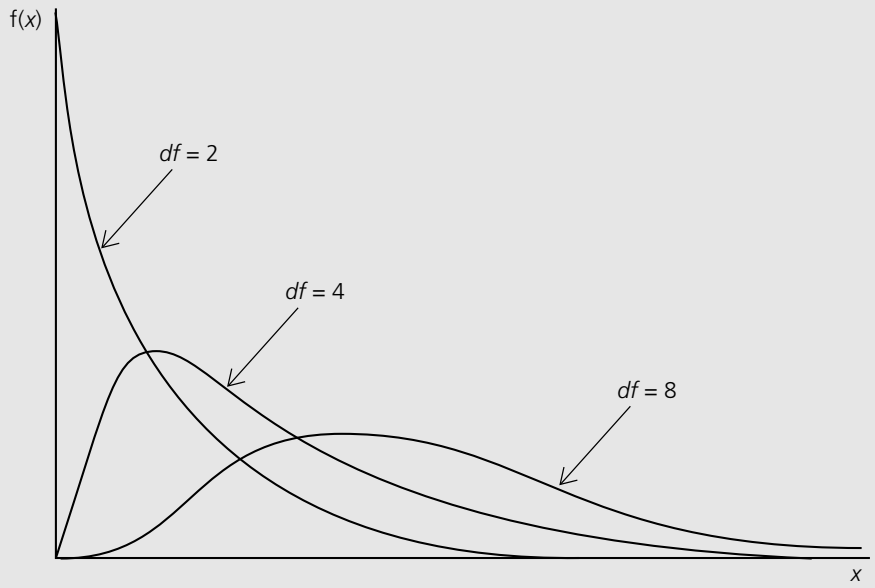
$$T = \frac{Z}{\sqrt{X/n}} \tag{B.42}$$

has a **t distribution** with n degrees of freedom. We will denote this by $T \sim t_n$. The t distribution gets its degrees of freedom from the chi-square random variable in the denominator of (B.42).

The pdf of the t distribution has a shape similar to that of the standard normal distribution, except that it is more spread out and therefore has more area in the tails. The

Figure B.9

The chi-square distribution with various degrees of freedom.



expected value of a t distributed random variable is zero (strictly speaking, the expected value exists only for $n > 1$), and the variance is $n/(n - 2)$ for $n > 2$. (The variance does not exist for $n \leq 2$ because the distribution is so spread out.) The pdf of the t distribution is plotted in Figure B.10 for various degrees of freedom. As the degrees of freedom gets large, the t distribution approaches the standard normal distribution.

The *F* Distribution

Another important distribution for statistics and econometrics is the F distribution. In particular, the F distribution will be used for testing hypotheses in the context of multiple regression analysis.

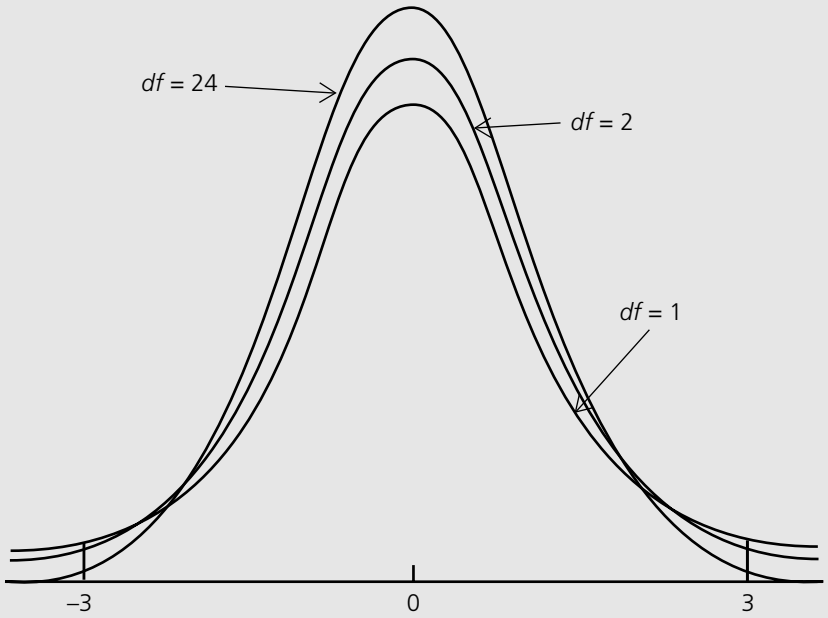
To define an F random variable, let $X_1 \sim \chi^2_{k_1}$ and $X_2 \sim \chi^2_{k_2}$ and assume that X_1 and X_2 are independent. Then, the random variable

$$F = \frac{(X_1/k_1)}{(X_2/k_2)} \tag{B.43}$$

has an **F distribution** with (k_1,k_2) degrees of freedom. We denote this as $F \sim F_{k_1,k_2}$. The pdf of the F distribution with different degrees of freedom is given in Figure B.11.

Figure B.10

The *t* distribution with various degrees of freedom.



The order of the degrees of freedom in F_{k_1,k_2} is critical. The integer k_1 is often called the *numerator degrees of freedom* because it is associated with the chi-square variable in the numerator. Likewise, the integer k_2 is called the *denominator degrees of freedom* because it is associated with the chi-square variable in the denominator. This can be a little tricky since (B.43) can also be written as $(X_1k_2)/(X_2k_1)$, so that k_1 appears in the denominator. Just remember that the numerator *df* is the integer associated with the chi-square variable in the numerator of (B.43), and similarly for the denominator *df*.

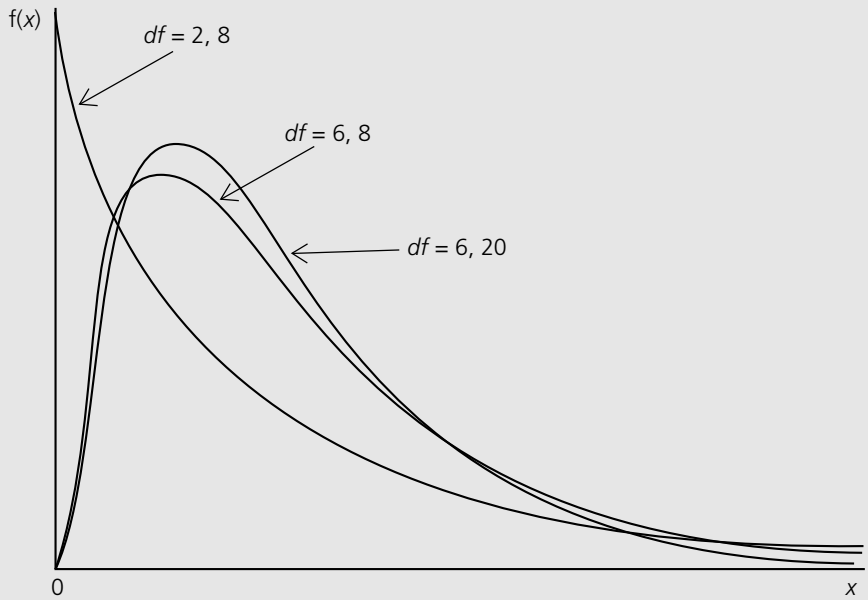
SUMMARY

In this appendix, we have reviewed the probability concepts that are needed in econometrics. Most of the concepts should be familiar from your introductory course in probability and statistics. Some of the more advanced topics, such as features of conditional expectations, do not need to be mastered now—there is time for that when these concepts arise in the context of regression analysis in Part 1.

In an introductory statistics course, the focus is on calculating means, variances, covariances, and so on, for particular distributions. In Part 1, we will not need such cal-

Figure B.11

The F_{k_1,k_2} distribution for various degrees of freedom, k_1 and k_2.



culations: we mostly rely on the *properties* of expectations, variances, and so on, that have been stated in this appendix.

KEY TERMS

Bernoulli (or Binary) Random Variable
Binomial Distribution
Chi-Square Distribution
Conditional Distribution
Conditional Expectation
Continuous Random Variable
Correlation Coefficient
Covariance
Cumulative Distribution Function (cdf)
Degrees of Freedom
Discrete Random Variable
Expected Value
Experiment
F Distribution
Independent Random Variables
Joint Distribution
Law of Iterated Expectations
Median
Normal Distribution
Pairwise Uncorrelated Random Variables
Probability Density Function (pdf)
Random Variable
Standard Deviation
Standard Normal Distribution
Standardized Random Variable
t Distribution
Variance

PROBLEMS

B.1 Suppose that a high school student is preparing to take the SAT exam. Explain why his or her eventual SAT score is properly viewed as a random variable.

B.2 Let X be a random variable distributed as Normal(5,4). Find the probabilities of the following events:

(i) $P(X \leq 6)$
(ii) $P(X > 4)$
(iii) $P(|X - 5| > 1)$

B.3 Much is made of the fact that certain mutual funds outperform the market year after year (that is, the return from holding shares in the mutual fund is higher than the return from holding a portfolio such as the S&P 500). For concreteness, consider a ten-year period and let the population be the 4,170 mutual funds reported in the *Wall Street Journal* on 1/6/95. By saying that performance relative to the market is random, we mean that each fund has a 50–50 chance of outperforming the market in any year and that performance is independent from year to year.

(i) If performance relative to the market is truly random, what is the probability that any particular fund outperforms the market in all 10 years?
(ii) Find the probability that *at least* one fund out of 4,170 funds outperforms the market in all 10 years. What do you make of your answer?
(iii) If you have a statistical package that computes binomial probabilities, find the probability that at least five funds outperform the market in all 10 years.

B.4 For a randomly selected county in the United States, let X represent the proportion of adults over age 65 who are employed, or the elderly employment rate. Then, X is restricted to a value between zero and one. Suppose that the cumulative distribution function for X is given by $F(x) = 3x^2 - 2x^3$ for $0 \leq x \leq 1$. Find the probability that the elderly employment rate is at least .6 (60%).

B.5 Just prior to jury selection for O. J. Simpson's murder trial in 1995, a poll found that about 20% of the adult population believed Simpson was innocent (after much of the physical evidence in the case had been revealed to the public). Ignore the fact that this 20% is an estimate based on a subsample from the population; for illustration, take it as the true percentage of people who thought Simpson was innocent prior to jury selection. Assume that the 12 jurors were selected randomly and independently from the population (although this turned out not to be true).

(i) Find the probability that the jury had at least one member who believed in Simpson's innocence prior to jury selection. (*Hint*: Define the Binomial(12,.20) random variable X to be the number of jurors believing in Simpson's innocence.)
(ii) Find the probability that the jury had at least two members who believed in Simpson's innocence. [*Hint*: $P(X \geq 2) = 1 - P(X \leq 1)$, and $P(X \leq 1) = P(X = 0) + P(X = 1)$.]

B.6 (Requires calculus) Let X denote the prison sentence, in years, for people convicted of auto theft in a particular state in the United States. Suppose that the pdf of X is given by

$$f(x) = (1/9)x^2, \ 0 < x < 3.$$

Use integration to find the expected prison sentence.

B.7 If a basketball player is a 74% free-throw shooter, then, on average, how many free throws will he or she make in a game with eight free-throw attempts?

B.8 Suppose that a college student is taking three courses: a two-credit course, a three-credit course, and a four-credit course. The expected grade in the two-credit course is 3.5, while the expected grade in the three- and four-credit courses is 3.0. What is the expected overall grade point average for the semester? (Remember that each course grade is weighted by its share of the total number of units.)

B.9 Let X denote the annual salary of university professors in the United States, measured in thousands of dollars. Suppose that the average salary is 52.3, with a standard deviation of 14.6. Find the mean and standard deviation when salary is measured in dollars.

B.10 Suppose that at a large university, college grade point average, *GPA*, and SAT score, *SAT*, are related by the conditional expectation $\mathrm{E}(GPA|SAT) = .70 + .002\,SAT$.

(i) Find the expected *GPA* when $SAT = 800$. Find $\mathrm{E}(GPA|SAT = 1{,}400)$. Comment on the difference.

(ii) If the average *SAT* in the university is 1,100, what is the average *GPA*? (*Hint*: Use Property CE.4.)

Appendix C

Fundamentals of Mathematical Statistics

C.1 POPULATIONS, PARAMETERS, AND RANDOM SAMPLING

Statistical inference involves learning something about a population given the availability of a sample from that population. By **population**, we mean any well-defined group of subjects, which could be individuals, firms, cities, or many other possibilities. By "learning," we can mean several things, which are broadly divided into the categories of *estimation* and *hypothesis testing*.

A couple of examples may help you understand these terms. In the population of all working adults in the United States, labor economists are interested in learning about the return to education, as measured by the average percentage increase in earnings given another year of education. It would be impractical and costly to obtain information on earnings and education for the entire working population in the United States, but we can obtain data on a subset of the population. Using the data collected, a labor economist may report that his or her best estimate of the return to another year of education is 7.5%. This is an example of a *point estimate*. Or, she or he may report a range, such as "the return to education is between 5.6% and 9.4%." This is an example of an *interval estimate*.

An urban economist might want to know whether neighborhood crime watch programs are associated with lower crime rates. After comparing crime rates of neighborhoods with and without such programs in a sample from the population, he or she can draw one of two conclusions: neighborhood watch programs do affect crime, or they do not. This example falls under the rubric of hypothesis testing.

The first step in statistical inference is to identify the population of interest. This may seem obvious, but it is important to be very specific. Once we have identified the population, we can specify a model for the population relationship of interest. Such models involve probability distributions or features of probability distributions, and these depend on unknown parameters. Parameters are simply constants that determine the directions and strengths of relationships among variables. In the labor economics example above, the parameter of interest is the return to education in the population.

Sampling

For reviewing statistical inference, we focus on the simplest possible setting. Let Y be a random variable representing a population with a probability density function $f(y;\theta)$, which depends on the single parameter θ. The probability density function (pdf) of Y is assumed to be known except for the value of θ; different values of θ imply different population distributions, and therefore we are interested in the value of θ. If we can obtain certain kinds of samples from the population, then we can learn something about θ. The easiest sampling scheme to deal with is random sampling.

RANDOM SAMPLING

If $Y_1, Y_2, \ldots, Y_n$ are independent random variables with a common probability density function $f(y;\theta)$, then $\{Y_1, \ldots, Y_n\}$ is said to be a **random sample** from $f(y;\theta)$ [or a random sample from the population represented by $f(y;\theta)$].

When $\{Y_1, \ldots, Y_n\}$ is a random sample from the density $f(y;\theta)$, we also say that the Y_i are *independent, identically distributed* (or *i.i.d.*) samples from $f(y;\theta)$. In some cases, we will not need to entirely specify what the common distribution is.

The random nature of $Y_1, Y_2, \ldots, Y_n$ in the definition of random sampling reflects the fact that many different outcomes are possible before the sampling is actually carried out. For example, if family income is obtained for a sample of $n = 100$ families in the United States, the incomes we observe will usually differ for each different sample of 100 families. Once a sample is obtained, we have a set of numbers, say $\{y_1, y_2, \ldots, y_n\}$, which constitute the data that we work with. Whether or not it is appropriate to assume the sample came from a random sampling scheme requires knowledge about the actual sampling process.

Random samples from a Bernoulli distribution are often used to illustrate statistical concepts, and they also arise in empirical applications. If $Y_1, Y_2, \ldots, Y_n$ are independent random variables and each is distributed as Bernoulli(θ), so that $P(Y_i = 1) = \theta$ and $P(Y_i = 0) = 1 - \theta$, then $\{Y_1, Y_2, \ldots, Y_n\}$ constitutes a random sample from the Bernoulli(θ) distribution. As an illustration, consider the airline reservation example carried along in Appendix B. Each Y_i denotes whether customer i shows up for his or her reservation; $Y_i = 1$ if passenger i shows up, and $Y_i = 0$ otherwise. Here, θ is the probability that a randomly drawn person from the population of all people who make airline reservations shows up for his or her reservation.

For many other applications, random samples can be assumed to be drawn from a normal distribution. If $\{Y_1, \ldots, Y_n\}$ is a random sample from the Normal(μ, σ^2) population, then the population is characterized by two parameters, the mean μ and the variance σ^2. Primary interest usually lies in μ, but σ^2 is of interest in its own right because making inferences about μ often requires learning about σ^2.

C.2 FINITE SAMPLE PROPERTIES OF ESTIMATORS

In this section, we study what are called finite sample properties of estimators. The term "finite sample" comes from the fact that the properties hold for a sample of any size, no matter how large or small. Sometimes, these are called small sample properties. In

Section C.3, we cover "asymptotic properties," which have to do with the behavior of estimators as the sample size grows without bound.

Estimators and Estimates

To study properties of estimators, we must define what we mean by an estimator. Given a random sample $\{Y_1, Y_2, \ldots, Y_n\}$ drawn from a population distribution that depends on an unknown parameter θ, an **estimator** of θ is a rule that assigns each possible outcome of the sample a value of θ. The rule is specified before any sampling is carried out; in particular, the rule is the same, regardless of the data actually obtained.

As an example of an estimator, let $\{Y_1, \ldots, Y_n\}$ be a random sample from a population with mean μ. A natural estimator of μ is the average of the random sample:

$$\bar{Y} = n^{-1} \sum_{i=1}^{n} Y_i. \qquad \textbf{(C.1)}$$

$\bar{Y}$ is called the **sample average** but, unlike in Appendix A where we defined the sample average of a set of numbers as a descriptive statistic, $\bar{Y}$ is now viewed as an estimator. Given any outcome of the random variables $Y_1, \ldots, Y_n$, we use the same rule to estimate μ: we simply average them. For actual data outcomes $\{y_1, \ldots, y_n\}$, the **estimate** is just the average in the sample: $\bar{y} = (y_1 + y_2 + \ldots + y_n)/n$.

EXAMPLE C.1
(City Unemployment Rates)

Suppose we obtain the following sample of unemployment rates for 10 cities in the United States:

City	Unemployment Rate
1	5.1
2	6.4
3	9.2
4	4.1
5	7.5
6	8.3
7	2.6

continued

City	Unemployment Rate
8	3.5
9	5.8
10	7.5

Our estimate of the average city unemployment rate in the United States is $\bar{y} = 6.0$. Each sample generally results in a different estimate. But the *rule* for obtaining the estimate is the same, regardless of which cities appear in the sample, or how many.

More generally, an estimator W of a parameter θ can be expressed as an abstract mathematical formula:

$$W = h(Y_1, Y_2, \ldots, Y_n), \tag{C.2}$$

for some known function h of the random variables Y_1, Y_2, …, Y_n. As with the special case of the sample average, W is a random variable because it depends on the random sample: as we obtain different random samples from the population, the value of W can change. When a particular set of numbers, say $\{y_1, y_2, \ldots, y_n\}$, is plugged into the function h, we obtain an *estimate* of θ, denoted $w = h(y_1, \ldots, y_n)$. Sometimes, W is called a point estimator and w a point estimate to distinguish these from *interval* estimators and estimates, which we will come to in Section C.4.

For evaluating estimation procedures, we study various properties of the probability distribution of the random variable W. The distribution of an estimator is often called its **sampling distribution**, since this distribution describes the likelihood of various outcomes of W across different random samples. Because there are unlimited rules for combining data to estimate parameters, we need some sensible criteria for choosing among estimators, or at least for eliminating some estimators from consideration. Therefore, we must leave the realm of descriptive statistics, where we compute things such as sample average to simply summarize a body of data. In mathematical statistics, we study the sampling distributions of estimators.

Unbiasedness

In principle, the entire sampling distribution of W can be obtained given the probability distribution of Y_i and the function h. It is usually easier to focus on a few features of the distribution of W in evaluating it as an estimator of θ. The first important property of an estimator involves its expected value.

UNBIASED ESTIMATOR

An estimator, W of θ, is *unbiased* if

$$\mathrm{E}(W) = \theta, \tag{C.3}$$

for all possible values of θ.

If an estimator is unbiased, then its probability distribution has an expected value equal to the parameter it is supposed to be estimating. **Unbiasedness** does *not* mean that the estimate we get with any particular sample is equal to θ, or even very close to θ. Rather, if we could *indefinitely* draw random samples on Y from the population, compute an estimate each time, and then average these estimates over all random samples, we would obtain θ. This thought experiment is abstract, because in most applications, we just have one random sample to work with.

For an estimator that is not unbiased, we define its bias as follows.

BIAS OF AN ESTIMATOR

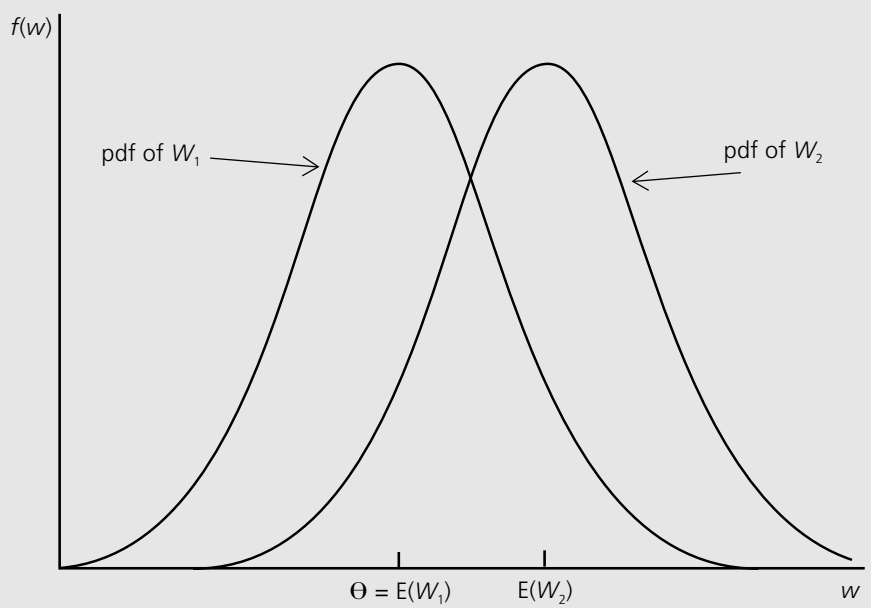
If W is an estimator of θ, its **bias** is defined as

$$\mathrm{Bias}(W) \equiv \mathrm{E}(W) - \theta. \tag{C.4}$$

Figure C.1 shows two estimators, the first of which is unbiased and the second of which has a positive bias.

Figure C.1

An unbiased estimator, W_1, and an estimator with positive bias, W_2.



The unbiasedness of an estimator and the size of any possible bias depend on the distribution of Y and on the function h. The distribution of Y is usually beyond our control (although we often choose a *model* for this distribution): it may be determined by nature or social forces. But the choice of the rule h is ours, and if we want an unbiased estimator, then we must choose h accordingly.

Some estimators can be shown to be unbiased quite generally. We now show that the sample average $\bar{Y}$ is an unbiased estimator of the population mean μ, regardless of the underlying population distribution. We use the properties of expected values (E.1 and E.2) that we covered in Section B.3:

$$\begin{aligned} \mathrm{E}(\bar{Y}) &= \mathrm{E}\left((1/n)\sum_{i=1}^{n} Y_i\right) = (1/n)\mathrm{E}\left(\sum_{i=1}^{n} Y_i\right) = (1/n)\left(\sum_{i=1}^{n} \mathrm{E}(Y_i)\right) \\ &= (1/n)\left(\sum_{i=1}^{n} \mu\right) = (1/n)(n\mu) = \mu. \end{aligned}$$

For hypothesis testing, we will need to estimate the variance σ^2 from a population with mean μ. Letting $\{Y_1, \ldots, Y_n\}$ denote the random sample from the population with $\mathrm{E}(Y) = \mu$ and $\mathrm{Var}(Y) = \sigma^2$, define the estimator as

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n} (Y_i - \bar{Y})^2, \qquad \textbf{(C.5)}$$

which is usually called the **sample variance**. It can be shown that S^2 is unbiased for σ^2: $\mathrm{E}(S^2) = \sigma^2$. The division by $n - 1$, rather than n, accounts for the fact that the mean μ is estimated rather than known. If μ were known, an unbiased estimator of σ^2 would be $n^{-1}\sum_{i=1}^{n} (Y_i - \mu)^2$, but μ is rarely known in practice.

Although unbiasedness has a certain appeal as a property for an estimator—indeed, its antonym, "biased", has decidedly negative connotations—it is not without its problems. One weakness of unbiasedness is that some reasonable, and even some very good estimators, are not unbiased. We will see an example shortly.

Another important weakness of unbiasedness is that unbiased estimators exist that are actually quite poor estimators. Consider estimating the mean μ from a population. Rather than using the sample average $\bar{Y}$ to estimate μ, suppose that, after collecting a sample of size n, we discard all of the observations except the first. That is, our estimator of μ is simply $W \equiv Y_1$. This estimator is unbiased since $\mathrm{E}(Y_1) = \mu$. Hopefully, you sense that ignoring all but the first observation is not a prudent approach to estimation: it throws out most of the information in the sample. For example, with $n = 100$, we obtain 100 outcomes of the random variable Y, but then we use only the first of these to estimate $\mathrm{E}(Y)$.

The Sampling Variance of Estimators

The example at the end of the previous subsection shows that we need additional criteria in order to evaluate estimators. Unbiasedness only ensures that the probability distribution of an estimator has a mean value equal to the parameter it is supposed to be estimating. This is fine, but we also need to know how spread out the distribution of an

estimator is. An estimator can be equal to θ, on average, but it can also be very far away with large probability. In Figure C.2, W_1 and W_2 are both unbiased estimators of θ. But the distribution of W_1 is more tightly centered about θ: the probability that W_1 is greater than any given distance from θ is less than the probability that W_2 is greater than that same distance from θ. Using W_1 as our estimator means that it is less likely that we will obtain a random sample that yields an estimate very far from θ.

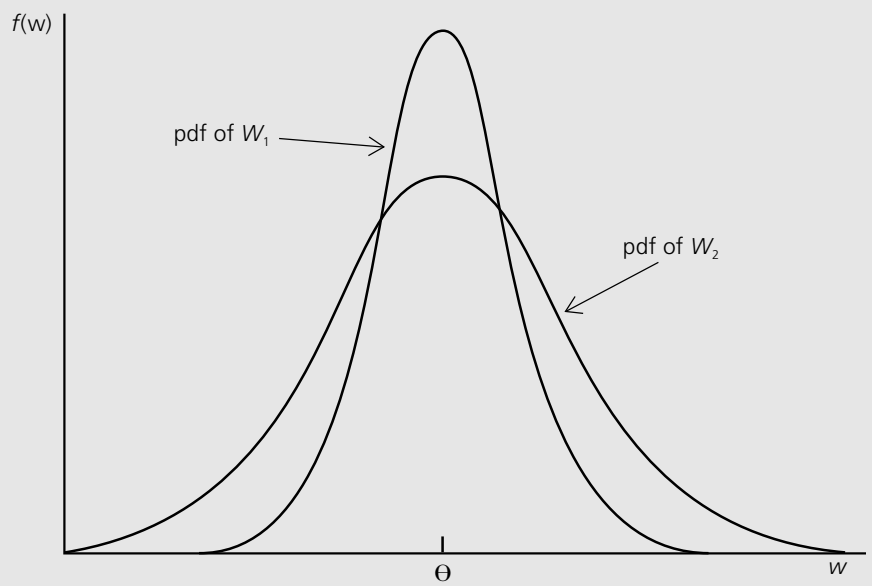
To summarize the situation shown in Figure C.2, we rely on the variance (or standard deviation) of an estimator. Recall that this gives a single measure of the dispersion in the distribution. The variance of an estimator is often called its **sampling variance**, since it is the variance associated with a sampling distribution. Remember, the sampling variance is not a random variable; it is a constant, but it might be unknown.

We now obtain the variance of the sample average for estimating the mean μ from a population:

$$\begin{aligned}\text{Var}(\bar{Y}) &= \text{Var}\left((1/n)\sum_{i=1}^{n} Y_i\right) = (1/n^2)\text{Var}\left(\sum_{i=1}^{n} Y_i\right) = (1/n^2)\left(\sum_{i=1}^{n} \text{Var}(Y_i)\right) \\ &= (1/n^2)\left(\sum_{i=1}^{n} \sigma^2\right) = (1/n^2)(n\sigma^2) = \sigma^2/n.\end{aligned} \tag{C.6}$$

Figure C.2

The sampling distributions of two unbiased estimators of θ.



Notice how we used the properties of variance from Sections B.3 and B.4 (VAR.2 and VAR.4), as well as the independence of the Y_i. To summarize: If $\{Y_i: i = 1,2,\ldots,n\}$ is a random sample from a population with mean μ and variance σ^2, then $\bar{Y}$ has the same mean as the population, but its sampling variance equals the population variance, σ^2, over the sample size.

An important implication of $\text{Var}(\bar{Y}) = \sigma^2/n$ is that it can be made very close to zero by increasing the sample size n. This is a key feature of a reasonable estimator, and we return to it in Section C.3.

As suggested by Figure C.2, among unbiased estimators, we prefer the estimator with the smallest variance. This allows us to eliminate certain estimators from consideration. For a random sample from a population with mean μ and variance σ^2, we know that $\bar{Y}$ is unbiased, and $\text{Var}(\bar{Y}) = \sigma^2/n$. What about the estimator Y_1, which is just the first observation drawn? Since Y_1 is a random draw from the population, $\text{Var}(Y_1) = \sigma^2$. Thus, the difference between $\text{Var}(Y_1)$ and $\text{Var}(\bar{Y})$ can be large even for small sample sizes. If $n = 10$, then $\text{Var}(Y_1)$ is ten times as large as $\text{Var}(\bar{Y}) = \sigma^2/10$. This gives us a formal way of excluding Y_1 as an estimator of μ.

To emphasize this point, Table C.1 contains the outcome of a small simulation study. Using the statistical package Stata, 20 random samples of size 10 were generated from a normal distribution, with $\mu = 2$ and $\sigma^2 = 1$; we are interested in estimating μ here. For each of the 20 random samples, we compute two estimates, y_1 and $\bar{y}$; these values are listed in Table C.1. As can be seen from the table, the values for y_1 are much more spread out than those for $\bar{y}$: y_1 ranges from -0.64 to 4.27, while $\bar{y}$ ranges only from 1.16 to 2.58. Further, in 16 out of 20 cases, $\bar{y}$ is closer than y_1 to $\mu = 2$. The average of y_1 across the simulations is about 1.89, while that for $\bar{y}$ is 1.96. The fact that these averages are close to 2 illustrates the unbiasedness of both estimators (and we could get these averages closer to 2 by doing more than 20 replications). But comparing just the average outcomes across random draws masks the fact that the sample average $\bar{Y}$ is far superior to Y_1 as an estimator of μ.

Table C.1

Simulation of Estimators for a Normal(μ,1) Distribution with $\mu = 2$

Replication	y_1	$\bar{y}$
1	−0.64	1.98
2	1.06	1.43
3	4.27	1.65
4	1.03	1.88
5	3.16	2.34

continued

Table C.1 (*concluded*)

Replication	y_1	$\bar{y}$
6	2.77	2.58
7	1.68	1.58
8	2.98	2.23
9	2.25	1.96
10	2.04	2.11
11	0.95	2.15
12	1.36	1.93
13	2.62	2.02
14	2.97	2.10
15	1.93	2.18
16	1.14	2.10
17	2.08	1.94
18	1.52	2.21
19	1.33	1.16
20	1.21	1.75

Efficiency

Comparing the variances of $\bar{Y}$ and Y_1 in the previous subsection is an example of a general approach to comparing different unbiased estimators.

RELATIVE EFFICIENCY

If W_1 and W_2 are two unbiased estimators of θ, W_1 is efficient relative to W_2 when $\text{Var}(W_1) \leq \text{Var}(W_2)$ for all θ, with strict inequality for at least one value of θ.

Earlier, we showed that, for estimating the population mean μ, $\text{Var}(\bar{Y}) < \text{Var}(Y_1)$ for any value of σ^2 whenever $n > 1$. Thus, $\bar{Y}$ is efficient relative to Y_1 for estimating μ. We can-

not always choose between unbiased estimators based on the smallest variance criterion: given two unbiased estimators of θ, one can have smaller variance from some values of θ, while the other can have smaller variance for other values of θ.

If we restrict our attention to a certain class of estimators, we can show that the sample average has the smallest variance. Problem C.2 asks you to show that $\bar{Y}$ has the smallest variance among all unbiased estimators that are also linear functions of Y_1, $Y_2, \ldots, Y_n$. The assumptions are that the Y_i have common mean and variance, and they are pairwise uncorrelated.

If we do not restrict our attention to unbiased estimators, then comparing variances is meaningless. For example, when estimating the population mean μ, we can use a trivial estimator that is equal to zero, regardless of the sample that we draw. Naturally, the variance of this estimator is zero (since it is the same value for every random sample). But the bias of this estimator is $-\mu$, and so it is a very poor estimator when $|\mu|$ is large.

One way to compare estimators that are not necessarily unbiased is to compute the **mean squared error (MSE)** of the estimators. If W is an estimator of θ, then the MSE of W is defined as $\text{MSE}(W) = \text{E}[(W - \theta)^2]$. The MSE measures how far, on average, the estimator is away from θ. It can be shown that $\text{MSE}(W) = \text{Var}(W) + [\text{Bias}(W)]^2$, so that MSE($W$) depends on the variance and bias (if any is present). This allows us to compare two estimators when one or both are biased.

C.3 ASYMPTOTIC OR LARGE SAMPLE PROPERTIES OF ESTIMATORS

In Section C.2, we encountered the estimator Y_1 for the population mean μ, and we saw that, even though it is unbiased, it is a poor estimator because its variance can be much larger than that of the sample mean. One notable feature of Y_1 is that it has the same variance for any sample size. It seems reasonable to require any estimation procedure to improve as the sample size increases. For estimating a population mean μ, $\bar{Y}$ improves in the sense that its variance gets smaller as n gets larger; Y_1 does not improve in this sense.

We can rule out certain silly estimators by studying the *asymptotic* or *large sample* properties of estimators. In addition, we can say something positive about estimators that are not unbiased and whose variances are not easily found.

Asymptotic analysis involves approximating the features of the sampling distribution of an estimator. These approximations depend on the size of the sample. Unfortunately, we are necessarily limited in what we can say about how "large" a sample size is needed for asymptotic analysis to be appropriate; this depends on the underlying population distribution. But large sample approximations have been known to work well for sample sizes as small as $n = 20$.

Consistency

The first asymptotic property of estimators concerns how far the estimator is likely to be from the parameter it is supposed to be estimating as we let the sample size increase indefinitely.

CONSISTENCY

Let W_n be an estimator of θ based on a sample $Y_1, Y_2, \ldots, Y_n$ of size n. Then, W_n is a **consistent estimator** of θ, if for every $\varepsilon > 0$,

$$P(|W_n - \theta| > \varepsilon) \to 0 \text{ as } n \to \infty. \qquad \textbf{(C.7)}$$

If W_n is not consistent for θ, then we say it is **inconsistent**.

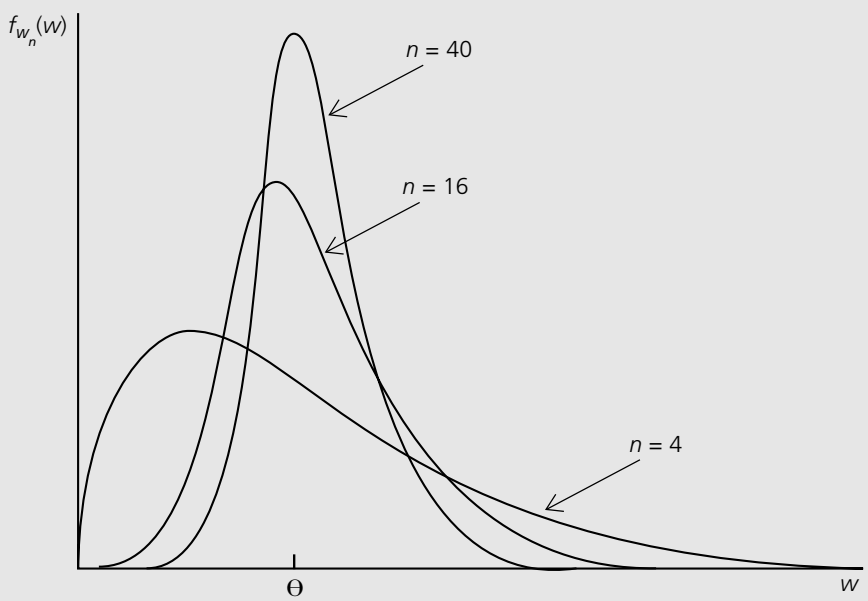
When W_n is consistent, we also say that θ is the **probability limit** of W_n, written as $\text{plim}(W_n) = \theta$.

