15 Estimation of Dynamic Causal Effects

In the 1983 movie *Trading Places*, the characters played by Dan Aykroyd and Eddie Murphy used inside information on how well Florida oranges had fared over the winter to make millions in the orange juice concentrate futures market, a market for contracts to buy or sell large quantities of orange juice concentrate at a specified price on a future date. In real life, traders in orange juice futures in fact do pay close attention to the weather in Florida: Freezes in Florida kill Florida oranges, the source of almost all frozen orange juice concentrate made in the United States, so its supply falls and the price rises. But precisely how much does the price rise when the weather in Florida turns sour? Does the price rise all at once, or are there delays; if so, for how long? These are questions that real-life traders in orange juice futures need to answer if they want to succeed.

This chapter takes up the problem of estimating the effect on Y now and in the future of a change in X, that is, the **dynamic causal effect** on Y of a change in X. What, for example, is the effect on the path of orange juice prices over time of a freezing spell in Florida? The starting point for modeling and estimating dynamic causal effects is the so-called distributed lag regression model, in which Y_t is expressed as a function of current and past values of X_t. Section 15.1 introduces the distributed lag model in the context of estimating the effect of cold weather in Florida on the price of orange juice concentrate over time. Section 15.2 takes a closer look at what, precisely, is meant by a dynamic causal effect.

One way to estimate dynamic causal effects is to estimate the coefficients of the distributed lag regression model using OLS. As discussed in Section 15.3, this estimator is consistent if the regression error has a conditional mean of zero given current and past values of X, a condition that (as in Chapter 12) is referred to as exogeneity. Because the omitted determinants of Y_t are correlated over time—that is, because they are serially correlated—the error term in the distributed lag model can be serially correlated. This possibility in turn requires "heteroskedasticity- and autocorrelation-consistent" (HAC) standard errors, the topic of Section 15.4.

A second way to estimate dynamic causal effects, discussed in Section 15.5, is to model the serial correlation in the error term as an autoregression and then to use this autoregressive model to derive an autoregressive distributed lag (ADL) model. Alternatively, the coefficients of the original distributed lag model can be estimated by generalized least squares (GLS). Both the ADL and GLS methods, however, require a stronger version of exogeneity than we have used so far: *strict* exogeneity, under which the regression errors have a conditional mean of zero given past, present, *and future* values of X.

Section 15.6 provides a more complete analysis of the relationship between orange juice prices and the weather. In this application, the weather is beyond human control and thus is exogenous (although, as discussed in Section 15.6, economic theory suggests that it is not necessarily strictly exogenous). Because exogeneity is necessary for estimating dynamic causal effects, Section 15.7 examines this assumption in several applications taken from macroeconomics and finance.

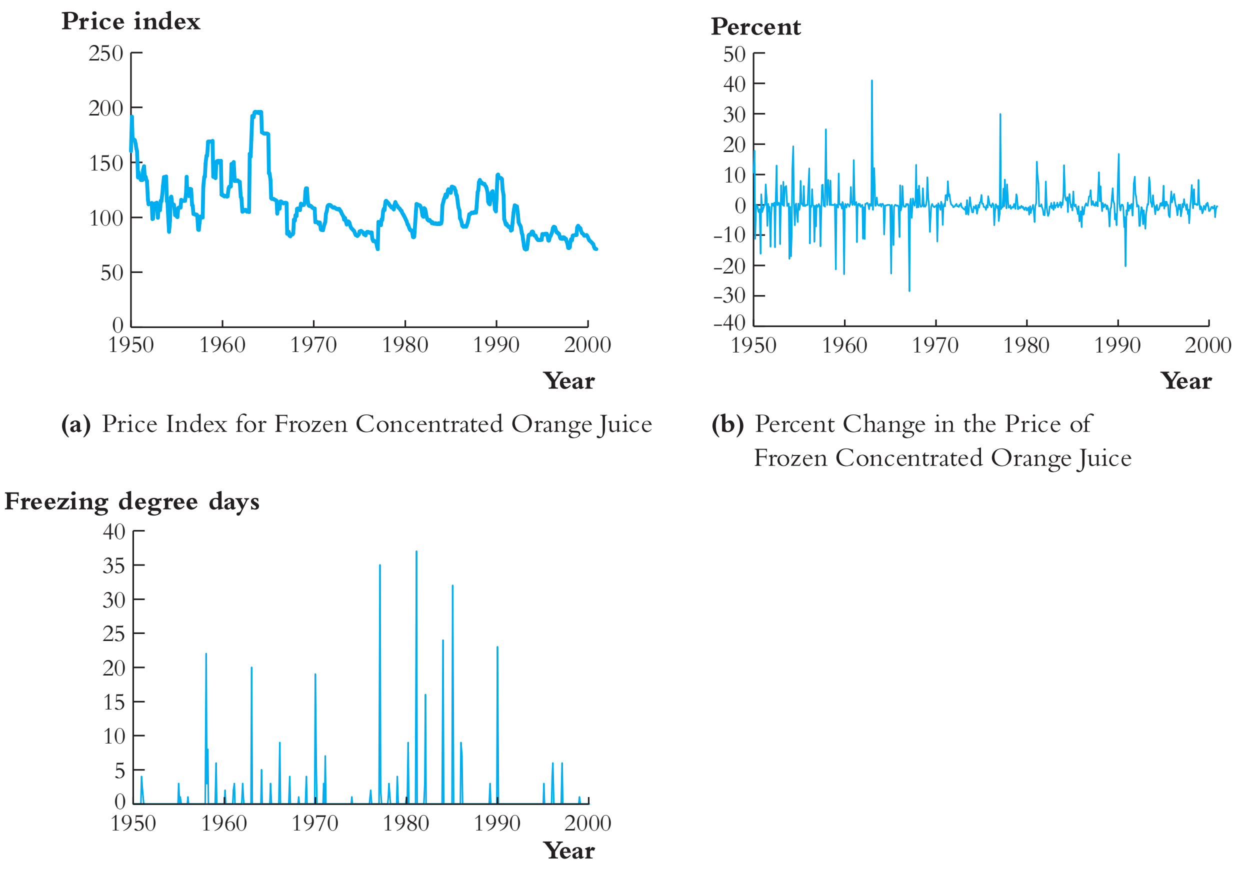
This chapter builds on the material in Sections 14.1 through 14.4 but, with the exception of a subsection (that can be skipped) of the empirical analysis in Section 15.6, does not require the material in Sections 14.5 through 14.7.

15.1 An Initial Taste of the Orange Juice Data

Orlando, the historical center of Florida's orange-growing region, is normally sunny and warm. But now and then there is a cold snap, and if temperatures drop below freezing for too long, the trees drop many of their oranges. If the cold snap is severe, the trees freeze. Following a freeze, the supply of orange juice concentrate falls and its price rises. The timing of the price increases is rather complicated, however. Orange juice concentrate is a "durable," or storable, commodity; that is, it can be stored in its frozen state, albeit at some cost (to run the freezer). Thus the price of orange juice concentrate depends not only on current supply but also on expectations of future supply. A freeze today means that future supplies of concentrate will be low, but because concentrate currently in storage can be used to meet either current or future demand, the price of existing concentrate rises today. But precisely how much does the price of concentrate rise when there is a freeze? The answer to this question is of interest not just to orange juice traders but more generally to economists interested in studying the operations of modern commodity markets. To learn how the price of orange juice changes in response to weather conditions, we must analyze data on orange juice prices and the weather.

Monthly data on the price of frozen orange juice concentrate, its monthly percentage change, and temperatures in the orange-growing region of Florida from January 1950 to December 2000 are plotted in Figure 15.1. The price, plotted in Figure 15.1a, is a measure of the average real price of frozen orange juice concentrate paid by wholesalers. This price was deflated by the overall producer price index for finished goods to eliminate the effects of overall price inflation.

FIGURE 15.1 Orange Juice Prices and Florida Weather, 1950–2000



(a) Price Index for Frozen Concentrated Orange Juice

(b) Percent Change in the Price of Frozen Concentrated Orange Juice

(c) Monthly Freezing Degree Days in Orlando, Florida

There have been large month-to-month changes in the price of frozen concentrated orange juice. Many of the large movements coincide with freezing weather in Orlando, home of many orange groves.

The percentage price change plotted in Figure 15.1b is the percent change in the price over the month. The temperature data plotted in Figure 15.1c are the number of “freezing degree days” at the Orlando, Florida, airport, calculated as the sum of the number of degrees Fahrenheit that the minimum temperature falls below freezing in a given day over all days in the month; for example, in November 1950 the airport temperature dropped below freezing twice, on the 25th (31°) and on the 29th (29°), for a total of 4 freezing degree days $[(32 - 31) + (32 - 29) = 4]$. (The data are described in more detail in Appendix 15.1.) As you can see by comparing the panels in Figure 15.1, the price of orange juice concentrate has large swings, some of which appear to be associated with cold weather in Florida.

We begin our quantitative analysis of the relationship between orange juice price and the weather by using a regression to estimate the amount by which orange juice prices rise when the weather turns cold. The dependent variable is the percentage change in the price over that month [$\%ChgP_t$, where $\%ChgP_t = 100 \times \Delta\ln(P_t^{OJ})$ and P_t^{OJ} is the real price of orange juice]. The regressor is the number of freezing degree days during that month (FDD_t). This regression is estimated using monthly data from January 1950 to December 2000 (as are all regressions in this chapter), for a total of $T = 612$ observations:

$$\widehat{\%ChgP_t} = \underset{(0.22)}{-0.40} + \underset{(0.13)}{0.47}\,FDD_t. \tag{15.1}$$

The standard errors reported in this section are not the usual OLS standard errors, but rather are heteroskedasticity- and autocorrelation-consistent (HAC) standard errors that are appropriate when the error term and regressors are autocorrelated. HAC standard errors are discussed in Section 15.4, and for now they are used without further explanation.

According to this regression, an additional freezing degree day during a month increases the price of orange juice concentrate over that month by 0.47%. In a month with 4 freezing degree days, such as November 1950, the price of orange juice concentrate is estimated to have increased by 1.88% (4 × 0.47% = 1.88%), relative to a month with no days below freezing.

Because the regression in Equation (15.1) includes only a contemporaneous measure of the weather, it does not capture any lingering effects of the cold snap on the orange juice price over the coming months. To capture these we need to consider the effect on prices of both contemporaneous and lagged values of FDD, which in turn can be done by augmenting the regression in Equation (15.1) with, for example, lagged values of FDD over the previous 6 months:

$$\begin{aligned}\widehat{\%ChgP_t} = &-\underset{(0.23)}{0.65} + \underset{(0.14)}{0.47}\,FDD_t + \underset{(0.08)}{0.14}\,FDD_{t-1} + \underset{(0.06)}{0.06}\,FDD_{t-2}\\ &+ \underset{(0.05)}{0.07}\,FDD_{t-3} + \underset{(0.03)}{0.03}\,FDD_{t-4} + \underset{(0.03)}{0.05}\,FDD_{t-5} + \underset{(0.04)}{0.05}\,FDD_{t-6}.\end{aligned} \tag{15.2}$$

Equation (15.2) is a distributed lag regression. The coefficient on FDD_t in Equation (15.2) estimates the percentage increase in prices over the course of the month in which the freeze occurs; an additional freezing degree day is estimated to increase prices that month by 0.47%. The coefficient on the first lag of FDD_t, FDD_{t-1}, estimates the percentage increase in prices arising from a freezing degree day in the preceding month, the coefficient on the second lag estimates the effect of a freezing degree day 2 months ago, and so forth. Equivalently, the coefficient on the first lag of *FDD* estimates the effect of a unit increase in *FDD* 1 month after the freeze occurs. Thus the estimated coefficients in Equation (15.2) are estimates of the effect of a unit increase in FDD_t on current and future values of *%ChgP*; that is, they are estimates of the dynamic effect of FDD_t on $\%ChgP_t$. For example, the 4 freezing degree days in November 1950 are estimated to have increased orange juice prices by 1.88% during November 1950, by an additional $0.56\%(= 4 \times 0.14)$ in December 1950, by an additional $0.24\%(= 4 \times 0.06)$ in January 1951, and so forth.