Unlike unbiasedness—which is a feature of an estimator for a given sample size—consistency involves the behavior of the sampling distribution of the estimator as the sample size n gets large. To emphasize this, we have indexed the estimator by the sample size in stating this definition, and we will continue with this convention throughout this section.

Equation (C.7) looks technical, and it can be rather difficult to establish based on fundamental probability principles. By contrast, interpreting (C.7) is straightforward. It means that the distribution of W_n becomes more and more concentrated about θ, which roughly means that for larger sample sizes, W_n is less and less likely to be very far from θ. This tendency is illustrated in Figure C.3.

Figure C.3

The sampling distributions of a consistent estimator for three sample sizes.



If an estimator is not consistent, then it does not help us to learn about θ, even with an unlimited amount of data. For this reason, consistency is a minimal requirement of an estimator used in statistics or econometrics. We will encounter estimators that are consistent under certain assumptions and inconsistent when those assumptions fail. When estimators are inconsistent, we can usually find their probability limits, and it will be important to know how far these probability limits are from θ.

As we noted earlier, unbiased estimators are not necessarily consistent, but those whose variances shrink to zero as the sample size grows *are* consistent. This can be stated formally: If W_n is an unbiased estimator of θ and $\text{Var}(W_n) \to 0$ as $n \to \infty$, then $\text{plim}(W_n) = \theta$. Unbiased estimators that use the entire data sample will usually have a variance that shrinks to zero as the sample size grows, thereby being consistent.

A good example of a consistent estimator is the average of a random sample drawn from a population with μ and variance σ^2. We have already shown that the sample average is unbiased for μ. In equation (C.6), we derived $\text{Var}(\bar{Y}_n) = \sigma^2/n$ for any sample size n. Therefore, $\text{Var}(\bar{Y}_n) \to 0$ as $n \to \infty$, and so $\bar{Y}_n$ is a consistent estimator of μ (in addition to being unbiased).

The conclusion that $\bar{Y}_n$ is consistent for μ holds even if $\text{Var}(\bar{Y}_n)$ does not exist. This classic result is known as the **law of large numbers (LLN)**.

LAW OF LARGE NUMBERS

Let $Y_1, Y_2, \ldots, Y_n$ be independent, identically distributed random variables with mean μ. Then,

$$\text{plim}(\bar{Y}_n) = \mu. \tag{C.8}$$

The law of large numbers means that, if we are interested in estimating the population average μ, we can get arbitrarily close to μ by choosing a sufficiently large sample. This fundamental result can be combined with basic properties of plims to show that fairly complicated estimators are consistent.

PROPERTY PLIM.1

Let θ be a parameter and define a new parameter, $\gamma = g(\theta)$, for some *continuous* function $g(\theta)$. Suppose that $\text{plim}(W_n) = \theta$. Define an estimator of γ by $G_n = g(W_n)$. Then,

$$\text{plim}(G_n) = \gamma. \tag{C.9}$$

This is often stated as

$$\text{plim } g(W_n) = g(\text{plim } W_n) \tag{C.10}$$

for a continuous function $g(\theta)$.

The assumption that $g(\theta)$ is continuous is a technical requirement that has often been described nontechnically as "a function that can be graphed without lifting your pencil from the paper." Since all of the functions we encounter in this text are continuous, we do not provide a formal definition of a continuous function. Examples of continuous

functions are $g(\theta) = a + b\theta$ for constants a and b, $g(\theta) = \theta^2$, $g(\theta) = 1/\theta$, $g(\theta) = \sqrt{\theta}$, $g(\theta) = \exp(\theta)$, and many variants on these. We will not need to mention the continuity assumption again.

As an important example of a consistent but biased estimator, consider estimating the standard deviation, σ, from a population with mean μ and variance σ^2. We already claimed that the sample variance $S_n^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ is unbiased for σ^2. Using the law of large numbers and some algebra, S_n^2 can also be shown to be consistent for σ^2. The natural estimator of $\sigma = \sqrt{\sigma^2}$ is $S_n = \sqrt{S_n^2}$ (where the square root is always the positive square root). S_n, which is called the **sample standard deviation**, is *not* an unbiased estimator because the expected value of the square root is *not* the square root of the expected value (see Section B.3). Nevertheless, by PLIM.1, plim $S_n = \sqrt{\text{plim } S_n^2} = \sqrt{\sigma^2} = \sigma$, so S_n is a consistent estimator of σ.

Here are some other useful properties of the probability limit:

PROPERTY PLIM.2

If plim$(T_n) = \alpha$ and plim$(U_n) = \beta$, then

(i) plim$(T_n + U_n) = \alpha + \beta$;

(ii) plim$(T_n U_n) = \alpha\beta$;

(iii) plim$(T_n/U_n) = \alpha/\beta$, provided $\beta \neq 0$.

These three facts about probability limits allow us to combine consistent estimators in a variety of ways to get other consistent estimators. For example, let $\{Y_1, \ldots, Y_n\}$ be a random sample of size n on annual earnings from the population of workers with a high school education and denote the population mean by μ_Y. Let $\{Z_1, \ldots, Z_n\}$ be a random sample on annual earnings from the population of workers with a college education and denote the population mean by μ_Z. We wish to estimate the percentage difference in annual earnings between the two groups, which is $\gamma = 100\cdot(\mu_Z - \mu_Y)/\mu_Y$. (This is the percent by which average earnings for college graduates differs from average earnings for high school graduates.) Since $\bar{Y}_n$ is consistent for μ_Y, and $\bar{Z}_n$ is consistent for μ_Z, it follows from PLIM.1 and part (iii) of PLIM.2 that

$$G_n \equiv 100\cdot(\bar{Z}_n - \bar{Y}_n)/\bar{Y}_n$$

is a consistent estimator of γ. G_n is just the percentage difference between $\bar{Z}_n$ and $\bar{Y}_n$ in the sample, so it is a natural estimator. G_n is not an unbiased estimator of γ, but it is still a good estimator unless n is small.

Asymptotic Normality

Consistency is a property of point estimators. While it does tell us that the distribution of the estimator is collapsing around the parameter as the sample size gets large, it tells us essentially nothing about the *shape* of that distribution for a given sample size. For constructing interval estimators and testing hypotheses, we need a way to approximate the distribution of our estimators. Most econometric estimators have distributions that are well-approximated by a normal distribution for large samples, which motivates the following definition.

ASYMPTOTIC NORMALITY

Let $\{Z_n: n = 1,2,\ldots\}$ be a sequence of random variables, such that for all numbers z,

$$P(Z_n \leq z) \rightarrow \Phi(z) \text{ as } n \rightarrow \infty, \tag{C.11}$$

where $\Phi(z)$ is the standard normal cumulative distribution function. Then, Z_n is said to have an *asymptotic standard normal distribution*. In this case, we often write $Z_n \overset{a}{\sim}$ Normal(0,1). (The "a" above the tilda stands for "asymptotically" or "approximately.")

Property (C.11) means that the cumulative distribution function for Z_n gets closer and closer to the cdf of the standard normal distribution, as the sample size n gets large. When **asymptotic normality** holds, for large n, we have the approximation $P(Z_n \leq z) \approx \Phi(z)$. Thus, probabilities concerning Z_n can be approximated by standard normal probabilities.

The **central limit theorem (CLT)** is one of the most powerful results in probability and statistics. It states that the average from a random sample for *any* population (with finite variance), when standardized, has an asymptotic standard normal distribution.

CENTRAL LIMIT THEOREM

Let $\{Y_1, Y_2, \ldots, Y_n\}$ be a random sample with mean μ and variance σ^2. Then,

$$Z_n = \frac{\bar{Y}_n - \mu}{\sigma/\sqrt{n}}, \tag{C.12}$$

has an asymptotic standard normal distribution.

The variable Z_n in (C.12) is the standardized version of $\bar{Y}_n$: we have subtracted off $E(\bar{Y}_n) = \mu$ and divided by $sd(\bar{Y}_n) = \sigma/\sqrt{n}$. Thus, regardless of the population distribution of Y, Z_n has mean zero and variance one, which coincides with the mean and variance of the standard normal distribution. Remarkably, the entire distribution of Z_n gets arbitrarily close to the standard normal distribution as n gets large.

Most estimators encountered in statistics and econometrics can be written as functions of sample averages, in which case, we can apply the law of large numbers and the central limit theorem. When two consistent estimators have asymptotic normal distributions, we choose the estimator with the smallest asymptotic variance.

In addition to the standardized sample average in (C.12), many other statistics that depend on sample averages turn out to be asymptotically normal. An important one is obtained by replacing σ with its consistent estimator S_n in equation (C.12):

$$\frac{\bar{Y}_n - \mu}{S_n/\sqrt{n}} \tag{C.13}$$

also has an approximate standard normal distribution for large n. The exact (finite sample) distributions of (C.12) and (C.13) are definitely not the same, but the difference is often small enough to be ignored for large n.

Throughout this section, each estimator has been subscripted by n to emphasize the nature of asymptotic or large sample analysis. Continuing this convention clutters the notation without providing additional insight, once the fundamentals of asymptotic analysis are understood. Henceforth, we drop the n subscript and rely on you to remember that estimators depend on the sample size, and properties such as consistency and asymptotic normality refer to the growth of the sample size without bound.

C.4 GENERAL APPROACHES TO PARAMETER ESTIMATION

Up to this point, we have used the sample average to illustrate the finite and large sample properties of estimators. It is natural to ask: Are there general approaches to estimation that produce estimators with good properties, such as unbiasedness, consistency, and efficiency?

The answer is yes. A detailed treatment of various approaches to estimation is beyond the scope of this text; here, we provide only an informal discussion. A thorough discussion is given in Larsen and Marx (1986, Chapter 5).

Method of Moments

Given a parameter θ appearing in a population distribution, there are usually many ways to obtain unbiased and consistent estimators of θ. Trying all different possibilities and comparing them on the basis of the criteria in Sections C.2 and C.3 is not practical. Fortunately, some methods have been shown to have good general properties, and for the most part, the logic behind them is intuitively appealing.

In the previous sections, we have seen some examples of **method of moments** procedures. Basically, method of moments estimation proceeds as follows. The parameter θ is shown to be related to some expected value in the distribution of Y, usually $\mathrm{E}(Y)$ or $\mathrm{E}(Y^2)$ (although more exotic choices are sometimes used). Suppose, for example, that the parameter of interest, θ, is related to the population mean as $\theta = g(\mu)$ for some function g. Since the sample average $\bar{Y}$ is an unbiased and consistent estimator of μ, it is natural to replace μ with $\bar{Y}$, which gives us the estimator $g(\bar{Y})$ of θ. The estimator $g(\bar{Y})$ is consistent for θ, and if $g(\mu)$ is a linear function of μ, then $g(\bar{Y})$ is unbiased as well. What we have done is replace the population moment, μ, with its sample counterpart, $\bar{Y}$. This is where the name "method of moments" comes from.

We cover two additional method of moments estimators that will be useful for our discussion of regression analysis. Recall that the covariance between two random variables X and Y is defined as $\sigma_{XY} = \mathrm{E}[(X - \mu_X)(Y - \mu_Y)]$. The method of moments suggests estimating σ_{XY} by $n^{-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$. This is a consistent estimator of σ_{XY}, but it turns out to be biased for essentially the same reason that the sample variance is biased if n, rather than $n - 1$, is used as the divisor. The **sample covariance** is defined as

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}). \qquad \textbf{(C.14)}$$

It can be shown that this is an unbiased estimator of σ_{XY} (and replacing n with $n - 1$ makes no difference as the sample size grows indefinitely, so this estimator is still consistent).

As we discussed in Section B.4, the covariance between two variables is often difficult to interpret. Usually, we are more interested in correlation. Since the population correlation is $\rho_{XY} = \sigma_{XY}/(\sigma_X\sigma_Y)$, the method of moments suggests estimating ρ_{XY} as

$$R_{XY} = \frac{S_{XY}}{S_X S_Y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right)^{1/2}\left(\sum_{i=1}^{n}(Y_i - \bar{Y})^2\right)^{1/2}}, \tag{C.15}$$

which is called the **sample correlation coefficient** (or sample correlation for short). Notice that we have canceled the division by $n - 1$ in the sample covariance and the sample standard deviations. In fact, we could divide each of these by n, and we would arrive at the same final formula.

It can be shown that the sample correlation coefficient is always in the interval $[-1,1]$, as it should be. Because S_{XY}, S_X, and S_Y are consistent for the corresponding population parameter, R_{XY} is a consistent estimator of the population correlation, ρ_{XY}. However, R_{XY} is a biased estimator for two reasons. First, S_X and S_Y are biased estimators of σ_X and σ_Y, respectively. Second, R_{XY} is a ratio of estimators, and so it would not be unbiased, even if S_X and S_Y were. For our purposes, this is not important, although the fact that no unbiased estimator of ρ_{XY} exists is a classical result in mathematical statistics.

Maximum Likelihood

Another general approach to estimation is the method of *maximum likelihood*, a topic covered in many introductory statistics courses. A brief summary in the simplest case will suffice here. Let $\{Y_1,Y_2,\ldots,Y_n\}$ be a random sample from the population distribution $f(y;\theta)$. Because of the random sampling assumption, the joint distribution of $\{Y_1,Y_2,\ldots,Y_n\}$ is simply the product of the densities: $f(y_1;\theta)f(y_2;\theta)\cdots f(y_n;\theta)$. In the discrete case, this is $P(Y_1 = y_1,Y_2 = y_2,\ldots,Y_n = y_n)$. Now, define the *likelihood function* as

$$L(\theta;Y_1,\ldots,Y_n) = f(Y_1;\theta)f(Y_2;\theta)\cdots f(Y_n;\theta), \tag{C.16}$$

which is a random variable because it depends on the outcome of the random sample $\{Y_1,Y_2,\ldots,Y_n\}$. The **maximum likelihood estimator** of θ, call it W, is the value of θ that maximizes the likelihood function (this is why we write L as a function of θ, followed by the random sample). Clearly, this value depends on the random sample. The maximum likelihood principle says that, out of all the possible values for θ, the value that makes the likelihood of the observed data largest should be chosen. Intuitively, this is a reasonable approach to estimating θ.

Maximum likelihood estimation (MLE) is usually consistent and sometimes unbiased. But so are many other estimators. The widespread appeal of MLE is that it is gen-

erally the most asymptotically efficient estimator when the population model $f(y;\theta)$ is correctly specified. In addition, the MLE is sometimes the **minimum variance unbiased estimator**; that is, it has the smallest variance among all unbiased estimators of θ. [See Larsen and Marx (1986, Chapter 5) for verification of these claims.] We only need to rely on MLE for some of the advanced topics in Part 3 of the text.

Least Squares

A third kind of estimator, and one that plays a major role throughout the text, is called a **least squares estimator**. We have already seen an example of least squares: the sample mean, $\bar{Y}$, is a least squares estimator of the population mean, μ. We already know $\bar{Y}$ is a method of moments estimator. What makes it a least squares estimator? It can be shown that the value of m which makes the sum of squared deviations

$$\sum_{i=1}^{n} (Y_i - m)^2$$

as small as possible is $m = \bar{Y}$. Showing this is not difficult, but we omit the algebra.

For some important distributions, including the normal and the Bernoulli, the sample average $\bar{Y}$ is also the maximum likelihood estimator of the population mean μ. Thus, the principles of least squares, method of moments, and maximum likelihood often result in the *same* estimator. In other cases, the estimators are similar but not identical.

C.5 INTERVAL ESTIMATION AND CONFIDENCE INTERVALS

The Nature of Interval Estimation

A point estimate obtained from a particular sample does not, by itself, provide enough information for testing economic theories or for informing policy discussions. A point estimate may be the researcher's best guess at the population value, but, by its nature, it provides no information about how close the estimate is "likely" to be to the population parameter. As an example, suppose a researcher reports, on the basis of a random sample of workers, that job training grants increase hourly wage by 6.4%. How are we to know whether or not this is close to the effect in the population of workers who could have been trained? Since we do not know the population value, we cannot know how close an estimate is for a particular sample. However, we can make statements involving probabilities, and this is where interval estimation comes in.

We already know one way of assessing the uncertainty in an estimator: find its sampling standard deviation. Reporting the standard deviation of the estimator, along with the point estimate, provides some information on the accuracy of our estimate. However, even if the problem of the standard deviation's dependence on unknown population parameters is ignored, reporting the standard deviation along with the point estimate makes no direct statement about where the population value is likely to lie in relation to the estimate. This limitation is overcome by constructing a **confidence interval**.

We illustrate the concept of a confidence interval with an example. Suppose the population has a Normal(μ,1) distribution and let $\{Y_1, \ldots, Y_n\}$ be a random sample from

this population. (We assume that the variance of the population is known and equal to unity for the sake of illustration; we then show what to do in the more realistic case that the variance is unknown.) The sample average, $\bar{Y}$, has a normal distribution with mean μ and variance $1/n$: $\bar{Y} \sim \text{Normal}(\mu,1/n)$. From this, we can standardize $\bar{Y}$, and since the standardized version of $\bar{Y}$ has a standard normal distribution, we have

$$\text{P}\left(-1.96 < \frac{\bar{Y} - \mu}{1/\sqrt{n}} < 1.96\right) = .95.$$

The event in parentheses is identical to the event $\bar{Y} - 1.96/\sqrt{n} < \mu < \bar{Y} + 1.96/\sqrt{n}$, and so

$$\text{P}(\bar{Y} - 1.96/\sqrt{n} < \mu < \bar{Y} + 1.96/\sqrt{n}) = .95. \qquad \textbf{(C.17)}$$

Equation (C.17) is interesting because it tells us that the probability that the random interval $[\bar{Y} - 1.96/\sqrt{n},\bar{Y} + 1.96/\sqrt{n}]$ contains the population mean μ is .95, or 95%. This information allows us to construct an *interval estimate* of μ, which is obtained by plugging in the sample outcome of the average, $\bar{y}$. Thus,

$$[\bar{y} - 1.96/\sqrt{n},\bar{y} + 1.96/\sqrt{n}] \qquad \textbf{(C.18)}$$

is an example of an interval estimate of μ. It is also called a 95% confidence interval. A shorthand notation for this interval is $\bar{y} \pm 1.96/\sqrt{n}$.

The confidence interval in equation (C.18) is easy to compute, once the sample data $\{y_1,y_2,\ldots,y_n\}$ are observed; $\bar{y}$ is the only factor that depends on the data. For example, suppose that $n = 16$ and the average of the 16 data points is 7.3. Then, the 95% confidence interval for μ is $7.3 \pm 1.96/\sqrt{16} = 7.3 \pm .49$, which we can write in interval form as [6.81,7.79]. By construction, $\bar{y} = 7.3$ is in the center of this interval.

Unlike its computation, the meaning of a confidence interval is more difficult to understand. When we say that equation (C.18) is a 95% confidence interval for μ, we mean that the *random* interval

$$[\bar{Y} - 1.96/\sqrt{n},\bar{Y} + 1.96/\sqrt{n}] \qquad \textbf{(C.19)}$$

contains μ with probability .95. In other words, *before* the random sample is drawn, there is a 95% chance that (C.19) contains μ. Equation (C.19) is an example of an **interval estimator**. It is a random interval, since the endpoints change with different samples.

A confidence interval is often interpreted as follows: "The probability that μ is in the interval (C.18) is .95." This is incorrect. Once the sample has been observed and $\bar{y}$ has been computed, the limits of the confidence interval are simply numbers (6.81 and 7.79 in the example just given). The population parameter, μ, while unknown, is also just some number. Therefore, μ either is or is not in the interval (C.18) (and we will never know with certainty which is the case). Probability plays no role, once the confidence interval is computed for the particular data at hand. The probabilistic interpretation comes from the fact that for 95% of all random samples, the constructed confidence interval will contain μ.

To emphasize the meaning of a confidence interval, Table C.2 contains calculations for 20 random samples (or replications) from the Normal(2,1) distribution with sample size $n = 10$. For each of the 20 samples, $\bar{y}$ is obtained, and (C.18) is computed as $\bar{y} \pm 1.96/\sqrt{10} = \bar{y} \pm .62$ (each rounded to two decimals). As you can see, the interval changes with each random sample. Nineteen of the 20 intervals contain the population value of μ. Only for replication number 19 is μ not in the confidence interval. In other words, 95% of the samples result in a confidence interval that contains μ. This did not have to be the case with only 20 replications, but it worked out that way for this particular simulation.

Table C.2

Simulated Confidence Intervals from a Normal (μ,1) Distribution with $\mu = 2$

Replication	$\bar{y}$	95% Interval	Contains μ?
1	1.98	(1.36,2.60)	Yes
2	1.43	(0.81,2.05)	Yes
3	1.65	(1.03,2.27)	Yes
4	1.88	(1.26,2.50)	Yes
5	2.34	(1.72,2.96)	Yes
6	2.58	(1.96,3.20)	Yes
7	1.58	(0.96,2.20)	Yes
8	2.23	(1.61,2.85)	Yes
9	1.96	(1.34,2.58)	Yes
10	2.11	(1.49,2.73)	Yes
11	2.15	(1.53,2.77)	Yes
12	1.93	(1.31,2.55)	Yes
13	2.02	(1.40,2.64)	Yes
14	2.10	(1.48,2.72)	Yes
15	2.18	(1.56,2.80)	Yes

continued

Table C.2 (*concluded*)

Replication	$\bar{y}$	95% Interval	Contains μ?
16	2.10	(1.48,2.72)	Yes
17	1.94	(1.32,2.56)	Yes
18	2.21	(1.59,2.83)	Yes
19	1.16	(0.54,1.78)	No
20	1.75	(1.13,2.37)	Yes

Confidence Intervals for the Mean from a Normally Distributed Population

The confidence interval derived in equation (C.18) helps illustrate how to construct and interpret confidence intervals. In practice, equation (C.18) is not very useful for the mean of a normal population because it assumes that the variance is known to be unity. It is easy to extend (C.18) to the case where the standard deviation σ is known to be any value: the 95% confidence interval is

$$[\bar{y} - 1.96\sigma/\sqrt{n}, \bar{y} + 1.96\sigma/\sqrt{n}]. \tag{C.20}$$

Therefore, provided σ is known, a confidence interval for μ is readily constructed. To allow for unknown σ, we must use an estimate. Let

$$s = \left(\frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 \right)^{1/2} \tag{C.21}$$

denote the sample standard deviation. Then, we obtain a confidence interval that depends entirely on the observed data by replacing σ in equation (C.20) with its estimate, s. Unfortunately, this does not preserve the 95% level of confidence because s depends on the particular sample. In other words, the random interval $[\bar{Y} \pm 1.96(S/\sqrt{n})]$ no longer contains μ with probability .95 because the constant σ has been replaced with the random variable S.

How should we proceed? Rather than using the standard normal distribution, we must rely on the t distribution. The t distribution arises from the fact that

$$\frac{\bar{Y} - \mu}{S/\sqrt{n}} \sim t_{n-1}, \tag{C.22}$$

where $\bar{Y}$ is the sample average, and S is the sample standard deviation of the random sample $\{Y_1, \ldots, Y_n\}$. We will not prove (C.22); a careful proof can be found in a variety of places [for example, Larsen and Marx (1988, Chapter 7)].

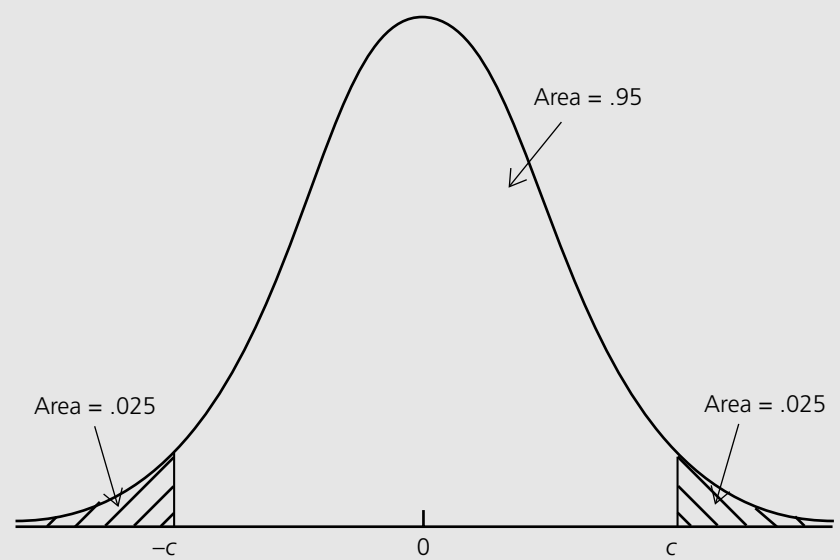
To construct a 95% confidence interval, let c denote the 97.5th percentile in the t_{n-1} distribution. In other words, c is the value such that 95% of the area in the t_{n-1} is between $-c$ and c: $P(-c < t_{n-1} < c) = .95$. (The value of c depends on the degrees of freedom $n - 1$, but we do not make this explicit.) The choice of c is illustrated in Figure C.4. Once c has been properly chosen, the random interval $[\bar{Y} - c \cdot S/\sqrt{n}, \bar{Y} + c \cdot S/\sqrt{n}]$ contains μ with probability .95. For a particular sample, the 95% confidence interval is calculated as

$$[\bar{y} - c \cdot s/\sqrt{n}, \bar{y} + c \cdot s/\sqrt{n}]. \qquad \textbf{(C.23)}$$

The values of c for various degrees of freedom can be obtained from Table G.2 in Appendix G. For example, if $n = 20$, so that the *df* is $n - 1 = 19$, then $c = 2.093$. Thus, the 95% confidence interval is $[\bar{y} \pm 2.093(s/\sqrt{20})]$, where $\bar{y}$ and s are the values obtained from the sample. Even if $s = \sigma$ (which is very unlikely), the confidence interval in (C.23) is wider than that in (C.20) because $c > 1.96$. For small degrees of freedom, (C.23) is much wider.

Figure C.4

The 97.5th percentile, *c*, in a *t* distribution.



More generally, let c_α denote the $100(1 - \alpha)$ percentile in the t_{n-1} distribution. Then, a $100(1 - \alpha)\%$ confidence interval is obtained as

$$[\bar{y} - c_{\alpha/2}s/\sqrt{n}, \bar{y} + c_{\alpha/2}s/\sqrt{n}]. \tag{C.24}$$

Obtaining $c_{\alpha/2}$ requires choosing α and knowing the degrees of freedom $n - 1$; then, Table G.2 can be used. For the most part, we will concentrate on 95% confidence intervals.

There is a simple way to remember how to construct a confidence interval for the mean of a normal distribution. Recall that $\text{sd}(\bar{Y}) = \sigma/\sqrt{n}$. Thus, $s/\sqrt{n}$ is the point estimate of $\text{sd}(\bar{Y})$. The associated random variable, $S/\sqrt{n}$, is sometimes called the **standard error** of Y. Since what shows up in formulas is the point estimate $s/\sqrt{n}$, we define the standard error of $\bar{y}$ as $\text{se}(\bar{y}) = s/\sqrt{n}$. Then, (C.24) can be written in shorthand as

$$[\bar{y} \pm c_{\alpha/2} \cdot \text{se}(\bar{y})]. \tag{C.25}$$

This equation shows why the notion of the standard error of an estimate plays an important role in econometrics.

EXAMPLE C.2

(Effect of Job Training Grants on Worker Productivity)

Holzer, Block, Cheatham, and Knott (1993) studied the effects of job training grants on worker productivity by collecting information on "scrap rates" for a sample of Michigan manufacturing firms receiving job training grants in 1988. Table C.3 lists the scrap rates—measured as number of items per 100 produced that are not usable and therefore need to be scrapped—for 20 firms. Each of these firms received a job training grant in 1988; there were no grants awarded in 1987. We are interested in contructing a confidence interval for the change in the scrap rate from 1987 to 1988 for the population of all manufacturing firms that could have received grants.

Table C.3

Scrap Rates for 20 Michigan Manufacturing Firms

Firm	1987	1988	Change
1	10	3	−7
2	1	1	0
3	6	5	−1
4	.45	.5	.05

continued

Table C.3 (*concluded*)

Firm	1987	1988	Change
5	1.25	1.54	.29
6	1.3	1.5	.2
7	1.06	.8	−.26
8	3	2	−1
9	8.18	.67	−7.51
10	1.67	1.17	−.5
11	.98	.51	−.47
12	1	.5	−.5
13	.45	.61	.16
14	5.03	6.7	1.67
15	8	4	−4
16	9	7	−2
17	18	19	1
18	.28	.2	−.08
19	7	5	−2
20	3.97	3.83	−.14
Average	4.38	3.23	−1.15

We assume that the change in scrap rates has a normal distribution. Since $n = 20$, a 95% confidence interval for the mean change in scrap rates μ is $[y \pm 2.093 \cdot \text{se}(\bar{y})]$, where $\text{se}(\bar{y}) = s/\sqrt{n}$. The value 2.093 is the 97.5$^{\text{th}}$ percentile in a t_{19} distribution. For the particular sample values, $\bar{y} = -1.15$ and $\text{se}(\bar{y}) = .54$ (each rounded to two decimals), and so the 95% confidence interval is $[-2.28, -.02]$. The value zero is excluded from this interval, so we conclude that, with 95% confidence, the average change in scrap rates in the population is not zero.

At this point, Example C.2 is mostly illustrative because it has some potentially serious flaws as an econometric analysis. Most importantly, it assumes that any systematic reduction in scrap rates is due to the job training grants. But many things can happen over the course of the year to change worker productivity. From this analysis, we have no way of knowing whether the fall in average scrap rates is attributable to the job training grants or if, at least partly, some external force is responsible.

A Simple Rule of Thumb for a 95% Confidence Interval

The confidence interval in (C.25) can be computed for any sample size and any confidence level. As we saw in Section B.4, the t distribution approaches the standard normal distribution as the degrees of freedom gets large. In particular, for $\alpha = .05$, $c_{\alpha/2} \rightarrow 1.96$ as $n \rightarrow \infty$, although $c_{\alpha/2}$ is always greater than 1.96 for each n. A *rule of thumb* for an approximate 95% confidence interval is

$$[\bar{y} \pm 2 \cdot \text{se}(\bar{y})]. \tag{C.26}$$

In other words, we obtain $\bar{y}$ and its standard error and then compute $\bar{y}$ plus and minus twice its standard error to obtain the confidence interval. This is slightly too wide for very large n, and it is too narrow for small n. As we can see from Example C.2, even for n as small as 20, (C.26) is in the ballpark for a 95% confidence interval for the mean from a normal distribution. This means we can get pretty close to a 95% confidence interval without having to refer to t tables.

Asymptotic Confidence Intervals for Nonnormal Populations

In some applications, the population is clearly nonnormal. A leading case is the Bernoulli distribution, where the random variable takes on only the values zero and one. In other cases, the nonnormal population has no standard distribution. This does not matter, provided the sample size is sufficiently large for the central limit theorem to give a good approximation for the distribution of the sample average $\bar{Y}$. For large n, an *approximate* 95% confidence interval is

$$[\bar{y} \pm 1.96 \cdot \text{se}(\bar{y})], \tag{C.27}$$

where the value 1.96 is the 97.5^{th} percentile in the standard normal distribution. Mechanically, computing an approximate confidence interval does not differ from the normal case. A slight difference is that the number multiplying the standard error comes from the standard normal distribution, rather than the t distribution, because we are using asymptotics. Because the t distribution approaches the standard normal as the df increases, equation (C.25) is also perfectly legitimate as an approximate 95% interval; some prefer this to (C.27) because the former is exact for normal populations.

EXAMPLE C.3
(Race Discrimination in Hiring)

The Urban Institute conducted a study in 1988 in Washington D.C. to examine the extent of race discrimination in hiring. Five pairs of people interviewed for several jobs. In each pair, one person was black, and the other person was white. They were given resumes indicating that they were virtually the same in terms of experience, education, and other factors that determine job qualification. The idea was to make individuals as similar as possible with the exception of race. Each person in a pair interviewed for the same job, and the researchers recorded which applicant received a job offer. This is an example of a *matched pairs analysis*, where each trial consists of data on two people (or two firms, two cities, and so on) that are thought to be similar in many respects but different in one important characteristic.

Let θ_B denote the probability that the black person is offered a job and let θ_W be the probability that the white person is offered a job. We are primarily interested in the difference, $\theta_B - \theta_W$. Let B_i denote a Bernoulli variable equal to one if the black person gets a job offer from employer i, and zero otherwise. Similarly, $W_i = 1$ if the white person gets a job offer from employer i, and zero otherwise. Pooling across the five pairs of people, there were a total of $n = 241$ trials (pairs of interviews with employees). Unbiased estimators of θ_B and θ_W are $\bar{B}$ and $\bar{W}$, the fractions of interviews for which blacks and whites were offered jobs, respectively.

To put this into the framework of computing a confidence interval for a population mean, define a new variable $Y_i = B_i - W_i$. Now, Y_i can take on three values: -1 if the black person did not get the job but the white person did, 0 if both people either did or did not get the job, and 1 if the black person got the job and the white person did not. Then, $\mu \equiv \mathrm{E}(Y_i) = \mathrm{E}(B_i) - \mathrm{E}(W_i) = \theta_B - \theta_W$.

The distribution of Y_i is certainly not normal—it is discrete and takes on only three values. Nevertheless, an approximate confidence interval for $\theta_B - \theta_W$ can be obtained by using large sample methods.

Using the 241 observed data points, $\bar{b} = .224$ and $\bar{w} = .357$, and so $\bar{y} = .224 - .357 = -.133$. Thus, 22.4% of black applicants were offered jobs, while 35.7% of white applicants were offered jobs. This is *prima facie* evidence of discrimination against blacks, but we can learn much more by computing a confidence interval for μ. To compute an approximate 95% confidence interval, we need the sample standard deviation. This turns out to be $s = .482$ [using equation (C.21)]. Using (C.27), we obtain a 95% CI for $\mu = \theta_B - \theta_W$ as $-.133 \pm 1.96(.482/\sqrt{241}) = -.133 \pm .031 = [-.164, -.102]$. The approximate 99% CI is $-.133 \pm 2.58(.482/\sqrt{241}) = [-.213, -.053]$. Naturally, this contains a wider range of values than the 95% CI. But even the 99% CI does not contain the value zero. Thus, we are very confident that the population difference $\theta_B - \theta_W$ is not zero.

One final comment needs to be made before we leave confidence intervals. Because the standard error for $\bar{y}$, $\mathrm{se}(\bar{y}) = s/\sqrt{n}$, shrinks to zero as the sample size grows, we see that—all else equal—a larger sample size means a smaller confidence interval. Thus, an important benefit of a large sample size is that it results in smaller confidence intervals.

C.6 HYPOTHESIS TESTING

So far, we have reviewed how to evaluate point estimators, and we have seen—in the case of a population mean—how to construct and interpret confidence intervals. But sometimes the question we are interested in has a definite yes or no answer. Here are some examples: (1) Does a job training program effectively increase average worker productivity? (see Example C.2); (2) Are blacks discriminated against in hiring? (see Example C.3); (3) Do stiffer state drunk driving laws reduce the number of drunk driving arrests? Devising methods for answering such questions, using a sample of data, is known as hypothesis testing.

Fundamentals of Hypothesis Testing

To illustrate the issues involved with hypothesis testing, consider an election example. Suppose there are two candidates in an election, Candidates A and B. Candidate A is reported to have received 42% of the popular vote, while Candidate B received 58%. These are supposed to represent the true percentages in the voting population, and we treat them as such.

Candidate A is convinced that more people must have voted for him, and so he would like to investigate whether the election was rigged. Knowing something about statistics, Candidate A hires a consulting agency to randomly sample 100 voters to record whether or not each person voted for him. Suppose that, for the sample collected, 53 people voted for Candidate A. This sample estimate of 53% clearly exceeds the reported population value of 42%. Should Candidate A conclude that the election was indeed a fraud?

While it appears that the votes for Candidate A were undercounted, we cannot be certain. Even if only 42% of the population voted for Candidate A, it is possible that, in a sample of 100, we observe 53 people who did vote for Candidate A. The question is: How *strong* is the sample evidence against the officially reported percentage of 42%?

One way to proceed is to set up a **hypothesis test**. Let θ denote the true proportion of the population voting for Candidate A. The hypothesis that the reported results are accurate can be stated as

$$H_0: \theta = .42. \tag{C.28}$$

This is an example of a **null hypothesis**. We always denote the null hypothesis by H_0. In hypothesis testing, the null hypothesis plays a role similar to that of a defendent on trial in many judicial systems: just as a defendent is presumed to be innocent until proven guilty, the null hypothesis is presumed to be true until the data strongly suggest otherwise. In the current example, Candidate A must present fairly strong evidence against (C.28) in order to win a recount.

The **alternative hypothesis** in the election example is that the true proportion voting for Candidate A in the election is greater than .42:

$$H_1: \theta > .42. \tag{C.29}$$

In order to conclude that H_0 is false and that H_1 is true, we must have evidence "beyond reasonable doubt" against H_0. How many votes out of 100 would be needed before we

feel the evidence is strongly against H_0? Most would agree that observing 43 votes out of a sample of 100 is not enough to overturn the original election results; such an outcome is well within the expected sampling variation. On the other hand, we do not need to observe 100 votes for Candidate A to cast doubt on H_0. Whether 53 out of 100 is enough to reject H_0 is much less clear. The answer depends on how we quantify "beyond reasonable doubt."

In hypothesis testing, we can make two kinds of mistakes. First, we can reject the null hypothesis when it is in fact true. This is called a **Type I error**. In the election example, a Type I occurs if we reject H_0 when the true proportion of people voting for Candidate A is in fact .42. The second kind of error is failing to reject H_0 when it is actually false. This is called a **Type II error**. In the election example, a Type II error occurs if $\theta > .42$ but we fail to reject H_0.

After we have made the decision of whether or not to reject the null hypothesis, we have either decided correctly or we have committed an error. We will never know with certainty whether an error was committed. However, we can compute the *probability* of making either a Type I or a Type II error. Hypothesis testing rules are constructed to make the probability of committing a Type I error fairly small. Generally, we define the **significance level** (or simply the *level*) of a test as the probability of a Type I error; it is typically denoted by α. Symbolically, we have

$$\alpha = \text{P(Reject } H_0 | H_0). \tag{C.30}$$

The right-hand side is read as: "The probability of rejecting H_0 given that H_0 is true."

Classical hypothesis testing requires that we initially specify a significance level for a test. When we specify a value for α, we are essentially quantifying our tolerance for a Type I error. Common values for α are .10, .05, and .01. If $\alpha = .05$, then the researcher is willing to falsely reject H_0 5% of the time, in order to detect deviations from H_0.

Once we have chosen the significance level, we would then like to minimize the probability of a Type II error. Alternatively, we would like to maximize the **power of a test** against all relevant alternatives. The power of a test is just one, minus the probability of a Type II error. Mathematically,

$$\pi(\theta) = \text{P(Reject } H_0 | \theta) = 1 - \text{P(Type II} | \theta),$$

where θ denotes the actual value of the parameter. Naturally, we would like the power to equal unity whenever the null hypothesis is false. But this is impossible to achieve while keeping the significance level small. Instead, we choose our tests to maximize the power for a given significance level.

Testing Hypotheses About the Mean in a Normal Population

In order to test a null hypothesis against an alternative, we need to choose a test statistic (or statistic, for short) and a critical value. The choices for the statistic and critical value are based on convenience and on the desire to maximize power given a significance level for the test. In this subsection, we review how to test hypotheses for the mean of a normal population.

A **test statistic**, denoted T, is some function of the random sample. When we compute the statistic for a particular outcome, we obtain an outcome of the test statistic, which we will denote t.

Given a test statistic, we can define a rejection rule that determines when H_0 is rejected in favor of H_1. In this text, all rejection rules are based on comparing the value of a test statistic, t, to a **critical value**, c. The values of t that result in rejection of the null hypothesis are collectively known as the **rejection region**. In order to determine the critical value, we must first decide on a significance level of the test. Then, given α, the critical value associated with α is determined by the distribution of T, *assuming* that H_0 is true. We will write this critical value as c, suppressing the fact that it depends on α.

Testing hypotheses about the mean μ from a Normal(μ,σ^2) population is straightforward. The null hypothesis is stated as

$$H_0: \mu = \mu_0, \tag{C.31}$$

where μ_0 is a value that we specify. In the majority of applications, $\mu_0 = 0$, but the general case is no more difficult.

The rejection rule we choose depends on the nature of the alternative hypothesis. The three alternatives of interest are

$$H_1: \mu > \mu_0, \tag{C.32}$$

$$H_1: \mu < \mu_0, \tag{C.33}$$

and

$$H_1: \mu \neq \mu_0. \tag{C.34}$$

Equation (C.32) gives a **one-sided alternative**, as does (C.33). When the alternative hypothesis is (C.32), the null is effectively $H_0: \mu \leq \mu_0$, since we reject H_0 only when $\mu > \mu_0$. This is appropriate when we are interested in the value of μ but only when μ is at least as large as μ_0. Equation (C.34) is a **two-sided alternative**. This is acceptable when we are interested in any departure from the null hypothesis.

Consider first the alternative in (C.32). Intuitively, we should reject H_0 in favor of H_1 when the value of the sample average, $\bar{y}$, is "sufficiently" greater than μ_0. But how should we determine when $\bar{y}$ is large enough for H_0 to be rejected at the chosen significance level? This requires knowing the probability of rejecting the null hypothesis when it is true. Rather than working directly with $\bar{y}$, we use its standardized version, where σ is replaced with the sample standard deviation, s:

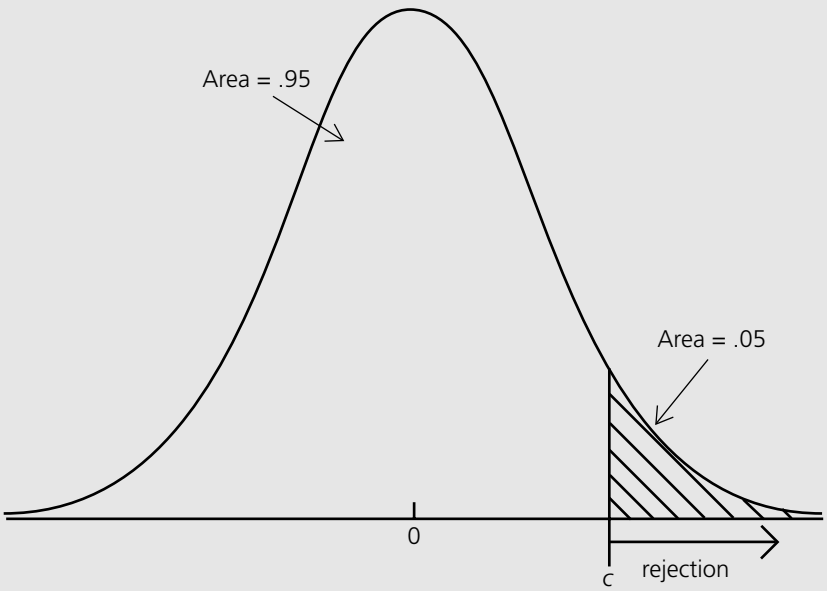
$$t = \sqrt{n}(\bar{y} - \mu_0)/s = (\bar{y} - \mu_0)/\text{se}(\bar{y}), \tag{C.35}$$

where $\text{se}(\bar{y}) = s/\sqrt{n}$ is the standard error of $\bar{y}$. Given the sample of data, it is easy to obtain t. The reason we work with t is that, under the null hypothesis, the random variable

$$T = \sqrt{n}(\bar{Y} - \mu_0)/S$$

Figure C.5

Rejection region for a 5% significance level test against the one-sided alternative $\mu > \mu_0$.



has a t_{n-1} distribution. Now, suppose we have settled on a 5% significance level. Then, the critical value c is chosen so that $P(T > c|H_0) = .05$; that is, the probability of a Type I error is 5%. Once we have found c, the rejection rule is

$$t > c, \tag{C.36}$$

where c is the $100(1 - \alpha)$ percentile in a t_{n-1} distribution; as a percent, the significance level is $100 \cdot \alpha\%$. This is an example of a **one-tailed test** because the rejection region is in one tail of the t distribution. For a 5% significance level, c is the 95th percentile in the t_{n-1} distribution; this is illustrated in Figure C.5. A different significance level leads to a different critical value.

The statistic in equation (C.35) is often called the ***t* statistic** for testing H_0: $\mu = \mu_0$. The t statistic measures the distance from $\bar{y}$ to μ_0 *relative* to the standard error of $\bar{y}$, se($\bar{y}$).

EXAMPLE C.4

(Effect of Enterprise Zones on Business Investments)

In the population of cities granted enterprise zones in a particular state [see Papke (1994) for Indiana], let Y denote the percentage change in investment from the year before to the

year after a city became an enterprise zone. Assume that Y has a Normal(μ,σ^2) distribution. The null hypothesis that enterprise zones have no effect on business investment is H_0: $\mu = 0$; the alternative that they have a positive effect is H_1: $\mu > 0$ (we assume that they do not have a negative effect). Suppose that we wish to test H_0 at the 5% level. The test statistic in this case is

$$t = \frac{\bar{y}}{s/\sqrt{n}} = \frac{\bar{y}}{\text{se}(\bar{y})}. \tag{C.37}$$

Suppose that we have a sample of 36 cities which are granted enterprise zones. Then, the critical value is $c = 1.69$ (see Table G.2), and we reject H_0 in favor of H_1 if $t > 1.69$. Suppose that the sample yields $\bar{y} = 8.2$ and $s = 23.9$. Then, $t \approx 2.06$, and H_0 is therefore rejected at the 5% level. Thus, we conclude that, at the 5% significance level, enterprise zones have an effect on average investment. The 1% critical value is 2.44, and so H_0 is not rejected at the 1% level. The same caveat holds here as in Example C.2: we have not controlled for other factors that might affect investment in cities over time, and so we cannot claim that the effect is causal.

The rejection rule is similar for the one-sided alternative (C.32). A test with a significance level of $100\cdot\alpha\%$ rejects H_0 against (C.33) whenever

$$t < -c; \tag{C.38}$$

in other words, we are looking for negative values of the t statistic—which implies $\bar{y} < \mu_0$—that are sufficiently far from zero to reject H_0.

For two-sided alternatives, we must be careful to choose the critical value so that the significance level of the test is still α. If H_1 is given by H_1: $\mu \neq \mu_0$, then we reject H_0 if $\bar{y}$ is far from μ_0 in *absolute value*: a $\bar{y}$ much larger or much smaller than μ_0 provides evidence against H_0 in favor of H_1. A $100\cdot\alpha\%$ level test is obtained from the rejection rule

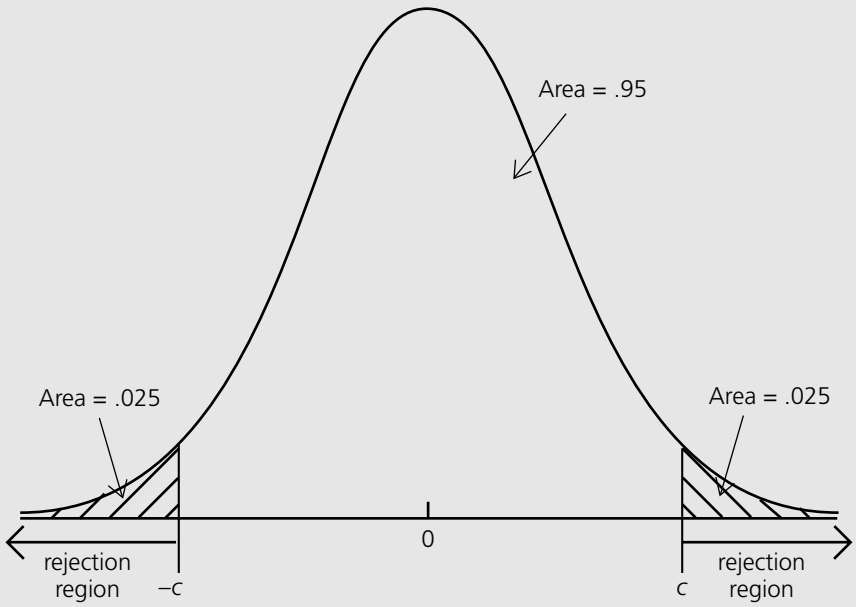
$$|t| > c, \tag{C.39}$$

where $|t|$ is the absolute value of the t statistic in (C.35). This gives a **two-tailed test**. We must now be careful in choosing the critical value: c is the $100(1 - \alpha/2)$ percentile in the t_{n-1} distribution. For example, if $\alpha = .05$, then the critical value is the 97.5$^{\text{th}}$ percentile in the t_{n-1} distribution. This ensures that H_0 is rejected only 5% of the time when it is true (see Figure C.6). For example, if $n = 22$, then the critical value is $c = 2.08$, the 97.5$^{\text{th}}$ percentile in a t_{21} distribution (see Table G.2). The absolute value of the t statistic must exceed 2.08 in order to reject H_0 against H_1 at the 5% level.

It is important to know the proper language of hypothesis testing. Sometimes, the appropriate phrase "we fail to reject H_0 in favor of H_1 at the 5% significance level" is replaced with "we accept H_0 at the 5% significance level." The latter wording is incorrect. With the same set of data there are usually many hypotheses that cannot be

Figure C.6

Rejection region for a 5% significance level test against the two-sided alternative H_1: $\mu \neq \mu_0$.



rejected. In the earlier election example, it would be logically inconsistent to say that H_0: $\theta = .42$ and H_0: $\theta = .43$ are both "accepted," since only one of these can be true. But it is entirely possible that neither of these hypotheses is rejected. For this reason, we always say "fail to reject H_0" rather than "accept H_0."

Asymptotic Tests for Nonnormal Populations

If the sample size is large enough to invoke the central limit theorem (see Section C.3), the mechanics of hypothesis testing for population means are the *same* whether or not the population distribution is normal. The theoretical justification comes from the fact that, under the null hypothesis,

$$T = \sqrt{n}(\bar{Y} - \mu_0)/S \overset{a}{\sim} \text{Normal}(0,1).$$

Therefore, with large n, we can compare the t statistic in (C.35) with the critical values from a standard normal distribution. Since the t_{n-1} distribution converges to the standard normal distribution as n gets large, the t and standard normal critical values will be very close for extremely large n. Since asymptotic theory is based on n increasing without bound, it cannot tell us whether the standard normal or t critical values are better. For moderate values of n, say between 30 and 60, it is traditional to use the t distribution because we know this is correct for normal populations. For $n > 120$, the choice

between the t and standard normal distributions is largely irrelevant because the critical values are practically the same.

Because the critical values chosen using either the standard normal or t distribution are only approximately valid for nonnormal populations, our chosen significance levels are also only approximate; thus, for nonnormal populations our significance levels are really *asymptotic* significance levels. Thus, if we choose a 5% significance level, but our population is nonnormal, then the actual significance level will be larger or smaller than 5% (and we cannot know which is the case). When the sample size is large, the actual significance level will be very close to 5%. Practically speaking, the distinction is not important, and so we will now drop the qualifier "asymptotic."

EXAMPLE C.5
(Race Discrimination in Hiring)

In the Urban Institute study of discrimination in hiring (see Example C.3), we are primarily interested in testing H_0: $\mu = 0$ against H_1: $\mu < 0$, where $\mu = \theta_B - \theta_W$ is the difference in probabilities that blacks and whites receive job offers. Recall that μ is the population mean of the variable $Y = B - W$, where B and W are binary indicators. Using the $n = 241$ paired comparisons, we obtained $\bar{y} = -.133$ and $\text{se}(\bar{y}) = .482/\sqrt{241} \approx .031$. The t statistic for testing H_0: $\mu = 0$ is $t = -.133/.031 \approx -4.29$. You will remember from Appendix B that the standard normal distribution is, for practical purposes, indistinguishable from the t distribution with 240 degrees of freedom. The value -4.29 is so far out in the left tail of the distribution that we reject H_0 at any reasonable significance level. In fact, the .005 (one-half of a percent) critical value (for the one-sided test) is about -2.58. A t value of -4.29 is *very* strong evidence against H_0 in favor of H_1. Thus, we conclude that there is discrimination in hiring.

Computing and Using *p*-Values

The traditional requirement of choosing a significance level ahead of time means that different researchers, using the same data and same procedure to test the same hypothesis, could wind up with different conclusions. Reporting the significance level at which we are carrying out the test solves this problem to some degree, but it does not completely remove the problem.

To provide more information, we can ask the following question: What is the *largest* significance level at which we could carry out the test and still fail to reject the null hypothesis? This value is known as the ***p*-value** of a test (sometimes called the *prob-value*). Compared with choosing a significance level ahead of time and obtaining a critical value, computing a p-value is somewhat more difficult. But with the advent of quick and inexpensive computing, p-values are now fairly easy to obtain.

As an illustration, consider the problem of testing H_0: $\mu = 0$ in a Normal(μ,σ^2) population. Our test statistic in this case is $T = \sqrt{n}\cdot\bar{Y}/S$, and we assume that n is large enough to treat T as having a standard normal distribution under H_0. Suppose that the observed value of T for our sample is $t = 1.52$ (note how we have skipped the step of choosing a significance level). Now that we have seen the value t, we can find the

largest significance level at which we would fail to reject H_0. This is the significance level associated with using t as our critical value. Since our test statistic T has a standard normal distribution under H_0, we have

$$p\text{-value} = P(T > 1.52|H_0) = 1 - \Phi(1.52) = .065, \tag{C.40}$$

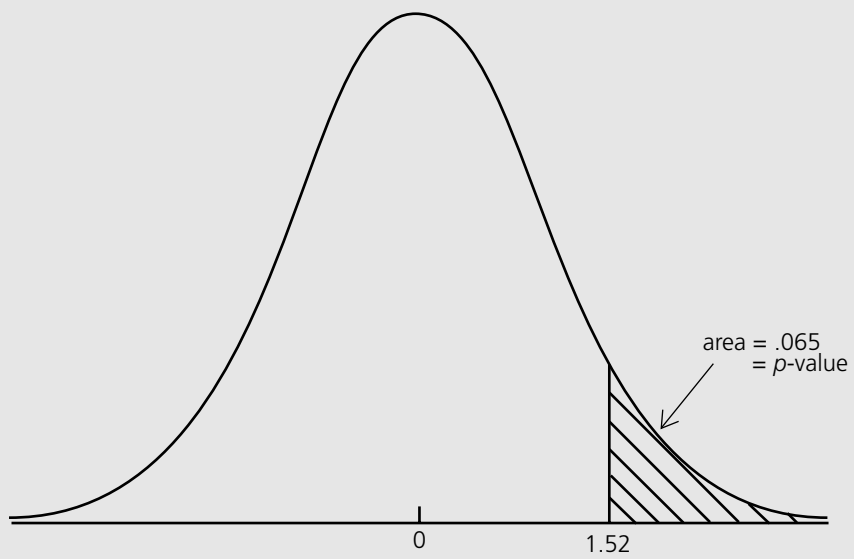
where $\Phi(\cdot)$ denotes the standard normal cdf. In other words, the p-value in this example is simply the area to the right of 1.52, the observed value of the test statistic, in a standard normal distribution. See Figure C.7 for illustration.

Since p-value $= .065$, the largest significance level at which we can carry out this test and fail to reject is 6.5%. If we carry out the test at a level below 6.5% (such as at 5%), we fail to reject H_0. If we carry out the test at a level larger than 6.5% (such as 10%), we reject H_0. With the p-value at hand, we can carry out the test at any level.

The p-value in this example has another useful interpretation: it is the probability that we observe a value of T as large as 1.52 when the null hypothesis is true. If the null hypothesis is actually true, we would observe a value of T as large as 1.52 due to chance only 6.5% of the time. Whether this is small enough to reject H_0 depends on our tolerance for a Type I error. The p-value has a similar interpretation in all other cases, as we will see.

Figure C.7

The p-value when $t = 1.52$ for the one-sided alternative $\mu > \mu_0$.



Generally, small *p*-values are evidence *against* H_0, since they indicate that the outcome of the data occurs with small probability if H_0 is true. In the previous example, if t had been a larger value, say $t = 2.85$, then the *p*-value would be $1 - \Phi(2.85) \approx .002$. This means that, if the null hypothesis were true, we would observe a value of T as large as 2.85 with probability .002. How do we interpret this? Either we obtained a very unusual sample or the null hypothesis is false. Unless we have a *very* small tolerance for Type I error, we would reject the null hypothesis. On the other hand, a large *p*-value is weak evidence against H_0. If we had gotten $t = .47$ in the previous example, then *p*-value $= 1 - \Phi(.47) = .32$. Observing a value of T larger than .47 happens with probability .32, even when H_0 is true; this is large enough so that there is insufficient doubt about H_0, unless we have a very high tolerance for Type I error.

For hypothesis testing about a population mean using the t distribution, we need detailed tables in order to compute *p*-values. Table G.2 only allows us to put bounds on *p*-values. Fortunately, many statistics and econometrics packages now compute *p*-values routinely, and they also provide calculation of cdfs for the t and other distributions used for computing *p*-values.

EXAMPLE C.6

(Effect of Job Training Grants on Worker Productivity)

Consider again the Holzer et al. (1993) data in Example C.2. From a policy perspective, there are two questions of interest. First, what is our best estimate of the mean change in scrap rates, μ? We have already obtained this for the sample of 20 firms listed in Table C.3: the sample average of the change in scrap rates is -1.15. Relative to the initial average scrap rate in 1987, this represents a fall in the scrap rate of about 26.3% ($-1.15/4.38 \approx -.263$), which is a nontrivial effect.

We would also like to know whether the sample provides strong evidence for an effect in the population of manufacturing firms that could have received grants. The null hypothesis is H_0: $\mu = 0$, and we test this against H_1: $\mu < 0$, where μ is the average change in scrap rates. Under the null, the job training grants have no effect on average scrap rates. The alternative states that there is an effect. We do not care about the alternative $\mu > 0$; the null hypothesis is effectively H_0: $\mu \geq 0$.

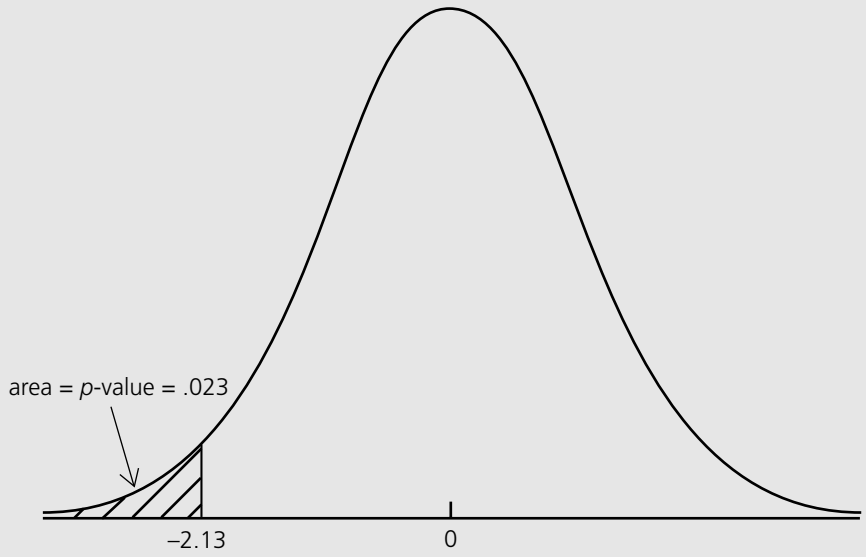
Since $\bar{y} = -1.15$ and $\text{se}(\bar{y}) = .54$, $t = -1.15/.54 = -2.13$. This is below the 5% critical value of -1.73 (from a t_{19} distribution) but above the 1% critical value, -2.54. The *p*-value in this case is computed as

$$p\text{-value} = \text{P}(T_{19} < -2.13), \tag{C.41}$$

where T_{19} represents a t distributed random variable with 19 degrees of freedom. The inequality is reversed from (C.40) because the alternative has the form (C.33), not (C.32). The probability in (C.41) is the area to the left of -2.13 in a t_{19} distribution (see Figure C.8).

Figure C.8

The *p*-value when $t = -2.13$ with 19 degrees of freedom for the one-sided alternative $\mu < 0$.



Using Table G.2, the most we can say is that the *p*-value is between .025 and .01, but it is closer to .025 (since the 97.5th percentile is about 2.09). Using a statistical package, such as Stata, we can compute the exact *p*-value. It turns out to be about .023, which is reasonable evidence against H_0. This is certainly enough evidence to reject the null hypothesis that the training grants had no effect at the 2.5% significance level (and therefore at the 5% level).

Computing a *p*-value for a two-sided test is similar, but we must account for the two-sided nature of the rejection rule. For *t* testing about population means, the *p*-value is computed as

$$P(|T_{n-1}| > |t|) = 2P(T_{n-1} > |t|), \tag{C.42}$$

where t is the value of the test statistic, and T_{n-1} is a t random variable. (For large n, replace T_{n-1} with a standard normal random variable.) Thus, to compute the absolute value of the t statistic, find the area to the right of this value in a t_{n-1} distribution and multiply the area by two.

For nonnormal populations, the exact p-value can be difficult to obtain. Nevertheless, we can find *asymptotic* p-values by using the same calculations. These p-values are valid for large sample sizes. For n larger than, say, 120, we might as well use the standard normal distribution. Table G.1 is detailed enough to get accurate p-values, but we can also use a statistics or econometrics program.

EXAMPLE C.7
(Race Discrimination in Hiring)

Using the matched pair data from the Urban Institute ($n = 241$), we obtained $t = -4.29$. If Z is a standard normal random variable, $P(Z < -4.29)$ is, for practical purposes, zero. In other words, the (asymptotic) p-value for this example is essentially zero. This is very strong evidence against H_0.

SUMMARY OF HOW TO USE *p*-VALUES

(i) Choose a test statistic T and decide on the nature of the alternative. This determines whether the rejection rule is $t > c$, $t < -c$, or $|t| > c$.

(ii) Use the observed value of the t statistic as the critical value and compute the corresponding significance level of the test. This is the p-value. If the rejection rule is of the form $t > c$, then p-value $= P(T > t)$. If the rejection rule is $t < -c$, then p-value $= P(T < t)$; if the rejection rule is $|t| > c$, then p-value $= P(|T| > |t|)$.

(iii) If a significance level α has been chosen, then we reject H_0 at the $100 \cdot \alpha\%$ level if p-value $< \alpha$. If p-value $\geq \alpha$, then we fail to reject H_0 at the $100 \cdot \alpha\%$ level. Thus, it is a small p-value that leads to rejection.

The Relationship Between Confidence Intervals and Hypothesis Testing

Since contructing confidence intervals and hypothesis tests both involve probability statements, it is natural to think that they are somehow linked. It turns out that they are. After a confidence interval has been constructed, we can carry out a variety of hypothesis tests.

The confidence intervals we have discussed are all two-sided by nature. (In this text, we will have no need to construct one-sided confidence intervals.) Thus, confidence intervals can be used to test against *two-sided* alternatives. In the case of a population mean, the null is given by (C.31), and the alternative is (C.34). Suppose we have constructed a 95% confidence interval for μ. Then, if the hypothesized value of μ under H_0, μ_0, is not in the confidence interval, then H_0: $\mu = \mu_0$ is rejected against H_1: $\mu \neq \mu_0$ at the 5% level. If μ_0 lies in this interval, then we fail to reject H_0 at the 5% level. Notice how any value for μ_0 can be tested once a confidence interval is constructed, and since a confidence interval contains more than one value, there are many null hypotheses that will not be rejected.

EXAMPLE C.8
(Training Grants and Worker Productivity)

In the Holzer et al. example, we constructed a 95% confidence interval for the mean change in scrap rate μ as $[-2.28, -.02]$. Since zero is excluded from this interval, we reject H_0: $\mu = 0$ against H_1: $\mu \neq 0$ at the 5% level. This 95% confidence interval also means that we fail to reject H_0: $\mu = -2$ at the 5% level. In fact, there is a continuum of null hypotheses that are not rejected given this confidence interval.

Practical Versus Statistical Significance

In the examples covered so far, we have produced three kinds of evidence concerning population parameters: point estimates, confidence intervals, and hypothesis tests. These tools for learning about population parameters are equally important. There is an understandable tendency for students to focus on confidence intervals and hypothesis tests because these are things to which we can attach confidence or significance levels. But in any study, we must also interpret the *magnitudes* of point estimates.

Statistical significance depends on the size of the t statistic and not just on the size of $\bar{y}$. For testing H_0: $\mu = 0$, $t = \bar{y}/\text{se}(\bar{y})$. Thus, statistical significance depends on the ratio of $\bar{y}$ to its standard error. A t statistic can be large either because $\bar{y}$ is large or because $\text{se}(\bar{y})$ is small.

EXAMPLE C.9
(Effect of Freeway Width on Commute Time)

Let Y denote the change in commute time, measured in minutes, for commuters in a metropolitan area from before a freeway was widened to after the freeway was widened. Assume that $Y \sim \text{Normal}(\mu, \sigma^2)$. The null hypothesis that the widening did not reduce average commute time is H_0: $\mu = 0$; the alternative that it reduced average commute time is H_1: $\mu < 0$. Suppose a random sample of commuters of size $n = 300$ is obtained to determine the effectiveness of the freeway project. The average change in commute time is computed to be $\bar{y} = -3.6$, and the sample standard deviation is $s = 18.7$; thus, $\text{se}(\bar{y}) = 18.7/\sqrt{300} \approx 1.08$. The t statistic is $t = -3.6/1.08 \approx -3.33$, which is very statistically significant; the p-value is essentially zero. Thus, we conclude that the freeway widening had a statistically significant effect on average commute time.

If the outcome of the hypothesis test is all that were reported from the study, it would be misleading. Reporting only statistical significance masks the fact that the estimated reduction in average commute time, 3.6 minutes, is pretty meager. To be up front, we should report the point estimate of -3.6, along with the significance test.

While the magnitude and sign of the t statistic determine statistical significance, the point estimate $\bar{y}$ determines what we might call **practical significance**. An estimate can be statistically significant without being especially large. We should always discuss the

practical significance along with the statistical significance of point estimates; this theme will arise often in the text.

Finding point estimates that are statistically significant without being practically significant often occurs when we are working with large samples. To discuss why this happens, it is useful to have the following definition.

TEST CONSISTENCY

A **consistent test** rejects H_0 with probability approaching one as the sample size grows, whenever H_1 is true.

Another way to say that a test is consistent is that, as the sample size tends to infinity, the power of the test gets closer and closer to unity, whenever H_1 is true. All of the tests we cover in this text have this property. In the case of testing hypotheses about a population mean, test consistency follows because the variance of $\bar{Y}$ converges to zero as the sample size gets large. The t statistic for testing H_0: $\mu = 0$ is $T = \bar{Y}/(S/\sqrt{n})$. Since $\text{plim}(\bar{Y}) = \mu$ and $\text{plim}(S) = \sigma$, it follows that if, say, $\mu > 0$, then T gets larger and larger (with high probability) as $n \to \infty$. In other words, no matter how close μ is to zero, we can be almost certain to reject H_0: $\mu = 0$, given a large enough sample size. This says nothing about whether μ is large in a practical sense.

C.7 REMARKS ON NOTATION

In our review of probability and statistics here and in Appendix B, we have been careful to use standard conventions to denote random variables, estimators, and test statistics. For example, we have used W to indicate an estimator (random variable) and w to denote a particular estimate (outcome of the random variable W). Distinguishing between an estimator and an estimate is important for understanding various concepts in estimation and hypothesis testing. However, making this distinction quickly becomes a burden in econometric analysis because the models are more complicated: many random variables and parameters will be involved, and being true to the usual conventions from probability and statistics requires many extra symbols.

In the main text, we use a simpler convention that is widely used in econometrics. If θ is a population parameter, the notation $\hat{\theta}$ ("theta hat") will be used to denote both an estimator and an estimate of θ. This notation is useful in that it provides a simple way of attaching an estimator to the population parameter it is supposed to be estimating. Thus, if the population parameter is β, then $\hat{\beta}$ denotes an estimator or estimate of β; if the parameter is σ^2, $\hat{\sigma}^2$ is an estimator or estimate of σ^2; and so on. Sometimes, we will discuss two estimators of the same parameter, in which case, we will need a different notation, such as $\tilde{\theta}$ ("theta tilda").

While dropping the conventions from probability and statistics to indicate estimators, random variables, and test statistics puts additional responsibility on you, it is not a big deal, once the difference between an estimator and an estimate is understood. If we are discussing *statistical* properties of $\hat{\theta}$—such as deriving whether or not it is unbiased or consistent—then we are necessarily viewing $\hat{\theta}$ as an estimator. On the other hand, if we write something like $\hat{\theta} = 1.73$, then we are clearly denoting a point estimate

from a given sample of data. The confusion that can arise by using $\hat{\theta}$ to denote both should be minimal, once you have a good understanding of probability and statistics.

SUMMARY

We have discussed topics from mathematical statistics that are heavily relied on in econometric analysis. The notion of an estimator, which is simply a rule for combining data to estimate a population parameter, is fundamental. We have covered various properties of estimators. The most important small sample properties are unbiasedness and efficiency, the latter of which depends on comparing variances when estimators are unbiased. Large sample properties concern the sequence of estimators obtained as the sample size grows, and they are also heavily relied on in econometrics. Any useful estimator is consistent. The central limit theorem implies that, in large samples, the sampling distribution of most estimators is approximately normal.

The sampling distribution of an estimator can be used to construct confidence intervals. We saw this for estimating the mean from a normal distribution and for computing approximate confidence intervals in nonnormal cases. Classical hypothesis testing, which requires specifying a null hypothesis, an alternative hypothesis, and a significance level, is carried out by comparing a test statistic to a critical value. Alternatively, a *p*-value can be computed that allows us to carry out a test at any significance level.

KEY TERMS

Alternative Hypothesis
Asymptotic Normality
Bias
Central Limit Theorem (CLT)
Confidence Interval
Consistent Estimator
Consistent Test
Critical Value
Estimate
Estimator
Hypothesis Test
Inconsistent
Interval Estimator
Law of Large Numbers (LLN)
Least Squares Estimator
Maximum Likelihood Estimator
Mean Squared Error (MSE)
Method of Moments
Minimum Variance Unbiased Estimator
Null Hypothesis
One-Sided Alternative
One-Tailed Test
Population
Power of a Test
Practical Significance
Probability Limit
p-Value
Random Sample
Rejection Region
Sample Average
Sample Correlation Coefficient
Sample Covariance
Sample Standard Deviation
Sample Variance
Sampling Distribution
Sampling Variance
Significance Level
Standard Error
t Statistic
Test Statistic
Two-Sided Alternative
Two-Tailed Test
Type I Error
Type II Error
Unbiasedness

PROBLEMS

C.1 Let Y_1, Y_2, Y_3, and Y_4 be independent, identically distributed random variables from a population with mean μ and variance σ^2. Let $\bar{Y} = \frac{1}{4}(Y_1 + Y_2 + Y_3 + Y_4)$ denote the average of these four random variables.

(i) What are the expected value and variance of $\bar{Y}$ in terms of μ and σ^2?

(ii) Now, consider a different estimator of μ:

$$W = \frac{1}{8}Y_1 + \frac{1}{8}Y_2 + \frac{1}{4}Y_3 + \frac{1}{2}Y_4.$$

This is an example of a *weighted* average of the Y_i. Show that W is also an unbiased estimator of μ. Find the variance of W.

(iii) Based on your answers to parts (i) and (ii), which estimator of μ do you prefer, $\bar{Y}$ or W?

(iv) Now, consider a more general estimator of μ, defined by

$$W_a = a_1Y_1 + a_2Y_2 + a_3Y_3 + a_4Y_4,$$

where the a_i are constants. What condition is needed on the a_i for W_a to be an unbiased estimator of μ?

(v) Compute the variance of the estimator W_a from part (iv).

C.2 This is a more general version of Problem C.1. Let $Y_1, Y_2, \ldots, Y_n$ be n pairwise uncorrelated random variables with common mean μ and common variance σ^2. Let $\bar{Y}$ denote the sample average.

(i) Define the class of *linear estimators* of μ by

$$W_a = a_1Y_1 + a_2Y_2 + \ldots + a_nY_n,$$

where the a_i are constants. What restriction on the a_i is needed for W_a to be an unbiased estimator of μ?

(ii) Find $\text{Var}(W_a)$.

(iii) For any numbers $a_1, a_2, \ldots, a_n$, the following inequality holds: $(a_1 + a_2 + \ldots + a_n)^2/n \leq a_1^2 + a_2^2 + \ldots + a_n^2$. Use this, along with parts (i) and (ii), to show that $\text{Var}(W_a) \geq \text{Var}(\bar{Y})$ whenever W_a is unbiased, so that $\bar{Y}$ is the *best linear unbiased estimator*. [*Hint*: What does the inequality become when the a_i satisfy the restriction from part (i)?]