15.2 Dynamic Causal Effects

Before learning more about the tools for estimating dynamic causal effects, we should spend a moment thinking about what, precisely, is meant by a dynamic causal effect. Having a clear idea about what a dynamic causal effect is leads to a clearer understanding of the conditions under which it can be estimated.

Causal Effects and Time Series Data

Section 1.2 defined a causal effect as the outcome of an ideal randomized controlled experiment: When a horticulturalist randomly applies fertilizer to some tomato plots but not others and then measures the yield, the expected difference in yield between the fertilized and unfertilized plots is the causal effect on tomato yield of the fertilizer. This concept of an experiment, however, is one in which there are multiple subjects (multiple tomato plots or multiple people), so the data are either cross-sectional (the tomato yield at the end of the harvest) or panel data (individual incomes before and after an experimental job training program). By having multiple subjects, it is possible to have both treatment and control groups and thereby to estimate the causal effect of the treatment.

In time series applications, this definition of causal effects in terms of an ideal randomized controlled experiment needs to be modified. To be concrete, consider an important problem of macroeconomics: estimating the effect of an unanticipated change in the short-term interest rate on the current and future economic activity in a given country, as measured by GDP. Taken literally, the randomized controlled experiment of Section 1.2 would entail randomly assigning different economies to treatment and control groups. The central banks in the treatment group would apply the treatment of a random interest rate change, while those in the control group would apply no such random changes; for both groups, economic activity (for example, GDP) would be measured over the next few years. But what if we are interested in estimating this effect for a specific country, say the United States? Then this experiment would entail having different "clones" of the United States as subjects and assigning some clone economies to the treatment group and some to the control group. Obviously, this "parallel universes" experiment is infeasible.

Instead, in time series data it is useful to think of a randomized controlled experiment consisting of the same subject (e.g., the U.S. economy) being given different treatments (randomly chosen changes in interest rates) at different points in time (the 1970s, the 1980s, and so forth). In this framework, the single subject at different times plays the role of both treatment and control group: Sometimes the Fed changes the interest rate, while at other times it does not. Because data are collected over time, it is possible to estimate the dynamic causal effect, that is, the time path of the effect on the outcome of interest of the treatment. For example, a surprise increase in the short-term interest rate of two percentage points, sustained for one quarter, might initially have a negligible effect on output; after two quarters GDP growth might slow, with the greatest slowdown after $1\frac{1}{2}$ years; then over the next 2 years, GDP growth might return to normal. This time path of causal effects is the dynamic causal effect on GDP growth of a surprise change in the interest rate.

As a second example, consider the causal effect on orange juice price changes of a freezing degree day. It is possible to imagine a variety of hypothetical experiments, each yielding a different causal effect. One experiment would be to change the weather in the Florida orange groves, holding weather constant elsewhere—for example, holding weather constant in the Texas grapefruit groves and in other citrus fruit regions. This experiment would measure a partial effect, holding other weather constant. A second experiment might change the weather in all the regions, where the "treatment" is application of overall weather patterns. If weather is correlated across regions for competing crops, then these two dynamic causal effects differ. In this chapter, we consider the causal effect in the latter experiment, that is, the causal effect of applying general weather patterns. This corresponds to measuring the dynamic effect on prices of a change in Florida weather, *not* holding weather constant in other agricultural regions.

Dynamic effects and the distributed lag model. Because dynamic effects necessarily occur over time, the econometric model used to estimate dynamic causal effects needs to incorporate lags. To do so, Y_t can be expressed as a distributed lag of current and r past values of X_t:

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 X_{t-1} + \beta_3 X_{t-2} + \cdots + \beta_{r+1} X_{t-r} + u_t, \qquad (15.3)$$

where u_t is an error term that includes measurement error in Y_t and the effect of omitted determinants of Y_t. The model in Equation (15.3) is called the **distributed lag model** relating X_t, and r of its lags, to Y_t.

As an illustration of Equation (15.3), consider a modified version of the tomato/fertilizer experiment: Because fertilizer applied today might remain in the ground in future years, the horticulturalist wants to determine the effect on tomato yield *over time* of applying fertilizer. Accordingly, she designs a 3-year experiment and randomly divides her plots into four groups: The first is fertilized in only the first year; the second is fertilized in only the second year; the third is fertilized in only the third year; and the fourth, the control group, is never fertilized. Tomatoes are grown annually in each plot, and the third-year harvest is weighed. The three treatment groups are denoted by the binary variables X_{t-2}, X_{t-1}, and X_t, where t represents the third year (the year in which the harvest is weighed), $X_{t-2} = 1$ if the plot is in the first group (fertilized two years earlier), $X_{t-1} = 1$ if the plot was fertilized 1 year earlier, and $X_t = 1$ if the plot was fertilized in the final year. In the context of Equation (15.3) (which applies to a single plot), the effect of being fertilized in the final year is β_1, the effect of being fertilized 1 year earlier is β_2, and the effect of being fertilized 2 years earlier is β_3. If the effect of fertilizer is greatest in the year it is applied, then β_1 would be larger than β_2 and β_3.

More generally, the coefficient on the contemporaneous value of X_t, β_1, is the contemporaneous or immediate effect of a unit change in X_t on Y_t. The coefficient on X_{t-1}, β_2, is the effect on Y_t of a unit change in X_{t-1} or, equivalently, the effect on Y_{t+1} of a unit change in X_t; that is, β_2 is the effect of a unit change in X on Y one period later. In general, the coefficient on X_{t-h} is the effect of a unit change in X on Y after h periods. The dynamic causal effect is the effect of a change in X_t on Y_t, Y_{t+1}, Y_{t+2}, and so forth; that is, it is the sequence of causal effects on current and future values of Y. Thus, in the context of the distributed lag model in Equation (15.3), the dynamic causal effect is the sequence of coefficients β_1, $\beta_2, \ldots, \beta_{r+1}$.

Implications for empirical time series analysis. This formulation of dynamic causal effects in time series data as the expected outcome of an experiment in which different treatment levels are repeatedly applied to the same subject has two implications for empirical attempts to measure the dynamic causal effect with observational time series data. The first implication is that the dynamic causal effect should not change over the sample on which we have data. This in turn is implied by the data being jointly stationary (Key Concept 14.5). As discussed in Section 14.7, the hypothesis that a population regression function is stable over time can be tested using the QLR test for a break, and it is possible to estimate the dynamic causal effect in different subsamples. The second implication is that X must be uncorrelated with the error term, and it is to this implication that we now turn.

Two Types of Exogeneity

Section 12.1 defined an "exogenous" variable as a variable that is uncorrelated with the regression error term and an "endogenous" variable as a variable that is correlated with the error term. This terminology traces to models with multiple equations, in which an "endogenous" variable is determined within the model while an "exogenous" variable is determined outside the model. Loosely speaking, if we are to estimate dynamic causal effects using the distributed lag model in Equation (15.3), the regressors (the X's) must be uncorrelated with the error term. Thus X must be exogenous. Because we are working with time series data, however, we need to refine the definitions of exogeneity. In fact, there are two different concepts of exogeneity that we use here.

The first concept of exogeneity is that the error term has a conditional mean of zero given current and all past values of X_t, that is, that $E(u_t|X_t, X_{t-1}, X_{t-2}, \ldots) = 0$. This modifies the standard conditional mean assumption for multiple regression with cross-sectional data (Assumption #1 in Key Concept 6.4), which requires only that u_t has a conditional mean of zero given the included regressors, that is, $E(u_t|X_t, X_{t-1}, \ldots, X_{t-r}) = 0$. Including all lagged values of X_t in the conditional expectation implies that all the more distant causal effects—all the causal effects beyond lag r—are zero. Thus, under this assumption, the r distributed lag coefficients in Equation (15.3) constitute all the nonzero dynamic causal effects. We can refer to this assumption—that $E(u_t|X_t, X_{t-1}, \ldots) = 0$—as *past and present exogeneity*, but because of the similarity of this definition and the definition of exogeneity in Chapter 12, we just use the term **exogeneity**.

The second concept of exogeneity is that the error term has mean zero, given all past, present, and *future* values of X_t, that is, that $E(u_t|\ldots, X_{t+2}, X_{t+1}, X_t, X_{t-1}, X_{t-2}, \ldots) = 0$. This is called **strict exogeneity**; for clarity, we also call it *past, present, and future exogeneity*. The reason for introducing the concept of strict exogeneity is that, when X is strictly exogenous, there are more efficient estimators of dynamic causal effects than the OLS estimators of the coefficients of the distributed lag regression in Equation (15.3).

The difference between exogeneity (past and present) and strict exogeneity (past, present, and future) is that strict exogeneity includes future values of X in the conditional expectation. Thus strict exogeneity implies exogeneity, but not the reverse. One way to understand the difference between the two concepts is to consider the implications of these definitions for correlations between X and u. If X is (past and present) exogenous, then u_t is uncorrelated with current and past values of X_t. If X is strictly exogenous, then in addition u_t is uncorrelated with *future* values of X_t. For example, if a change in Y_t causes *future* values of X_t to change, then X_t is not strictly exogenous even though it might be (past and present) exogenous.

As an illustration, consider the hypothetical multiyear tomato/fertilizer experiment described following Equation (15.3). Because the fertilizer is randomly applied in the hypothetical experiment, it is exogenous. Because tomato yield today does not depend on the amount of fertilizer applied in the future, the fertilizer time series is also strictly exogenous.

As a second illustration, consider the orange juice price example, in which Y_t is the monthly percentage change in orange juice prices and X_t is the number of freezing degree days in that month. From the perspective of orange juice markets, we can think of the weather—the number of freezing degree days—as if it were randomly assigned, in the sense that the weather is outside human control. If the effect of FDD is linear and if it has no effect on prices after r months, then it follows that the weather is exogenous. But is the weather *strictly* exogenous? If the conditional mean of u_t given future FDD is nonzero, then FDD is not strictly exogenous. Answering this question requires thinking carefully about what, precisely, is contained in u_t. In particular, if OJ market participants use forecasts of FDD when they decide how much they will buy or sell at a given price, then OJ prices, and thus the error term u_t, could incorporate information about future FDD that would make u_t a useful predictor of FDD. This means that u_t will be correlated with future values of FDD_t. According to this logic, because u_t includes forecasts of future Florida weather, FDD would be (past and present) exogenous but not *strictly* exogenous. The difference between this and the tomato/fertilizer example is that, while tomato plants are unaffected by future fertilization, OJ market participants *are* influenced by forecasts of future Florida weather. We return to the question of whether FDD is strictly exogenous when we analyze the orange juice price data in more detail in Section 15.6.

The two definitions of exogeneity are summarized in Key Concept 15.1.

15.3 Estimation of Dynamic Causal Effects with Exogenous Regressors

If X is exogenous, then its dynamic causal effect on Y can be estimated by OLS estimation of the distributed lag regression in Equation (15.4). This section summarizes the conditions under which these OLS estimators lead to valid statistical inferences and introduces dynamic multipliers and cumulative dynamic multipliers.