C.3 Let Y denote the sample average from a random sample with mean μ and variance σ^2. Consider two alternative estimators of μ: $W_1 = [(n - 1)/n]\bar{Y}$ and $W_2 = \bar{Y}/2$.

(i) Show that W_1 and W_2 are both biased estimators of μ and find the biases. What happens to the biases as $n \to \infty$? Comment on any important differences in bias for the two estimators as the sample size gets large.

(ii) Find the probability limits of W_1 and W_2. {*Hint*: Use properties PLIM.1 and PLIM.2; for W_1, note that plim $[(n - 1)/n] = 1$.} Which estimator is consistent?

(iii) Find $\text{Var}(W_1)$ and $\text{Var}(W_2)$.

(iv) Argue that W_1 is a better estimator than $\bar{Y}$ if μ is "close" to zero. (Consider both bias and variance.)

C.4 For positive random variables X and Y, suppose the expected value of Y given X is $E(Y|X) = \theta X$. The unknown parameter θ shows how the expected value of Y changes with X.

(i) Define the random variable $Z = Y/X$. Show that $E(Z) = \theta$. [*Hint*: Use Property CE.2 along with the law of iterated expectations, Property CE.4. In particular, first show that $E(Z|X) = \theta$ and then use CE.4.]

(ii) Use part (i) to prove that the estimator $W = n^{-1}\sum_{i=1}^{n}(Y_i/X_i)$ is unbiased for W, where $\{(X_i,Y_i): i = 1,2,\ldots,n\}$ is a random sample.

(iii) The following table contains data on corn yields for several counties in Iowa. The USDA predicts the number of hectares of corn in each county based on satellite photos. Researchers count the number of "pixels" of corn in the satellite picture (as opposed to, for example, the number of pixels of soybeans or of uncultivated land) and use these to predict the actual number of hectares. To develop a prediction equation to be used for counties in general, the USDA surveyed farmers in selected counties to obtain corn yields in hectares. Let Y_i = corn yield in county i and let X_i = number of corn pixels in the satellite picture for county i. There are $n = 17$ observations for eight counties. Use this sample to compute the estimate of θ devised in part (ii).

Plot	Corn Yield	Corn Pixels
1	165.76	374
2	96.32	209
3	76.08	253
4	185.35	432
5	116.43	367
6	162.08	361
7	152.04	288
8	161.75	369
9	92.88	206

continued

Plot	Corn Yield	Corn Pixels
10	149.94	316
11	64.75	145
12	127.07	355
13	133.55	295
14	77.70	223
15	206.39	459
16	108.33	290
17	118.17	307

C.5 Let Y denote a Bernoulli(θ) random variable with $0 < \theta < 1$. Suppose we are interested in estimating the *odds ratio*, $\gamma = \theta/(1 - \theta)$, which is the probability of success over the probability of failure. Given a random sample $\{Y_1, \ldots, Y_n\}$, we know that an unbiased and consistent estimator of θ is $\bar{Y}$, the proportion of successes in n trials. A natural estimator of γ is $G = \{\bar{Y}/(1 - \bar{Y})\}$, the proportion of successes over the proportion of failures in the sample.

(i) Why is G not an unbiased estimator of γ?

(ii) Use PLIM.2(iii) to show that G is a consistent estimator of γ.

C.6 You are hired by the governor to study whether a tax on liquor has decreased average liquor consumption in your state. You are able to obtain, for a sample of individuals selected at random, the difference in liquor consumption (in ounces) for the years before and after the tax. For person i who is sampled randomly from the population, Y_i denotes the change in liquor consumption. Treat these as a random sample from a Normal(μ,σ^2) distribution.

(i) The null hypothesis is that there was no change in average liquor consumption. State this formally in terms of μ.

(ii) The alternative is that there was a decline in liquor consumption; state the alternative in terms of μ.

(iii) Now, suppose your sample size is $n = 900$ and you obtain the estimates $\bar{y} = -32.8$ and $s = 466.4$. Calculate the t statistic for testing H_0 against H_1; obtain the p-value for the test. (Because of the large sample size, just use the standard normal distribution tabulated in Table G.1.) Do you reject H_0 at the 5% level? at the 1% level?

(iv) Would you say that the estimated fall in consumption is large in magnitude? Comment on the practical versus statistical significance of this estimate.

(v) What has been implicitly assumed in your analysis about other determinants of liquor consumption over the two-year period in order to infer causality from the tax change to liquor consumption?

C.7 The new management at a bakery claims that workers are now more productive than they were under old management, which is why wages have "generally increased." Let W_i^b be Worker i's wage under the old management and let W_i^a be Worker i's wage after the change. The difference is $D_i \equiv W_i^a - W_i^b$. Assume that the D_i are a random sample from a Normal(μ,σ^2) distribution.

(i) Using the following data on 15 workers, construct an exact 95% confidence interval for μ.

(ii) Formally state the null hypothesis that there has been no change in average wages. In particular, what is $E(D_i)$ under H_0? If you are hired to examine the validity of the new management's claim, what is the relevant alternative hypothesis in terms of $\mu = E(D_i)$?

(iii) Test the null hypothesis from part (ii) against the stated alternative at the 5% and 1% levels.

(iv) Obtain the p-value for the test in part (iii).

Worker	Wage Before	Wage After
1	8.30	9.25
2	9.40	9.00
3	9.00	9.25
4	10.50	10.00
5	11.40	12.00
6	8.75	9.50
7	10.00	10.25
8	9.50	9.50
9	10.80	11.50
10	12.55	13.10
11	12.00	11.50
12	8.65	9.00

continued

Worker	Wage Before	Wage After
13	7.75	7.75
14	11.25	11.50
15	12.65	13.00

C.8 The *New York Times* (2/5/90) reported three-point shooting performance for the top ten three-point shooters in the NBA. The following table summarizes these data:

Player	FGA-FGM
Mark Price	429-188
Trent Tucker	833-345
Dale Ellis	1,149-472
Craig Hodges	1,016-396
Danny Ainge	1,051-406
Byron Scott	676-260
Reggie Miller	416-159
Larry Bird	1,206-455
Jon Sundvold	440-166
Brian Taylor	417-157

Note: *FGA* = field goals attempted and *FGM* = field goals made.

For a given player, the outcome of a particular shot can be modeled as a Bernoulli (zero-one) variable: if Y_i is the outcome of shot i, then $Y_i = 1$ if the shot is made, and $Y_i = 0$ if the shot is missed. Let θ denote the probability of making any particular three-point shot attempt. The natural estimator of θ is $\bar{Y} = FGM/FGA$.

(i) Estimate θ for Mark Price.

(ii) Find the standard deviation of the estimator $\bar{Y}$ in terms of θ and the number of shot attempts, n.

(iii) The asymptotic distribution of $(\bar{Y} - \theta)/\text{se}(\bar{Y})$ is standard normal, where $\text{se}(\bar{Y}) = \sqrt{\bar{Y}(1 - \bar{Y})/n}$. Use this fact to test H_0: $\theta = .5$ against H_1: $\theta < .5$ for Mark Price. Use a 1% significance level.

C.9 Suppose that a military dictator in an unnamed country holds a plebiscite (a yes/no vote of confidence) and claims that he was supported by 65% of the voters. A human rights group suspects foul play and hires you to test the validity of the dictator's claim. You have a budget that allows you to randomly sample 200 voters from the country.

(i) Let X be the number of yes votes obtained from a random sample of 200 out of the entire voting population. What is the expected value of X if, in fact, 65% of all voters supported the dictator?

(ii) What is the standard deviation of X, again assuming that the true fraction voting yes in the plebiscite is .65?

(iii) Now, you collect your sample of 200, and you find that 115 people actually voted yes. Use the CLT to approximate the probability that you would find 115 or fewer yes votes from a random sample of 200 if, in fact, 65% of the entire population voted yes.

(iv) How would you explain the relevance of the number in part (iii) to someone who does not having training in statistics?

C.10 Before a strike prematurely ended the 1994 major league baseball season, Tony Gwynn of the San Diego Padres had 165 hits in 419 at bats, for a .394 batting average. There was discussion about whether Gwynn was a potential .400 hitter that year. This issue can be couched in terms of Gwynn's probability of getting a hit on a particular at bat, call it θ. Let Y_i be the Bernoulli(θ) indicator equal to unity if Gwynn gets a hit during his i^{th} at bat, and zero otherwise. Then, $Y_1, Y_2, \ldots, Y_n$ is a random sample from a Bernoulli(θ) distribution, where θ is the probability of success, and $n = 419$.

Our best point estimate of θ is Gwynn's batting average, which is just the proportion of successes: $\bar{y} = .394$. Using the fact that $\text{se}(\bar{y}) = \sqrt{\bar{y}(1 - \bar{y})/n}$, construct an approximate 95% confidence interval for θ, using the standard normal distribution. Would you say there is strong evidence against Gwynn's being a potential .400 hitter? Explain.

Appendix D

Summary of Matrix Algebra

This appendix summarizes the matrix algebra concepts, including the algebra of probability, needed for the study of multiple linear regression models using matrices in Appendix E. None of this material is used in the main text.

D.1 BASIC DEFINITIONS

DEFINITION D.1 (Matrix)
A **matrix** is a rectangular array of numbers. More precisely, an $m \times n$ matrix has m rows and n columns. The positive integer m is called the *row dimension*, and n is called the *column dimension*.

We use uppercase boldface letters to denote matrices. We can write an $m \times n$ matrix generically as

$$\mathbf{A} = [a_{ij}] = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ \vdots & & & & \\ a_{m1} & a_{m2} & a_{m3} & \dots & a_{mn} \end{bmatrix}$$

where a_{ij} represents the element in the i^{th} row and the j^{th} column. For example, a_{25} stands for the number in the second row and the fifth column of $\mathbf{A}$. A specific example of a 2×3 matrix is

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & 7 \\ -4 & 5 & 0 \end{bmatrix} \tag{D.1}$$

where $a_{13} = 7$. The shorthand $\mathbf{A} = [a_{ij}]$ is often used to define matrix operations.

DEFINITION D.2 (Square Matrix)
A **square matrix** has the same number of rows and columns. The dimension of a square matrix is its number of rows and columns.

DEFINITION D.3 (Vectors)
(i) A $1 \times m$ matrix is called a **row vector** (of dimension m) and can be written as $\boldsymbol{x} \equiv (x_1, x_2, \ldots, x_m)$.

(ii) An $n \times 1$ matrix is called a **column vector** and can be written as

$$\boldsymbol{y} \equiv \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}.$$

DEFINITION D.4 (Diagonal Matrix)
A square matrix $\boldsymbol{A}$ is a **diagonal matrix** when all of its diagonal elements are zero, that is, $a_{ij} = 0$ for all $i \neq j$. We can always write a diagonal matrix as

$$\mathbf{A} = \begin{bmatrix} a_{11} & 0 & 0 & \ldots & 0 \\ 0 & a_{22} & 0 & \ldots & 0 \\ \vdots & & & & \\ 0 & 0 & 0 & \ldots & a_{nn} \end{bmatrix}.$$

DEFINITION D.5 (Identity and Zero Matrices)
(i) The $n \times n$ **identity matrix**, denoted $\mathbf{I}$, or sometimes $\mathbf{I}_n$ to emphasize its dimension, is the diagonal matrix with unity (one) in each diagonal position, and zero elsewhere:

$$\mathbf{I} \equiv \mathbf{I}_n \equiv \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & \ldots & 0 \\ \vdots & & & & \vdots \\ 0 & 0 & 0 & \ldots & 1 \end{bmatrix}.$$

(ii) The $m \times n$ **zero matrix**, denoted $\mathbf{0}$, is the $m \times n$ matrix with zero for all entries. This need not be a square matrix.

D.2 MATRIX OPERATIONS

Matrix Addition

Two matrices $\mathbf{A}$ and $\mathbf{B}$, each having dimension $m \times n$, can be added element by element: $\mathbf{A} + \mathbf{B} = [a_{ij} + b_{ij}]$. More precisely,

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \ldots & a_{1n} + b_{1n} \\ a_{21} + b_{21} & a_{22} + b_{22} & \ldots & a_{2n} + b_{2n} \\ \vdots & & & \\ a_{m1} + b_{m1} & a_{m2} + b_{m2} & \ldots & a_{mn} + b_{mn} \end{bmatrix}.$$

For example,

$$\begin{bmatrix} 2 & -1 & 7 \\ -4 & 5 & 0 \end{bmatrix} + \begin{bmatrix} 1 & 0 & -4 \\ 4 & 2 & 3 \end{bmatrix} = \begin{bmatrix} 3 & -1 & 3 \\ 0 & 7 & 3 \end{bmatrix}.$$

Matrices of different dimensions cannot be added.

Scalar Multiplication

Given any real number γ (often called a scalar), **scalar multiplication** is defined as $\gamma\mathbf{A} \equiv [\gamma a_{ij}]$, or

$$\gamma\mathbf{A} = \begin{bmatrix} \gamma a_{11} & \gamma a_{12} & \cdots & \gamma a_{1n} \\ \gamma a_{21} & \gamma a_{22} & \cdots & \gamma a_{2n} \\ \vdots & & & \\ \gamma a_{m1} & \gamma a_{m2} & \cdots & \gamma a_{mn} \end{bmatrix}.$$

For example, if $\gamma = 2$ and $\mathbf{A}$ is the matrix in equation (D.1), then

$$\gamma\mathbf{A} = \begin{bmatrix} 4 & -2 & 14 \\ -8 & 10 & 0 \end{bmatrix}.$$

Matrix Multiplication

To multiply matrix $\mathbf{A}$ by matrix $\mathbf{B}$ to form the product $\mathbf{AB}$, the *column* dimension of $\mathbf{A}$ must equal the *row* dimension of $\mathbf{B}$. Therefore, let $\mathbf{A}$ be an $m \times n$ matrix and let $\mathbf{B}$ be an $n \times p$ matrix. Then **matrix multiplication** is defined as

$$\mathbf{AB} = \left[\sum_{k=1}^{n} a_{ik} b_{kj} \right].$$

In other words, the $(i,j)^{\text{th}}$ element of the new matrix $\mathbf{AB}$ is obtained by multiplying each element in the i^{th} row of $\mathbf{A}$ by the corresponding element in the j^{th} column of $\mathbf{B}$ and adding these n products together. A schematic may help make this process more transparent:

$$\begin{array}{ccccc} & \mathbf{A} & & \mathbf{B} & \mathbf{AB} \\ i^{\text{th}}\text{ row} \rightarrow & \begin{bmatrix} a_{i1}\, a_{i2}\, a_{i3} \ldots a_{in} \end{bmatrix} & \begin{bmatrix} b_{1j} \\ b_{2j} \\ b_{3j} \\ \vdots \\ b_{nj} \end{bmatrix} & = & \begin{bmatrix} \sum_{k=1}^{n} a_{ik} b_{kj} \end{bmatrix}, \\ & & \uparrow & & \uparrow \\ & & j^{\text{th}}\text{ column} & & (i,j)^{\text{th}}\text{ element} \end{array}$$

where, by the definition of the summation operator in Appendix A,

$$\sum_{k=1}^{n} a_{ik}b_{kj} = a_{i1}b_{1j} + a_{i2}b_{2j} + \ldots + a_{in}b_{nj}.$$

For example,

$$\begin{bmatrix} 2 & -1 & 0 \\ -4 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 & 6 & 0 \\ -1 & 2 & 0 & 1 \\ 3 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 12 & -1 \\ -1 & -2 & -24 & 1 \end{bmatrix}.$$

We can also multiply a matrix and a vector. If $\mathbf{A}$ is an $n \times m$ matrix and $\mathbf{y}$ is an $m \times 1$ vector, then $\mathbf{Ay}$ is an $n \times 1$ vector. If $\mathbf{x}$ is a $1 \times n$ vector, then $\mathbf{xA}$ is a $1 \times m$ vector.

Matrix addition, scalar multiplication, and matrix multiplication can be combined in various ways, and these operations satisfy several rules that are familiar from basic operations on numbers. In the following list of properties, $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ are matrices with appropriate dimensions for applying each operation, and α and β are real numbers. Most of these properties are easy to illustrate from the definitions.

PROPERTIES OF MATRIX MULTIPLICATION: (1) $(\alpha + \beta)\mathbf{A} = \alpha\mathbf{A} + \beta\mathbf{A}$; (2) $\alpha(\mathbf{A} + \mathbf{B}) = \alpha\mathbf{A} + \alpha\mathbf{B}$; (3) $(\alpha\beta)\mathbf{A} = \alpha(\beta\mathbf{A})$; (4) $\alpha(\mathbf{AB}) = (\alpha\mathbf{A})\mathbf{B}$; (5) $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$; (6) $(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C})$; (7) $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$; (8) $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$; (9) $(\mathbf{A} + \mathbf{B})\mathbf{C} = \mathbf{AC} + \mathbf{BC}$; (10) $\mathbf{IA} = \mathbf{AI} = \mathbf{A}$; (11) $\mathbf{A} + \mathbf{0} = \mathbf{0} + \mathbf{A} = \mathbf{A}$; (12) $\mathbf{A} - \mathbf{A} = \mathbf{0}$; (13) $\mathbf{A0} = \mathbf{0A} = \mathbf{0}$; (14) $\mathbf{AB} \neq \mathbf{BA}$, even when both products are defined.

The last property deserves further comment. If $\mathbf{A}$ is $n \times m$ and $\mathbf{B}$ is $m \times p$, then $\mathbf{AB}$ is defined, but $\mathbf{BA}$ is defined only if $n = p$ (the row dimension of $\mathbf{A}$ equals the column dimension of $\mathbf{B}$). If $\mathbf{A}$ is $m \times n$ and $\mathbf{B}$ is $n \times m$, then $\mathbf{AB}$ and $\mathbf{BA}$ are both defined, but they are not usually the same; in fact, they have different dimensions, unless $\mathbf{A}$ and $\mathbf{B}$ are both square matrices. Even when $\mathbf{A}$ and $\mathbf{B}$ are both square, $\mathbf{AB} \neq \mathbf{BA}$, except under special circumstances.

Transpose

DEFINITION D.6 (Transpose)

Let $\mathbf{A} = [a_{ij}]$ be an $m \times n$ matrix. The **transpose** of $\mathbf{A}$, denoted $\mathbf{A}'$ (called $\mathbf{A}$ *prime*), is the $n \times m$ matrix obtained by interchanging the rows and columns of $\mathbf{A}$. We can write this as $\mathbf{A}' \equiv [a_{ji}]$.

For example,

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & 7 \\ -4 & 5 & 0 \end{bmatrix}, \quad \mathbf{A}' = \begin{bmatrix} 2 & -4 \\ -1 & 5 \\ 7 & 0 \end{bmatrix}.$$

PROPERTIES OF TRANSPOSE: (1) $(\mathbf{A}')' = \mathbf{A}$; (2) $(\alpha\mathbf{A})' = \alpha\mathbf{A}'$ for any scalar α; (3) $(\mathbf{A} + \mathbf{B})' = \mathbf{A}' + \mathbf{B}'$; (4) $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$, where $\mathbf{A}$ is $m \times n$ and $\mathbf{B}$ is $n \times k$; (5) $\mathbf{x}'\mathbf{x} = \sum_{i=1}^{n} x_i^2$,

where $\mathbf{x}$ is an $n \times 1$ vector; (6) If $\mathbf{A}$ is an $n \times k$ matrix with *rows* given by the $1 \times k$ vectors $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n$, so that we can write

$$\mathbf{A} = \begin{bmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_2 \\ \vdots \\ \boldsymbol{a}_n \end{bmatrix},$$

then $\mathbf{A}' = (\boldsymbol{a}_1' \; \boldsymbol{a}_2' \ldots \boldsymbol{a}_n')$.

DEFINITION D.7 (Symmetric Matrix)
A square matrix $\mathbf{A}$ is a **symmetric matrix** if and only if $\mathbf{A}' = \mathbf{A}$.

If $\mathbf{X}$ is any $n \times k$ matrix, then $\mathbf{X}'\mathbf{X}$ is always defined and is a symmetric matrix, as can be seen by applying the first and fourth transpose properties (see Problem D.3).

Partitioned Matrix Multiplication

Let $\mathbf{A}$ be an $n \times k$ matrix with rows given by the $1 \times k$ vectors $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n$, and let $\mathbf{B}$ be an $n \times m$ matrix with rows given by $1 \times m$ vectors $\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_n$:

$$\mathbf{A} = \begin{bmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_2 \\ \vdots \\ \boldsymbol{a}_n \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \\ \vdots \\ \boldsymbol{b}_n \end{bmatrix}.$$

Then,

$$\mathbf{A}'\mathbf{B} = \sum_{i=1}^{n} \boldsymbol{a}_i' \boldsymbol{b}_i,$$

where for each i, $\boldsymbol{a}_i' \boldsymbol{b}_i$ is a $k \times m$ matrix. Therefore, $\mathbf{A}'\mathbf{B}$ can be written as the sum of n matrices, each of which is $k \times m$. As a special case, we have

$$\mathbf{A}'\mathbf{A} = \sum_{i=1}^{n} \boldsymbol{a}_i' \boldsymbol{a}_i,$$

where $\boldsymbol{a}_i' \boldsymbol{a}_i$ is a $k \times k$ matrix for all i.

Trace

The trace of a matrix is a very simple operation defined only for *square* matrices.

DEFINITION D.8 (Trace)
For any $n \times n$ matrix $\mathbf{A}$, the **trace of a matrix A**, denoted tr($\mathbf{A}$), is the sum of its diagonal elements. Mathematically,

$$\text{tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}.$$

PROPERTIES OF TRACE: (1) $\text{tr}(\mathbf{I}_n) = n$; (2) $\text{tr}(\mathbf{A}') = \text{tr}(\mathbf{A})$; (3) $\text{tr}(\mathbf{A} + \mathbf{B}) = \text{tr}(\mathbf{A}) + \text{tr}(\mathbf{B})$; (4) $\text{tr}(\alpha\mathbf{A}) = \alpha\text{tr}(\mathbf{A})$, for any scalar α; (5) $\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})$, where $\mathbf{A}$ is $m \times n$ and $\mathbf{B}$ is $n \times m$.

Inverse

The notion of a matrix inverse is very important for square matrices.

DEFINITION D.9 (Inverse)
An $n \times n$ matrix $\mathbf{A}$ has an **inverse**, denoted $\mathbf{A}^{-1}$, provided that $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_n$ and $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}_n$. In this case, $\mathbf{A}$ is said to be *invertible* or *nonsingular.* Otherwise, it is said to be *noninvertible* or *singular*.

PROPERTIES OF INVERSE: (1) If an inverse exists, it is unique; (2) $(\alpha\mathbf{A})^{-1} = (1/\alpha)\mathbf{A}^{-1}$, if $\alpha \neq 0$ and $\mathbf{A}$ is invertible; (3) $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$, if $\mathbf{A}$ and $\mathbf{B}$ are both $n \times n$ and invertible; (4) $(\mathbf{A}')^{-1} = (\mathbf{A}^{-1})'$.

We will not be concerned with the mechanics of calculating the inverse of a matrix. Any matrix algebra text contains detailed examples of such calculations.

D.3 LINEAR INDEPENDENCE. RANK OF A MATRIX

For a set of vectors having the same dimension, it is important to know whether one vector can be expressed as a linear combination of the remaining vectors.

DEFINITION D.10 (Linear Independence)
Let $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_r\}$ be a set of $n \times 1$ vectors. These are **linearly independent vectors** if and only if

$$\alpha_1\boldsymbol{x}_1 + \alpha_2\boldsymbol{x}_2 + \ldots + \alpha_r\boldsymbol{x}_r = \mathbf{0} \qquad \textbf{(D.2)}$$

implies that $\alpha_1 = \alpha_2 = \ldots = \alpha_r = 0$. If (D.2) holds for a set of scalars that are not all zero, then $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_r\}$ is *linearly dependent*.

The statement that $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_r\}$ is linearly dependent is equivalent to saying that at least one vector in this set can be written as a linear combination of the others.

DEFINITION D.11 (Rank)
(i) Let $\mathbf{A}$ be an $n \times m$ matrix. The **rank of a matrix A**, denoted rank($\mathbf{A}$), is the maximum number of linearly independent columns of $\mathbf{A}$.

(ii) If $\mathbf{A}$ is $n \times m$ and $\text{rank}(\mathbf{A}) = m$, then $\mathbf{A}$ has *full column rank*.

If $\mathbf{A}$ is $n \times m$, its rank can be at most m. A matrix has full column rank if its columns form a linearly independent set. For example, the 3×2 matrix

$$\begin{bmatrix} 1 & 3 \\ 2 & 6 \\ 0 & 0 \end{bmatrix}$$

can have at most rank two. In fact, its rank is only one because the second column is three times the first column.

PROPERTIES OF RANK: (1) $\text{rank}(\mathbf{A}') = \text{rank}(\mathbf{A})$; (2) If $\mathbf{A}$ is $n \times k$, then $\text{rank}(\mathbf{A}) \leq \min(n,k)$; (3) If $\mathbf{A}$ is $k \times k$ and $\text{rank}(\mathbf{A}) = k$, then $\mathbf{A}$ is nonsingular.

D.4 QUADRATIC FORMS AND POSITIVE DEFINITE MATRICES

DEFINITION D.12 (Quadratic Form)

Let $\mathbf{A}$ be an $n \times n$ *symmetric* matrix. The **quadratic form** associated with the matrix $\mathbf{A}$ is the real-valued function defined for all $n \times 1$ vectors $\boldsymbol{x}$:

$$f(\boldsymbol{x}) = \boldsymbol{x}'\mathbf{A}\boldsymbol{x} = \sum_{i=1}^{n} a_{ii}x_i^2 + 2\sum_{i=1}^{n}\sum_{j>i} a_{ij}x_i x_j.$$

DEFINITION D.13 (Positive Definite and Positive Semi-Definite)

(i) A symmetric matrix $\mathbf{A}$ is said to be **positive definite** (p.d.) if

$$\boldsymbol{x}'\mathbf{A}\boldsymbol{x} > 0 \text{ for all } n \times 1 \text{ vectors } \boldsymbol{x} \text{ except } \boldsymbol{x} = \mathbf{0}.$$

(ii) A symmetric matrix $\mathbf{A}$ is **positive semi-definite** (p.s.d.) if

$$\boldsymbol{x}'\mathbf{A}\boldsymbol{x} \geq 0 \text{ for all } n \times 1 \text{ vectors.}$$

If a matrix is positive definite or positive semi-definite, it is automatically assumed to be symmetric.

PROPERTIES OF POSITIVE DEFINITE AND POSITIVE SEMI-DEFINITE MATRICES: (1) A positive definite matrix has diagonal elements that are strictly positive, while a p.s.d. matrix has nonnegative diagonal elements; (2) If $\mathbf{A}$ is p.d., then $\mathbf{A}^{-1}$ exists and is p.d.; (3) If $\mathbf{X}$ is $n \times k$, then $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}\mathbf{X}'$ are p.s.d.; (4) If $\mathbf{X}$ is $n \times k$ and $\text{rank}(\mathbf{X}) = k$, then $\mathbf{X}'\mathbf{X}$ is p.d. (and therefore nonsingular).

D.5 IDEMPOTENT MATRICES

DEFINITION D.14 (Idempotent Matrix)

Let $\mathbf{A}$ be an $n \times n$ symmetric matrix. Then $\mathbf{A}$ is said to be an **idempotent matrix** if and only if $\mathbf{A}\mathbf{A} = \mathbf{A}$.

For example,

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

is an idempotent matrix, as direct multiplication verifies.

PROPERTIES OF IDEMPOTENT MATRICES: Let $\mathbf{A}$ be an $n \times n$ idempotent matrix. (1) $\text{rank}(\mathbf{A}) = \text{tr}(\mathbf{A})$; (2) $\mathbf{A}$ is positive semi-definite.

We can construct idempotent matrices very generally. Let $\mathbf{X}$ be an $n \times k$ matrix with $\text{rank}(\mathbf{X}) = k$. Define

$$\mathrm{P} \equiv \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$
$$\mathrm{M} \equiv \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{I}_n - \mathbf{P}.$$

Then $\mathbf{P}$ and $\mathbf{M}$ are symmetric, idempotent matrices with $\text{rank}(\mathbf{P}) = k$ and $\text{rank}(\mathbf{M}) = n - k$. The ranks are most easily obtained by using Property 1: $\text{tr}(\mathbf{P}) = \text{tr}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}]$ (from Property 5 for trace) $= \text{tr}(\mathbf{I}_k) = k$ (by Property 1 for trace). It easily follows that $\text{tr}(\mathbf{M}) = \text{tr}(\mathbf{I}_n) - \text{tr}(\mathbf{P}) = n - k$.

D.6 DIFFERENTIATION OF LINEAR AND QUADRATIC FORMS

For a given $n \times 1$ vector $\boldsymbol{a}$, consider the linear function defined by

$$f(\boldsymbol{x}) = \boldsymbol{a}'\boldsymbol{x},$$

for all $n \times 1$ vectors $\boldsymbol{x}$. The derivative of f with respect to $\boldsymbol{x}$ is the $1 \times n$ vector of partial derivatives, which is simply

$$\partial f(\boldsymbol{x})/\partial \boldsymbol{x} = \boldsymbol{a}'.$$

For an $n \times n$ symmetric matrix $\boldsymbol{A}$, define the quadratic form

$$g(\boldsymbol{x}) = \boldsymbol{x}'\mathbf{A}\boldsymbol{x}.$$

Then,

$$\partial g(\boldsymbol{x})/\partial \boldsymbol{x} = 2\boldsymbol{x}'\mathbf{A},$$

which is a $1 \times n$ vector.

D.7 MOMENTS AND DISTRIBUTIONS OF RANDOM VECTORS

In order to derive the expected value and variance of the OLS estimators using matrices, we need to define the expected value and variance of a **random vector**. As its name suggests, a random vector is simply a vector of random variables. We also need to define the multivariate normal distribution. These concepts are simply extensions of those covered in Appendix B.

Expected Value

DEFINITION D.15 (Expected Value)
(i) If $\boldsymbol{y}$ is an $n \times 1$ random vector, the **expected value** of $\boldsymbol{y}$, denoted $\mathrm{E}(\boldsymbol{y})$, is the vector of expected values: $\mathrm{E}(\boldsymbol{y}) = [\mathrm{E}(y_1), \mathrm{E}(y_2), \ldots, \mathrm{E}(y_n)]'$.

(ii) If $\mathbf{Z}$ is an $n \times m$ random matrix, $\mathrm{E}(\mathbf{Z})$ is the $n \times m$ matrix of expected values: $\mathrm{E}(\mathbf{Z}) = [\mathrm{E}(z_{ij})]$.

PROPERTIES OF EXPECTED VALUE: (1) If $\mathbf{A}$ is an $m \times n$ matrix and $\boldsymbol{b}$ is an $n \times 1$ vector, where both are nonrandom, then $\mathrm{E}(\mathbf{A}\boldsymbol{y} + \boldsymbol{b}) = \mathbf{A}\mathrm{E}(\boldsymbol{y}) + \boldsymbol{b}$; (2) If $\mathbf{A}$ is $p \times n$ and $\mathbf{B}$ is $m \times k$, where both are nonrandom, then $\mathrm{E}(\mathbf{AZB}) = \mathbf{A}\mathrm{E}(\mathbf{Z})\mathbf{B}$.

Variance-Covariance Matrix

DEFINITION D.16 (Variance-Covariance Matrix)
If $\boldsymbol{y}$ is an $n \times 1$ random vector, its **variance-covariance matrix**, denoted $\mathrm{Var}(\boldsymbol{y})$, is defined as

$$\mathrm{Var}(\boldsymbol{y}) = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \ldots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \ldots & \sigma_{2n} \\ \vdots & & & \\ \sigma_{n1} & \sigma_{n2} & \ldots & \sigma_n^2 \end{bmatrix},$$

where $\sigma_j^2 = \mathrm{Var}(y_j)$ and $\sigma_{ij} = \mathrm{Cov}(y_i, y_j)$. In other words, the variance-covariance matrix has the variances of each element of $\boldsymbol{y}$ down its diagonal, with covariance terms in the off diagonals. Because $\mathrm{Cov}(y_i, y_j) = \mathrm{Cov}(y_j, y_i)$, it immediately follows that a variance-covariance matrix is symmetric.

PROPERTIES OF VARIANCE: (1) If $\boldsymbol{a}$ is an $n \times 1$ nonrandom vector, then $\mathrm{Var}(\boldsymbol{a}'\boldsymbol{y}) = \boldsymbol{a}'[\mathrm{Var}(\mathbf{y})]\boldsymbol{a} \geq 0$; (2) If $\mathrm{Var}(\boldsymbol{a}'\boldsymbol{y}) > 0$ for all $\boldsymbol{a} \neq \mathbf{0}$, $\mathrm{Var}(\boldsymbol{y})$ is positive definite; (3) $\mathrm{Var}(\boldsymbol{y}) = \mathrm{E}[(\boldsymbol{y} - \boldsymbol{\mu})(\boldsymbol{y} - \boldsymbol{\mu})']$, where $\boldsymbol{\mu} = \mathrm{E}(\boldsymbol{y})$; (4) If the elements of $\boldsymbol{y}$ are uncorrelated, $\mathrm{Var}(\boldsymbol{y})$ is a diagonal matrix. If, in addition, $\mathrm{Var}(y_j) = \sigma^2$ for $j = 1,2,\ldots,n$, then $\mathrm{Var}(\boldsymbol{y}) = \sigma^2 \boldsymbol{I}_n$; (5) If $\boldsymbol{A}$ is an $m \times n$ nonrandom matrix and $\boldsymbol{b}$ is an $n \times 1$ nonrandom vector, then $\mathrm{Var}(\mathbf{A}\boldsymbol{y} + \boldsymbol{b}) = \boldsymbol{A}[\mathrm{Var}(\boldsymbol{y})]\boldsymbol{A}'$.

Multivariate Normal Distribution

The normal distribution for a random variable was discussed at some length in Appendix B. We need to extend the normal distribution to random vectors. We will not provide an expression for the probability distribution function, as we do not need it. It is important to know that a multivariate normal random vector is completely characterized by its mean and its variance-covariance matrix. Therefore, if $\boldsymbol{y}$ is an $n \times 1$ multivariate normal random vector with mean $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$, we write $\boldsymbol{y} \sim \mathrm{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We now state several useful properties of the **multivariate normal distribution**.

PROPERTIES OF THE MULTIVARIATE NORMAL DISTRIBUTION: (1) If $\boldsymbol{y} \sim \text{Normal}(\boldsymbol{\mu},\boldsymbol{\Sigma})$, then each element of $\boldsymbol{y}$ is normally distributed; (2) If $\boldsymbol{y} \sim \text{Normal}(\boldsymbol{\mu},\boldsymbol{\Sigma})$, then y_i and y_j, any two elements of $\boldsymbol{y}$, are independent if and only if they are uncorrelated, that is, $\sigma_{ij} = 0$; (3) If $\boldsymbol{y} \sim \text{Normal}(\boldsymbol{\mu},\boldsymbol{\Sigma})$, then $\mathbf{A}\boldsymbol{y} + \boldsymbol{b} \sim \text{Normal}(\mathbf{A}\boldsymbol{\mu} + \boldsymbol{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$, where $\mathbf{A}$ and $\boldsymbol{b}$ are nonrandom; (4) If $\boldsymbol{y} \sim \text{Normal}(\mathbf{0},\boldsymbol{\Sigma})$, then, for nonrandom matrices $\mathbf{A}$ and $\mathbf{B}$, $\mathbf{A}\boldsymbol{y}$ and $\mathbf{B}\boldsymbol{y}$ are independent if and only if $\mathbf{A}\boldsymbol{\Sigma}\mathbf{B}' = \mathbf{0}$. In particular, if $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}_n$, then $\mathbf{A}\mathbf{B}' = \mathbf{0}$ is necessary and sufficient for independence of $\mathbf{A}\boldsymbol{y}$ and $\mathbf{B}\boldsymbol{y}$; (5) If $\boldsymbol{y} \sim \text{Normal}(\mathbf{0},\sigma^2\mathbf{I}_n)$, $\mathbf{A}$ is a $k \times n$ nonrandom matrix, and $\mathbf{B}$ is an $n \times n$ symmetric, idempotent matrix, then $\mathbf{A}\boldsymbol{y}$ and $\boldsymbol{y}'\mathbf{B}\boldsymbol{y}$ are independent if and only if $\mathbf{A}\mathbf{B} = \mathbf{0}$; (6) If $\boldsymbol{y} \sim \text{Normal}(\mathbf{0},\sigma^2\mathbf{I}_n)$ and $\mathbf{A}$ and $\mathbf{B}$ are nonrandom symmetric, idempotent matrices, then $\boldsymbol{y}'\mathbf{A}\boldsymbol{y}$ and $\boldsymbol{y}'\mathbf{B}\boldsymbol{y}$ are independent if and only if $\mathbf{A}\mathbf{B} = \mathbf{0}$.

Chi-Square Distribution

In Appendix B, we defined a **chi-square random variable** as the sum of *squared* independent standard normal random variables. In vector notation, if $\boldsymbol{u} \sim \text{Normal}(\mathbf{0},\mathbf{I}_n)$, then $\boldsymbol{u}'\boldsymbol{u} \sim \chi^2_n$.

PROPERTIES OF THE CHI-SQUARE DISTRIBUTION: (1) If $\boldsymbol{u} \sim \text{Normal}(\mathbf{0},\mathbf{I}_n)$ and $\mathbf{A}$ is an $n \times n$ symmetric, idempotent matrix with $\text{rank}(\mathbf{A}) = q$, then $\boldsymbol{u}'\mathbf{A}\boldsymbol{u} \sim \chi^2_q$; (2) If $\boldsymbol{u} \sim \text{Normal}(\mathbf{0},\mathbf{I}_n)$ and $\mathbf{A}$ and $\mathbf{B}$ are $n \times n$ symmetric, idempotent matrices such that $\mathbf{A}\mathbf{B} = \mathbf{0}$, then $\boldsymbol{u}'\mathbf{A}\boldsymbol{u}$ and $\boldsymbol{u}'\mathbf{B}\boldsymbol{u}$ are independent, chi-square random variables.

t Distribution

We also defined the ***t* distribution** in Appendix B. Now we add an important property.

PROPERTY OF THE *t* DISTRIBUTION: If $\boldsymbol{u} \sim \text{Normal}(\mathbf{0},\mathbf{I}_n)$, $\boldsymbol{c}$ is an $n \times 1$ nonrandom vector, $\mathbf{A}$ is a nonrandom $n \times n$ symmetric, idempotent matrix with rank q, and $\mathbf{A}\boldsymbol{c} = \mathbf{0}$, then $\{\boldsymbol{c}'\boldsymbol{u}/(\boldsymbol{c}'\boldsymbol{c})^{1/2}\}/(\boldsymbol{u}'\mathbf{A}\boldsymbol{u})^{1/2} \sim t_q$.

F Distribution

Recall that an ***F* random variable** is obtained by taking two *independent* chi-square random variables and finding the ratio of each standardized by degrees of freedom.

PROPERTY OF THE *F* DISTRIBUTION: If $\boldsymbol{u} \sim \text{Normal}(\mathbf{0},\mathbf{I}_n)$ and $\mathbf{A}$ and $\mathbf{B}$ are $n \times n$ nonrandom symmetric, idempotent matrices with $\text{rank}(\mathbf{A}) = k_1$, $\text{rank}(\mathbf{B}) = k_2$, and $\mathbf{A}\mathbf{B} = \mathbf{0}$, then $(\boldsymbol{u}'\mathbf{A}\boldsymbol{u}/k_1)/(\boldsymbol{u}'\mathbf{B}\boldsymbol{u}/k_2) \sim F_{k_1,k_2}$.

SUMMARY

This appendix contains a condensed form of the background information needed to study the classical linear model using matrices. While the material here is self-contained, it is primarily intended as a review for readers who are familiar with matrix algebra and multivariate statistics, and it will be used extensively in Appendix E.

KEY TERMS

Chi-Square Random Variable
Column Vector
Diagonal Matrix
Expected Value
F Random Variable
Idempotent Matrix
Identity Matrix
Inverse
Linearly Independent Vectors
Matrix
Matrix Multiplication
Multivariate Normal Distribution
Positive Definite
Positive Semi-Definite
Quadratic Form
Random Vector
Rank of a Matrix
Row Vector
Scalar Multiplication
Square Matrix
Symmetric Matrix
t Distribution
Trace of a Matrix
Transpose
Variance-Covariance Matrix
Zero Matrix

PROBLEMS

D.1 (i) Find the product **AB** using

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & 7 \\ -4 & 5 & 0 \end{bmatrix}, \ \mathbf{B} = \begin{bmatrix} 0 & 1 & 6 \\ 1 & 8 & 0 \\ 3 & 0 & 0 \end{bmatrix}.$$

(ii) Does **BA** exist?

D.2 If **A** and **B** are $n \times n$ diagonal matrices, show that $\mathbf{AB} = \mathbf{BA}$.

D.3 Let **X** be any $n \times k$ matrix. Show that $\mathbf{X}'\mathbf{X}$ is a symmetric matrix.

D.4 (i) Use the properties of trace to argue that $\text{tr}(\mathbf{A}'\mathbf{A}) = \text{tr}(\mathbf{A}\mathbf{A}')$ for any $n \times m$ matrix **A**.

(ii) For $\mathbf{A} = \begin{bmatrix} 2 & 0 & -1 \\ 0 & 3 & 0 \end{bmatrix}$, verify that $\text{tr}(\mathbf{A}'\mathbf{A}) = \text{tr}(\mathbf{A}\mathbf{A}')$.

D.5 (i) Use the definition of inverse to prove the following: if **A** and **B** are $n \times n$ nonsingular matrices, then $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$.

(ii) If **A**, **B**, and **C** are all $n \times n$ nonsingular matrices, find $(\mathbf{ABC})^{-1}$ in terms of $\mathbf{A}^{-1}$, $\mathbf{B}^{-1}$, and $\mathbf{C}^{-1}$.

D.6 (i) Show that if **A** is an $n \times n$ symmetric, positive definite matrix, then **A** must have strictly positive diagonal elements.

(ii) Write down a 2×2 symmetric matrix with strictly positive diagonal elements that is *not* positive definite.

D.7 Let **A** be an $n \times n$ symmetric, positive definite matrix. Show that if **P** is any $n \times n$ nonsingular matrix, then $\mathbf{P}'\mathbf{AP}$ is positive definite.

D.8 Prove Property 5 of variances for vectors, using Property 3.

Appendix E

The Linear Regression Model in Matrix Form

This appendix derives various results for ordinary least squares estimation of the multiple linear regression model using matrix notation and matrix algebra (see Appendix D for a summary). The material presented here is much more advanced than that in the text.

E.1 THE MODEL AND ORDINARY LEAST SQUARES ESTIMATION

Throughout this appendix, we use the t subscript to index observations and an n to denote the sample size. It is useful to write the multiple linear regression model with k parameters as follows:

$$y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \ldots + \beta_k x_{tk} + u_t, \; t = 1,2,\ldots,n, \qquad \textbf{(E.1)}$$

where y_t is the dependent variable for observation t, and x_{tj}, $j = 2,3,\ldots,k$, are the independent variables. Notice how our labeling convention here differs from the text: we call the intercept β_1 and let $\beta_2, \ldots, \beta_k$ denote the slope parameters. This relabeling is not important, but it simplifies the matrix approach to multiple regression.

For each t, define a $1 \times k$ vector, $\boldsymbol{x}_t = (1, x_{t2}, \ldots, x_{tk})$, and let $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_k)'$ be the $k \times 1$ vector of all parameters. Then, we can write (E.1) as

$$y_t = \boldsymbol{x}_t\boldsymbol{\beta} + u_t, \; t = 1,2,\ldots,n. \qquad \textbf{(E.2)}$$

[Some authors prefer to define $\boldsymbol{x}_t$ as a column vector, in which case, $\boldsymbol{x}_t$ is replaced with $\boldsymbol{x}_t'$ in (E.2). Mathematically, it makes more sense to define it as a row vector.] We can write (E.2) in full matrix notation by appropriately defining data vectors and matrices. Let $\boldsymbol{y}$ denote the $n \times 1$ vector of observations on y: the t^{th} element of $\boldsymbol{y}$ is y_t. Let $\boldsymbol{X}$ be the $n \times k$ vector of observations on the explanatory variables. In other words, the t^{th} row of $\mathbf{X}$ consists of the vector $\boldsymbol{x}_t$. Equivalently, the $(t,j)^{\text{th}}$ element of $\mathbf{X}$ is simply x_{tj}:

$$\underset{n \times k}{\mathbf{X}} \equiv \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_n \end{bmatrix} = \begin{bmatrix} 1 & x_{12} & x_{13} & \dots & x_{1k} \\ 1 & x_{22} & x_{23} & \dots & x_{2k} \\ \vdots & & & & \\ 1 & x_{n2} & x_{n3} & \dots & x_{nk} \end{bmatrix}.$$

Finally, let $\boldsymbol{u}$ be the $n \times 1$ vector of unobservable disturbances. Then, we can write (E.2) for all n observations in **matrix notation**:

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{u}. \tag{E.3}$$

Remember, because $\mathbf{X}$ is $n \times k$ and $\boldsymbol{\beta}$ is $k \times 1$, $\mathbf{X}\boldsymbol{\beta}$ is $n \times 1$.

Estimation of $\boldsymbol{\beta}$ proceeds by minimizing the sum of squared residuals, as in Section 3.2. Define the sum of squared residuals function for any possible $k \times 1$ parameter vector $\boldsymbol{b}$ as

$$\text{SSR}(\boldsymbol{b}) \equiv \sum_{t=1}^{n} (y_t - \boldsymbol{x}_t\boldsymbol{b})^2.$$

The $k \times 1$ vector of ordinary least squares estimates, $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k)'$, minimizes $\text{SSR}(\boldsymbol{b})$ over all possible $k \times 1$ vectors $\boldsymbol{b}$. This is a problem in multivariable calculus. For $\hat{\boldsymbol{\beta}}$ to minimize the sum of squared residuals, it must solve the **first order condition**

$$\partial \text{SSR}(\hat{\boldsymbol{\beta}})/\partial \boldsymbol{b} \equiv 0. \tag{E.4}$$

Using the fact that the derivative of $(y_t - \boldsymbol{x}_t\boldsymbol{b})^2$ with respect to $\boldsymbol{b}$ is the $1 \times k$ vector $-2(y_t - \boldsymbol{x}_t\boldsymbol{b})\boldsymbol{x}_t$, (E.4) is equivalent to

$$\sum_{t=1}^{n} \boldsymbol{x}_t'(y_t - \boldsymbol{x}_t\hat{\boldsymbol{\beta}}) \equiv \mathbf{0}. \tag{E.5}$$

(We have divided by -2 and taken the transpose.) We can write this first order condition as

$$\begin{aligned} \sum_{t=1}^{n} (y_t - \hat{\beta}_1 - \hat{\beta}_2 x_{t2} - \dots - \hat{\beta}_k x_{tk}) &= 0 \\ \sum_{t=1}^{n} x_{t2}(y_t - \hat{\beta}_1 - \hat{\beta}_2 x_{t2} - \dots - \hat{\beta}_k x_{tk}) &= 0 \\ \vdots & \\ \sum_{t=1}^{n} x_{tk}(y_t - \hat{\beta}_1 - \hat{\beta}_2 x_{t2} - \dots - \hat{\beta}_k x_{tk}) &= 0, \end{aligned}$$

which, apart from the different labeling convention, is identical to the first order conditions in equation (3.13). We want to write these in matrix form to make them more useful. Using the formula for partitioned multiplication in Appendix D, we see that (E.5) is equivalent to

$$\mathbf{X}'(\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{0} \tag{E.6}$$

or

$$(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}'\boldsymbol{y}. \tag{E.7}$$

It can be shown that (E.7) always has at least one solution. Multiple solutions do not help us, as we are looking for a unique set of OLS estimates given our data set. Assuming that the $k \times k$ symmetric matrix $\mathbf{X}'\mathbf{X}$ is nonsingular, we can premultiply both sides of (E.7) by $(\mathbf{X}'\mathbf{X})^{-1}$ to solve for the OLS estimator $\hat{\boldsymbol{\beta}}$:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{y}. \tag{E.8}$$

This is the critical formula for matrix analysis of the multiple linear regression model. The assumption that $\mathbf{X}'\mathbf{X}$ is invertible is equivalent to the assumption that $\text{rank}(\mathbf{X}) = k$, which means that the columns of $\mathbf{X}$ must be linearly independent. This is the matrix version of MLR.4 in Chapter 3.

Before we continue, (E.8) warrants a word of warning. It is tempting to simplify the formula for $\hat{\boldsymbol{\beta}}$ as follows:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{y} = \mathbf{X}^{-1}(\mathbf{X}')^{-1}\mathbf{X}'\boldsymbol{y} = \mathbf{X}^{-1}\boldsymbol{y}.$$

The flaw in this reasoning is that $\mathbf{X}$ is usually not a square matrix, and so it cannot be inverted. In other words, we cannot write $(\mathbf{X}'\mathbf{X})^{-1} = \mathbf{X}^{-1}(\mathbf{X}')^{-1}$ unless $n = k$, a case that virtually never arises in practice.

The $n \times 1$ vectors of OLS fitted values and residuals are given by

$$\hat{\boldsymbol{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{u}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{\beta}}.$$

From (E.6) and the definition of $\hat{\boldsymbol{u}}$, we can see that the first order condition for $\hat{\boldsymbol{\beta}}$ is the same as

$$\mathbf{X}'\hat{\boldsymbol{u}} = \mathbf{0}. \tag{E.9}$$

Because the first column of $\mathbf{X}$ consists entirely of ones, (E.9) implies that the OLS residuals always sum to zero when an intercept is included in the equation and that the sample covariance between each independent variable and the OLS residuals is zero. (We discussed both of these properties in Chapter 3.)

The sum of squared residuals can be written as

$$\text{SSR} = \sum_{t=1}^{n} \hat{u}_t^2 = \hat{\boldsymbol{u}}'\hat{\boldsymbol{u}} = (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{\beta}}). \tag{E.10}$$

All of the algebraic properties from Chapter 3 can be derived using matrix algebra. For example, we can show that the total sum of squares is equal to the explained sum of squares plus the sum of squared residuals [see (3.27)]. The use of matrices does not provide a simpler proof than summation notation, so we do not provide another derivation.

The matrix approach to multiple regression can be used as the basis for a geometrical interpretation of regression. This involves mathematical concepts that are even more advanced than those we covered in Appendix D. [See Goldberger (1991) or Greene (1997).]

E.2 FINITE SAMPLE PROPERTIES OF OLS

Deriving the expected value and variance of the OLS estimator $\hat{\boldsymbol{\beta}}$ is facilitated by matrix algebra, but we must show some care in stating the assumptions.

ASSUMPTION E.1 (LINEAR IN PARAMETERS)
The model can be written as in (E.3), where $\boldsymbol{y}$ is an observed $n \times 1$ vector, $\mathbf{X}$ is an $n \times k$ observed matrix, and $\boldsymbol{u}$ is an $n \times 1$ vector of unobserved errors or disturbances.

ASSUMPTION E.2 (ZERO CONDITIONAL MEAN)
Conditional on the entire matrix $\mathbf{X}$, each error u_t has zero mean: $\mathrm{E}(u_t|\mathbf{X}) = 0$, $t = 1,2,\ldots,n$. In vector form,

$$\mathrm{E}(\boldsymbol{u}|\mathbf{X}) = \mathbf{0}. \tag{E.11}$$

This assumption is implied by MLR.3 under the random sampling assumption, MLR.2. In time series applications, Assumption E.2 imposes strict exogeneity on the explanatory variables, something discussed at length in Chapter 10. This rules out explanatory variables whose future values are correlated with u_t; in particular, it eliminates lagged dependent variables. Under Assumption E.2, we can condition on the x_{tj} when we compute the expected value of $\hat{\boldsymbol{\beta}}$.

ASSUMPTION E.3 (NO PERFECT COLLINEARITY)
The matrix $\mathbf{X}$ has rank k.

This is a careful statement of the assumption that rules out linear dependencies among the explanatory variables. Under Assumption E.3, $\mathbf{X}'\mathbf{X}$ is nonsingular, and so $\hat{\boldsymbol{\beta}}$ is unique and can be written as in (E.8).

THEOREM E.1 (UNBIASEDNESS OF OLS)
Under Assumptions E.1, E.2, and E.3, the OLS estimator $\hat{\boldsymbol{\beta}}$ is unbiased for $\boldsymbol{\beta}$.

PROOF: Use Assumptions E.1 and E.3 and simple algebra to write

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{u}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{u} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{u},\end{aligned} \tag{E.12}$$

where we use the fact that $(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X}) = \mathbf{I}_k$. Taking the expectation conditional on $\mathbf{X}$ gives

$$\begin{aligned} \mathrm{E}(\hat{\boldsymbol{\beta}}|\mathbf{X}) &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{E}(\boldsymbol{u}|\mathbf{X}) \\ &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{0} = \boldsymbol{\beta}, \end{aligned}$$

because $\mathrm{E}(\boldsymbol{u}|\mathbf{X}) = \mathbf{0}$ under Assumption E.2. This argument clearly does not depend on the value of $\boldsymbol{\beta}$, so we have shown that $\hat{\boldsymbol{\beta}}$ is unbiased.

To obtain the simplest form of the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$, we impose the assumptions of homoskedasticity and no serial correlation.

ASSUMPTION E.4 (HOMOSKEDASTICITY AND NO SERIAL CORRELATION)

(i) $\mathrm{Var}(u_t|\mathbf{X}) = \sigma^2$, $t = 1,2,\ldots,n$. (ii) $\mathrm{Cov}(u_t,u_s|\mathbf{X}) = 0$, for all $t \neq s$. In matrix form, we can write these two assumptions as

$$\mathrm{Var}(\boldsymbol{u}|\mathbf{X}) = \sigma^2\mathbf{I}_n, \tag{E.13}$$

where $\mathbf{I}_n$ is the $n \times n$ identity matrix.

Part (i) of Assumption E.4 is the homoskedasticity assumption: the variance of u_t cannot depend on any element of $\mathbf{X}$, and the variance must be constant across observations, t. Part (ii) is the no serial correlation assumption: the errors cannot be correlated across observations. Under random sampling, and in any other cross-sectional sampling schemes with independent observations, part (ii) of Assumption E.4 automatically holds. For time series applications, part (ii) rules out correlation in the errors over time (both conditional on $\mathbf{X}$ and unconditionally).

Because of (E.13), we often say that $\boldsymbol{u}$ has **scalar variance-covariance matrix** when Assumption E.4 holds. We can now derive the **variance-covariance matrix of the OLS estimator**.

THEOREM E.2 (VARIANCE-COVARIANCE MATRIX OF THE OLS ESTIMATOR)

Under Assumptions E.1 through E.4,

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}. \tag{E.14}$$

PROOF: From the last formula in equation (E.12), we have

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) = \mathrm{Var}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{u}|\mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[\mathrm{Var}(\boldsymbol{u}|\mathbf{X})]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}.$$

Now, we use Assumption E.4 to get

$$\begin{aligned} \mathrm{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2\mathbf{I}_n)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}. \end{aligned}$$

Formula (E.14) means that the variance of $\hat{\beta}_j$ (conditional on $\mathbf{X}$) is obtained by multiplying σ^2 by the j^{th} diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$. For the slope coefficients, we gave an interpretable formula in equation (3.51). Equation (E.14) also tells us how to obtain the covariance between any two OLS estimates: multiply σ^2 by the appropriate off diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$. In Chapter 4, we showed how to avoid explicitly finding covariances for obtaining confidence intervals and hypotheses tests by appropriately rewriting the model.

The Gauss-Markov Theorem, in its full generality, can be proven.

THEOREM E.3 (GAUSS-MARKOV THEOREM)
Under Assumptions E.1 through E.4, $\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimator.

PROOF: Any other linear estimator of $\boldsymbol{\beta}$ can be written as

$$\tilde{\boldsymbol{\beta}} = \mathbf{A}'\mathbf{y}, \tag{E.15}$$

where $\mathbf{A}$ is an $n \times k$ matrix. In order for $\tilde{\boldsymbol{\beta}}$ to be unbiased conditional on $\mathbf{X}$, $\mathbf{A}$ can consist of nonrandom numbers and functions of $\mathbf{X}$. (For example, $\mathbf{A}$ cannot be a function of $\mathbf{y}$.) To see what further restrictions on $\mathbf{A}$ are needed, write

$$\tilde{\boldsymbol{\beta}} = \mathbf{A}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) = (\mathbf{A}'\mathbf{X})\boldsymbol{\beta} + \mathbf{A}'\mathbf{u}. \tag{E.16}$$

Then,

$$\begin{aligned} \mathrm{E}(\tilde{\boldsymbol{\beta}}|\mathbf{X}) &= \mathbf{A}'\mathbf{X}\boldsymbol{\beta} + \mathrm{E}(\mathbf{A}'\mathbf{u}|\mathbf{X}) \\ &= \mathbf{A}'\mathbf{X}\boldsymbol{\beta} + \mathbf{A}'\mathrm{E}(\mathbf{u}|\mathbf{X}) \text{ since } \mathbf{A} \text{ is a function of } \mathbf{X} \\ &= \mathbf{A}'\mathbf{X}\boldsymbol{\beta} \text{ since } \mathrm{E}(\mathbf{u}|\mathbf{X}) = \mathbf{0}. \end{aligned}$$

For $\tilde{\boldsymbol{\beta}}$ to be an unbiased estimator of $\boldsymbol{\beta}$, it must be true that $\mathrm{E}(\tilde{\boldsymbol{\beta}}|\mathbf{X}) = \boldsymbol{\beta}$ for all $k \times 1$ vectors $\boldsymbol{\beta}$, that is,

$$\mathbf{A}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta} \text{ for all } k \times 1 \text{ vectors } \boldsymbol{\beta}. \tag{E.17}$$

Because $\mathbf{A}'\mathbf{X}$ is a $k \times k$ matrix, (E.17) holds if and only if $\mathbf{A}'\mathbf{X} = \mathbf{I}_k$. Equations (E.15) and (E.17) characterize the class of linear, unbiased estimators for $\boldsymbol{\beta}$.

Next, from (E.16), we have

$$\mathrm{Var}(\tilde{\boldsymbol{\beta}}|\mathbf{X}) = \mathbf{A}'[\mathrm{Var}(\mathbf{u}|\mathbf{X})]\mathbf{A} = \sigma^2\mathbf{A}'\mathbf{A},$$

by Assumption E.4. Therefore,

$$\begin{aligned} \mathrm{Var}(\tilde{\boldsymbol{\beta}}|\mathbf{X}) - \mathrm{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) &= \sigma^2[\mathbf{A}'\mathbf{A} - (\mathbf{X}'\mathbf{X})^{-1}] \\ &= \sigma^2[\mathbf{A}'\mathbf{A} - \mathbf{A}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{A}] \text{ because } \mathbf{A}'\mathbf{X} = \mathbf{I}_k \\ &= \sigma^2\mathbf{A}'[\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{A} \\ &\equiv \sigma^2\mathbf{A}'\mathbf{M}\mathbf{A}, \end{aligned}$$

where $\mathbf{M} \equiv \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Because $\mathbf{M}$ is symmetric and idempotent, $\mathbf{A}'\mathbf{M}\mathbf{A}$ is positive semi-definite for any $n \times k$ matrix $\mathbf{A}$. This establishes that the OLS estimator $\hat{\boldsymbol{\beta}}$ is BLUE. How

is this significant? Let $\boldsymbol{c}$ be any $k \times 1$ vector and consider the linear combination $\boldsymbol{c}'\boldsymbol{\beta} = c_1\beta_1 + c_2\beta_2 + \ldots + c_k\beta_k$, which is a scalar. The unbiased estimators of $\boldsymbol{c}'\boldsymbol{\beta}$ are $\boldsymbol{c}'\hat{\boldsymbol{\beta}}$ and $\boldsymbol{c}'\tilde{\boldsymbol{\beta}}$. But

$$\text{Var}(\boldsymbol{c}'\tilde{\boldsymbol{\beta}}|\mathbf{X}) - \text{Var}(\boldsymbol{c}'\hat{\boldsymbol{\beta}}|\mathbf{X}) = \boldsymbol{c}'[\text{Var}(\tilde{\boldsymbol{\beta}}|\mathbf{X}) - \text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X})]\boldsymbol{c} \geq 0,$$

because $[\text{Var}(\tilde{\boldsymbol{\beta}}|\mathbf{X}) - \text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X})]$ is p.s.d. Therefore, when it is used for estimating any linear combination of $\boldsymbol{\beta}$, OLS yields the smallest variance. In particular, $\text{Var}(\hat{\beta}_j|\mathbf{X}) \leq \text{Var}(\tilde{\beta}_j|\mathbf{X})$ for any other linear, unbiased estimator of β_j.

The unbiased estimator of the error variance σ^2 can be written as

$$\hat{\sigma}^2 = \hat{\boldsymbol{u}}'\hat{\boldsymbol{u}}/(n-k),$$

where we have labeled the explanatory variables so that there are k total parameters, including the intercept.

THEOREM E.4 (UNBIASEDNESS OF $\hat{\sigma}^2$)
Under Assumptions E.1 through E.4, $\hat{\sigma}^2$ is unbiased for σ^2: $\text{E}(\hat{\sigma}^2|\mathbf{X}) = \sigma^2$ for all $\sigma^2 > 0$.

PROOF: Write $\hat{\boldsymbol{u}} = \boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \boldsymbol{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{y} = \mathbf{M}\boldsymbol{y} = \mathbf{M}\boldsymbol{u}$, where $\mathbf{M} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, and the last equality follows because $\mathbf{MX} = \mathbf{0}$. Because $\mathbf{M}$ is symmetric and idempotent,

$$\hat{\boldsymbol{u}}'\hat{\boldsymbol{u}} = \boldsymbol{u}'\mathbf{M}'\mathbf{M}\boldsymbol{u} = \boldsymbol{u}'\mathbf{M}\boldsymbol{u}.$$

Because $\boldsymbol{u}'\mathbf{M}\boldsymbol{u}$ is a scalar, it equals its trace. Therefore,

$$\begin{aligned}\text{E}(\boldsymbol{u}'\mathbf{M}\boldsymbol{u}|\mathbf{X}) &= \text{E}[\text{tr}(\boldsymbol{u}'\mathbf{M}\boldsymbol{u})|\mathbf{X}] = \text{E}[\text{tr}(\mathbf{M}\boldsymbol{u}\boldsymbol{u}')|\mathbf{X}] \\ &= \text{tr}[\text{E}(\mathbf{M}\boldsymbol{u}\boldsymbol{u}'|\mathbf{X})] = \text{tr}[\mathbf{M}\text{E}(\boldsymbol{u}\boldsymbol{u}'|\mathbf{X})] \\ &= \text{tr}(\mathbf{M}\sigma^2\mathbf{I}_n) = \sigma^2\text{tr}(\mathbf{M}) = \sigma^2(n-k).\end{aligned}$$

The last equality follows from $\text{tr}(\mathbf{M}) = \text{tr}(\mathbf{I}_n) - \text{tr}[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] = n - \text{tr}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}] = n - \text{tr}(\mathbf{I}_k) = n - k$. Therefore,

$$\text{E}(\hat{\sigma}^2|\mathbf{X}) = \text{E}(\boldsymbol{u}'\mathbf{M}\boldsymbol{u}|\mathbf{X})/(n-k) = \sigma^2.$$

E.3 STATISTICAL INFERENCE

When we add the final classical linear model assumption, $\hat{\boldsymbol{\beta}}$ has a multivariate normal distribution, which leads to the t and F distributions for the standard test statistics covered in Chapter 4.

ASSUMPTION E.5 (NORMALITY OF ERRORS)
Conditional on $\mathbf{X}$, the u_t are independent and identically distributed as Normal$(0,\sigma^2)$. Equivalently, $\boldsymbol{u}$ given $\mathbf{X}$ is distributed as multivariate normal with mean zero and variance-covariance matrix $\sigma^2\mathbf{I}_n$: $\boldsymbol{u} \sim \text{Normal}(\mathbf{0},\sigma^2\mathbf{I}_n)$.

Under Assumption E.5, each u_t is independent of the explanatory variables for all t. In a time series setting, this is essentially the strict exogeneity assumption.

THEOREM E.5 (NORMALITY OF $\hat{\boldsymbol{\beta}}$)
Under the classical linear model Assumptions E.1 through E.5, $\hat{\boldsymbol{\beta}}$ conditional on $\mathbf{X}$ is distributed as multivariate normal with mean $\boldsymbol{\beta}$ and variance-covariance matrix $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

Theorem E.5 is the basis for statistical inference involving $\boldsymbol{\beta}$. In fact, along with the properties of the chi-square, t, and F distributions that we summarized in Appendix D, we can use Theorem E.5 to establish that t statistics have a t distribution under Assumptions E.1 through E.5 (under the null hypothesis) and likewise for F statistics. We illustrate with a proof for the t statistics.

THEOREM E.6
Under Assumptions E.1 through E.5,

$$(\hat{\beta}_j - \beta_j)/\text{se}(\hat{\beta}_j) \sim t_{n-k}, j = 1,2,\ldots,k.$$

PROOF: The proof requires several steps; the following statements are initially conditional on $\mathbf{X}$. First, by Theorem E.5, $(\hat{\beta}_j - \beta_j)/\text{sd}(\hat{\beta}_j) \sim \text{Normal}(0,1)$, where $\text{sd}(\hat{\beta}_j) = \sigma\sqrt{c_{jj}}$, and c_{jj} is the j^{th} diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$. Next, under Assumptions E.1 through E.5, conditional on $\mathbf{X}$,

$$(n - k)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-k}. \tag{E.18}$$

This follows because $(n - k)\hat{\sigma}^2/\sigma^2 = (\boldsymbol{u}/\sigma)'\mathbf{M}(\boldsymbol{u}/\sigma)$, where $\mathbf{M}$ is the $n \times n$ symmetric, idempotent matrix defined in Theorem E.4. But $\boldsymbol{u}/\sigma \sim \text{Normal}(\mathbf{0},\mathbf{I}_n)$ by Assumption E.5. It follows from Property 1 for the chi-square distribution in Appendix D that $(\boldsymbol{u}/\sigma)'\mathbf{M}(\boldsymbol{u}/\sigma) \sim \chi^2_{n-k}$ (because $\mathbf{M}$ has rank $n - k$).

We also need to show that $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are independent. But $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{u}$, and $\hat{\sigma}^2 = \boldsymbol{u}'\mathbf{M}\boldsymbol{u}/(n - k)$. Now, $[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{M} = \mathbf{0}$ because $\mathbf{X}'\mathbf{M} = \mathbf{0}$. It follows, from Property 5 of the multivariate normal distribution in Appendix D, that $\hat{\boldsymbol{\beta}}$ and $\mathbf{M}\boldsymbol{u}$ are independent. Since $\hat{\sigma}^2$ is a function of $\mathbf{M}\boldsymbol{u}$, $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are also independent.

Finally, we can write

$$(\hat{\beta}_j - \beta_j)/\text{se}(\hat{\beta}_j) = [(\hat{\beta}_j - \beta_j)/\text{sd}(\hat{\beta}_j)]/(\hat{\sigma}^2/\sigma^2)^{1/2},$$

which is the ratio of a standard normal random variable and the square root of a $\chi^2_{n-k}/(n - k)$ random variable. We just showed that these are independent, and so, by definition of a t random variable, $(\hat{\beta}_j - \beta_j)/\text{se}(\hat{\beta}_j)$ has the t_{n-k} distribution. Because this distribution does not depend on $\mathbf{X}$, it is the unconditional distribution of $(\hat{\beta}_j - \beta_j)/\text{se}(\hat{\beta}_j)$ as well.

From this theorem, we can plug in any hypothesized value for β_j and use the t statistic for testing hypotheses, as usual.

Under Assumptions E.1 through E.5, we can compute what is known as the *Cramer-Rao* lower bound for the variance-covariance matrix of unbiased estimators of $\boldsymbol{\beta}$ (again

conditional on $\mathbf{X}$) [see Greene (1997, Chapter 4)]. This can be shown to be $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, which is exactly the variance-covariance matrix of the OLS estimator. This implies that $\hat{\boldsymbol{\beta}}$ is the **minimum variance unbiased** estimator of $\boldsymbol{\beta}$ (conditional on $\mathbf{X}$): $\text{Var}(\tilde{\boldsymbol{\beta}}|\mathbf{X}) - \text{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X})$ is positive semi-definite for any other unbiased estimator $\tilde{\boldsymbol{\beta}}$; we no longer have to restrict our attention to estimators linear in $\boldsymbol{y}$.

It is easy to show that the OLS estimator is in fact the maximum likelihood estimator of $\boldsymbol{\beta}$ under Assumption E.5. For each t, the distribution of y_t given $\mathbf{X}$ is $\text{Normal}(\boldsymbol{x}_t\boldsymbol{\beta},\sigma^2)$. Because the y_t are independent conditional on $\mathbf{X}$, the likelihood function for the sample is obtained from the product of the densities:

$$\prod_{t=1}^{n} (2\pi\sigma^2)^{-1/2}\exp[-(y_t - \boldsymbol{x}_t\boldsymbol{\beta})^2/(2\sigma^2)].$$

Maximizing this function with respect to $\boldsymbol{\beta}$ and σ^2 is the same as maximizing its natural logarithm:

$$\sum_{t=1}^{n} [-(1/2)\log(2\pi\sigma^2) - (y_t - \boldsymbol{x}_t\boldsymbol{\beta})^2/(2\sigma^2)].$$

For obtaining $\hat{\boldsymbol{\beta}}$, this is the same as minimizing $\sum_{t=1}^{n} (y_t - \boldsymbol{x}_t\boldsymbol{\beta})^2$—the division by $2\sigma^2$ does not affect the optimization—which is just the problem that OLS solves. The estimator of σ^2 that we have used, $\text{SSR}/(n - k)$, turns out not to be the MLE of σ^2; the MLE is SSR/n, which is a biased estimator. Because the unbiased estimator of σ^2 results in t and F statistics with exact t and F distributions under the null, it is always used instead of the MLE.

SUMMARY

This appendix has provided a brief discussion of the linear regression model using matrix notation. This material is included for more advanced classes that use matrix algebra, but it is not needed to read the text. In effect, this appendix proves some of the results that we either stated without proof, proved only in special cases, or proved through a more cumbersome method of proof. Other topics—such as asymptotic properties, instrumental variables estimation, and panel data models—can be given concise treatments using matrices. Advanced texts in econometrics, including Davidson and MacKinnon (1993), Greene (1997), and Wooldridge (1999), can be consulted for details.

KEY TERMS

First Order Condition
Matrix Notation
Minimum Variance Unbiased
Scalar Variance-Covariance Matrix
Variance-Covariance Matrix of the OLS Estimator

PROBLEMS

E.1 Let $\boldsymbol{x}_t$ be the $1 \times k$ vector of explanatory variables for observation t. Show that the OLS estimator $\hat{\boldsymbol{\beta}}$ can be written as

$$\hat{\boldsymbol{\beta}} = \left(\sum_{t=1}^{n} \boldsymbol{x}_t' \boldsymbol{x}_t\right)^{-1} \left(\sum_{t=1}^{n} \boldsymbol{x}_t' y_t\right).$$

Dividing each summation by n shows that $\hat{\boldsymbol{\beta}}$ is a function of sample averages.

E.2 Let $\hat{\boldsymbol{\beta}}$ be the $k \times 1$ vector of OLS estimates.

(i) Show that for any $k \times 1$ vector $\boldsymbol{b}$, we can write the sum of squared residuals as

$$\text{SSR}(\boldsymbol{b}) = \hat{\boldsymbol{u}}'\hat{\boldsymbol{u}} + (\hat{\boldsymbol{\beta}} - \boldsymbol{b})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{b}).$$

[*Hint*: Write $(\boldsymbol{y} - \mathbf{X}\boldsymbol{b})'(\boldsymbol{y} - \mathbf{X}\boldsymbol{b}) = [\hat{\boldsymbol{u}} + \mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{b})]'[\hat{\boldsymbol{u}} + \mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{b})]$ and use the fact that $\mathbf{X}'\hat{\boldsymbol{u}} = \mathbf{0}$.]

(ii) Explain how the expression for SSR($\boldsymbol{b}$) in part (i) proves that $\hat{\boldsymbol{\beta}}$ uniquely minimizes SSR($\boldsymbol{b}$) over all possible values of $\boldsymbol{b}$, assuming $\mathbf{X}$ has rank k.