KEY CONCEPT 15.1

The Distributed Lag Model and Exogeneity

In the distributed lag model

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 X_{t-1} + \beta_3 X_{t-2} + \cdots + \beta_{r+1} X_{t-r} + u_t, \qquad (15.4)$$

there are two different types of exogeneity, that is, two different exogeneity conditions:

Past and present exogeneity (exogeneity):

$$E(u_t | X_t, X_{t-1}, X_{t-2}, \ldots) = 0; \qquad (15.5)$$

Past, present, and future exogeneity (strict exogeneity):

$$E(u_t | \ldots, X_{t+2}, X_{t+1}, X_t, X_{t-1}, X_{t-2}, \ldots) = 0. \qquad (15.6)$$

If X is strictly exogenous, it is exogenous, but exogeneity does not imply strict exogeneity.

The Distributed Lag Model Assumptions

The four assumptions of the distributed lag regression model are similar to the four assumptions for the cross-sectional multiple regression model (Key Concept 6.4), modified for time series data.

The first assumption is that X is exogenous, which extends the zero conditional mean assumption for cross-sectional data to include all lagged values of X. As discussed in Section 15.2, this assumption implies that the r distributed lag coefficients in Equation (15.3) constitute all the nonzero dynamic causal effects. In this sense, the population regression function summarizes the entire dynamic effect on Y of a change in X.

The second assumption has two parts: Part (a) requires that the variables have a stationary distribution, and part (b) requires that they become independently distributed when the amount of time separating them becomes large. This assumption is the same as the corresponding assumption for the ADL model (the second assumption in Key Concept 14.6), and the discussion of this assumption in Section 14.4 applies here as well.

The third assumption is that large outliers are unlikely, made mathematically precise by assuming that the variables have more than eight nonzero, finite moments.

The Distributed Lag Model Assumptions

KEY CONCEPT

15.2

The distributed lag model is given in Key Concept 15.1 [Equation (15.4)], where

1. X is exogenous, that is, $E(u_t|X_t, X_{t-1}, X_{t-2}, \ldots) = 0$;
2. (a) The random variables Y_t and X_t have a stationary distribution, and
   (b) (Y_t, X_t) and (Y_{t-j}, X_{t-j}) become independent as j gets large;
3. Large outliers are unlikely: Y_t and X_t have more than eight nonzero, finite moments; and
4. There is no perfect multicollinearity.

This is stronger than the assumption of four finite moments that is used elsewhere in this book. As discussed in Section 15.4, this stronger assumption is used in the mathematics behind the HAC variance estimator.

The fourth assumption, which is the same as in the cross-sectional multiple regression model, is that there is no perfect multicollinearity.

The distributed lag regression model and assumptions are summarized in Key Concept 15.2.

Extension to additional X's. The distributed lag model extends directly to multiple X's: The additional X's and their lags are simply included as regressors in the distributed lag regression, and the assumptions in Key Concept 15.2 are modified to include these additional regressors. Although the extension to multiple X's is conceptually straightforward, it complicates the notation, obscuring the main ideas of estimation and inference in the distributed lag model. For this reason, the case of multiple X's is not treated explicitly in this chapter but is left as a straightforward extension of the distributed lag model with a single X.

Autocorrelated u_t, Standard Errors, and Inference

In the distributed lag regression model, the error term u_t can be autocorrelated; that is, u_t can be correlated with its lagged values. This autocorrelation arises because, in time series data, the omitted factors included in u_t can themselves be serially correlated. For example, suppose that the demand for orange juice also depends on income, so one factor that influences the price of orange juice is income, specifically, the aggregate income of potential orange juice consumers. Then aggregate income is an omitted variable in the distributed lag regression of orange juice price changes against freezing degree days. Aggregate income, however, is serially correlated: Income tends to fall in recessions and rise in expansions. Thus, income is serially correlated, and, because it is part of the error term, u_t will be serially correlated. This example is typical: Because omitted determinants of Y are themselves serially correlated, in general u_t in the distributed lag model will be serially correlated.

The autocorrelation of u_t does not affect the consistency of OLS, nor does it introduce bias. If, however, the errors are autocorrelated, then in general the usual OLS standard errors are inconsistent and a different formula must be used. Thus serial correlation of the errors is analogous to heteroskedasticity: The homoskedasticity-only standard errors are "wrong" when the errors are in fact heteroskedastic, in the sense that using homoskedasticity-only standard errors results in misleading statistical inferences when the errors are heteroskedastic. Similarly, when the errors are serially correlated, standard errors predicated upon i.i.d. errors are "wrong" in the sense that they result in misleading statistical inferences. The solution to this problem is to use heteroskedasticity- and autocorrelation-consistent (HAC) standard errors, the topic of Section 15.4.

Dynamic Multipliers and Cumulative Dynamic Multipliers

Another name for the dynamic causal effect is the dynamic multiplier. The cumulative dynamic multipliers are the cumulative causal effects, up to a given lag; thus the cumulative dynamic multipliers measure the cumulative effect on Y of a change in X.

Dynamic multipliers. The effect of a unit change in X on Y after h periods, which is β_{h+1} in Equation (15.4), is called the h-period **dynamic multiplier**. Thus the dynamic multipliers relating X to Y are the coefficients on X_t and its lags in Equation (15.4). For example, β_2 is the one-period dynamic multiplier, β_3 is the two-period dynamic multiplier, and so forth. In this terminology, the zero-period (or contemporaneous) dynamic multiplier, or **impact effect**, is β_1, the effect on Y of a change in X in the same period.

Because the dynamic multipliers are estimated by the OLS regression coefficients, their standard errors are the HAC standard errors of the OLS regression coefficients.

Cumulative dynamic multipliers. The h-period **cumulative dynamic multiplier** is the cumulative effect of a unit change in X on Y over the next h periods. Thus the cumulative dynamic multipliers are the cumulative sum of the dynamic multipliers. In terms of the coefficients of the distributed lag regression in Equation (15.4), the zero-period cumulative multiplier is β_1, the one-period cumulative multiplier is $\beta_1 + \beta_2$, and the h-period cumulative dynamic multiplier is $\beta_1 + \beta_2 + \cdots + \beta_{h+1}$. The sum of all the individual dynamic multipliers, $\beta_1 + \beta_2 + \cdots + \beta_{r+1}$, is the cumulative long-run effect on Y of a change in X and is called the **long-run cumulative dynamic multiplier**.

For example, consider the regression in Equation (15.2). The immediate effect of an additional freezing degree day is that the price of orange juice concentrate rises by 0.47%. The cumulative effect of a price change over the next month is the sum of the impact effect and the dynamic effect one month ahead; thus the cumulative effect on prices is the initial increase of 0.47% plus the subsequent smaller increase of 0.14% for a total of 0.61%. Similarly, the cumulative dynamic multiplier over 2 months is 0.47% + 0.14% + 0.06% = 0.67%.

The cumulative dynamic multipliers can be estimated directly using a modification of the distributed lag regression in Equation (15.4). This modified regression is

$$Y_t = \delta_0 + \delta_1 \Delta X_t + \delta_2 \Delta X_{t-1} + \delta_3 \Delta X_{t-2} + \cdots + \delta_r \Delta X_{t-r+1} + \delta_{r+1} X_{t-r} + u_t. \tag{15.7}$$

The coefficients in Equation (15.7), $\delta_1, \delta_2, \ldots, \delta_{r+1}$, are in fact the cumulative dynamic multipliers. This can be shown by a bit of algebra (Exercise 15.5), which demonstrates that the population regressions in Equations (15.7) and (15.4) are equivalent, where $\delta_0 = \beta_0$, $\delta_1 = \beta_1$, $\delta_2 = \beta_1 + \beta_2$, $\delta_3 = \beta_1 + \beta_2 + \beta_3$, and so forth. The coefficient on X_{t-r}, δ_{r+1}, is the long-run cumulative dynamic multiplier; that is, $\delta_{r+1} = \beta_1 + \beta_2 + \beta_3 + \cdots + \beta_{r+1}$. Moreover, the OLS estimators of the coefficients in Equation (15.7) are the same as the corresponding cumulative sum of the OLS estimators in Equation (15.4). For example, $\hat{\delta}_2 = \hat{\beta}_1 + \hat{\beta}_2$. The main benefit of estimating the cumulative dynamic multipliers using the specification in Equation (15.7) is that, because the OLS estimators of the regression coefficients are estimators of the cumulative dynamic multipliers, the HAC standard errors of the coefficients in Equation (15.7) are the HAC standard errors of the cumulative dynamic multipliers.

15.4 Heteroskedasticity- and Autocorrelation-Consistent Standard Errors

If the error term u_t is autocorrelated, then OLS coefficient estimators are consistent, but in general the usual OLS standard errors for cross-sectional data are not. This means that conventional statistical inferences—hypothesis tests and confidence intervals—based on the usual OLS standard errors will, in general, be misleading.

For example, confidence intervals constructed as the OLS estimator ±1.96 conventional standard errors need not contain the true value in 95% of repeated samples, even if the sample size is large. This section begins with a derivation of the correct formula for the variance of the OLS estimator with autocorrelated errors, then turns to heteroskedasticity- and autocorrelation-consistent (HAC) standard errors.

This section covers HAC standard errors for regression with time series data. Chapter 10 introduced a type of HAC standard errors, clustered standard errors, which are appropriate for panel data. Although clustered standard errors for panel data and HAC standard errors for time series data have the same goal, the different data structures lead to different formulas. This section is self-contained, and Chapter 10 is not a prerequisite.

Distribution of the OLS Estimator with Autocorrelated Errors

To keep things simple, consider the OLS estimator $\hat{\beta}_1$ in the distributed lag regression model with no lags, that is, the linear regression model with a single regressorX_t:

$$Y_t = \beta_0 + \beta_1 X_t + u_t, \tag{15.8}$$

where the assumptions of Key Concept 15.2 are satisfied. This section shows that the variance of $\hat{\beta}_1$ can be written as the product of two terms: the expression for var($\hat{\beta}_1$), applicable if u_t is not serially correlated, multiplied by a correction factor that arises from the autocorrelation in u_t or, more precisely, the autocorrelation in $(X_t - \mu_X)u_t$.