E.3 Let $\hat{\boldsymbol{\beta}}$ be the OLS estimate from the regression of $\boldsymbol{y}$ on $\mathbf{X}$. Let $\mathbf{A}$ be a $k \times k$ nonsingular matrix and define $\boldsymbol{z}_t \equiv \boldsymbol{x}_t\mathbf{A}$, $t = 1, \ldots, n$. Therefore, $\boldsymbol{z}_t$ is $1 \times k$ and is a nonsingular linear combination of $\boldsymbol{x}_t$. Let $\mathbf{Z}$ be the $n \times k$ matrix with rows $\boldsymbol{z}_t$. Let $\tilde{\boldsymbol{\beta}}$ denote the OLS estimate from a regression of $\boldsymbol{y}$ on $\mathbf{Z}$.

(i) Show that $\tilde{\boldsymbol{\beta}} = \mathbf{A}^{-1}\hat{\boldsymbol{\beta}}$.

(ii) Let $\hat{y}_t$ be the fitted values from the original regression and let $\tilde{y}_t$ be the fitted values from regressing $\boldsymbol{y}$ on $\mathbf{Z}$. Show that $\tilde{y}_t = \hat{y}_t$, for all $t = 1,2, \ldots, n$. How do the residuals from the two regressions compare?

(iii) Show that the estimated variance matrix for $\tilde{\boldsymbol{\beta}}$ is $\hat{\sigma}^2\mathbf{A}^{-1}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}^{-1'}$, where $\hat{\sigma}^2$ is the usual variance estimate from regressing $\boldsymbol{y}$ on $\mathbf{X}$.

(iv) Let the $\hat{\beta}_j$ be the OLS estimates from regressing y_t on $1, x_{t2}, \ldots, x_{tk}$, and let the $\tilde{\beta}_j$ be the OLS estimates from the regression of y_t on $1, a_2x_{t2}, \ldots, a_kx_{tk}$, where $a_j \neq 0$, $j = 2, \ldots, k$. Use the results from part (i) to find the relationship between the $\tilde{\beta}_j$ and the $\hat{\beta}_j$.

(v) Assuming the setup of part (iv), use part (iii) to show that $\text{se}(\tilde{\beta}_j) = \text{se}(\hat{\beta}_j)/|a_j|$.

(vi) Assuming the setup of part (iv), show that the absolute values of the t statistics for $\tilde{\beta}_j$ and $\hat{\beta}_j$ are identical.

Appendix F

Answers to Chapter Questions

CHAPTER 2

QUESTION 2.1

When student ability, motivation, age, and other factors in u are not related to attendance, (2.6) would hold. This seems unlikely to be the case.

QUESTION 2.2

About \$9.64. To see this, from the average wages measured in 1976 and 1997 dollars, we can get the CPI deflator as $16.64/5.90 \approx 2.82$. When we multiply 3.42 by 2.82, we obtain about 9.64.

QUESTION 2.3

59.26, as can be seen by plugging $shareA = 60$ into equation (2.28). This is not unreasonable: if Candidate A spends 60% of the total money spent, he or she is predicted to receive just over 59% of the vote.

QUESTION 2.4

The equation will be $\widehat{salaryhun} = 9{,}631.91 + 185.01\ roe$, as is easily seen by multiplying equation (2.39) by 10.

QUESTION 2.5

Equation (2.58) can be written as $\text{Var}(\hat{\beta}_0) = (\sigma^2 n^{-1}) \left(\sum_{i=1}^{n} x_i^2 \right) \Big/ \left(\sum_{i=1}^{n} (x_i - \bar{x})^2 \right)$, where the term multiplying $\sigma^2 n^{-1}$ is greater than or equal to one, but it is equal to one if and only if $\bar{x} = 0$. In this case, the variance is as small as it can possibly be: $\text{Var}(\hat{\beta}_0) = \sigma^2/n$.

CHAPTER 3

QUESTION 3.1

Just a few factors include age and gender distribution, size of the police force (or, more generally, resources devoted to crime fighting), population, and general historical factors. These factors certainly might be correlated with *prbconv* and *avgsen*, which means (3.5) would not hold. For example, size of the police force is possibly correlated with

both *prbcon* and *avgsen*, as some cities put more effort into crime prevention and enforcement. We should try to bring as many of these factors into the equation as possible.

QUESTION 3.2

We use the third property of OLS concerning predicted values and residuals: when we plug the average values of all independent variables into the OLS regression line, we obtain the average value of the dependent variable. So $\overline{colGPA} = 1.29 + .453\,\overline{hsGPA} + .0094\,\overline{ACT} = 1.29 + .453(3.4) + .0094(24.2) \approx 3.06$. You can check the average of *colGPA* in GPA1.RAW to verify this to the second decimal place.

QUESTION 3.3

No. The variable *shareA* is not an exact linear function of *expendA* and *expendB*, even though it is an exact *nonlinear* function: $shareA = 100 \cdot [expendA/(expendA + expendB)]$. Therefore, it is legitimate to have *expendA*, *expendB*, and *shareA* as explanatory variables.

QUESTION 3.4

As we discussed in Section 3.4, if we are interested in the effect of x_1 on y, correlation among the other explanatory variables (x_2, x_3, and so on) does not affect $\text{Var}(\hat{\beta}_1)$. These variables are included as controls, and we do not have to worry about this kind of collinearity. Of course, we are controlling for them primarily because we think they are correlated with attendance, but this is necessary to perform a ceteris paribus analysis.

CHAPTER 4

QUESTION 4.1

Under these assumptions, the Gauss-Markov assumptions are satisfied: u is independent of the explanatory variables, so $\text{E}(u|x_1, \ldots, x_k) = \text{E}(u)$, and $\text{Var}(u|x_1, \ldots, x_k) = \text{Var}(u)$. Further, it is easily seen that $\text{E}(u) = 0$. Therefore, MLR.3 and MLR.5 hold. The classical linear model assumptions are not satisfied, because u is not normally distributed (which is a violation of MLR.6).

QUESTION 4.2

H_0: $\beta_1 = 0$, H_1: $\beta_1 < 0$.

QUESTION 4.3

Because $\hat{\beta}_1 = .56 > 0$ and we are testing against H_1: $\beta_1 > 0$, the one-sided p-value is one-half of the two-sided p-value, or .043.

QUESTION 4.4

H_0: $\beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$. $k = 8$ and $q = 4$. The restricted version of the model is

$$score = \beta_0 + \beta_1 classize + \beta_2 expend + \beta_3 tchcomp + \beta_4 enroll + u.$$

QUESTION 4.5

The F statistic for testing exclusion of *ACT* is $[(.291 - .183)/(1 - .291)](680 - 3) \approx 103.13$. Therefore, the absolute value of the t statistic is about 10.16. The t statistic on *ACT* is negative, because $\hat{\beta}_{ACT}$ is negative, so $t_{ACT} = -10.16$.

QUESTION 4.6

Not by much. The F test for joint significance of *droprate* and *gradrate* is easily computed from the R-squareds in the table: $F = [(.361 - .353)/(1 - .361)](402/2) \approx 2.52$. The 10% critical value is obtained from Table G.3(a) as 2.30, while the 5% critical value from Table G.3(b) is 3. The p-value is about .082. Thus, *droprate* and *gradrate* are jointly significant at the 10% level, but not at the 5% level. In any case, controlling for these variables has a minor effect on the *b*/*s* coefficient.

CHAPTER 5

QUESTION 5.1

This requires some assumptions. It seems reasonable to assume that $\beta_2 > 0$ (*score* depends positively on *priGPA*) and $\text{Cov}(skipped, priGPA) < 0$ (*skipped* and *priGPA* are negatively correlated). This means that $\beta_2\delta_1 < 0$, which means that plim $\tilde{\beta}_1 < \beta_1$. Because β_1 is thought to be negative (or at least nonpositive), a simple regression is likely to overestimate the importance of skipping classes.

QUESTION 5.2

$\hat{\beta}_j \pm 1.96\text{se}(\hat{\beta}_j)$ is the asymptotic 95% confidence interval. Or, we can replace 1.96 with 2.

CHAPTER 6

QUESTION 6.1

Because $fincdol = 1{,}000 \cdot faminc$, the coefficient on *fincdol* will be the coefficient on *faminc* divided by 1,000, or .0927/1,000 = .0000927. The standard error also drops by a factor of 1,000, and so the t statistic does not change, nor do any of the other OLS statistics. For readability, it is better to measure family income in thousands of dollars.

QUESTION 6.2

We can do this generally. The equation is

$$\log(y) = \beta_0 + \beta_1\log(x_1) + \beta_2 x_2 + \ldots,$$

where x_2 is a proportion rather than a percentage. Then, ceteris paribus, $\Delta\log(y) = \beta_2\Delta x_2$, $100 \cdot \Delta\log(y) = \beta_2(100 \cdot \Delta x_2)$, or $\%\Delta y \approx \beta_2(100 \cdot \Delta x_2)$. Now, because Δx_2 is the change in the proportion, $100 \cdot \Delta x_2$ is a percentage point change. In particular, if $\Delta x_2 = .01$, then $100 \cdot \Delta x_2 = 1$, which corresponds to a one percentage point change. But then β_2 is the percentage change in y when $100 \cdot \Delta x_2 = 1$.

QUESTION 6.3

The new model would be $stndfnl = \beta_0 + \beta_1 atndrte + \beta_2 priGPA + \beta_3 ACT + \beta_4 priGPA^2 + \beta_5 ACT^2 + \beta_6 priGPA \cdot atndrte + \beta_7 ACT \cdot atndrte + u$. Therefore, the partial effect of *atndrte* on *stndfnl* is $\beta_1 + \beta_6 priGPA + \beta_7 ACT$. This is what we multiply by $\Delta atndrte$ to obtain the ceteris paribus change in *stndfnl*.

QUESTION 6.4

From equation (6.21), $\bar{R}^2 = 1 - \hat{\sigma}^2/[\text{SST}/(n-1)]$. For a given sample and a given dependent variable, $\text{SST}/(n-1)$ is fixed. When we use different sets of explanatory variables, only $\hat{\sigma}^2$ changes. As $\hat{\sigma}^2$ decreases, $\bar{R}^2$ increases. If we make $\hat{\sigma}$, and therefore $\hat{\sigma}^2$, as small as possible, we are making $\bar{R}^2$ as large as possible.

QUESTION 6.5

One possibility is to collect data on annual earnings for a sample of actors, along with profitability of the movies in which they each appeared. In a simple regression analysis, we could relate earnings to profitability. But we should probably control for other factors that may affect salary, such as age, gender, and the kinds of movies in which the actors performed. Methods for including qualitative factors in regression models are considered in Chapter 7.

CHAPTER 7

QUESTION 7.1

No, because it would not be clear when *party* is one and when it is zero. A better name would be something like *Dem*, which is one for Democratic candidates, and zero for Republicans. Or, *Rep*, which is one for Republicans, and zero for Democrats.

QUESTION 7.2

With *outfield* as the base group, we would include the dummy variables *frstbase*, *scndbase*, *thrdbase*, *shrtstop*, and *catcher*.

QUESTION 7.3

The null in this case is H_0: $\delta_1 = \delta_2 = \delta_3 = \delta_4 = 0$, so that there are four restrictions. As usual, we would use an F test (where $q = 4$ and k depends on the number of other explanatory variables).

QUESTION 7.4

Because *tenure* appears as a quadratic, we should allow separate quadratics for men and women. That is, we would add the explanatory variables $female \cdot tenure$ and $female \cdot tenure^2$.

QUESTION 7.5

We plug $pcnv = 0$, $avgsen = 0$, $tottime = 0$, $ptime86 = 0$, $qemp86 = 0$, $black = 1$, and $hispan = 0$ into (7.31): $\widehat{arr86} = .380 - .038(4) + .170 = .398$, or almost .4. It is hard to know whether this is "reasonable." For someone with no prior convictions who was

employed throughout the year, this estimate might seem high, but remember that the population consists of men who were already arrested at least once prior to 1986.

CHAPTER 8

QUESTION 8.1

This statement is clearly false. For example, in equation (8.7), the usual standard error for *black* is .147, while the heteroskedasticity-robust standard error is .118.

QUESTION 8.2

The F test would be obtained by regressing $\hat{u}^2$ on *marrmale*, *marrfem*, and *singfem* (*singmale* is the base group). With $n = 526$ and three independent variables in this regression, the *df* are 3 and 522.

QUESTION 8.3

Not really. Because this is a simple regression model, heteroskedasticity only matters if it is related to *inc*. But the Breusch-Pagan test in this case is equivalent to a t statistic in regressing $\hat{u}^2$ on *inc*. A t statistic of .96 is not large enough to reject the homoskedasticity assumption.

QUESTION 8.4

We can use weighted least squares but compute the heteroskedasticity-robust standard errors. In equation (8.26), if our variance model is incorrect, we still have heteroskedasticity. Thus, we can make a guess at the form of heteroskedasticity and perform WLS, but our analysis can be made robust to incorrect forms of heteroskedasticity. Unfortunately, we probably have to explicity obtain the transformed variables.

CHAPTER 9

QUESTION 9.1

These are binary variables, and squaring them has no effect: $black^2 = black$, and $hispan^2 = hispan$.

QUESTION 9.2

When $educ \cdot IQ$ is in the equation, the coefficient on *educ*, say β_1, measures the effect of *educ* on log(*wage*) when $IQ = 0$. (The partial effect of education is $\beta_1 + \beta_9 IQ$.) There is no one in the population of interest with an IQ close to zero. At the average population IQ, which is 100, the estimated return to education from column (3) is .018 + .00034(100) = .052, which is almost what we obtain as the coefficient on *educ* in column (2).

QUESTION 9.3

No. If *educ** is an integer—which means someone has no education past the previous grade completed—the measurement error is zero. If *educ** is not an integer, *educ* <

*educ**, and so the measurement error is negative. At a minimum, e_1 cannot have zero mean, and e_1 and *educ** are probably correlated.

QUESTION 9.4

An incumbent's decision not to run may be systematically related to how he or she expects to do in the election. Therefore, we may only have a sample of incumbents who are stronger, on average, than all possible incumbents who could run. This results in a sample selection problem if the population of interest includes all incumbents. If we are only interested in the effects of campaign expenditures on election outcomes for incumbents who seek reelection, there is no sample selection problem.

CHAPTER 10

QUESTION 10.1

The impact propensity is .48, while the long-run propensity is $.48 - .15 + .32 = .65$.

QUESTION 10.2

The explanatory variables are $x_{t1} = z_t$ and $x_{t2} = z_{t-1}$. The absence of perfect collinearity means that these cannot be constant, and there cannot be an exact linear relationship between them in the sample. This rules out the possibility that all the $z_1, \ldots, z_n$ take on the same value or that the $z_0, z_1, \ldots, z_{n-1}$ take on the same value. But it eliminates other patterns as well. For example, if $z_t = a + bt$ for constants a and b, then $z_{t-1} = a + b(t - 1) = (a + bt) - b = z_t - b$, which is a perfect linear function of z_t.

QUESTION 10.3

If $\{z_t\}$ is slowly moving over time—as is the case for the levels or logs of many economic time series—then z_t and z_{t-1} can be highly correlated. For example, the correlation between $unem_t$ and $unem_{t-1}$ in PHILLIPS.RAW is .74.

QUESTION 10.4

No, because a linear time trend with $\alpha_1 < 0$ becomes more and more negative as t gets large. Since *gfr* cannot be negative, a linear time trend with a negative trend coefficient cannot represent *gfr* in all future time periods.

QUESTION 10.5

The intercept for March is $\beta_0 + \delta_2$. Seasonal dummy variables are strictly exogenous because they follow a deterministic pattern. For example, the months do not change based upon whether either the explanatory variables or the dependent variable change.

CHAPTER 11

QUESTION 11.1

(i) No, because $E(y_t) = \delta_0 + \delta_1 t$ depends on t. (ii) Yes, because $y_t - E(y_t) = e_t$ is an i.i.d. sequence.

QUESTION 11.2

We plug $inf_t^e = (1/2)inf_{t-1} + (1/2)inf_{t-2}$ into $inf_t - inf_t^e = \beta_1(unem_t - \mu_0) + e_t$ and rearrange: $inf_t - (1/2)(inf_{t-1} + inf_{t-2}) = \beta_0 + \beta_1 unem_t + e_t$, where $\beta_0 = -\beta_1\mu_0$, as before. Therefore, we would regress y_t on $unem_t$, where $y_t = inf_t - (1/2)(inf_{t-1} + inf_{t-2})$. Note that we lose the first two observations in constructing y_t.

QUESTION 11.3

No, because u_t and u_{t-1} are correlated. In particular, $\text{Cov}(u_t,u_{t-1}) = \text{E}[(e_t + \alpha_1 e_{t-1})(e_{t-1} + \alpha_1 e_{t-2})] = \alpha_1\text{E}(e_{t-1}^2) = \alpha_1\sigma_e^2 \neq 0$ if $\alpha_1 \neq 0$. If the errors are serially correlated, the model cannot be dynamically complete.

CHAPTER 12

QUESTION 12.1

We use equation (12.4). Now, only adjacent terms are correlated. In particular, the covariance between $x_t u_t$ and $x_{t+1}u_{t+1}$ is $x_t x_{t+1}\text{Cov}(u_t,u_{t+1}) = x_t x_{t+1}\alpha\sigma_e^2$. Therefore, the formula is

$$\begin{aligned}\text{Var}(\hat{\beta}_1) &= \text{SST}_x^{-2}\left(\sum_{t=1}^{n} x_t^2\text{Var}(u_t) + 2\sum_{t=1}^{n-1} x_t x_{t+1}\text{E}(u_t u_{t+1})\right)\\ &= \sigma^2/\text{SST}_x + (2/\text{SST}_x^2)\sum_{t=1}^{n-1}\alpha\sigma_e^2 x_t x_{t+1}\\ &= \sigma^2/\text{SST}_x + \alpha\sigma_e^2(2/\text{SST}_x^2)\sum_{t=1}^{n-1} x_t x_{t+1}\end{aligned}$$

where $\sigma^2 = \text{Var}(u_t) = \sigma_e^2 + \alpha_1^2\sigma_e^2 = \sigma_e^2(1 + \alpha_1^2)$. Unless x_t and x_{t+1} are uncorrelated in the sample, the second term is nonzero whenever $\alpha \neq 0$. Notice that if x_t and x_{t+1} are positively correlated and $\alpha < 0$, the true variance is actually *smaller* than the usual variance. When the equation is in levels (as opposed to being differenced), the typical case is $\alpha > 0$, with positive correlation between x_t and x_{t+1}.

QUESTION 12.2

$\hat{\rho} \pm 1.96\text{se}(\hat{\rho})$, where $\text{se}(\hat{\rho})$ is the standard error reported in the regression. Or, we could use the heteroskedasticity-robust standard error. Showing that this is asymptotically valid is complicated because the OLS residuals depend on $\hat{\beta}_j$, but it can be done.

QUESTION 12.3

The model we have in mind is $u_t = \rho_1 u_{t-1} + \rho_4 u_{t-4} + e_t$, and we want to test H_0: $\rho_1 = 0$, $\rho_4 = 0$ against the alternative that H_0 is false. We would run the regression of $\hat{u}_t$ on $\hat{u}_{t-1}$ and $\hat{u}_{t-4}$ to obtain the usual F statistic for joint significance of the two lags. (We are testing two restrictions.)

QUESTION 12.4

We would probably estimate the equation using first differences, as $\hat{\rho} = .92$ is close enough to one to raise questions about the levels regression. See Chapter 18 for more discussion.

QUESTION 12.5

Because there is only one explanatory variable, the White test is easy to compute. Simply regress $\hat{u}_t^2$ on $return_{t-1}$ and $return_{t-1}^2$ (with an intercept, as always) and compute the F test for joint significance of $return_{t-1}$ and $return_{t-1}^2$. If these are jointly significant at a small enough significance level, we reject the null of homoskedasticity.

CHAPTER 13

QUESTION 13.1

Yes, assuming that we have controlled for all relevant factors. The coefficient on *black* is 1.076, and, with a standard error of .174, it is not statistically different from one. The 95% confidence interval is from about .735 to 1.417.

QUESTION 13.2

The coefficient on *highearn* shows that, in the absence of any change in the earnings cap, high earners spend much more time—on the order of 29.2% on average [because $\exp(.256) - 1 \approx .292$]—on workers' compensation.

QUESTION 13.3

First, $E(v_{i1}) = E(a_i + u_{i1}) = E(a_i) + E(v_{i1}) = 0$. Similarly, $E(v_{i2}) = 0$. Therefore, the covariance between v_{i1} and v_{i2} is simply $E(v_{i1}v_{i2}) = E[(a_i + u_{i1})(a_i + u_{i2})] = E(a_i^2) + E(a_iu_{i1}) + E(a_iu_{i2}) + E(u_{i1}u_{i2}) = E(a_i^2)$, because all of the covariance terms are zero by assumption. But $E(a_i^2) = Var(a_i)$, because $E(a_i) = 0$. This causes positive serial correlation across time in the errors within each i, which biases the usual OLS standard errors in a pooled cross-sectional regression.

QUESTION 13.4

Because $\Delta admn = admn_{90} - admn_{85}$ is the difference in binary indicators, it can be -1 if and only if $admn_{90} = 0$ and $admn_{85} = 1$. In other words, Washington state had an administrative per se law in 1985 but it was repealed by 1990.

QUESTION 13.5

No, just as it does not cause bias and inconsistency in a time series regression with strictly exogenous explanatory variables. There are two reasons it is a concern. First, serial correlation in the errors in any equation generally biases the usual OLS standard errors and test statistics. Second, it means that pooled OLS is not as efficient as estimators that account for the serial correlation (as in Chapter 12).

CHAPTER 14

QUESTION 14.1

Whether we use first differencing or the within transformation, we will have trouble estimating the coefficient on $kids_{it}$. For example, using the within transformation, if $kids_{it}$ does not vary for family i, then $\ddot{kids}_{it} = kids_{it} - \overline{kids}_i = 0$ for $t = 1,2,3$. As long as some families have variation in $kids_{it}$, then we can compute the fixed effects estima-

tor, but the kids coefficient could be very imprecisely estimated. This is a form of multicollinearity in fixed effects estimation (or first-differencing estimation).

QUESTION 14.2

If a firm did not receive a grant in the first year, it may or may not receive a grant in the second year. But if a firm did receive a grant in the first year, it could not get a grant in the second year. That is, if $grant_{-1} = 1$, then $grant = 0$. This induces a negative correlation between *grant* and $grant_{-1}$. We can verify this by computing a regression of *grant* on $grant_{-1}$, using the data in JTRAIN.RAW for 1989. Using all firms in the sample, we get

$$\widehat{grant} = \underset{(.035)}{.248} - \underset{(.072)}{.248}\ grant_{-1}.$$
$$n = 157,\ R^2 = .070.$$

The coefficient on $grant_{-1}$ must be the negative of the intercept, because $\widehat{grant} = 0$ when $grant_{-1} = 1$.

QUESTION 14.3

It suggests that the unobserved effect a_i is positively correlated with $union_{it}$. Remember, pooled OLS leaves a_i in the error term, while fixed effects removes a_i. By definition, a_i has a positive effect on log(*wage*). By the standard omitted variables analysis (see Chapter 3), OLS has an upward bias when the explanatory variable (*union*) is positively correlated with the omitted variable (a_i). Thus, belonging to a union appears to be positively related to time-constant, unobserved factors that affect wage.

QUESTION 14.4

Not if all sisters within a family have the same mother and father. Then, because the parents' race variables would not change by sister, they would be differenced away in (14.13).

CHAPTER 15

QUESTION 15.1

Probably not. In the simple equation (15.18), years of education is part of the error term. If some men who were assigned low draft lottery numbers obtained additional schooling, then lottery number and education are negatively correlated, which violates the first requirement for an instrumental variable in equation (15.4).

QUESTION 15.2

(i) For (15.27), we require that high school peer group effects carry over to college. Namely, for a given SAT score, a student who went to a high school where smoking marijuana was more popular would smoke more marijuana in college. Even if the identification condition (15.27) holds, the link might be weak.

(ii) We have to assume that percent of students using marijuana at a student's high school is not correlated with unobserved factors that affect college grade point average.

While we are somewhat controlling for high school quality by including *SAT* in the equation, this might not be enough. Perhaps high schools that did a better job of preparing students for college also had fewer students smoking marijuana. Or, marijuana usage could be correlated with average income levels. These are, of course, empirical questions that we may or may not be able to answer.

QUESTION 15.3

While prevalence of the NRA and subscribers to gun magazines are probably correlated with the presence of gun control legislation, it is not obvious that they are uncorrelated with unobserved factors that affect the violent crime rate. In fact, we might argue that a population interested in guns is a reflection of high crime rates, and controlling for economic and demographic variables is not sufficient to capture this. It would be hard to argue persuasively that these are truly exogenous in the violent crime equation.

QUESTION 15.4

As usual, there are two requirements. First, it should be the case that growth in government spending is systematically related to the party of the president, after netting out the investment rate and growth in the labor force. In other words, the instrument must be partially correlated with the endogenous explanatory variable. While we might think that government spending grows more slowly under Republican presidents, this certainly has not always been true in the United States and would have to be tested using the t statistic on REP_{t-1} in the reduced form $gGOV_t = \pi_0 + \pi_1 REP_{t-1} + \pi_2 INVRAT_t + \pi_3 gLAB_t + v_t$. We must assume that the party of the president has no separate effect on *gGDP*. This would be violated if, for example, monetary policy differs systematically by presidential party and has a separate effect on GDP growth.

CHAPTER 16

QUESTION 16.1

Probably not. It is because firms choose price and advertising expenditures jointly that we are not interested in the experiment where, say, advertising changes exogenously and we want to know the effect on price. Instead, we would model price and advertising each as a function of demand and cost variables. This is what falls out of the economic theory.

QUESTION 16.2

We must assume two things. First, money supply growth should appear in equation (16.22), so that it is partially correlated with *inf*. Second, we must assume that money supply growth does not appear in equation (16.23). If we think we must include money supply growth in equation (16.23), then we are still short an instrument for *inf*. Of course, the assumption that money supply growth is exogenous can also be questioned.

QUESTION 16.3

Use the Hausman test from Chapter 15. In particular, let $\hat{v}_2$ be the OLS residuals from the reduced form regression of *open* on log(*pcinc*) and log(*land*). Then, use an OLS

regression of *inf* on *open*, log(*pcinc*), and $\hat{v}_2$ and compute the *t* statistic for significance of $\hat{v}_2$. If $\hat{v}_2$ is significant, the 2SLS and OLS estimates are statistically different.

QUESTION 16.4

The demand equation looks like

$$\log(fish_t) = \beta_0 + \beta_1 \log(prcfish_t) + \beta_2 \log(inc_t) + \beta_3 \log(prcchick_t) + \beta_4 \log(prcbeef_t) + u_{t1},$$

where logarithms are used so that all elasticities are constant. By assumption, the demand function contains no seasonality, so the equation does not contain monthly dummy variables (say feb_t, $mar_t, \ldots, dec_t$, with January as the base month). Also, by assumption, the supply of fish is seasonal, which means that the supply function does depend on at least some of the monthly dummy variables. Even without solving the reduced form for log(*prcfish*), we conclude that it depends on the monthly dummy variables. Since these are exogenous, they can be used as instruments for log(*prcfish*) in the demand equation. Therefore, we can estimate the demand-for-fish equation using monthly dummies as the IVs for log(*prcfish*). Identification requires that at least one monthly dummy variable appears with a nonzero coefficient in the reduced form for log(*prcfish*).

CHAPTER 17

QUESTION 17.1

H_0: $\beta_4 = \beta_5 = \beta_6 = 0$, so that there are three restrictions and therefore three *df* in the *LR* or Wald test.

QUESTION 17.2

We need the partial derivative of $\Phi(\hat{\beta}_0 + \hat{\beta}_1 nwifeinc + \hat{\beta}_2 educ + \hat{\beta}_3 exper + \hat{\beta}_4 exper^2 + \ldots)$ with respect to *exper*, which is $\phi(\cdot)(\hat{\beta}_3 + 2\hat{\beta}_4 exper)$, where $\phi(\cdot)$ is evaluated at the given values and the initial level of experience. Therefore, we need to evaluate the standard normal probability density at $.270 - .012(20.13) + .131(12.3) + .123(10) - .0019(10^2) - .053(42.5) - .868(0) + .036(1) \approx .463$, where we plug in the initial level of experience (10). But $\phi(.463) = (2\pi)^{-1/2}\exp[-(.463^2)/2] \approx .358$. Next, we multiply this by $\hat{\beta}_3 + 2\hat{\beta}_4 exper$, which is evaluated at $exper = 10$. The partial effect using the calculus approximation is $.358[.123 - 2(.0019)(10)] \approx .030$. In other words, at the given values of the explanatory variables and starting at $exper = 10$, the next year of experience increases the probability of labor force participation by about .03.

QUESTION 17.3

No. The number of extramarital affairs is a nonnegative integer, which presumably takes on zero or small numbers for a substantial fraction of the population. It is not realistic to use a Tobit model, which, while allowing a pileup at zero, treats *y* as being continuously distributed over positive values. Formally, assuming that $y = \max(0,y^*)$, where y^* is normally distributed, is at odds with the discreteness of the number of extramarital affairs when $y > 0$.

QUESTION 17.4

The adjusted standard errors are the usual Poisson MLE standard errors multiplied by $\hat{\sigma} = \sqrt{2} \approx 1.41$, so the adjusted standard errors will be about 41% higher. The quasi-*LR* statistic is the usual *LR* statistic divided by $\hat{\sigma}^2$, so it will be one-half of the usual *LR* statistic.

QUESTION 17.5

By assumption, $mvp_i = \beta_0 + \boldsymbol{x}_i\boldsymbol{\beta} + u_i$, where, as usual, $\boldsymbol{x}_i\boldsymbol{\beta}$ denotes a linear function of the exogenous variables. Now, observed wage is the largest of the minimum wage and the marginal value product, so $wage_i = \max(minwage_i, mvp_i)$, which is very similar to equation (17.34), except that the max operator has replaced the min operator.

CHAPTER 18

QUESTION 18.1

We can plug these values directly into equation (18.1) and take expectations. First, because $z_s = 0$, for all $s < 0$, $y_{-1} = \alpha + u_{-1}$. Then, $z_0 = 1$, so $y_0 = \alpha + \delta_0 + u_0$. For $h \geq 1$, $y_h = \alpha + \delta_{h-1} + \delta_h + u_h$. Because the errors have zero expected values, $\mathrm{E}(y_{-1}) = \alpha$, $\mathrm{E}(y_0) = \alpha + \delta_0$, and $\mathrm{E}(y_h) = \alpha + \delta_{h-1} + \delta_h$, for all $h \geq 1$. As $h \to \infty$, $\delta_h \to 0$. It follows that $\mathrm{E}(y_h) \to \alpha$ as $h \to \infty$, that is, the expected value of y_h returns to the expected value before the increase in z, at time zero. This makes sense: while the increase in z lasted for two periods, it is still a temporary increase.

QUESTION 18.2

Under the described setup, Δy_t and Δx_t are i.i.d. sequences that are independent of one another. In particular, Δy_t and Δx_t are uncorrelated. If $\hat{\gamma}_1$ is the slope coefficient from regressing Δy_t on Δx_t, $t = 1,2,\ldots,n$, then plim $\hat{\gamma}_1 = 0$. This is as it should be, as we are regressing one I(0) process on another I(0) process, and they are uncorrelated. We write the equation $\Delta y_t = \gamma_0 + \gamma_1 \Delta x_t + e_t$, where $\gamma_0 = \gamma_1 = 0$. Because $\{e_t\}$ is independent of $\{\Delta x_t\}$, the strict exogeneity assumption holds. Moreover, $\{e_t\}$ is serially uncorrelated and homoskedastic. By Theorem 11.2 in Chapter 11, the t statistic for $\hat{\gamma}_1$ has an approximate standard normal distribution. If e_t is normally distributed, the classical linear model assumptions hold, and the t statistic has an exact t distribution.

QUESTION 18.3

Write $x_t = x_{t-1} + a_t$, where $\{a_t\}$ is I(0). By assumption, there is a linear combination, say $s_t = y_t - \beta x_t$, which is I(0). Now, $y_t - \beta x_{t-1} = y_t - \beta(x_t - a_t) = s_t + \beta a_t$. Because s_t and a_t are I(0) by assumption, so is $s_t + \beta a_t$.

QUESTION 18.4

Just use the sum of squared residuals form of the *F* test and assume homoskedasticity. The restricted SSR is obtained by regressing $\Delta hy6_t - \Delta hy3_{t-1} + (hy6_{t-1} - hy3_{t-2})$ on a constant. Notice that α_0 is the only parameter to estimate in $\Delta hy6_t = \alpha_0 + \gamma_0 \Delta hy3_{t-1} + \delta(hy6_{t-1} - hy3_{t-2})$ when the restrictions are imposed. The unrestricted sum of squared residuals is obtained from equation (18.39).

QUESTION 18.5

We are fitting two equations: $\hat{y}_t = \hat{\alpha} + \hat{\beta}t$ and $\hat{y}_t = \hat{\gamma} + \hat{\delta}year_t$. We can obtain the relationship between the parameters by noting that $year_t = t + 49$. Plugging this into the second equation gives $\hat{y}_t = \hat{\gamma} + \hat{\delta}(t + 49) = (\hat{\gamma} + 49\hat{\delta}) + \hat{\delta}t$. Matching the slope and intercept with the first equation gives $\hat{\delta} = \hat{\beta}$—so that the slopes on t and $year_t$ are identical—and $\hat{\alpha} = \hat{\gamma} + 49\hat{\delta}$. Generally, when we use *year* rather than t, the intercept will change, but the slope will not. (You can verify this by using one of the time series data sets, such as HSEINV.RAW or INVEN.RAW.) Whether we use t or some measure of year does not change fitted values, and, naturally, it does not change forecasts of future values. The intercept simply adjusts appropriately to different ways of including a trend in the regression.

Appendix G

Statistical Tables

TABLE G.1

Cumulative Areas Under the Standard Normal Distribution

z	0	1	2	3	4	5	6	7	8	9
−3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
−2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
−2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
−2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
−2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
−2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
−2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
−2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
−2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
−2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
−2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
−1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
−1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
−1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
−1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
−1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
−1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
−1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
−1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
−1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
−1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
−0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
−0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
−0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148

continued

TABLE G.1 (*concluded*)

z	0	1	2	3	4	5	6	7	8	9
−0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
−0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
−0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
−0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
−0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
−0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
−0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Examples: If $Z \sim$ Normal(0,1) then $P(Z \leq -1.32) = .0934$ and $P(Z \leq 1.84) = .9671$.
Source: This table was generated using the Stata® function normd.

TABLE G.2

Critical Values of the *t* Distribution

		Significance Level				
1-Tailed:		**.10**	**.05**	**.025**	**.01**	**.005**
2-Tailed:		**.20**	**.10**	**.05**	**.02**	**.01**
Degrees of Freedom	1	3.078	6.314	12.706	31.821	63.657
	2	1.886	2.920	4.303	6.965	9.925
	3	1.638	2.353	3.182	4.541	5.841
	4	1.533	2.132	2.776	3.747	4.604
	5	1.476	2.015	2.571	3.365	4.032
	6	1.440	1.943	2.447	3.143	3.707
	7	1.415	1.895	2.365	2.998	3.499
	8	1.397	1.860	2.306	2.896	3.355
	9	1.383	1.833	2.262	2.821	3.250
	10	1.372	1.812	2.228	2.764	3.169
	11	1.363	1.796	2.201	2.718	3.106
	12	1.356	1.782	2.179	2.681	3.055
	13	1.350	1.771	2.160	2.650	3.012
	14	1.345	1.761	2.145	2.624	2.977
	15	1.341	1.753	2.131	2.602	2.947
	16	1.337	1.746	2.120	2.583	2.921
	17	1.333	1.740	2.110	2.567	2.898
	18	1.330	1.734	2.101	2.552	2.878
	19	1.328	1.729	2.093	2.539	2.861
	20	1.325	1.725	2.086	2.528	2.845
	21	1.323	1.721	2.080	2.518	2.831
	22	1.321	1.717	2.074	2.508	2.819
	23	1.319	1.714	2.069	2.500	2.807
	24	1.318	1.711	2.064	2.492	2.797
	25	1.316	1.708	2.060	2.485	2.787
	26	1.315	1.706	2.056	2.479	2.779
	27	1.314	1.703	2.052	2.473	2.771
	28	1.313	1.701	2.048	2.467	2.763
	29	1.311	1.699	2.045	2.462	2.756
	30	1.310	1.697	2.042	2.457	2.750
	40	1.303	1.684	2.021	2.423	2.704
	60	1.296	1.671	2.000	2.390	2.660
	90	1.291	1.662	1.987	2.368	2.632
	120	1.289	1.658	1.980	2.358	2.617
	∞	1.282	1.645	1.960	2.326	2.576

Examples: The 1% critical value for a one-tailed test with 25 *df* is 2.485. The 5% critical for a two-tailed test with large (> 120) *df* is 1.96.