As shown in Appendix 4.3, the formula for the OLS estimator $\hat{\beta}_1$ in Key Concept 4.2 can be rewritten as

$$\hat{\beta}_1 = \beta_1 + \frac{\frac{1}{T}\sum_{t=1}^{T}(X_t - \overline{X})u_t}{\frac{1}{T}\sum_{t=1}^{T}(X_t - \overline{X})^2}, \tag{15.9}$$

where Equation (15.9) is Equation (4.30) with a change of notation so that i and n are replaced by t and T. Because $\overline{X} \xrightarrow{p} \mu_X$ and $\frac{1}{T}\sum_{t=1}^{T}(X_t - \overline{X})^2 \xrightarrow{p} \sigma_X^2$, in large samples $\hat{\beta}_1 - \beta_1$ is approximately given by

$$\hat{\beta}_1 - \beta_1 \cong \frac{\frac{1}{T}\sum_{t=1}^{T}(X_t - \mu_X)u_t}{\sigma_X^2} = \frac{\frac{1}{T}\sum_{t=1}^{T}v_t}{\sigma_X^2} = \frac{\overline{v}}{\sigma_X^2}, \tag{15.10}$$

where $v_t = (X_t - \mu_X)u_t$ and $\overline{v} = \frac{1}{T}\Sigma_{t=1}^{T} v_t$. Thus

$$\text{var}(\hat{\beta}_1) = \text{var}\left(\frac{\overline{v}}{\sigma_X^2}\right) = \frac{\text{var}(\overline{v})}{(\sigma_X^2)^2}. \tag{15.11}$$

If v_t is i.i.d.—as assumed for cross-sectional data in Key Concept 4.3—then $\text{var}(\overline{v}) = \text{var}(v_t)/T$ and the formula for the variance of $\hat{\beta}_1$ from Key Concept 4.4 applies. If, however, u_t and X_t are not independently distributed over time, then in general v_t will be serially correlated, so $\text{var}(\overline{v}) \neq \text{var}(v_t)/T$ and Key Concept 4.4 does not apply. Instead, if v_t is serially correlated, the variance of $\overline{v}$ is given by

$$\begin{aligned}
\text{var}(\overline{v}) &= \text{var}\,[(v_1 + v_2 + \cdots + v_T)/T] \\
&= [\text{var}(v_1) + \text{cov}(v_1, v_2) + \cdots + \text{cov}(v_1, v_T) \\
&\quad + \text{cov}(v_2, v_1) + \text{var}(v_2) + \cdots + \text{var}(v_T)]/T^2 \\
&= [T\text{var}(v_t) + 2(T-1)\text{cov}(v_t, v_{t-1}) \\
&\quad + 2(T-2)\text{cov}(v_t, v_{t-2}) + \cdots + 2\text{cov}(v_t, v_{t-T+1})]/T^2 \\
&= \frac{\sigma_v^2}{T} f_T,
\end{aligned} \tag{15.12}$$

where

$$f_T = 1 + 2\sum_{j=1}^{T-1}\left(\frac{T-j}{T}\right)\rho_j, \tag{15.13}$$

where $\rho_j = \text{corr}(v_t, v_{t-j})$. In large samples, f_T tends to the limit, $f_T \longrightarrow f_\infty = 1 + 2\Sigma_{j=1}^{\infty}\rho_j$.

Combining the expressions in Equation (15.10) for $\hat{\beta}_1$ and Equation (15.12) for $\text{var}(\overline{v})$ gives the formula for the variance of $\hat{\beta}_1$ when v_t is autocorrelated:

$$\text{var}(\hat{\beta}_1) = \left[\frac{1}{T}\frac{\sigma_v^2}{(\sigma_X^2)^2}\right] f_T, \tag{15.14}$$

where f_T is given in Equation (15.13).

Equation (15.14) expresses the variance of $\hat{\beta}_1$ as the product of two terms. The first, in square brackets, is the formula for the variance of $\hat{\beta}_1$ given in Key Concept 4.4, which applies in the absence of serial correlation. The second is the factor f_T, which adjusts this formula for serial correlation. Because of this additional factor

f_T in Equation (15.14), the usual OLS standard error computed using Equation (5.4) is incorrect if the errors are serially correlated: If $v_t = (X_t - \mu_X)u_t$ is serially correlated, the estimator of the variance is off by the factor f_T.

HAC Standard Errors

If the factor f_T, defined in Equation (15.13), was known, then the variance of $\hat{\beta}_1$ could be estimated by multiplying the usual cross-sectional estimator of the variance by f_T. This factor, however, depends on the unknown autocorrelations of v_t, so it must be estimated. The estimator of the variance of $\hat{\beta}_1$ that incorporates this adjustment is consistent whether or not there is heteroskedasticity and whether or not v_t is autocorrelated. Accordingly, this estimator is called the **heteroskedasticity- and autocorrelation-consistent (HAC)** estimator of the variance of $\hat{\beta}_1$, and the square root of the HAC variance estimator is the **HAC standard error** of $\hat{\beta}_1$.

The HAC variance formula. The heteroskedasticity- and autocorrelation-consistent estimator of the variance of $\hat{\beta}_1$ is

$$\tilde{\sigma}^2_{\hat{\beta}_1} = \hat{\sigma}^2_{\hat{\beta}_1} \hat{f}_T, \tag{15.15}$$

where $\hat{\sigma}^2_{\hat{\beta}_1}$ is the estimator of the variance of $\hat{\beta}_1$ in the absence of serial correlation, given in Equation (5.4), and where $\hat{f}_T$ is an estimator of the factor f_T in Equation (15.13).

The task of constructing a consistent estimator $\hat{f}_T$ is challenging. To see why, consider two extremes. At one extreme, given the formula in Equation (15.13), it might seem natural to replace the population autocorrelations ρ_j with the sample autocorrelations $\hat{\rho}_j$ [defined in Equation (14.6)], yielding the estimator $1 + 2\sum_{j=1}^{T-1}(\frac{T-j}{T})\hat{\rho}_j$. But this estimator contains so many estimated autocorrelations that it is inconsistent. Intuitively, because each of the estimated autocorrelations contains an estimation error, by estimating so many autocorrelations the estimation error in this estimator of f_T remains large even in large samples. At the other extreme, one could imagine using only a few sample autocorrelations, for example, only the first sample autocorrelation, and ignoring all the higher autocorrelations. Although this estimator eliminates the problem of estimating too many autocorrelations, it has a different problem: It is inconsistent because it ignores the additional autocorrelations that appear in Equation (15.13). In short, using too many sample autocorrelations makes the estimator have a large variance, but using too few autocorrelations ignores the autocorrelations at higher lags, so in either of these extreme cases the estimator is inconsistent.

Estimators of f_T used in practice strike a balance between these two extreme cases by choosing the number of autocorrelations to include in a way that depends on the sample size T. If the sample size is small, only a few autocorrelations are used, but if the sample size is large, more autocorrelations are included (but still far fewer than T). Specifically, let $\hat{f}_T$ be given by

$$\hat{f}_T = 1 + 2\sum_{j=1}^{m-1}\left(\frac{m-j}{m}\right)\widetilde{\rho}_j, \tag{15.16}$$

where $\widetilde{\rho}_j = \sum_{t=j+1}^{T} \hat{v}_t\hat{v}_{t-j}/\sum_{t=1}^{T}\hat{v}_t^2$, where $\hat{v}_t = (X_t - \overline{X})\hat{u}_t$ (as in the definition of $\hat{\sigma}^2_{\hat{\beta}_1}$). The parameter m in Equation (15.16) is called the **truncation parameter** of the HAC estimator because the sum of autocorrelations is shortened, or truncated, to include only $m - 1$ autocorrelations instead of the $T - 1$ autocorrelations appearing in the population formula in Equation (15.13).

For $\hat{f}_T$ to be consistent, m must be chosen so that it is large in large samples, although still much less than T. One guideline for choosing m in practice is to use the formula

$$m = 0.75T^{1/3}, \tag{15.17}$$

rounded to an integer. This formula, which is based on the assumption that there is a moderate amount of autocorrelation in v_t, gives a benchmark rule for determining m as a function of the number of observations in the regression.[1]

The value of the truncation parameter m resulting from Equation (15.17) can be modified using your knowledge of the series at hand. On the one hand, if there is a great deal of serial correlation in v_t, then you could increase m beyond the value from Equation (15.17). On the other hand, if v_t has little serial correlation, you could decrease m. Because of the ambiguity associated with the choice of m, it is good practice to try one or two alternative values of m for at least one specification to make sure your results are not sensitive to m.

The HAC estimator in Equation (15.15), with $\hat{f}_T$ given in Equation (15.16), is called the **Newey–West variance estimator**, after the econometricians Whitney Newey and Kenneth West, who proposed it. They showed that, when used along with a rule like that in Equation (15.17), under general assumptions this estimator is a consistent estimator of the variance of $\hat{\beta}_1$ (Newey and West, 1987). Their

[1] Equation (15.17) gives the "best" choice of m if u_t and X_t are first-order autoregressive processes with first autocorrelation coefficients 0.5, where "best" means the estimator that minimizes $E(\widetilde{\sigma}^2_{\hat{\beta}_1} - \sigma^2_{\hat{\beta}_1})^2$. Equation (15.17) is based on a more general formula derived by Andrews [1991, Equation (5.3)].

proofs (and those in Andrews, 1991) assume that v_t has more than four moments, which in turn is implied by X_t and u_t having more than eight moments, and this is the reason that the third assumption in Key Concept 15.2 is that X_t and u_t have more than eight moments.

Other HAC estimators. The Newey–West variance estimator is not the only HAC estimator. For example, the weights $(m - j)/m$ in Equation (15.16) can be replaced by different weights. If different weights are used, then the rule for choosing the truncation parameter in Equation (15.17) no longer applies and a different rule, developed for those weights, should be used instead. Discussion of HAC estimators using other weights goes beyond the scope of this book. For more information on this topic, see Hayashi (2000, Section 6.6).

Extension to multiple regression. All the issues discussed in this section generalize to the distributed lag regression model in Key Concept 15.1 with multiple lags and, more generally, to the multiple regression model with serially correlated errors. In particular, if the error term is serially correlated, then the usual OLS standard errors are an unreliable basis for inference and HAC standard errors should be used instead. If the HAC variance estimator used is the Newey–West estimator [the HAC variance estimator based on the weights $(m - j)/m$], then the truncation parameter m can be chosen according to the rule in Equation (15.17) whether there is a single regressor or multiple regressors. The formula for HAC standard errors in multiple regression is incorporated into modern regression software designed for use with time series data. Because this formula involves matrix algebra, we omit it here and instead refer the reader to Hayashi (2000, Section 6.6) for the mathematical details.

HAC standard errors are summarized in Key Concept 15.3.

15.5 Estimation of Dynamic Causal Effects with Strictly Exogenous Regressors

When X_t is strictly exogenous, two alternative estimators of dynamic causal effects are available. The first such estimator involves estimating an autoregressive distributed lag (ADL) model instead of a distributed lag model and calculating the dynamic multipliers from the estimated ADL coefficients. This method can entail estimating fewer coefficients than OLS estimation of the distributed lag model, thus potentially reducing estimation error. The second method is to estimate the coefficients of the distributed lag model, using **generalized least squares (GLS)**

HAC Standard Errors

KEY CONCEPT 15.3

The problem: The error term u_t in the distributed lag regression model in Key Concept 15.1 can be serially correlated. If so, the OLS coefficient estimators are consistent but in general the usual OLS standard errors are not, resulting in misleading hypothesis tests and confidence intervals.

The solution: Standard errors should be computed using a heteroskedasticity- and autocorrelation-consistent (HAC) estimator of the variance. The HAC estimator involves estimates of $m - 1$ autocovariances as well as the variance; in the case of a single regressor, the relevant formulas are given in Equations (15.15) and (15.16).

In practice, using HAC standard errors entails choosing the truncation parameter m. To do so, use the formula in Equation (15.17) as a benchmark, then increase or decrease m depending on whether your regressors and errors have high or low serial correlation.

instead of OLS. Although the same number of coefficients in the distributed lag model are estimated by GLS as by OLS, the GLS estimator has a smaller variance. To keep the exposition simple, these two estimation methods are initially laid out and discussed in the context of a distributed lag model with a single lag and AR(1) errors. The potential advantages of these two estimators are greatest, however, when many lags appear in the distributed lag model, so these estimators are then extended to the general distributed lag model with higher-order autoregressive errors.

The Distributed Lag Model with AR(1) Errors

Suppose that the causal effect on Y of a change in X lasts for only two periods; that is, it has an initial impact effect β_1 and an effect in the next period of β_2, but no effect thereafter. Then the appropriate distributed lag regression model is the distributed lag model with only current and past values of X_{t-1}:

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 X_{t-1} + u_t. \tag{15.18}$$

As discussed in Section 15.2, in general the error term u_t in Equation (15.18) is serially correlated. One consequence of this serial correlation is that, if the distributed lag coefficients are estimated by OLS, then inference based on the usual OLS standard errors can be misleading. For this reason, Sections 15.3 and 15.4

emphasized the use of HAC standard errors when β_1 and β_2 in Equation (15.18) are estimated by OLS.