Source: This table was generated using the Stata® function invt.

TABLE G.3a

10% Critical Values of the *F* Distribution

		Numerator Degrees of Freedom									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Denominator Degrees of Freedom	10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32
	11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25
	12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19
	13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14
	14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10
	15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06
	16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03
	17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00
	18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98
	19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96
	20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94
	21	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92
	22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90
	23	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89
	24	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88
	25	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87
	26	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86
	27	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85
	28	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84
	29	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83
	30	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82
	40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76
	60	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71
	90	2.76	2.36	2.15	2.01	1.91	1.84	1.78	1.74	1.70	1.67
	120	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65
	∞	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60

Example: The 10% critical value for numerator $df = 2$ and denominator $df = 40$ is 2.44.
Source: This table was generated using the Stata® function invfprob.

TABLE G.3b

5% Critical Values of the *F* Distribution

		Numerator Degrees of Freedom									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Denominator Degrees of Freedom	10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
	11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
	12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
	13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
	14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
	15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
	16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
	17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
	18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
	19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
	20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
	21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32
	22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
	23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27
	24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
	25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24
	26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
	27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20
	28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19
	29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18
	30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16
	40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08
	60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99
	90	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94
	120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91
	∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83

Example: The 5% critical value for numerator $df = 4$ and large denominator df (∞) is 2.37.
Source: This table was generated using the Stata® function invfprob.

TABLE G.3c

1% Critical Values of the *F* Distribution

		Numerator Degrees of Freedom									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Denominator Degrees of Freedom	10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85
	11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54
	12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30
	13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10
	14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94
	15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80
	16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69
	17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59
	18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51
	19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43
	20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37
	21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31
	22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26
	23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21
	24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17
	25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13
	26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09
	27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06
	28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03
	29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00
	30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98
	40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80
	60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63
	90	6.93	4.85	4.01	3.54	3.23	3.01	2.84	2.72	2.61	2.52
	120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47
	∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32

Example: The 1% critical value for numerator $df = 3$ and denominator $df = 60$ is 4.13.
Source: This table was generated using the Stata® function invfprob.

TABLE G.4

Critical Values of the Chi-Square Distribution

		Significance Level		
		.10	.05	.01
Degrees of Freedom	1	2.71	3.84	6.63
	2	4.61	5.99	9.21
	3	6.25	7.81	11.34
	4	7.78	9.49	13.28
	5	9.24	11.07	15.09
	6	10.64	12.59	16.81
	7	12.02	14.07	18.48
	8	13.36	15.51	20.09
	9	14.68	16.92	21.67
	10	15.99	18.31	23.21
	11	17.28	19.68	24.72
	12	18.55	21.03	26.22
	13	19.81	22.36	27.69
	14	21.06	23.68	29.14
	15	22.31	25.00	30.58
	16	23.54	26.30	32.00
	17	24.77	27.59	33.41
	18	25.99	28.87	34.81
	19	27.20	30.14	36.19
	20	28.41	31.41	37.57
	21	29.62	32.67	38.93
	22	30.81	33.92	40.29
	23	32.01	35.17	41.64
	24	33.20	36.42	42.98
	25	34.38	37.65	44.31
	26	35.56	38.89	45.64
	27	36.74	40.11	46.96
	28	37.92	41.34	48.28
	29	39.09	42.56	49.59
	30	40.26	43.77	50.89

Example: The 5% critical value with $df = 8$ is 15.51.
Source: This table was generated using the Stata® function invchi.

REFERENCES

Angrist, J. D. (1990), "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records," *American Economic Review* 80, 313–336.

Angrist, J. D., and A. B. Krueger (1991), "Does Compulsory School Attendance Affect Schooling and Earnings?" *Quarterly Journal of Economics* 106, 979–1014.

Ashenfelter, O., and A. B. Krueger (1994), "Estimates of the Economic Return to Schooling from a New Sample of Twins," *American Economic Review* 84, 1157–1173.

Averett, S., and S. Korenman (1996), "The Economic Reality of the Beauty Myth," *Journal of Human Resources* 31, 304–330.

Ayers, I., and S. D. Levitt (1998), "Measuring Positive Externalities from Unobservable Victim Precaution: An Empirical Analysis of Lojack," *Quarterly Journal of Economics* 108, 43–77.

Banerjee, A. , J. Dolado, J. W. Galbraith, and D. F. Hendry (1993), *Co-Integration, Error-Correction, and the Econometric Analysis of Non-Stationary Data*. Oxford: Oxford University Press.

Bartik, T. J. (1991), "The Effects of Property Taxes and Other Local Public Policies on the Intrametropolitan Pattern of Business Location," in *Industry Location and Public Policy.* Ed. H. W. Herzog and A. M. Schlottmann, 57–80. Knoxville: University of Tennessee Press.

Becker, G. S. (1968), "Crime and Punishment: An Economic Approach," *Journal of Political Economy* 76, 169–217.

Belsley, D., E. Kuh, and R. Welsch (1980), *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: Wiley.

Berk, R. A. (1990), "A Primer on Robust Regression," in *Modern Methods of Data Analysis*. Ed. J. Fox and J. S. Long, 292–324. Newbury Park, CA: Sage Publications.

Betts, J. R. (1995), "Does School Quality Matter? Evidence from the National Longitudinal Survey of Youth," *Review of Economics and Statistics* 77, 231–250.

Biddle, J. E., and D. S. Hamermesh (1990), "Sleep and the Allocation of Time," *Journal of Political Economy* 98, 922–943.

Biddle, J. E., and D. S. Hamermesh (1998), "Beauty, Productivity, and Discrimination: Lawyers' Looks and Lucre," *Journal of Labor Economics* 16, 172–201.

Blackburn, M., and S. Korenman (1994), "The Declining Marital-Status Earnings Differential," *Journal of Population Economics* 7, 247–270.

Blackburn, M., and D. Neumark (1992), "Unobserved Ability, Efficiency Wages, and Interindustry Wage Differentials," *Quarterly Journal of Economics* 107, 1421–1436.

Blömstrom, M. , R. E. Lipsey, and M. Zejan (1996), "Is Fixed Investment the Key to Economic Growth?" *Quarterly Journal of Economics* 111, 269–276.

Bollerslev, T. , R. Y. Chou, and K. F. Kroner (1992), "ARCH Modeling in Finance: A Review of the Theory and Empirical Evidence," *Journal of Econometrics* 52, 5–59.

Bollerslev, T. , R. F. Engle, and D. B. Nelson (1994), "ARCH Models," Chapter 49 in *Handbook of Econometrics*, Volume 4. Ed. R. F. Engle and D. L. McFadden, 2959–3038. Amsterdam: North-Holland.

Bound, J. , D. A. Jaeger, and R. M. Baker (1995), "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and Endogenous Explanatory Variables is Weak," *Journal of the American Statistical Association* 90, 443–450.

Breusch, T. S., and A. R. Pagan (1979), "A Simple Test for Heteroskedasticity and Random Coefficient Variation," *Econometrica* 50, 987–1007.

Cameron, A. C., and P. K. Trivedi (1998), *Regression Analysis of Count Data*. Cambridge: Cambridge University Press.

Campbell, J. Y., and N. G. Mankiw (1990), "Permanent Income, Current Income, and Consumption," *Journal of Business and Economic Statistics* 8, 265–279.

Card, D. (1995), "Using Geographic Variation in College Proximity to Estimate the Return to Schooling," in *Aspects of Labour Market Behavior: Essays in Honour of John Vanderkamp*. Ed. L. N. Christophides, E. K. Grant, and R. Swidinsky, 201–222. Toronto: University of Toronto Press.

Card, D., and A. Krueger (1992), "Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States," *Journal of Political Economy* 100, 1–40.

Castillo-Freeman, A. J., and R. B. Freeman (1992), "When the Minimum Wage Really Bites: The Effect of the U.S.-Level Minimum on Puerto Rico," in *Immigration and the Work Force*. Ed. G. J. Borjas and R. B. Freeman, 177–211. Chicago: University of Chicago Press.

Clark, K. B. (1984), "Unionization and Firm Performance: The Impact on Profits, Growth, and Productivity," *American Economic Review* 74, 893–919.

Cloninger, D. O. (1991), "Lethal Police Response as a Crime Deterrent: 57-City Study Suggests a Decrease in Certain Crimes," *American Journal of Economics and Sociology* 50, 59–69.

Cloninger, D. O., and L. C. Sartorius (1979), "Crime Rates, Clearance Rates and Enforcement Effort: The Case of Houston, Texas," *American Journal of Economics and Sociology* 38, 389–402.

Cochrane, J. H. (1997), "Where is the Market Going? Uncertain Facts and Novel Theories," *Economic Perspectives* 21, Federal Reserve Bank of Chicago, 3–37.

Cornwell, C., and W. N. Trumbull (1994), "Estimating the Economic Model of Crime Using Panel Data," *Review of Economics and Statistics* 76, 360–366.

Currie, J. (1995), *Welfare and the Well-Being of Children*. Chur, Switzerland: Harwood Academic Publishers.

Currie, J., and N. Cole (1993), "Welfare and Child Health: The Link Between AFDC Participation and Birth Weight," *American Economic Review* 83, 971–983.

Currie, J., and D. Thomas (1995), "Does Head Start Make a Difference?" *American Economic Review* 85, 341–364.

Davidson, R., and J. G. MacKinnon (1981), "Several Tests of Model Specification in the Presence of Alternative Hypotheses," *Econometrica* 49, 781–793.

Davidson, R., and J. G. MacKinnon (1993), *Estimation and Inference in Econometrics*. New York: Oxford University Press.

De Long, J. B., and L. H. Summers (1991), "Equipment Investment and Economic Growth," *Quarterly Journal of Economics* 106, 445–502.

Dickey, D. A., and W. A. Fuller (1979), "Distributions of the Estimators for Autoregressive Time Series with a Unit Root," *Journal of the American Statistical Association* 74, 427–431.

Diebold, F. X. (1998), *Elements of Forecasting*. Cincinnati, OH: South-Western.

Downes, T. A., and S. M. Greenstein (1996), "Understanding the Supply Decisions of Nonprofits: Modeling the Location of Private Schools," *Rand Journal of Economics* 27, 365–390.

Draper, N., and H. Smith (1981), *Applied Regression Analysis*. 2d ed. New York: Wiley.

Durbin, J. (1970), "Testing for Serial Correlation in Least Squares Regressions When Some of the Regressors are Lagged Dependent Variables," *Econometrica* 38, 410–421.

Durbin, J., and G. S. Watson (1950), "Testing for Serial Correlation in Least Squares Regressions I," *Biometrika* 37, 409–428.

Eicker, F. (1967), "Limit Theorems for Regressions with Unequal and Dependent Errors," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* 1, 59–82. Berkeley: University of California Press.

Eide, E. (1994), *Economics of Crime: Deterrence and the Rational Offender*. Amsterdam: North Holland.

Engle, R. F. (1982), "Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of United Kingdom Inflation," *Econometrica* 50, 987–1007.

Engle, R. F., and C. W. J. Granger (1987), "Co-integration and Error Correction: Representation, Estimation, and Testing," *Econometrica* 55, 251–276.

Evans, W. N., and R. M. Schwab (1995), "Finishing High School and Starting College: Do Catholic Schools Make a Difference?" *Quarterly Journal of Economics* 110, 941–974.

Fair, R. C. (1996), "Econometrics and Presidential Elections," *Journal of Economic Perspectives* 10, 89–102.

Friedman, B. M., and K. N. Kuttner (1992), "Money, Income, Prices, and Interest Rates," *American Economic Review* 82, 472–492.

Garen, J. E. (1994), "Executive Compensation and Principal-Agent Theory," *Journal of Political Economy* 102, 175–1199.

Geronimus, A. T., and S. Korenman (1992), "The Socioeconomic Consequences of Teen Childbearing Reconsidered," *Quarterly Journal of Economics* 107, 1187–1214.

Goldberger, A. S. (1991), *A Course in Econometrics*. Cambridge, MA: Harvard University Press.

Granger, C. W. J., and P. Newbold (1974), "Spurious Regressions in Econometrics," *Journal of Econometrics* 2, 111–120.

Greene, W. (1997), *Econometric Analysis.* 3rd edition. New York: MacMillan.

Griliches, Z. (1957), "Specification Bias in Estimates of Production Functions," *Journal of Farm Economics* 39, 8–20.

Grogger, J. (1990), "The Deterrent Effect of Capital Punishment: An Analysis of Daily Homicide Counts," *Journal of the American Statistical Association* 410, 295–303.

Grogger, J. (1991), "Certainty vs. Severity of Punishment," *Economic Inquiry* 29, 297–309.

Hall, R. J. (1988), "The Relation Between Price and Marginal Cost in U. S. Industry," *Journal of Political Economy* 96, 921–948.

Hamermesh, D. S., and J. E. Biddle (1994), "Beauty and the Labor Market," *American Economic Review* 84, 1174–1194.

Hamilton, J. D. (1994), *Time Series Analysis*. Princeton, NJ: Princeton University Press.

Hanushek, E. (1986), "The Economics of Schooling: Production and Efficiency in Public Schools," *Journal of Economic Literature*, 1141–1177.

Harvey, A. (1990), *The Econometric Analysis of Economic Time Series*. 2d ed. Cambridge, MA: MIT Press.

Hausman, J. A. (1978), "Specification Tests in Econometrics," *Econometrica* 46, 1251–1271.

Hausman, J. A., and D. A. Wise (1977), "Social Experimentation, Truncated Distributions, and Efficient Estimation," *Econometrica* 45, 319–339.

Herrnstein, R. J., and C. Murray (1994), *The Bell Curve: Intelligence and Class Structure in American Life*. New York: Free Press.

Hersch, J., and L. S. Stratton (1997), "Housework, Fixed Effects, and Wages of Married Workers," *Journal of Human Resources* 32, 285–307.

Hines, J. R. (1996), "Altered States: Taxes and the Location of Foreign Direct Investment in America," *American Economic Review* 86, 1076–1094.

Holzer, H. (1991), "The Spatial Mismatch Hypothesis: What Has the Evidence Shown?" *Urban Studies* 28, 105–122.

Holzer, H. , R. Block, M. Cheatham, and J. Knott (1993), "Are Training Subsidies Effective? The Michigan Experience," *Industrial and Labor Relations Review* 46, 625–636.

Hoxby, C. M. (1994), "Do Private Schools Provide Competition for Public Schools?" National Bureau of Economic Research Working Paper Number 4978.

Huber, P. J. (1967), "The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* 1, 221–233. Berkeley: University of California Press.

Hunter, W. C., and M. B. Walker (1996), "The Cultural Affinity Hypothesis and Mortgage Lending Decisions," *Journal of Real Estate Finance and Economics* 13, 57–70.

Hylleberg, S. (1991), *Modelling Seasonality*. Oxford: Oxford University Press.

Kane, T. J., and C. E. Rouse (1995), "Labor-Market Returns to Two- and Four-Year Colleges," *American Economic Review* 85, 600–614.

Kiel, K. A., and K. T. McClain (1995), "House Prices During Siting Decision Stages: The Case of an Incinerator from Rumor Through Operation," *Journal of Environmental Economics and Management* 28, 241–255.

Kleck, G., and E. B. Patterson (1993), "The Impact of Gun Control Ownership Levels on Violence Rates," *Journal of Quantitative Criminology* 9, 249–287.

Koenker, R. (1981), "A Note on Studentizing a Test for Heteroskedasticity," *Journal of Econometrics* 17, 107–112.

Korenman, S., and D. Neumark (1991), "Does Marriage Really Make Men More Productive?" *Journal of Human Resources* 26, 282–307.

Korenman, S., and D. Neumark (1992), "Marriage, Motherhood, and Wages," *Journal of Human Resources* 27, 233–255.

Krueger, A. B. (1993), "How Computers Have Changed the Wage Structure: Evidence from Microdata, 1984–1989," *Quarterly Journal of Economics* 108, 33–60.

Krupp, C. M., and P. S. Pollard (1996), "Market Responses to Antidumping Laws: Some Evidence from the U. S. Chemical Industry," *Canadian Journal of Economics* 29, 199–227.

Kwiatkowski, D. , P. C. B. Phillips, P. Schmidt, and Y. Shin (1992), "Testing the Null Hypothesis of

Stationarity Against the Alternative of a Unit Root: How Sure Are We that Economic Time Series Have a Unit Root?" *Journal of Econometrics* 54, 159–178.

Larsen, R. J., and M. L. Marx (1986), *An Introduction to Mathematical Statistics and Its Applications.* 2nd edition. Englewood Cliffs, NJ: Prentice-Hall.

Leamer, E. E. (1983), "Let's Take the Con Out of Econometrics," *American Economic Review* 73, 31–43.

Levine, P. B. , A. B. Trainor, and D. J. Zimmerman (1996), "The Effect of Medicaid Abortion Funding Restrictions on Abortions, Pregnancies, and Births," *Journal of Health Economics* 15, 555–578.

Levine, P. B., and D. J. Zimmerman (1995), "The Benefit of Additional High-School Math and Science Classes for Young Men and Women," *Journal of Business and Economics Statistics* 13, 137–149.

Levitt, S. D. (1994), "Using Repeat Challengers to Estimate the Effect of Campaign Spending on Election Outcomes in the U. S. House," *Journal of Political Economy* 102, 777–798.

Levitt, S. D. (1996), "The Effect of Prison Population Size on Crime Rates: Evidence from Prison Overcrowding Legislation," *Quarterly Journal of Economics* 111, 319–351.

Low, S. A., and L. R. McPheters (1983), "Wage Differentials and the Risk of Death: An Empirical Analysis," *Economic Inquiry* 21, 271–280.

Lynch, L. M. (1992), "Private Sector Training and the Earnings of Young Workers," *American Economic Review* 82, 299–312.

MacKinnon, J. G., and H. White (1985), "Some Heteroskedasticity Consistent Covariance Matrix Estimators with Improved Finite Sample Properties," *Journal of Econometrics* 29, 305–325.

Maddala, G. S. (1983), *Limited Dependent and Qualitative Variables in Econometrics.* Cambridge: Cambridge University Press.

Maloney, M. T., and R. E. McCormick (1993), "An Examination of the Role that Intercollegiate Athletic Participation Plays in Academic Achievement: Athletes' Feats in the Classroom," *Journal of Human Resources* 28, 555–570.

Mankiw, N. G. (1994), *Macroeconomics*. 2d ed. New York: Worth.

McCarthy, P. S. (1994), "Relaxed Speed Limits and Highway Safety: New Evidence From California," *Economics Letters* 46, 173–179.

McClain, K. T., and J. M. Wooldridge (1995), "A Simple Test for the Consistency of Dynamic Linear Regression in Rational Distributed Lag Models," *Economics Letters* 48, 235–240.

McCormick, R. E., and M. Tinsley (1987), "Athletics versus Academics: Evidence from SAT Scores," *Journal of Political Economy* 95, 1103–1116.

McFadden, D. L. (1974), "Conditional Logit Analysis of Qualitative Choice Analysis," in *Frontiers in Econometrics.* Ed. P. Zarembka, 105–142. New York: Academic Press.

Meyer, B. D. (1995), "Natural and Quasi-Experiments in Economics," *Journal of Business and Economic Statistics* 13, 151–161.

Meyer, B. D., W. K. Viscusi, and D. L. Durbin (1995), "Workers' Compensation and Injury Duration: Evidence from a Natural Experiment," *American Economic Review* 85, 322–340.

Mizon, G. E., and J. F. Richard (1986), "The Encompassing Principle and Its Application to Testing Nonnested Hypotheses," *Econometrica* 54, 657–678.

Mroz, T. A. (1987), "The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions," *Econometrica* 55, 765–799.

Mullahy, J., and P. R. Portney (1990), "Air Pollution, Cigarette Smoking, and the Production of Respiratory Health," *Journal of Health Economics* 9, 193–205.

Mullahy, J., and J. L. Sindelar (1994), "Do Drinkers Know When to Say When? An Empirical Analysis of Drunk Driving," *Economic Inquiry* 32, 383–394.

Netzer, D. (1992), "Differences in Reliance on User Charges by American State and Local Governments," *Public Finance Quarterly* 20, 499–511.

Neumark, D., and W. Wascher (1995), "Minimum Wage Effects on Employment and School Enrollment," *Journal of Business and Economic Statistics* 13, 199–206.

Newey, W. K., and K. D. West (1987), "A Simple, Positive Semi-Definite Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica* 55, 703–708.

Papke, L. E. (1987), "Subnational Taxation and Capital Mobility: Estimates of Tax-Price Elasticities," *National Tax Journal* 40, 191–203.

Papke, L. E. (1994), "Tax Policy and Urban Development: Evidence from the Indiana Enterprise Zone Program," *Journal of Public Economics* 54, 37–49.

Papke, L. E. (1995), "Participation in and Contributions to 401(k) Pension Plans: Evidence from Plan Data," *Journal of Human Resources* 30, 311–325.

Papke, L. E. (1999), "Are 401(k) Plans Replacing Other Employer-Provided Pensions? Evidence from Panel Data," *Journal of Human Resources*, 34, 346–368.

Park, R. (1966), "Estimation with Heteroskedastic Error Terms," *Econometrica* 34, 888.

Pavlik, E. L., and A. Belkaoui (1991), *Determinants of Executive Compensation*. New York: Quorum Books.

Peek, J. (1982), "Interest Rates, Income Taxes, and Anticipated Inflation," *American Economic Review* 72, 980–991.

Pindyck, R. S., and D. L. Rubinfeld (1992), *Microeconomics*. 2d ed. New York: MacMillan.

Ram, R. (1986), "Government Size and Economic Growth: A New Framework and Some Evidence from Cross-Section and Time-Series Data," *American Economics Review* 76, 191–203.

Ramanathan, R. (1995), *Introductory Econometrics with Applications*. 3d ed. Fort Worth: Dryden Press.

Ramey, V. (1991), "Nonconvex Costs and the Behavior of Inventories," *Journal of Political Economy* 99, 306–334.

Ramsey, J. B. (1969), "Tests for Specification Errors in Classical Linear Least-Squares Analysis, *Journal of the Royal Statistical Association*, Series B, 71, 350–371.

Romer, D. (1993), "Openness and Inflation: Theory and Evidence," *Quarterly Journal of Economics* 108, 869–903.

Rose, N. L. (1985), "The Incidence of Regulatory Rents in the Motor Carrier Industry," *Rand Journal of Economics* 16, 299–318.

Rose, N. L., and A. Shepard (1997), "Firm Diversification and CEO Compensation: Managerial Ability or Executive Entrenchment?" *Rand Journal of Economics* 28, 489–514.

Rouse, C. E. (1998), "Private School Vouchers and Student Achievement: An Evaluation of the Milwaukee Parental Choice Program," *Quarterly Journal of Economics* 113, 553–602.

Sander, W. (1992), "The Effect of Women's Schooling on Fertility," *Economic Letters* 40, 229–233.

Savin, N. E., and K. J. White (1977), "The Durbin-Watson Test for Serial Correlation with Extreme Sample Sizes or Many Regressors," *Econometrica* 45, 1989–1996.

Shea, J. (1993), "The Input-Output Approach to Instrument Selection," *Journal of Business and Economic Statistics* 11, 145–155.

Shugart, W. F., and R. D. Tollison (1984), "The Random Character of Merger Activity," *Rand Journal of Economics* 15, 500–509.

Solon, G. (1985), "The Minimum Wage and Teen-age Employment: A Re-analysis with Attention to Serial Correlation and Seasonality," *Journal of Human Resources* 20, 292–297.

Stigler, S. M. (1986), *The History of Statistics*. Cambridge, MA: Harvard University Press.

Stock, J. H., and M. W. Watson (1989), "Interpreting the Evidence on Money-Income Causality," *Journal of Econometrics* 40, 161–181.

Stock, J. H., and M. W. Watson (1993), "A Simple Estimator of Cointegrating Vectors in Higher Order Integrated Systems," *Econometrica* 61, 783–820.

Vella, F., and M. Verbeek (1998), "Whose Wages Do Unions Raise? A Dynamic Model of Unionism and Wage Rate Determination for Young Men," *Journal of Applied Econometrics* 13, 163–183.

Wald, A. (1940), "The Fitting of Straight Lines if Both Variables Are Subject to Error," *Annals of Mathematical Statistics* 11, 284–300.

Wallis, K. F. (1972), "Testing for Fourth-Order Autocorrelation in Quarterly Regression Equations," *Econometrica* 40, 617–636.

White, H. (1980), "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica* 48, 817–838.

White, H. (1984), *Asymptotic Theory for Econometricians*. Orlando: Academic Press.

White, M. J. (1986), "Property Taxes and Firm Location: Evidence from Proposition 13," in *Studies in State and Local Public Finance*. Ed. H. S. Rosen, 83–112. Chicago: University of Chicago Press.

Whittington, L. A. , J. Alm, and H. E. Peters (1990), "Fertility and the Personal Exemption: Implicit Pronatalist Policy in the United States," *American Economic Review* 80, 545–556.

Wooldridge, J. M. (1989), "A Computationally Simple Heteroskedasticity and Serial Correlation-Robust Standard Error for the Linear Regression Model," *Economics Letters* 31, 239–243.

Wooldridge, J. M. (1991a), "A Note on Computing *R*-Squared and Adjusted *R*-Squared for Trending and Seasonal Data," *Economics Letters* 36, 49–54.

Wooldridge, J. M. (1991b), "On the Application of Robust, Regression-Based Diagnostics to Models of Conditional Means and Conditional Variances," *Journal of Econometrics* 47, 5–46.

Wooldridge, J. M. (1994a), "A Simple Specification Test for the Predictive Ability of Transformation Models," *Review of Economics and Statistics* 76, 59–65.

Wooldridge, J. M. (1994b), "Estimation and Inference for Dependent Processes," Chapter 45 in *Handbook of Econometrics*, Volume 4. Ed. R. F. Engle and D. L. McFadden, 2639–2738. Amsterdam: North-Holland.

Wooldridge, J. M. (1995), "Score Diagnostics for Linear Models Estimated by Two Stage Least Squares," in *Advances in Econometrics and Quantitative Economics*. Ed. G. S. Maddala, P. C. B. Phillips, and T. N. Srinivasan, 66–87. Oxford: Blackwell.

Wooldridge, J. M. (1999), *Econometric Analysis of Cross Section and Panel Data*. Forthcoming, Cambridge, MA: MIT Press.

G L O S S A R Y

A

Adjusted *R*-Squared: A goodness-of-fit measure in multiple regression analysis that penalizes additional explanatory variables by using a degrees of freedom adjustment in estimating the error variance.

Alternative Hypothesis: The hypothesis against which the null hypothesis is tested.

AR(1) Serial Correlation: The errors in a time series regression model follow an AR(1) model.

Asymptotic Bias: *See* inconsistency.

Asymptotic Confidence Interval: A confidence interval that is approximately valid in large sample sizes.

Asymptotic Normality: The sampling distribution of a properly normalized estimator converges to the standard normal distribution.

Asymptotic Properties: Properties of estimators and test statistics that apply when the sample size grows without bound.

Asymptotic Standard Error: A standard error that is valid in large samples.

Asymptotic *t* Statistic: A *t* statistic that has an approximate standard normal distribution in large samples.

Asymptotic Variance: The square of the value we must divide an estimator by in order to obtain an asymptotic standard normal distribution.

Asymptotically Efficient: For consistent estimators with asymptotically normal distributions, the estimator with the smallest asymptotic variance.

Asymptotically Uncorrelated: A time series process in which the correlation between random variables at two points in time tends to zero as the time interval between them increases. (*See also* weakly dependent.)

Attenuation Bias: Bias in an estimator that is always toward zero; thus, the expected value of an estimator with attenuation bias is less in magnitude than the absolute value of the parameter.

Augmented Dickey-Fuller Test: A test for a unit root that includes lagged changes of the variable as regressors.

Autocorrelation: *See* serial correlation.

Autoregressive Conditional Heteroskedasticity (ARCH): A model of dynamic heteroskedasticity where the variance of the error term, given past information, depends linearly on the past squared errors.

Autoregressive Process of Order One [AR(1)]: A time series model whose current value depends linearly on its most recent value plus an unpredictable disturbance.

Auxiliary Regression: A regression used to compute a test statistic—such as the test statistics for heteroskedasticity and serial correlation—or any other regression that does not estimate the model of primary interest.

Average: The sum of n numbers divided by n.

B

Balanced Panel: A panel data set where all years (or periods) of data are available for all cross-sectional units.

Base Group: The group represented by the overall intercept in a multiple regression model that includes dummy explanatory variables.

Base Period: For index numbers, such as price or production indices, the period against which all other time periods are measured.

Base Value: The value assigned to the base period for constructing an index number; usually the base value is one or 100.

Benchmark Group: *See* base group.

Bernoulli Random Variable: A random variable that takes on the values zero or one.

Best Linear Unbiased Estimator (BLUE): Among all linear, unbiased estimators, the estimator with the smallest variance. OLS is BLUE, conditional on the sample values of the explanatory variables, under the Gauss-Markov assumptions.

Beta Coefficients: *See* standardized coefficients.

Bias: The difference between the expected and the population parameter values of an estimator.

Biased Estimator: An estimator whose expectation, or sampling mean, is different from the population value it is supposed to be estimating.

Biased Towards Zero: A description of an estimator whose expectation in absolute value is less than the absolute value of the population parameter.

Binary Response Model: A model for a binary (dummy) dependent variable.

Binary Variable: *See* dummy variable.

Binomial Distribution: The probability distribution of the number of successes out of n independent Bernoulli trials, where each trial has the same probability of success.

Bivariate Regression Model: *See* simple linear regression model.

BLUE: *See* best linear unbiased estimator.

Breusch-Godfrey Test: An asymptotically justified test for AR(p) serial correlation, with AR(1) being the most popular; the test allows for lagged dependent variables as well as other regressors that are not strictly exogenous.

Breusch-Pagan Test: A test for heteroskedasticity where the squared OLS residuals are regressed on the explanatory variables in the model.

C

Causal Effect: A ceteris paribus change in one variable has an effect on another variable.

Censored Regression Model: A multiple regression model where the dependent variable has been censored above or below some known threshold.

Central Limit Theorem: A key result from probability theory which implies that the sum of independent random variables, or even weakly dependent random variables, when standardized by its standard deviation, has a distribution that tends to standard normal as the sample size grows.

Ceteris Paribus: All other relevant factors are held fixed.

Chi-Square Distribution: A probability distribution obtained by adding the squares of independent standard normal random variables. The number of terms in the sum equals the degrees of freedom in the distribution.

Chow Statistic: An F statistic for testing the equality of regression parameters across different groups (say, men and women) or time periods (say, before and after a policy change).

Classical Errors-in-Variables (CEV): A measurement error model where the observed measure equals the actual variable plus an independent, or at least an uncorrelated, measurement error.

Classical Linear Model: The multiple linear regression model under the full set of classical linear model assumptions.

Classical Linear Model (CLM) Assumptions: The ideal set of assumptions for multiple regression analysis: for cross-sectional analysis, Assumptions MLR.1 through MLR.6 and for time series analysis, Assumptions TS.1 through TS.6. The assumptions include linearity in the parameters, no perfect collinearity, the zero conditional mean assumption, homoskedasticity, no serial correlation, and normality of the errors.

Cluster Effect: An unobserved effect that is common to all units, usually people, in the cluster.

Cluster Sample: A sample of natural clusters or groups which usually consist of people.

Cochrane-Orcutt (CO) Estimation: A method of estimating a multiple linear regression model with AR(1) errors and strictly exogenous explanatory variables; unlike Prais-Winsten, Cochrane-Orcutt does not use the equation for the first time period.

Coefficient of Determination: *See* R-squared.

Cointegration: The notion that a linear combination of two series, each of which is integrated of order one, is integrated of order zero.

Composite Error: In a panel data model, the sum of the time constant unobserved effect and the idiosyncratic error.

Conditional Distribution: The probability distribution of one random variable, given the values of one or more other random variables.

Conditional Expectation: The expected or average value of one random variable, called the dependent or explained variable, that depends on the values of one or more other variables, called the independent or explanatory variables.

Conditional Forecast: A forecast that assumes the future values of some explanatory variables are known with certainty.

Conditional Variance: The variance of one random variable, given one or more other random variables.

Confidence Interval (CI): A rule used to construct a random interval so that a certain percentage of all data sets, determined by the confidence level, yields an interval that contains the population value.

Confidence Level: The percentage of samples in which we want our confidence interval to contain the population value; 95% is the most common confidence level, but 90% and 99% are also used.

Consistent Estimator: An estimator that converges in probability to the population parameter as the sample size grows without bound.

Consistent Test: A test where, under the alternative hypothesis, the probability of rejecting the null hypothesis converges to one as the sample size grows without bound.

Constant Elasticity Model: A model where the elasticity of the dependent variable, with respect to an explanatory variable, is constant; in multiple regression, both variables appear in logarithmic form.

Contemporaneously Exogenous Regressor: In time series or panel data applications, a regressor that is uncorrelated with the error term in the same time period, but not necessarily in other time periods.

Continuous Random Variable: A random variable that takes on any particular value with probability zero.

Control Group: In program evaluation, the group that does not participate in the program.

Control Variable: *See* explanatory variable.

Corner Solution: A nonnegative dependent variable that is roughly continuous over strictly positive values but takes on the value zero with some regularity.

Correlation Coefficient: A measure of linear dependence between two random variables that does not depend on units of measurement and is bounded between -1 and 1.

Count Variable: A variable that takes on nonnegative integer values.

Covariance: A measure of linear dependence between two random variables.

Covariance Stationary: A time series process with constant mean and variance where the covariance between any two random variables in the sequence depends only on the distance between them.

Covariate: *See* explanatory variable.

Critical Value: In hypothesis testing, the value against which a test statistic is compared to determine whether or not the null hypothesis is rejected.

Cross-Sectional Data Set: A data set collected from a population at a given point in time.

Cumulative Distribution Function (cdf): A function that gives the probability of a random variable being less than or equal to any specified real number.

D

Data Censoring: A situation that arises when we do not always observe the outcome on the dependent variable because at an upper (or lower) threshold we only know that the outcome was above (or below) the threshold. (*See also* censored regression model.)

Data Frequency: The interval at which time series data are collected. Yearly, quarterly, and monthly are the most common data frequencies.

Data Mining: The practice of using the same data set to estimate numerous models in a search to find the "best" model.

Davidson-MacKinnon Test: A test that is used for testing a model against a nonnested alternative; it can be implemented as a t test on the fitted values from the competing model.

Degrees of Freedom (*df*): In multiple regression analysis, the number of observations minus the number of estimated parameters.

Denominator Degrees of Freedom: In an F test, the degrees of freedom in the unrestricted model.

Dependent Variable: The variable to be explained in a multiple regression model (and a variety of other models).

Descriptive Statistic: A statistic used to summarize a set of numbers; the sample average, sample median, and sample standard deviation are the most common.

Deseasonalizing: The removing of the seasonal components from a monthly or quarterly time series.

Detrending: The practice of removing the trend from a time series.

Dickey-Fuller Distribution: The limiting distribution of the t statistic in testing the null hypothesis of a unit root.

Dickey-Fuller (DF) Test: A t test of the unit root null hypothesis in an AR(1) model. (*See also* augmented Dickey-Fuller test.)

Difference in Slopes: A description of a model where some slope parameters may differ by group or time period.

Difference-in-Differences Estimator: An estimator that arises in policy analysis with data for two time periods. One version of the estimator applies to independently pooled cross sections and another to panel data sets.

Diminishing Marginal Effect: The marginal effect of an explanatory variable becomes smaller as the value of the explanatory variable increases.

Discrete Random Variable: A random variable that takes on at most a finite or countably infinite number of values.

Distributed Lag Model: A time series model that relates the dependent variable to current and past values of an explanatory variable.

Disturbance: *See* error term.

Downward Bias: The expected value of an estimator is below the population value of the parameter.

Dummy Dependent Variable: *See* binary response model.

Dummy Variable: A variable that takes on the value zero or one.

Dummy Variable Regression: In a panel data setting, the regression that includes a dummy variable for each cross-sectional unit, along with the remaining explanatory variables. It produces the fixed effects estimator.

Dummy Variable Trap: The mistake of including too many dummy variables among the independent variables; it occurs when an overall intercept is in the model and a dummy variable is included for each group.

Duration Analysis: An application of the censored regression model, where the dependent variable is time elapsed until a certain event occurs, such as the time before an unemployed person becomes reemployed.

Durbin-Watson (DW) Statistic: A statistic used to test for first order serial correlation in the errors of a time series regression model under the classical linear model assumptions.

Dynamically Complete Model: A time series model where no further lags of either the dependent variable or the explanatory variables help to explain the mean of the dependent variable.

E

Econometric Model: An equation relating the dependent variable to a set of explanatory variables and unobserved disturbances, where unknown population parameters determine the ceteris paribus effect of each explanatory variable.

Economic Model: A relationship derived from economic theory or less formal economic reasoning.

Economic Significance: *See* practical significance.

Elasticity: The percent change in one variable given a 1% ceteris paribus increase in another variable.

Empirical Analysis: A study that uses data in a formal econometric analysis to test a theory, estimate a relationship, or determine the effectiveness of a policy.

Endogeneity: A term used to describe the presence of an endogenous explanatory variable.

Endogenous Explanatory Variable: An explanatory variable in a multiple regression model that is correlated with the error term, either because of an omitted variable, measurement error, or simultaneity.

Endogenous Sample Selection: Nonrandom sample selection where the selection is related to the dependent variable, either directly or through the error term in the equation.

Endogenous Variables: In simultaneous equations models, variables that are determined by the equations in the system.

Engle-Granger Two-Step Procedure: A two-step method for estimating error correction models whereby the cointegrating parameter is estimated in the first stage, and the error correction parameters are estimated in the second.

Error Correction Model: A time series model in first differences that also contains an error correction term, which works to bring two I(1) series back into long-run equilibrium.

Error Term: The variable in a simple or multiple regression equation that contains unobserved factors that affect the dependent variable. The error term may also include measurement errors in the observed dependent or independent variables.

Error Variance: The variance of the error term in a multiple regression model.

Errors-in-Variables: A situation where either the dependent variable or some independent variables are measured with error.

Estimate: The numerical value taken on by an estimator for a particular sample of data.

Estimator: A rule for combining data to produce a numerical value for a population parameter; the form of the rule does not depend on the particular sample obtained.

Event Study: An econometric analysis of the effects of an event, such as a change in government regulation or economic policy, on an outcome variable.

Excluding a Relevant Variable: In multiple regression analysis, leaving out a variable that has a nonzero partial effect on the dependent variable.

Exclusion Restrictions: Restrictions which state that certain variables are excluded from the model (or have zero population coefficients).

Exogenous Explanatory Variable: An explanatory variable that is uncorrelated with the error term.

Exogenous Sample Selection: Sample selection that either depends on exogenous explanatory variables or is independent of the error term in the equation of interest.

Exogenous Variable: Any variable that is uncorrelated with the error term in the model of interest.

Expected Value: A measure of central tendency in the distribution of a random variable, including an estimator.

Experiment: In probability, a general term used to denote an event whose outcome is uncertain. In econometric analysis, it denotes a situation where data are collected by randomly assigning individuals to control and treatment groups.

Experimental Data: Data that have been obtained by running a controlled experiment.

Experimental Group: *See* treatment group.

Explained Sum of Squares (SSE): The total sample variation of the fitted values in a multiple regression model.

Explained Variable: *See* dependent variable.

Explanatory Variable: In regression analysis, a variable that is used to explain variation in the dependent variable.

Exponential Function: A mathematical function defined for all values that have an increasing slope but a constant proportionate change.

Exponential Smoothing: A simple method of forecasting a variable that involves a weighting of all previous outcomes on that variable.

Exponential Trend: A trend with a constant growth rate.

F

***F* Distribution:** The probability distribution obtained by forming the ratio of two independent chi-square random variables, where each has been divided by its degrees of freedom.

***F* Statistic:** A statistic used to test multiple hypotheses about the parameters in a multiple regression model.

Feasible GLS (FGLS) Estimator: A GLS procedure where variance or correlation parameters are unknown and therefore must first be estimated. (*See also* generalized least squares estimator.)

Finite Distributed Lag (FDL) Model: A dynamic model where one or more explanatory variables are allowed to have lagged effects on the dependent variable.

First Difference: A transformation on a time series constructed by taking the difference of adjacent time periods, where the earlier time period is subtracted from the later time period.

First-Differenced Equation: In time series or panel data models, an equation where the dependent and independent variables have all been first-differenced.

First-Differenced Estimator: In a panel data setting, the pooled OLS estimator applied to first differences of the data across time.

First Order Conditions: The set of linear equations used to solve for the OLS estimates.

Fitted Values: The estimated values of the dependent variable when the values of the independent variables for each observation are plugged into the OLS regression line.

Fixed Effect: *See* unobserved effect.

Fixed Effects Estimator: For the unobserved effects panel data model, the estimator obtained by applying pooled OLS to a time-demeaned equation.

Fixed Effects Transformation: For panel data, the time-demeaned data.

Forecast Error: The difference between the actual outcome and the forecast of the outcome.

Forecast Interval: In forecasting, a confidence interval for a yet unrealized future value of a time series variable. (*See also* prediction interval.)

Functional Form Misspecification: A problem that occurs when a model has omitted functions of the explanatory variables (such as quadratics) or uses the wrong functions of either the dependent variable or some explanatory variables.

G

Gauss-Markov Assumptions: The set of assumptions (Assumptions MLR.1 through MLR.5 or TS.1 through TS.5) under which OLS is BLUE.

Gauss-Markov Theorem: The theorem which states that, under the five Gauss-Markov assumptions (for cross-sectional or time series models), the OLS estimator is BLUE (conditional on the sample values of the explanatory variables).

Generalized Least Squares (GLS) Estimator: An estimator that accounts for a known structure of the error variance (heteroskedasticity), serial correlation pattern in the errors, or both, via a transformation of the original model.

Geometric (or Koyck) Distributed Lag: An infinite distributed lag model where the lag coefficients decline at a geometric rate.

Goodness-of-Fit Measure: A statistic that summarizes how well a set of explanatory variables explains a dependent or response variable.

Granger Causality: A limited notion of causality where past values of one series (x_t) are useful for predicting future values of another series (y_t), after past values of y_t have been controlled for.

Growth Rate: The proportionate change in a time series from the previous period. It may be approximated as the difference in logs or reported in percentage form.

H

Heckit Method: An econometric procedure used to correct for sample selection bias due to incidental truncation or some other form of nonrandomly missing data.

Heterogeneity Bias: The bias in OLS due to omitted heterogeneity (or omitted variables).

Heteroskedasticity: The variance of the error term, given the explanatory variables, is not constant.

Heteroskedasticity of Unknown Form: Heteroskedasticity that may depend on the explanatory variables in an unknown, arbitrary fashion.

Heteroskedasticity-Robust *F* Statistic: An *F*-type statistic that is (asymptotically) robust to heteroskedasticity of unknown form.

Heteroskedasticity-Robust *LM* Statistic: An *LM* statistic that is robust to heteroskedasticity of unknown form.

Heteroskedasticity-Robust Standard Error: A standard error that is (asymptotically) robust to heteroskedasticity of unknown form.

Heteroskedasticity-Robust *t* Statistic: A *t* statistic that is (asymptotically) robust to heteroskedasticity of unknown form.

Highly Persistent Process: A time series process where outcomes in the distant future are highly correlated with current outcomes.

Homoskedasticity: The errors in a regression model have constant variance, conditional on the explanatory variables.

Hypothesis Test: A statistical test of the null, or maintained, hypothesis against an alternative hypothesis.

I

Identified Equation: An equation whose parameters can be consistently estimated, especially in models with endogenous explanatory variables.

Idiosyncratic Error: In panel data models, the error that changes over time as well as across units (say, individuals, firms, or cities).

Impact Elasticity: In a distributed lag model, the immediate percentage change in the dependent variable given a 1% increase in the independent variable.

Impact Multiplier: *See* impact propensity.

Impact Propensity: In a distributed lag model, the immediate change in the dependent variable given a one-unit increase in the independent variable.

Incidental Truncation: A sample selection problem whereby one variable, usually the dependent variable, is only observed for certain outcomes of another variable.

Inclusion of an Irrelevant Variable: The including of an explanatory variable in a regression model that has a zero population parameter in estimating an equation by OLS.

Inconsistency: The difference between the probability limit of an estimator and the parameter value.

Independent Random Variables: Random variables whose joint distribution is the product of the marginal distributions.

Independent Variable: *See* explanatory variable.

Independently Pooled Cross Section: A data set obtained by pooling independent random samples from different points in time.

Index Number: A statistic that aggregates information on economic activity, such as production or prices.

Infinite Distributed Lag (IDL) Model: A distributed lag model where a change in the explanatory variable can have an impact on the dependent variable into the indefinite future.

Influential Observations: *See* outliers.

Information Set: In forecasting, the set of variables that we can observe prior to forming our forecast.

In-Sample Criteria: Criteria for choosing forecasting models that are based on goodness-of-fit within the sample used to obtain the parameter estimates.

Instrumental Variable (IV): In an equation with an endogenous explanatory variable, an IV is a variable that does not appear in the equation, is uncorrelated with the error in the equation, and is (partially) correlated with the endogenous explanatory variable.

Instrumental Variables (IV) Estimator: An estimator in a linear model used when instrumental variables are available for one or more endogenous explanatory variables.

Integrated of Order One [I(1)]: A time series process that needs to be first-differenced in order to produce an I(0) process.

Integrated of Order Zero [I(0)]: A stationary, weakly dependent time series process that, when used in regression analysis, satisfies the law of large numbers and the central limit theorem.

Interaction Effect: In multiple regression, the partial effect of one explanatory variable depends on the value of a different explanatory variable.

Interaction Term: An independent variable in a regression model that is the product of two explanatory variables.

Intercept Parameter: The parameter in a multiple linear regression model that gives the expected value of the dependent variable when all the independent variables equal zero.

Intercept Shift: The intercept in a regression model differs by group or time period.

Internet: A global computer network that can be used to access information and download data bases.

Interval Estimator: A rule that uses data to obtain lower and upper bounds for a population parameter. (*See also* confidence interval.)

Inverse Mills Ratio: A term that can be added to a multiple regression model to remove sample selection bias.

J

Joint Distribution: The probability distribution determining the probabilities of outcomes involving two or more random variables.

Joint Hypothesis Test: A test involving more than one restriction on the parameters in a model.

Jointly Statistically Significant: The null hypothesis that two or more explanatory variables have zero population coefficients is rejected at the chosen significance level.

Just Identified Equation: For models with endogenous explanatory variables, an equation that is identified but would not be identfied with one fewer instrumental variable.

L

Lag Distribution: In a finite or infinite distributed lag model, the lag coefficients graphed as a function of the lag length.

Lagged Dependent Variable: An explanatory variable that is equal to the dependent variable from an earlier time period.

Lagged Endogenous Variable: In a simultaneous equations model, a lagged value of one of the endogenous variables.

Lagrange Multiplier Statistic: A test statistic with large sample justification that can be used to test for omitted variables, heteroskedasticity, and serial correlation, among other model specification problems.

Large Sample Properties: *See* asymptotic properties.

Latent Variable Model: A model where the observed dependent variable is assumed to be a function of an underlying latent, or unobserved, variable.

Law of Iterated Expectations: A result from probability that relates unconditional and conditional expectations.

Law of Large Numbers (LLN): A theorem which says that the average from a random sample converges in probability to the population average; the LLN also holds for stationary and weakly dependent time series.

Leads and Lags Estimator: An estimator of a cointegrating parameter in a regression with I(1) variables, where the current, some past, and some future first differences in the explanatory variable are included as regressors.

Level-Level Model: A regression model where the dependent variable and the independent variables are in level (or original) form.

Level-Log Model: A regression model where the dependent variable is in level form and (at least some of) the independent variables are in logarithmic form.

Likelihood Ratio Statistic: A statistic that can be used to test single or multiple hypotheses when the constrained and unconstrained models have been estimated by maximum likelihood. The statistic is twice the difference in the unconstrained and constrained log-likelihoods.

Limited Dependent Variable: A dependent or response variable whose range is restricted in some important way.

Linear Function: A function where the change in the dependent variable, given a one-unit change in an independent variable, is constant.

Linear Probability Model (LPM): A binary response model where the response probability is linear in its parameters.

Linear Time Trend: A trend that is a linear function of time.

Linear Unbiased Estimator: In multiple regression analysis, an unbiased estimator that is a linear function of the outcomes on the dependent variable.

Logarithmic Function: A mathematical function defined for positive arguments that has a positive, but diminishing, slope.

Log-Level Model: A regression model where the dependent variable is in logarithmic form and the independent variables are in level (or original) form.

Log-Log Model: A regression model where the dependent variable and (at least some of) the explanatory variables are in logarithmic form.

Logit Model: A model for binary response where the response probability is the logit function evaluated at a linear function of the explanatory variables.

Log-Likelihood Function: The sum of the log-likelihoods, where the log-likelihood for each observation is the log of the density of the dependent variable given the explanatory variables; the log-likelihood function is viewed as a function of the parameters to be estimated.

Long-Run Elasticity: The long-run propensity in a distributed lag model with the dependent and independent variables in logarithmic form; thus, the long-run elasticity is the eventual percentage increase in the explained variable, given a permanent 1% increase in the explanatory variable.

Long-Run Multiplier: *See* long-run propensity.

Long-Run Propensity: In a distributed lag model, the eventual change in the dependent variable given a permanent, one-unit increase in the independent variable.

Longitudinal Data: *See* panel data.

Loss Function: A function that measures the loss when a forecast differs from the actual outcome; the most common examples are absolute value loss and squared loss.

M

Marginal Effect: The effect on the dependent variable that results from changing an independent variable by a small amount.

Martingale: A time series process whose expected value, given all past outcomes on the series, simply equals the most recent value.

Martingale Difference Sequence: The first difference of a martingale. It is unpredictable (or has a zero mean), given past values of the sequence.

Matched Pairs Sample: A sample where each observation is matched with another, as in a sample consisting of a husband and wife or a set of two siblings.

Matrix: An array of numbers.

Matrix Notation: A convenient mathematical notation, grounded in matrix algebra, for expressing and manipulating the multiple regression model.

Maximum Likelihood Estimation (MLE): A broadly applicable estimation method where the parameter estimates are chosen to maximize the log-likelihood function.

Mean: *See* expected value.

Mean Absolute Error (MAE): A performance measure in forecasting, computed as the average of the absolute values of the forecast errors.

Mean Squared Error: The expected squared distance that an estimator is from the population value; it equals the variance plus the square of any bias.

Measurement Error: The difference between an observed variable and the variable that belongs in a multiple regression equation.

Median: In a probability distribution, it is the value where there is a 50% chance of being below the value and a 50% chance of being above it. In a sample of numbers, it is the middle value after the numbers have been ordered.

Method of Moments Estimator: An estimator obtained by using the sample analog of population moments; ordinary least squares and two stage least squares are both method of moments estimators.

Micronumerosity: A term introduced by Arthur Goldberger to describe properties of econometric estimators with small sample sizes.

Minimum Variance Unbiased Estimator: An estimator with the smallest variance in the class of all unbiased estimators.

Missing Data: A data problem that occurs when we do not observe values on some variables for certain observations (individuals, cities, time periods, and so on) in the sample.

Moving Average Process of Order One [MA(1)]: A time series process generated as a linear function of the current value and one lagged value of a zero-mean, constant variance, uncorrelated stochastic process.

Multicollinearity: A term that refers to correlation among the independent variables in a multiple regression model; it is usually invoked when some correlations are "large," but an actual magnitude is not well-defined.

Multiple Hypothesis Test: A test of a null hypothesis involving more than one restriction on the parameters.

Multiple Linear Regression (MLR) Model: A model linear in its parameters, where the dependent variable is a function of independent variables plus an error term.

Multiple Regression Analysis: A type of analysis that is used to describe estimation of and inference in the multiple linear regression model.

Multiple Restrictions: More than one restriction on the parameters in an econometric model.

Multiple Step-Ahead Forecast: A time series forecast of more than one period into the future.

Multiplicative Measurement Error: Measurement error where the observed variable is the product of the true unobserved variable and a positive measurement error.

N

***n*-*R*-Squared Statistic:** *See* Lagrange multiplier statistic.

Natural Experiment: A situation where the economic environment—sometimes summarized by an explanatory variable—exogenously changes, perhaps inadvertently, due to a policy or institutional change.

Natural Logarithm: *See* logarithmic function.

Nominal Variable: A variable measured in nominal or current dollars.

Nonexperimental Data: Data that have not been obtained through a controlled experiment.

Nonlinear Function: A function whose slope is not constant.

Nonnested Models: Two (or more) models where no model can be written as a special case of the other by imposing restrictions on the parameters.

Nonrandom Sample Selection: A sample selection process that cannot be characterized as drawing randomly from the population of interest.

Nonstationary Process: A time series process whose joint distributions are not constant across different epochs.

Normal Distribution: A probability distribution commonly used in statistics and econometrics for modeling a population. Its probability distribution function has a bell shape.

Normality Assumption: The classical linear model assumption which states that the error (or dependent variable) has a normal distribution, conditional on the explanatory variables.

Null Hypothesis: In classical hypothesis testing, we take this hypothesis as true and require the data to provide substantial evidence against it.

Numerator Degrees of Freedom: In an F test, the number of restrictions being tested.

O

Observational Data: *See* nonexperimental data.

OLS: *See* ordinary least squares.

OLS Intercept Estimate: The intercept in an OLS regression line.

OLS Regression Line: The equation relating the predicted value of the dependent variable to the independent variables, where the parameter estimates have been obtained by OLS.

OLS Slope Estimate: A slope in an OLS regression line.

Omitted Variable Bias: The bias that arises in the OLS estimators when a relevant variable is omitted from the regression.

Omitted Variables: One or more variables, which we would like to control for, have been omitted in estimating a regression model.

One-Sided Alternative: An alternative hypothesis which states that the parameter is greater than (or less than) the value hypothesized under the null.

One-Step-Ahead Forecast: A time series forecast one period into the future.

One-Tailed Test: A hypothesis test against a one-sided alternative.

On-Line Data Bases: Data bases that can be accessed via a computer network.

On-Line Search Services: Computer software that allows the Internet or data bases on the Internet to be searched by topic, name, title, or key words.

Order Condition: A necessary condition for identifying the parameters in a model with one or more endogenous explanatory variables: the total number of exogenous variables must be at least as great as the total number of explanatory variables.

Ordinal Variable: A variable where the ordering of the values conveys information but the magnitude of the values does not.

Ordinary Least Squares (OLS): A method for estimating the parameters of a multiple linear regression model. The ordinary least squares estimates are obtained by minimizing the sum of squared residuals.

Outliers: Observations in a data set that are substantially different from the bulk of the data, perhaps because of errors or because some data are generated by a different model than most of the other data.

Out-of-Sample Criteria: Criteria used for choosing forecasting models that are based on a part of the sample that was not used in obtaining parameter estimates.

Overall Significance of a Regression: A test of the joint significance of all explanatory variables appearing in a multiple regression equation.

Overdispersion: In modeling a count variable, the variance is larger than the mean.

Overidentified Equation: In models with endogenous explanatory variables, an equation where the number of instrumental variables is strictly greater than the number of endogenous explanatory variables.

Overidentifying Restrictions: The extra moment conditions that come from having more instrumental variables than endogenous explanatory variables in a linear model.

Overspecifying a Model: *See* inclusion of an irrelevant variable.

P

***p*-value:** The smallest significance level at which the null hypothesis can be rejected. Equivalently, the largest significance level at which the null hypothesis cannot be rejected.

Panel Data: A data set constructed from repeated cross sections over time. With a *balanced* panel, the same units appear in each time period. With an *unbalanced* panel, some units do not appear in each time period, often due to attrition.

Pairwise Uncorrelated Random Variables: A set of two or more random variables where each pair is uncorrelated.

Parameter: An unknown value that describes a population relationship.

Parsimonious Model: A model with as few parameters as possible for capturing any desired features.

Partial Effect: The effect of an explanatory variable on the dependent variable, holding other factors in the regression model fixed.

Percent Correctly Predicted: In a binary response model, the percentage of times the prediction of zero or one coincides with the actual outcome.

Percentage Change: The porportionate change in a variable, multiplied by 100.

Percentage Point Change: The change in a variable that is measured as a percent.

Perfect Collinearity: In multiple regression, one independent variable is an exact linear function of one or more other independent variables.

Plug-In Solution to the Omitted Variables Problem: A proxy variable is substituted for an unobserved omitted variable in an OLS regression.

Point Forecast: The forecasted value of a future outcome.

Poisson Distribution: A probability distribution for count variables.

Poisson Regression Model: A model for a count dependent variable where the dependent variable, conditional on the explanatory variables, is nominally assumed to have a Poisson distribution.

Policy Analysis: An empirical analysis that uses econometric methods to evaluate the effects of a certain policy.

Pooled Cross Section: A data configuration where independent cross sections, usually collected at different points in time, are combined to produce a single data set.

Pooled OLS Estimation: OLS estimation with independently pooled cross sections, panel data, or cluster samples, where the observations are pooled across time (or group) as well as across the cross-sectional units.

Population: A well-defined group (of people, firms, cities, and so on) that is the focus of a statistical or econometric analysis.

Population Model: A model, especially a multiple linear regression model, that describes a population.

Population *R*-Squared: In the population, the fraction of the variation in the dependent variable that is explained by the explanatory variables.

Population Regression Function: *See* conditional expectation.

Power of a Test: The probability of rejecting the null hypothesis when it is false; the power depends on the values of the population parameters under the alternative.

Practical Significance: The practical or economic importance of an estimate, which is measured by its sign and magnitude, as opposed to its statistical significance.

Prais-Winsten (PW) Estimation: A method of estimating a multiple linear regression model with AR(1) errors and strictly exogenous explanatory variables; unlike Cochrane-Orcutt, Prais-Winsten uses the equation for the first time period in estimation.

Predetermined Variable: In a simultaneous equations model, either a lagged endogenous variable or a lagged exogenous variable.

Predicted Variable: *See* dependent variable.

Prediction: The estimate of an outcome obtained by plugging specific values of the explanatory variables into an estimated model, usually a multiple regression model.

Prediction Error: The difference between the actual outcome and a prediction of that outcome.

Prediction Interval: A confidence interval for an

unknown outcome on a dependent variable in a multiple regression model.

Predictor Variable: *See* explanatory variable.

Probability Density Function (pdf): A function that, for discrete random variables, gives the probability that the random variable takes on each value; for continuous random variables, the area under the pdf gives the probability of various events.

Probability Limit: The value to which an estimator converges as the sample size grows without bound.

Probit Model: A model for binary responses where the response probability is the standard normal cdf evaluated at a linear function of the explanatory variables.

Program Evaluation: An analysis of a particular private or public program using econometric methods to obtain the causal effect of the program.

Proportionate Change: The change in a variable relative to its initial value; mathematically, the change divided by the initial value.

Proxy Variable: An observed variable that is related but not identical to an unobserved explanatory variable in multiple regression analysis.

Q

Quadratic Functions: Functions that contain squares of one or more explanatory variables; they capture diminishing or increasing effects on the dependent variable.

Qualitative Variable: A variable describing a nonquantitative feature of an individual, a firm, a city, and so on.

Quasi-Demeaned Data: In random effects estimation for panel data, it is the original data in each time period minus a fraction of the time average; these calculations are done for each cross-sectional observation.

Quasi-Differenced Data: In estimating a regression model with AR(1) serial correlation, it is the difference between the current time period and a multiple of the previous time period, where the multiple is the parameter in the AR(1) model.

Quasi-Experiment: *See* natural experiment.

Quasi-Likelihood Ratio Statistic: A modification of the likelihood ratio statistic that accounts for possible distributional misspecification, as in a Poisson regression model.

Quasi-Maximum Likelihood Estimation: Maximum likelihood estimation but where the log-likelihood function may not correspond to the actual conditional distribution of the dependent variable.

R

***R*-Bar Squared:** *See* adjusted *R*-squared.

***R*-Squared:** In a multiple regression model, the proportion of the total sample variation in the dependent variable that is explained by the independent variable.

***R*-Squared Form of the *F* Statistic:** The *F* statistic for testing exclusion restrictions expressed in terms of the *R*-squareds from the restricted and unrestricted models.

Random Effects Estimator: A feasible GLS estimator in the unobserved effects model where the unobserved effect is assumed to be uncorrelated with the explanatory variables in each time period.

Random Effects Model: The unobserved effects panel data model where the unobserved effect is assumed to be uncorrelated with the explanatory variables in each time period.

Random Sampling: A sampling scheme whereby each observation is drawn at random from the population. In particular, no unit is more likely to be selected than any other unit, and each draw is independent of all other draws.

Random Variable: A variable whose outcome is uncertain.

Random Walk: A time series process where next period's value is obtained as this period's value, plus an independent (or at least an uncorrelated) error term.

Random Walk with Drift: A random walk that has a constant (or drift) added in each period.

Rank Condition: A sufficient condition for identification of a model with one or more endogenous explanatory variables.

Rational Distributed Lag (RDL) Model: A type of infinite distributed lag model where the lag distribution depends on relatively few parameters.

Real Variable: A monetary value measured in terms of a base period.

Reduced Form Equation: A linear equation where an endogenous variable is a function of exogenous variables and unobserved errors.

Reduced Form Error: The error term appearing in a reduced form equation.

Reduced Form Parameters: The parameters appearing in a reduced form equation.

Regressand: *See* dependent variable.

Regression Error Specification Test (RESET): A general test for functional form in a multiple regression model; it is an *F* test of joint significance of the

squares, cubes, and perhaps higher powers of the fitted values from the initial OLS estimation.

Regression Through the Origin: Regression analysis where the intercept is set to zero; the slopes are obtained by minimizing the sum of squared residuals, as usual.

Regressor: *See* explanatory variable.

Rejection Region: The set of values of a test statistic that leads to rejecting the null hypothesis.

Rejection Rule: In hypothesis testing, the rule that determines when the null hypothesis is rejected in favor of the alternative hypothesis.

Residual: The difference between the actual value and the fitted (or predicted) value; there is a residual for each observation in the sample used to obtain an OLS regression line.

Residual Analysis: A type of analysis that studies the sign and size of residuals for particular observations after a multiple regression model has been estimated.

Residual Sum of Squares: *See* sum of squared residuals.

Response Probability: In a binary response model, the probability that the dependent variable takes on the value one, conditional on explanatory variables.

Response Variable: *See* dependent variable.

Restricted Model: In hypothesis testing, the model obtained after imposing all of the restrictions required under the null.

Root Mean Squared Error (RMSE): Another name for the standard error of the regression in multiple regression analysis.

S

Sample Average: The sum of n numbers divided by n; a measure of central tendency.

Sample Correlation: For outcomes on two random variables, the sample covariance divided by the product of the sample standard deviations.

Sample Covariance: An unbiased estimator of the population covariance between two random variables.

Sample Regression Function: *See* OLS regression line.

Sample Selection Bias: Bias in the OLS estimator which is induced by using data that arise from endogenous sample selection.

Sample Standard Deviation: A consistent estimator of the population standard deviation.

Sample Variance: An unbiased, consistent estimator of the population variance.

Sampling Distribution: The probability distribution of an estimator over all possible sample outcomes.

Sampling Variance: The variance in the sampling distribution of an estimator; it measures the spread in the sampling distribution.

Score Statistic: *See* Lagrange multiplier statistic.

Seasonal Dummy Variables: A set of dummy variables used to denote the quarters or months of the year.

Seasonality: A feature of monthly or quarterly time series where the average value differs systematically by season of the year.

Seasonally Adjusted: Monthly or quarterly time series data where some statistical procedure—possibly regression on seasonal dummy variables—has been used to remove the seasonal component.

Selected Sample: A sample of data obtained not by random sampling but by selecting on the basis of some observed or unobserved characteristic.

Semi-Elasticity: The percentage change in the dependent variable given a one-unit increase in an independent variable.

Sensitivity Analysis: The process of checking whether the estimated effects and statistical significance of key explanatory variables are sensitive to inclusion of other explanatory variables, functional form, dropping of potentially outlying observations, or different methods of estimation.

Serial Correlation: In a time series or panel data model, correlation between the errors in different time periods.

Serial Correlation-Robust Standard Error: A standard error for an estimator that is (asymptotically) valid whether or not the errors in the model are serially correlated.

Serially Uncorrelated: The errors in a time series or panel data model are pairwise uncorrelated across time.

Short-Run Elasticity: The impact propensity in a distributed lag model when the dependent and independent variables are in logarithmic form.

Significance Level: The probability of Type I error in hypothesis testing.

Simple Linear Regression Model: A model where the dependent variable is a linear function of a single independent variable, plus an error term.

Simultaneity: A term that means at least one explanatory variable in a multiple linear regression model is determined jointly with the dependent variable.

Simultaneity Bias: The bias that arises from using OLS to estimate an equation in a simultaneous equations model.

Simultaneous Equations Model (SEM): A model that jointly determines two or more endogenous variables,

where each endogenous variable can be a function of other endogenous variables as well as of exogenous variables and an error term.

Slope Parameter: The coefficient on an independent variable in a multiple regression model.

Spreadsheet: Computer software used for entering and manipulating data.

Spurious Correlation: A correlation between two variables that is not due to causality, but perhaps to the dependence of the two variables on another unobserved factor.

Spurious Regression Problem: A problem that arises when regression analysis indicates a relationship between two or more unrelated time series processes simply because each has a trend, is an integrated time series (such as a random walk), or both.

Stable AR(1) Process: An AR(1) process where the parameter on the lag is less than one in absolute value. The correlation between two random variables in the sequence declines to zero at a geometric rate as the distance between the random variables increases, and so a stable AR(1) process is weakly dependent.

Standard Deviation: A common measure of spread in the distribution of a random variable.

Standard Deviation of $\hat{\beta}_j$: A common measure of spread in the sampling distribution of $\hat{\beta}_j$.

Standard Error of $\hat{\beta}_j$: An estimate of the standard deviation in the sampling distribution of $\hat{\beta}_j$.

Standard Error of the Estimate: *See* standard error of the regression.

Standard Error of the Regression (SER): In multiple regression analysis, the estimate of the standard deviation of the population error, obtained as the square root of the sum of squared residuals over the degrees of freedom.

Standard Normal Distribution: The normal distribution with mean zero and variance one.

Standardized Coefficient: A regression coefficient that measures the standard deviation change in the dependent variable given a one standard deviation increase in an independent variable.

Standardized Random Variable: A random variable transformed by subtracting off its expected value and dividing the result by its standard deviation; the new random variable has mean zero and standard deviation one.

Static Model: A time series model where only contemporaneous explanatory variables affect the dependent variable.

Stationary Process: A time series process where the marginal and all joint distributions are invariant across time.

Statistical Inference: The act of testing hypotheses about population parameters.

Statistically Different from Zero: *See* statistically significant.

Statistically Insignificant: Failure to reject the null hypothesis that a population parameter is equal to zero, at the chosen significance level.

Statistically Significant: Rejecting the null hypothesis that a parameter is equal to zero against the specified alternative, at the chosen significance level.

Stochastic Process: A sequence of random variables indexed by time.

Strict Exogeneity: An assumption that holds in a time series or panel data model when the explanatory variables are strictly exogenous.

Strictly Exogenous: A feature of explanatory variables in a time series or panel data model where the error term at any time period has zero expectation, conditional on the explanatory variables in all time periods; a less restrictive version is stated in terms of zero correlations.

Strongly Dependent: *See* highly persistent process.

Structural Equation: An equation derived from economic theory or from less formal economic reasoning.

Structural Error: The error term in a structural equation, which could be one equation in a simultaneous equations model.

Structural Parameters: The parameters appearing in a structural equation.

Sum of Squared Residuals: In multiple regression analysis, the sum of the squared OLS residuals across all observations.

Summation Operator: A notation, denoted by Σ, used to define the summing of a set of numbers.

T

***t* Distribution:** The distribution of the ratio of a standard normal random variable and the square root of an independent chi-square random variable, where the chi-square random variable is first divided by its *df*.

***t* Ratio:** *See t* statistic.

***t* Statistic:** The statistic used to test a single hypothesis about the parameters in an econometric model.

Test Statistic: A rule used for testing hypotheses where each sample outcome produces a numerical value.

Text Editor: Computer software that can be used to edit text files.

Text (ASCII) File: A universal file format that can be transported across numerous computer platforms.

Time-Demeaned Data: Panel data where, for each cross-sectional unit, the average over time is subtracted from the data in each time period.

Time Series Data: Data collected over time on one or more variables.

Time Series Process: *See* stochastic process.

Time Trend: A function of time that is the expected value of a trending time series process.

Tobit Model: A model for a dependent variable that takes on the value zero with positive probability but is roughly continuously distributed over strictly positive values. (*See also* corner solution.)

Top Coding: A form of data censoring where the value of a variable is not reported when it is above a given threshold; we only know that it is at least as large as the threshold.

Total Sum of Squares (SST): The total sample variation in a dependent variable about its sample average.

Treatment Group: In program evaluation, the group that participates in the program. (*See also* experimental group.)

Trending Process: A time series process whose expected value is an increasing or decreasing function of time.

Trend-Stationary Process: A process that is stationary once a time trend has been removed; it is usually implicit that the detrended series is weakly dependent.

Truncated Regression Model: A classical linear regression model for cross-sectional data in which the sampling scheme entirely excludes, on the basis of outcomes on the dependent variable, part of the population.

True Model: The actual population model relating the dependent variable to the relevant independent variables, plus a disturbance, where the zero conditional mean assumption holds.

Two Stage Least Squares (2SLS) Estimator: An instrumental variables estimator where the IV for an endogenous explanatory variable is obtained as the fitted value from regressing the endogenous explanatory variable on all exogenous variables.

Two-Sided Alternative: An alternative where the population parameter can be either less than or greater than the value stated under the null hypothesis.

Two-Tailed Test: A test against a two-sided alternative.

Type I Error: A rejection of the null hypothesis when it is true.

Type II Error: The failure to reject the null hypothesis when it is false.

U

Unbalanced Panel: A panel data set where certain years (or periods) of data are missing for some cross-sectional units.

Unbiased Estimator: An estimator whose expected value (or mean of its sampling distribution) equals the population value (regardless of the population value).

Unconditional Forecast: A forecast that does not rely on knowing, or assuming values for, future explanatory variables.

Uncorrelated Random Variables: Random variables that are not linearly related.

Underspecifying a Model: *See* excluding a relevant variable.

Unidentified Equation: An equation with one or more endogenous explanatory variables where sufficient instrumental variables do not exist to identify the parameters.

Unit Root Process: A highly persistent time series process where the current value equals last period's value, plus a weakly dependent disturbance.

Unobserved Effect: In a panel data model, an unobserved variable in the error term that does not change over time. For cluster samples, an unobserved variable that is common to all units in the cluster.

Unobserved Effects Model: A model for panel data or cluster samples where the error term contains an unobserved effect.

Unobserved Heterogeneity: *See* unobserved effect.

Unrestricted Model: In hypothesis testing, the model that has no restrictions placed on its parameters.

Upward Bias: The expected value of an estimator is greater than the population parameter value.

V

Variance: A measure of spread in the distribution of a random variable.

Variance of the Prediction Error: The variance in the error that arises when predicting a future value of the dependent variable based on an estimated multiple regression equation.

Vector Autoregressive (VAR) Model: A model for two or more time series where each variable is modeled as a linear function of past values of all variables, plus disturbances that have zero means given all past values of the observed variables.

W

Weakly Dependent: A term that describes a time series process where some measure of dependence between random variables at two points in time—such as correlation—diminishes as the interval between the two points in time increases.

Weighted Least Squares (WLS) Estimator: An estimator used to adjust for a known form of heteroskedasticity, where each squared residual is weighted by the inverse of the (estimated) variance of the error.

White Test: A test for heteroskedasticity that involves regressing the squared OLS residuals on the OLS fitted values and on the squares of the fitted values; in its most general form, the squared OLS residuals are regressed on the explanatory variables, the squares of the explanatory variables, and all the nonredundant cross products of the explanatory variables.

Within Estimator: *See* fixed effects estimator.

Within Transformation: *See* fixed effects transformation.

Y

Year Dummy Variables: For data sets with a time series component, dummy (binary) variables equal to one in the relevant year and zero in all other years.

Z

Zero Conditional Mean Assumption: A key assumption used in multiple regression analysis which states that, given any values of the explanatory variables, the expected value of the error equals zero. (*See* Assumptions MLR.3, TS.2, and TS.2′.)