In this section, we take a different approach toward the serial correlation in u_t. This approach, which is possible if X_t is strictly exogenous, involves adopting an autoregressive model for the serial correlation in u_t, then using this AR model to derive some estimators that can be more efficient than the OLS estimator in the distributed lag model.

Specifically, suppose that u_t follows the AR(1) model

$$u_t = \phi_1 u_{t-1} + \widetilde{u}_t, \tag{15.19}$$

where ϕ_1 is the autoregressive parameter, $\widetilde{u}_t$ is serially uncorrelated, and no intercept is needed because $E(u_t) = 0$. Equations (15.18) and (15.19) imply that the distributed lag model with a serially correlated error can be rewritten as an autoregressive distributed lag model with a serially uncorrelated error. To do so, lag each side of Equation (15.18) and subtract ϕ_1 multiplied by this lag from each side:

$$\begin{aligned} Y_t - \phi_1 Y_{t-1} &= (\beta_0 + \beta_1 X_t + \beta_2 X_{t-1} + u_t) - \phi_1(\beta_0 + \beta_1 X_{t-1} + \beta_2 X_{t-2} + u_{t-1}) \\ &= \beta_0 + \beta_1 X_t + \beta_2 X_{t-1} - \phi_1\beta_0 - \phi_1\beta_1 X_{t-1} - \phi_1\beta_2 X_{t-2} + \widetilde{u}_t, \quad (15.20) \end{aligned}$$

where the second equality uses $\widetilde{u}_t = u_t - \phi_1 u_{t-1}$. Collecting terms in Equation (15.20), we have that

$$Y_t = \alpha_0 + \phi_1 Y_{t-1} + \delta_0 X_t + \delta_1 X_{t-1} + \delta_2 X_{t-2} + \widetilde{u}_t, \tag{15.21}$$

where

$$\alpha_0 = \beta_0(1 - \phi_1),\ \delta_0 = \beta_1,\ \delta_1 = \beta_2 - \phi_1\beta_1, \text{ and } \delta_2 = -\phi_1\beta_2, \tag{15.22}$$

where β_0, β_1, and β_2 are the coefficients in Equation (15.18) and ϕ_1 is the autocorrelation coefficient in Equation (15.19).

Equation (15.21) is an ADL model that includes a contemporaneous value of X and two of its lags. We will refer to Equation (15.21) as the ADL representation of the distributed lag model with autoregressive errors given in Equations (15.18) and (15.19).

The terms in Equation (15.20) can be reorganized differently to obtain an expression that is equivalent to Equations (15.21) and (15.22). Let $\widetilde{Y}_t = Y_t - \phi_1 Y_{t-1}$ be the **quasi-difference** of Y_t ("quasi" because it is not the first difference, the difference between Y_t and Y_{t-1}; rather, it is the difference between Y_t and $\phi_1 Y_{t-1}$).

Similarly, let $\widetilde{X}_t = X_t - \phi_1 X_{t-1}$ be the quasi-difference of X_t. Then Equation (15.20) can be written

$$\widetilde{Y}_t = \alpha_0 + \beta_1 \widetilde{X}_t + \beta_2 \widetilde{X}_{t-1} + \widetilde{u}_t. \tag{15.23}$$

We will refer to Equation (15.23) as the quasi-difference representation of the distributed lag model with autoregressive errors given in Equations (15.18) and (15.19).

The ADL model Equation (15.21) [with the parameter restrictions in Equation (15.22)] and the quasi-difference model in Equation (15.23) are equivalent. In both models, the error term, $\widetilde{u}_t$, is serially uncorrelated. The two representations, however, suggest different estimation strategies. But before discussing those strategies, we turn to the assumptions under which they yield consistent estimators of the dynamic multipliers, β_1 and β_2.

The conditional mean zero assumption in the ADL(1,2) and quasi-difference models. Because Equations (15.21) [with the restrictions in Equation (15.22)] and (15.23) are equivalent, the conditions for their estimation are the same, so for convenience we consider Equation (15.23).

The quasi-difference model in Equation (15.23) is a distributed lag model involving the quasi-differenced variables with a serially uncorrelated error. Accordingly, the conditions for OLS estimation of the coefficients in Equation (15.23) are the least squares assumptions for the distributed lag model in Key Concept 15.2, expressed in terms of $\widetilde{u}_t$ and $\widetilde{X}_t$. The critical assumption here is the first assumption, which, applied to Equation (15.23), is that $\widetilde{X}_t$ is exogenous; that is,

$$E(\widetilde{u}_t | \widetilde{X}_t, \widetilde{X}_{t-1}, \dots) = 0, \tag{15.24}$$

where letting the conditional expectation depend on distant lags of $\widetilde{X}_t$ ensures that no additional lags of $\widetilde{X}_t$, other than those appearing in Equation (15.23), enter the population regression function.

Because $\widetilde{X}_t = X_t - \phi_1 X_{t-1}$, so $X_t = \widetilde{X}_t + \phi_1 X_{t-1}$, conditioning on $\widetilde{X}_t$ and all of its lags is equivalent to conditioning on X_t and all of its lags. Thus the conditional expectation condition in Equation (15.24) is equivalent to the condition that $E(\widetilde{u}_t | X_t, X_{t-1}, \dots) = 0$. Furthermore, because $\widetilde{u}_t = u_t - \phi_1 u_{t-1}$, this condition in turn implies that

$$\begin{aligned} 0 &= E(\widetilde{u}_t | X_t, X_{t-1}, \dots) \\ &= E(u_t - \phi_1 u_{t-1} | X_t, X_{t-1}, \dots) \\ &= E(u_t | X_t, X_{t-1}, \dots) - \phi_1 E(u_{t-1} | X_t, X_{t-1}, \dots). \end{aligned} \tag{15.25}$$

For the equality in Equation (15.25) to hold for general values of ϕ_1, it must be the case that both $E(u_t|X_t, X_{t-1}, \ldots) = 0$ and $E(u_{t-1}|X_t, X_{t-1}, \ldots) = 0$. By shifting the time subscripts forward one time period, the condition that $E(u_{t-1}|X_t, X_{t-1}, \ldots) = 0$ can be rewritten as

$$E(u_t|X_{t+1}, X_t, X_{t-1}, \ldots) = 0, \tag{15.26}$$

which (by the law of iterated expectations) implies that $E(u_t|X_t, X_{t-1}, \ldots) = 0$. In summary, having the zero conditional mean assumption in Equation (15.24) hold for general values of ϕ_1 is equivalent to having the condition in Equation (15.26) hold.

The condition in Equation (15.26) is implied by X_t being strictly exogenous, but it is *not* implied by X_t being (past and present) exogenous. Thus the least squares assumptions for estimation of the distributed lag model in Equation (15.23) hold if X_t is strictly exogenous, but it is not enough that X_t be (past and present) exogenous.

Because the ADL representation [Equations (15.21) and (15.22)] is equivalent to the quasi-differenced representation [Equation (15.23)], the conditional mean assumption needed to estimate the coefficients of the quasi-differenced representation [that $E(u_t|X_{t+1}, X_t, X_{t-1}, \ldots) = 0$] is also the conditional mean assumption for consistent estimation of the coefficients of the ADL representation.

We now turn to the two estimation strategies suggested by these two representations: estimation of the ADL coefficients and estimation of the coefficients of the quasi-difference model.

OLS Estimation of the ADL Model

The first strategy is to use OLS to estimate the coefficients in the ADL model in Equation (15.21). As the derivation leading to Equation (15.21) shows, including the lag of Y and the extra lag of X as regressors makes the error term serially uncorrelated (under the assumption that the error follows a first order autoregression). Thus the usual OLS standard errors can be used; that is, HAC standard errors are not needed when the ADL model coefficients in Equation (15.21) are estimated by OLS.

The estimated ADL coefficients are not themselves estimates of the dynamic multipliers, but the dynamic multipliers can be computed from the ADL coefficients. A general way to compute the dynamic multipliers is to express the estimated regression function as a function of current and past values of X_t, that is, to eliminate Y_t from the estimated regression function. To do so, repeatedly substitute expressions for lagged values of Y_t into the estimated regression function. Specifically, consider the estimated regression function

$$\hat{Y}_t = \hat{\phi}_1 Y_{t-1} + \hat{\delta}_0 X_t + \hat{\delta}_1 X_{t-1} + \hat{\delta}_2 X_{t-2}, \tag{15.27}$$

where the estimated intercept has been omitted because it does not enter any expression for the dynamic multipliers. Lagging both sides of Equation (15.27) yields $\hat{Y}_{t-1} = \hat{\phi}_1 Y_{t-2} + \hat{\delta}_0 X_{t-1} + \hat{\delta}_1 X_{t-2} + \hat{\delta}_2 X_{t-3}$, so replacing $\hat{Y}_{t-1}$ in Equation (15.27) by this expression for $\hat{Y}_{t-1}$ and collecting terms yields

$$\begin{aligned}\hat{Y}_t &= \hat{\phi}_1(\hat{\phi}_1 Y_{t-2} + \hat{\delta}_0 X_{t-1} + \hat{\delta}_1 X_{t-2} + \hat{\delta}_2 X_{t-3}) + \hat{\delta}_0 X_t + \hat{\delta}_1 X_{t-1} + \hat{\delta}_2 X_{t-2} \\ &= \hat{\delta}_0 X_t + (\hat{\delta}_1 + \hat{\phi}_1 \hat{\delta}_0) X_{t-1} + (\hat{\delta}_2 + \hat{\phi}_1 \hat{\delta}_1) X_{t-2} + \hat{\phi}_1 \hat{\delta}_2 X_{t-3} + \hat{\phi}_1^2 Y_{t-2}. \end{aligned} \tag{15.28}$$

Repeating this process by repeatedly substituting expressions for Y_{t-2}, Y_{t-3}, and so forth yields

$$\begin{aligned}\hat{Y}_t = {} & \hat{\delta}_0 X_t + (\hat{\delta}_1 + \hat{\phi}_1 \hat{\delta}_0) X_{t-1} + (\hat{\delta}_2 + \hat{\phi}_1 \hat{\delta}_1 + \hat{\phi}_1^2 \hat{\delta}_0) X_{t-2} \\ & + \hat{\phi}_1(\hat{\delta}_2 + \hat{\phi}_1 \hat{\delta}_1 + \hat{\phi}_1^2 \hat{\delta}_0) X_{t-3} + \hat{\phi}_1^2(\hat{\delta}_2 + \hat{\phi}_1 \hat{\delta}_1 + \hat{\phi}_1^2 \hat{\delta}_0) X_{t-4} + \cdots . \end{aligned} \tag{15.29}$$

The coefficients in Equation (15.29) are the estimators of the dynamic multipliers, computed from the OLS estimators of the coefficients in the ADL model in Equation (15.21). If the restrictions on the coefficients in Equation (15.22) were to hold exactly for the *estimated* coefficients, then the dynamic multipliers beyond the second (that is, the coefficients on X_{t-2}, X_{t-3}, and so forth) would all be zero.[2] However, under this estimation strategy those restrictions will not hold exactly, so the estimated multipliers beyond the second in Equation (15.29) will generally be nonzero.

GLS Estimation

The second strategy for estimating the dynamic multipliers when X_t is strictly exogenous is to use generalized least squares (GLS), which entails estimating Equation (15.23). To describe the GLS estimator, we initially assume that ϕ_1 is known. Because in practice it is unknown, this estimator is infeasible, so it is called the infeasible GLS estimator. The infeasible GLS estimator, however, can be modified using an estimator of ϕ_1, which yields a feasible version of the GLS estimator.

[2] Substitute the equalities in Equation (15.22) to show that, if those equalities hold, then $\delta_2 + \phi_1\delta_1 + \phi_1^2\delta_0 = 0$.

Infeasible GLS. Suppose that ϕ_1 were known; then the quasi-differenced variables $\widetilde{X}_t$ and $\widetilde{Y}_t$ could be computed directly. As discussed in the context of Equations (15.24) and (15.26), if X_t is strictly exogenous, then $E(\widetilde{u}_t|\widetilde{X}_t, \widetilde{X}_{t-1}, \ldots) = 0$. Thus, if X_t is strictly exogenous and if ϕ_1 is known, the coefficients α_0, β_1, and β_2 in Equation (15.23) can be estimated by the OLS regression of $\widetilde{Y}_t$ on $\widetilde{X}_t$ and $\widetilde{X}_{t-1}$ (including an intercept). The resulting estimator of β_1 and β_2—that is, the OLS estimator of the slope coefficients in Equation (15.23) when ϕ_1 is known—is the **infeasible GLS estimator**. This estimator is infeasible because ϕ_1 is unknown, so $\widetilde{X}_t$ and $\widetilde{Y}_t$ cannot be computed and thus these OLS estimators cannot actually be computed.

Feasible GLS. The **feasible GLS estimator** modifies the infeasible GLS estimator by using a preliminary estimator of ϕ_1, $\hat{\phi}_1$, to compute the estimated quasi-differences. Specifically, the feasible GLS estimators of β_1 and β_2 are the OLS estimators of β_1 and β_2 in Equation (15.23), computed by regressing $\hat{\widetilde{Y}}_t$ on $\hat{\widetilde{X}}_t$ and $\hat{\widetilde{X}}_{t-1}$ (with an intercept), where $\hat{\widetilde{X}}_t = X_t - \hat{\phi}_1 X_{t-1}$ and $\hat{\widetilde{Y}}_t = Y_t - \hat{\phi}_1 Y_{t-1}$.

The preliminary estimator, $\hat{\phi}_1$, can be computed by first estimating the distributed lag regression in Equation (15.18) by OLS, then using OLS to estimate ϕ_1 in Equation (15.19) with the OLS residuals $\hat{u}_t$ replacing the unobserved regression errors u_t. This version of the GLS estimator is called the Cochrane–Orcutt (1949) estimator.

An extension of the Cochrane–Orcutt method is to continue this process iteratively: Use the GLS estimator of β_1 and β_2 to compute revised estimators of u_t; use these new residuals to re-estimate ϕ_1; use this revised estimator of ϕ_1 to compute revised estimated quasi-differences; use these revised estimated quasi-differences to re-estimate β_1 and β_2; and continue this process until the estimators of β_1 and β_2 converge. This is referred to as the iterated Cochrane–Orcutt estimator.

A nonlinear least squares interpretation of the GLS estimator. An equivalent interpretation of the GLS estimator is that it estimates the ADL model in Equation (15.21), imposing the parameter restrictions in Equation (15.22). These restrictions are nonlinear functions of the original parameters β_0, β_1, β_2, and ϕ_1, so this estimation cannot be performed using OLS. Instead, the parameters can be estimated by nonlinear least squares (NLLS). As discussed in Appendix 8.1, NLLS minimizes the sum of squared mistakes made by the estimated regression function, recognizing that the regression function is a nonlinear function of the parameters being estimated. In general, NLLS estimation can require sophisticated algorithms for minimizing nonlinear functions of unknown parameters.

In the special case at hand, however, those sophisticated algorithms are not needed; rather, the NLLS estimator can be computed using the algorithm described previously for the iterated Cochrane–Orcutt estimator. Thus the iterated Cochrane–Orcutt GLS estimator is in fact the NLLS estimator of the ADL coefficients, subject to the nonlinear constraints in Equation (15.22).

Efficiency of GLS. The virtue of the GLS estimator is that when X is strictly exogenous and the transformed errors $\widetilde{u}_t$ are homoskedastic, it is efficient among linear estimators, at least in large samples. To see this, first consider the infeasible GLS estimator. If $\widetilde{u}_t$ is homoskedastic, if ϕ_1 is known (so that $\widetilde{X}_t$ and $\widetilde{Y}_t$ can be treated as if they are observed), and if X_t is strictly exogenous, then the Gauss–Markov theorem implies that the OLS estimator of α_0, β_1, and β_2 in Equation (15.23) is efficient among all linear conditionally unbiased estimators based on $\widetilde{X}_t$ and $\widetilde{Y}_t$, for $t = 2, \ldots, T$, where the first observation ($t = 1$) is lost because of quasi-differencing. That is, the OLS estimator of the coefficients in Equation (15.23) is the best linear unbiased estimator, or BLUE (Section 5.5). Because the OLS estimator of Equation (15.23) is the infeasible GLS estimator, this means that the infeasible GLS estimator is BLUE. The feasible GLS estimator is similar to the infeasible GLS estimator, except that ϕ_1 is estimated. Because the estimator of ϕ_1 is consistent and its variance is inversely proportional to T, the feasible and infeasible GLS estimators have the same variances in large samples, and the loss of information from the first observation ($t = 1$) is negligible when T is large. In this sense, if X is strictly exogenous, then the feasible GLS estimator is BLUE in large samples. In particular, if X is strictly exogenous, then GLS is more efficient than the OLS estimator of the distributed lag coefficients discussed in Section 15.3.

The Cochrane–Orcutt and iterated Cochrane–Orcutt estimators presented here are special cases of GLS estimation. In general, GLS estimation involves transforming the regression model so that the errors are homoskedastic and serially uncorrelated, then estimating the coefficients of the transformed regression model by OLS. In general, the GLS estimator is consistent and BLUE in large samples if X is strictly exogenous, but is not consistent if X is only (past and present) exogenous. The mathematics of GLS involve matrix algebra, so they are postponed to Section 18.6.

The Distributed Lag Model with Additional Lags and AR(p) Errors

The foregoing discussion of the distributed lag model in Equations (15.18) and (15.19), which has a single lag of X_t and an AR(1) error term, carries over to the general distributed lag model with multiple lags and an AR(p) error term.

The general distributed lag model with autoregressive errors. The general distributed lag model with r lags and an AR(p) error term is

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 X_{t-1} + \cdots + \beta_{r+1} X_{t-r} + u_t, \tag{15.30}$$

$$u_t = \phi_1 u_{t-1} + \phi_2 u_{t-2} + \cdots + \phi_p u_{t-p} + \tilde{u}_t, \tag{15.31}$$

where $\beta_1, \ldots, \beta_{r+1}$ are the dynamic multipliers and $\phi_1, \ldots, \phi_p$ are the autoregressive coefficients of the error term. Under the AR(p) model for the errors, $\tilde{u}_t$ is serially uncorrelated.

Algebra of the sort that led to the ADL model in Equation (15.21) shows that Equations (15.30) and (15.31) imply that Y_t can be written in ADL form:

$$Y_t = \alpha_0 + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \delta_0 X_t + \delta_1 X_{t-1} + \cdots + \delta_q X_{t-q} + \tilde{u}_t, \tag{15.32}$$

where $q = r + p$ and $\delta_0, \ldots, \delta_q$ are functions of the β's and ϕ's in Equations (15.30) and (15.31). Equivalently, the model of Equations (15.30) and (15.31) can be written in quasi-difference form as

$$\tilde{Y}_t = \alpha_0 + \beta_1 \tilde{X}_t + \beta_2 \tilde{X}_{t-1} + \cdots + \beta_{r+1} \tilde{X}_{t-r} + \tilde{u}_t, \tag{15.33}$$

where $\tilde{Y}_t = Y_t - \phi_1 Y_{t-1} - \cdots - \phi_p Y_{t-p}$ and $\tilde{X}_t = X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p}$.

Conditions for estimation of the ADL coefficients. The foregoing discussion of the conditions for consistent estimation of the ADL coefficients in the AR(1) case extends to the general model with AR(p) errors. The conditional mean zero assumption for Equation (15.33) is that

$$E(\tilde{u}_t | \tilde{X}_t, \tilde{X}_{t-1}, \ldots) = 0. \tag{15.34}$$

Because $\tilde{u}_t = u_t - \phi_1 u_{t-1} - \phi_2 u_{t-2} - \cdots - \phi_p u_{t-p}$ and $\tilde{X}_t = X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p}$, this condition is equivalent to

$$\begin{aligned} E(u_t | X_t, X_{t-1}, \ldots) &- \phi_1 E(u_{t-1} | X_t, X_{t-1}, \ldots) \\ &- \cdots - \phi_p E(u_{t-p} | X_t, X_{t-1}, \ldots) = 0. \end{aligned} \tag{15.35}$$

For Equation (15.35) to hold for general values of $\phi_1, \ldots, \phi_p$, it must be the case that each of the conditional expectations in Equation (15.35) is zero; equivalently, it must be the case that

$$E(u_t | X_{t+p}, X_{t+p-1}, X_{t+p-2}, \ldots) = 0. \tag{15.36}$$

This condition is not implied by X_t being (past and present) exogenous, but it is implied by X_t being strictly exogenous. In fact, in the limit when p is infinite (so that the error term in the distributed lag model follows an infinite-order autoregression), the condition in Equation (15.36) becomes the condition in Key Concept 15.1 for strict exogeneity.

Estimation of the ADL model by OLS. As in the distributed lag model with a single lag and an AR(1) error term, the dynamic multipliers can be estimated from the OLS estimators of the ADL coefficients in Equation (15.32). The general formulas are similar to, but more complicated than, those in Equation (15.29) and are best expressed using lag multiplier notation; these formulas are given in Appendix 15.2. In practice, modern regression software designed for time series regression analysis does these computations for you.

Estimation by GLS. Alternatively, the dynamic multipliers can be estimated by (feasible) GLS. This entails OLS estimation of the coefficients of the quasi-differenced specification in Equation (15.33), using estimated quasi-differences. The estimated quasi-differences can be computed using preliminary estimators of the autoregressive coefficients $\phi_1, \ldots, \phi_p$, as in the AR(1) case. The GLS estimator is asymptotically BLUE, in the sense discussed earlier for the AR(1) case.

Estimation of dynamic multipliers under strict exogeneity is summarized in Key Concept 15.4.

Which to use: ADL or GLS? The two estimation options, OLS estimation of the ADL coefficients and GLS estimation of the distributed lag coefficients, have both advantages and disadvantages.

The advantage of the ADL approach is that it can reduce the number of parameters needed for estimating the dynamic multipliers, compared to OLS estimation of the distributed lag model. For example, the estimated ADL model in Equation (15.27) led to the infinitely long estimated distributed lag representation in Equation (15.29). To the extent that a distributed lag model with only r lags is really an approximation to a longer-lagged distributed lag model, the ADL model can provide a simple way to estimate those many longer lags using only a few unknown parameters. Thus in practice it might be possible to estimate the ADL model in Equation (15.39) with values of p and q much smaller than the value of r needed for OLS estimation of the distributed lag coefficients in Equation (15.37). In other words, the ADL specification can provide a compact, or parsimonious, summary of a long and complex distributed lag (see Appendix 15.2 for additional discussion).

KEY CONCEPT 15.4

Estimation of Dynamic Multipliers Under Strict Exogeneity

The general distributed lag model with r lags and AR(p) error term is

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 X_{t-1} + \cdots + \beta_{r+1} X_{t-r} + u_t \tag{15.37}$$

$$u_t = \phi_1 u_{t-1} + \phi_2 u_{t-2} + \cdots + \phi_p u_{t-p} + \tilde{u}_t. \tag{15.38}$$

If X_t is strictly exogenous, then the dynamic multipliers $\beta_1, \ldots, \beta_{r+1}$ can be estimated by first using OLS to estimate the coefficients of the ADL model

$$Y_t = \alpha_0 + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \delta_0 X_t + \delta_1 X_{t-1} + \cdots + \delta_q X_{t-q} + \tilde{u}_t, \tag{15.39}$$

where $q = r + p$, and then computing the dynamic multipliers using regression software. Alternatively, the dynamic multipliers can be estimated by estimating the distributed lag coefficients in Equation (15.37) by GLS.

The advantage of the GLS estimator is that, for a given lag length r in the distributed lag model, the GLS estimator of the distributed lag coefficients is more efficient than the ADL estimator, at least in large samples. In practice, then, the advantage of using the ADL approach arises because the ADL specification can permit estimating fewer parameters than are estimated by GLS.

15.6 Orange Juice Prices and Cold Weather

This section uses the tools of time series regression to squeeze additional insights from our data on Florida temperatures and orange juice prices. First, how long lasting is the effect of a freeze on the price? Second, has this dynamic effect been stable or has it changed over the 51 years spanned by the data and, if so, how?

We begin this analysis by estimating the dynamic causal effects using the method of Section 15.3, that is, by OLS estimation of the coefficients of a distributed lag regression of the percentage change in prices ($\%ChgP_t$) on the number of freezing degree days in that month (FDD_t) and its lagged values. For the distributed lag estimator to be consistent, *FDD* must be (past and present) exogenous. As discussed in Section 15.2, this assumption is reasonable here. Humans cannot influence the weather, so treating the weather as if it were randomly assigned experimentally is appropriate. Because *FDD* is exogenous, we can estimate the dynamic causal effects by OLS estimation of the coefficients in the distributed lag model of Equation (15.4) in Key Concept 15.1.

As discussed in Sections 15.3 and 15.4, the error term can be serially correlated in distributed lag regressions, so it is important to use HAC standard errors, which adjust for this serial correlation. For the initial results, the truncation parameter for the Newey–West standard errors (m in the notation of Section 15.4) was chosen using the rule in Equation (15.17): Because there are 612 monthly observations, according to that rule $m = 0.75T^{1/3} = 0.75 \times 612^{1/3} = 6.37$, but because m must be an integer, this was rounded up to $m = 7$; the sensitivity of the standard errors to this choice of truncation parameter is investigated below.

The results of OLS estimation of the distributed lag regression of $\%ChgP_t$ on FDD_t, $FDD_{t-1}, \ldots, FDD_{t-18}$ are summarized in column (1) of Table 15.1. The coefficients of this regression (only some of which are reported in the table) are estimates of the dynamic causal effect on orange juice price changes (in percent) for the first 18 months following a unit increase in the number of freezing degree days in a month. For example, a single freezing degree day is estimated to increase prices by 0.50% over the month in which the freezing degree day occurs. The subsequent effect on price in later months of a freezing degree day is less: After 1 month the estimated effect is to increase the price by a further 0.17%, and after 2 months the estimated effect is to increase the price by an additional 0.07%. The R^2 from this regression is 0.12, indicating that much of the monthly variation in orange juice prices is not explained by current and past values of *FDD*.

Plots of dynamic multipliers can convey information more effectively than tables such as Table 15.1. The dynamic multipliers from column (1) of Table 15.1 are plotted in Figure 15.2a along with their 95% confidence intervals, computed as the estimated coefficient ± 1.96 HAC standard errors. After the initial sharp price rise, subsequent price rises are less, although prices are estimated to rise slightly in each of the first 6 months after the freeze. As can be seen from Figure 15.2a, for months other than the first the dynamic multipliers are not statistically significantly different from zero at the 5% significance level, although they are estimated to be positive through the seventh month.

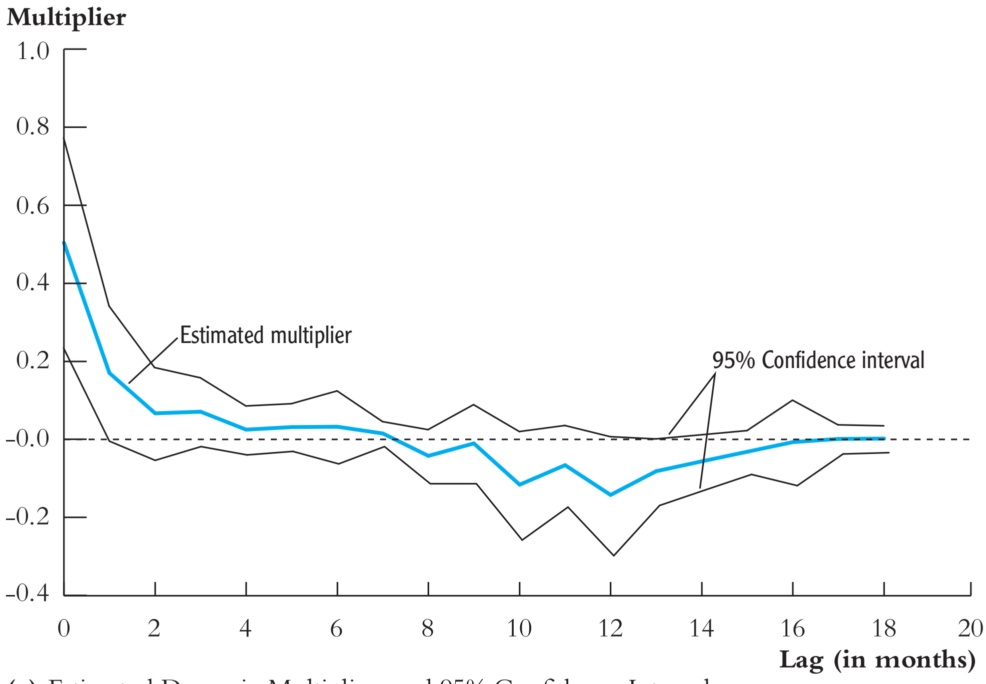
Column (2) of Table 15.1 contains the cumulative dynamic multipliers for this specification, that is, the cumulative sum of the dynamic multipliers reported in

TABLE 15.1 The Dynamic Effect of a Freezing Degree Day (*FDD*) on the Price of Orange Juice: Selected Estimated Dynamic Multipliers and Cumulative Dynamic Multipliers

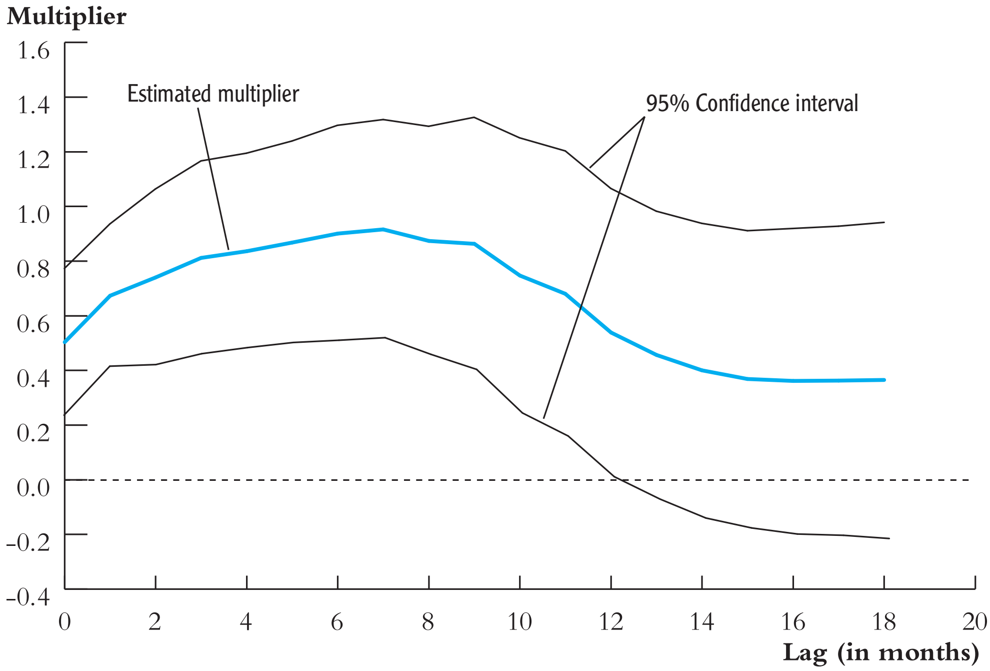
Lag Number	(1) Dynamic Multipliers	(2) Cumulative Multipliers	(3) Cumulative Multipliers	(4) Cumulative Multipliers
0	0.50 (0.14)	0.50 (0.14)	0.50 (0.14)	0.51 (0.15)
1	0.17 (0.09)	0.67 (0.14)	0.67 (0.13)	0.70 (0.15)
2	0.07 (0.06)	0.74 (0.17)	0.74 (0.16)	0.76 (0.18)
3	0.07 (0.04)	0.81 (0.18)	0.81 (0.18)	0.84 (0.19)
4	0.02 (0.03)	0.84 (0.19)	0.84 (0.19)	0.87 (0.20)
5	0.03 (0.03)	0.87 (0.19)	0.87 (0.19)	0.89 (0.20)
6 . . .	0.03 (0.05)	0.90 (0.20)	0.90 (0.21)	0.91 (0.21)
12 . . .	−0.14 (0.08)	0.54 (0.27)	0.54 (0.28)	0.54 (0.28)
18	0.00 (0.02)	0.37 (0.30)	0.37 (0.31)	0.37 (0.30)
Monthly indicators?	No	No	No	Yes $F = 1.01$ $(p = 0.43)$
HAC standard error truncation parameter (m)	7	7	14	7

All regressions were estimated by OLS using monthly data (described in Appendix 15.1) from January 1950 to December 2000, for a total of $T = 612$ monthly observations. The dependent variable is the monthly percentage change in the price of orange juice ($\%ChgP_t$). Regression (1) is the distributed lag regression with the monthly number of freezing degree days and 18 of its lagged values, that is, $FDD_t, FDD_{t-1}, \ldots, FDD_{t-18}$, and the reported coefficients are the OLS estimates of the dynamic multipliers. The cumulative multipliers are the cumulative sum of estimated dynamic multipliers. All regressions include an intercept, which is not reported. Newey–West HAC standard errors, computed using the truncation number given in the final row, are reported in parentheses.

FIGURE 15.2 The Dynamic Effect of a Freezing Degree Day (*FDD*) on the Price of Orange Juice



(a) Estimated Dynamic Multipliers and 95% Confidence Interval



(b) Estimated Cumulative Dynamic Multipliers and 95% Confidence Interval

The estimated dynamic multipliers show that a freeze leads to an immediate increase in prices. Future price rises are much smaller than the initial impact. The cumulative multiplier shows that freezes have a persistent effect on the level of orange juice prices, with prices peaking seven months after the freeze.

column (1). These dynamic multipliers are plotted in Figure 15.2b along with their 95% confidence intervals. After 1 month, the cumulative effect of the freezing degree day is to increase prices by 0.67%, after 2 months the price is estimated to have risen by 0.74%, and after 6 months the price is estimated to have risen by 0.90%. As can be seen in Figure 15.2b, these cumulative multipliers increase through the seventh month, because the individual dynamic multipliers are positive for the first 7 months. In the eighth month, the dynamic multiplier is negative, so the price of orange juice begins to fall slowly from its peak. After 18 months, the cumulative increase in prices is only 0.37%; that is, the long-run cumulative dynamic multiplier is only 0.37%. This long-run cumulative dynamic multiplier is not statistically significantly different from zero at the 10% significance level ($t = 0.37/0.30 = 1.23$).

Sensitivity analysis. As in any empirical analysis, it is important to check whether these results are sensitive to changes in the details of the empirical analysis. We therefore examine three aspects of this analysis: sensitivity to the computation of the HAC standard errors; an alternative specification that investigates potential omitted variable bias; and an analysis of the stability over time of the estimated multipliers.

First, we investigate whether the standard errors reported in the second column of Table 15.1 are sensitive to different choices of the HAC truncation parameter m. In column (3), results are reported for $m = 14$, twice the value used in column (2). The regression specification is the same as in column (2), so the estimated coefficients and dynamic multipliers are identical; only the standard errors differ but, as it happens, not by much. We conclude that the results are insensitive to changes in the HAC truncation parameter.

Second, we investigate a possible source of omitted variable bias. Freezes in Florida are not randomly assigned throughout the year, but rather occur in the winter (of course). If demand for orange juice is seasonal (is demand for orange juice greater in the winter than the summer?), then the seasonal patterns in orange juice demand could be correlated with FDD, resulting in omitted variable bias. The quantity of oranges sold for juice is endogenous: Prices and quantities are simultaneously determined by the forces of supply and demand. Thus, as discussed in Section 9.2, including quantity would lead to simultaneity bias. Nevertheless, the seasonal component of demand can be captured by including seasonal variables as regressors. The specification in column (4) of Table 15.1 therefore includes 11 monthly binary variables, one indicating whether the month is January, one indicating February, and so forth (as usual one binary variable must be omitted to prevent perfect multicollinearity with the intercept). These monthly indicator variables are not jointly statistically significant at the 10% level ($p = 0.43$), and the estimated cumulative dynamic multipliers are essentially the same as for the specifications excluding the monthly indicators. In summary, seasonal fluctuations in demand are not an important source of omitted variable bias.

Have the dynamic multipliers been stable over time?[3] To assess the stability of the dynamic multipliers, we need to check whether the distributed lag regression coefficients have been stable over time. Because we do not have a specific break date in mind, we test for instability in the regression coefficients using the Quandt likelihood ratio (QLR) statistic (Key Concept 14.9). The QLR statistic (with 15% trimming and HAC variance estimator), computed for the regression of column (1) with all coefficients interacted, has a value of 21.19, with $q = 20$ degrees of freedom (the coefficients on FDD_t, its 18 lags, and the intercept). The 1% critical value in Table 14.5 is 2.43, so the QLR statistic rejects at the 1% significance level. These QLR regressions have 40 regressors, a large number; recomputing them for six lags only (so that there are 16 regressors and $q = 8$) also results in rejection at the 1% level. Thus the hypothesis that the dynamic multipliers are stable is rejected at the 1% significance level.

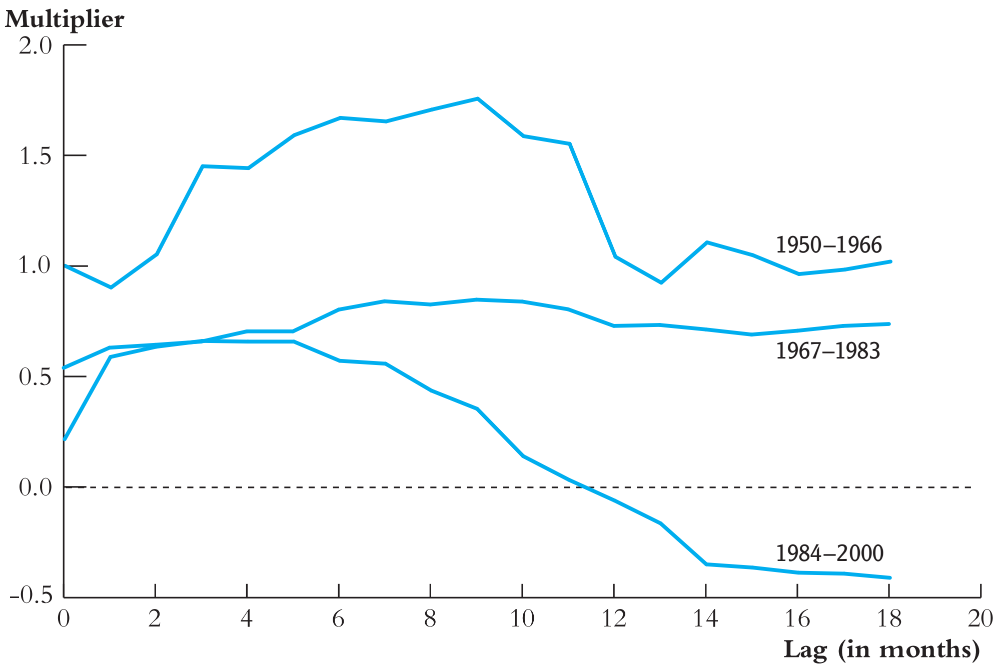
One way to see how the dynamic multipliers have changed over time is to compute them for different parts of the sample. Figure 15.3 plots the estimated cumulative dynamic multipliers for the first third (1950–1966), middle third (1967–1983), and final third (1984–2000) of the sample, computed by running separate regressions on each subsample. These estimates show an interesting and noticeable pattern. In the 1950s and early 1960s, a freezing degree day had a large and persistent effect on the price. The magnitude of the effect on price of a freezing degree day diminished in the 1970s, although it remained highly persistent. In the late 1980s and 1990s, the short-run effect of a freezing degree day was the same as in the 1970s, but it became much less persistent and was essentially eliminated after a year. These estimates suggest that the dynamic causal effect on orange juice prices of a Florida freeze became smaller and less persistent over the second half of the twentieth century. The box "Orange Trees on the March" discusses one possible explanation for the instability of the dynamic causal effects.

ADL and GLS estimates. As discussed in Section 15.5, if the error term in the distributed lag regression is serially correlated and FDD is strictly exogenous, it is possible to estimate the dynamic multipliers more efficiently than by OLS

[3] The discussion of stability in this subsection draws on material from Section 14.7 and can be skipped if that material has not been covered.

FIGURE 15.3 Estimated Cumulative Dynamic Multipliers from Different Sample Periods

The dynamic effect on orange juice prices of freezes changed significantly over the second half of the twentieth century. A freeze had a larger impact on prices during 1950–1966 than later, and the effect of a freeze was less persistent during 1984–2000 than earlier.



estimation of the distributed lag coefficients. Before using either the GLS estimator or the estimator based on the ADL model, however, we need to consider whether *FDD* is in fact strictly exogenous. True, humans cannot affect the daily weather, but does that mean that the weather is *strictly* exogenous? Does the error term u_t in the distributed lag regression have conditional mean zero, given past, present, and *future* values of *FDD*?

The error term in the population counterpart of the distributed lag regression in column (1) of Table 15.1 is the discrepancy between the price and its population prediction based on the past 18 months of weather. This discrepancy might arise for many reasons, one of which is that traders use forecasts of the weather in Orlando. For example, if an especially cold winter is forecasted, then traders would incorporate this into the price, so the price would be above its predicted value based on the population regression; that is, the error term would be positive. If this forecast is accurate, then in fact future weather would turn out to be cold. Thus future freezing degree days would be positive ($X_{t+1} > 0$) when the current price is unusually high ($u_t > 0$), so $\text{corr}(X_{t+1}, u_t)$ is positive. Stated more simply, although orange juice traders cannot influence the weather, they can—and do—predict it (see the box). Consequently, the error term in the price/weather regression

Orange Trees on the March

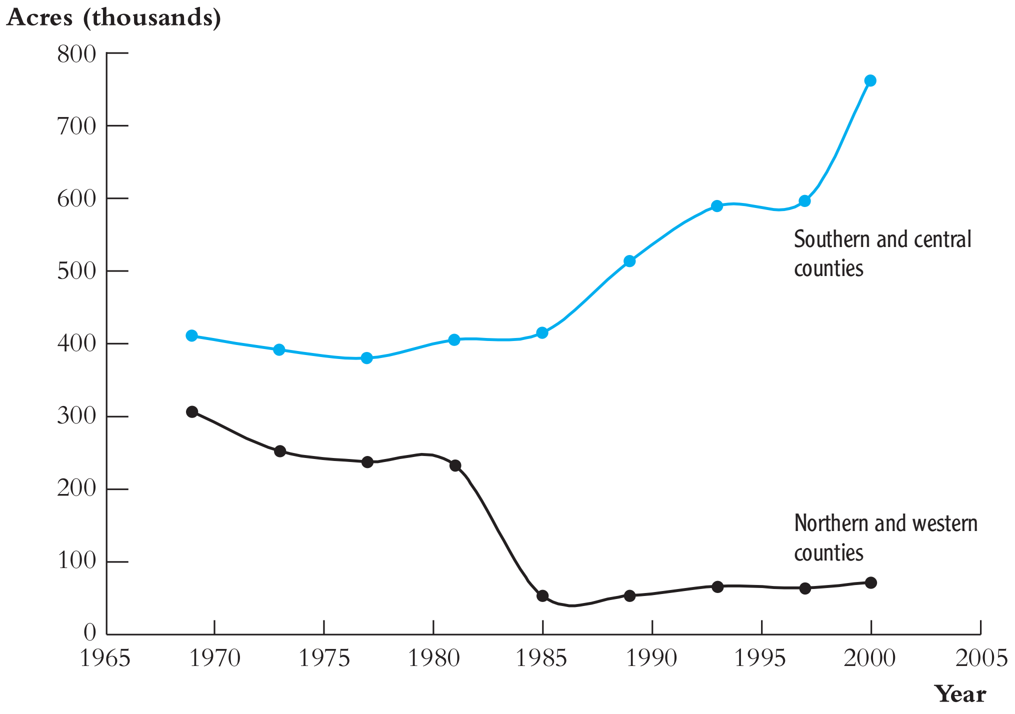
Why do the dynamic multipliers in Figure 15.3 vary over time? One possible explanation is changes in markets, but another is that the trees moved south.

According to the Florida Department of Citrus, the severe freezes in the 1980s, which are visible in Figure 15.1(c), spurred citrus growers to seek a warmer climate. As shown in Figure 15.4, the number of acres of orange trees in the more frost-prone northern and western counties fell from 232,000 acres in 1981 to 53,000 acres in 1985, and orange acreage in southern and central counties subsequently increased from 413,000 in 1985 to 588,000 in 1993. With the groves farther south, northern frosts damage a smaller fraction of the crop, and—as indicated by the dynamic multipliers in Figure 15.3—price becomes less sensitive to temperatures in the more northern city of Orlando.

OK, the orange trees themselves might not have been on the march—that can be left to *MacBeth*—but southern migration of the orange groves does give new meaning to the term "nonstationarity."[4]

[4] We are grateful to Professor James Cobbe of Florida State University for telling us about the southern movement of the orange groves.

FIGURE 15.4 Orange Grove Acreage in Regions of Florida



is correlated with future weather. In other words, *FDD* is exogenous, but if this reasoning is true, it is not strictly exogenous, and the GLS and ADL estimators will not be consistent estimators of the dynamic multipliers. These estimators therefore are not used in this application.

15.7 Is Exogeneity Plausible? Some Examples

As in regression with cross-sectional data, the interpretation of the coefficients in a distributed lag regression as causal dynamic effects hinges on the assumption that X is exogenous. If X_t or its lagged values are correlated with u_t, then the conditional mean of u_t will depend on X_t or its lags, in which case X is not (past and present) exogenous. Regressors can be correlated with the error term for several reasons, but with economic time series data a particularly important concern is that there could be simultaneous causality, which (as discussed in Sections 9.2 and 12.1) results in endogenous regressors. In Section 15.6, we discussed the assumptions of exogeneity and strict exogeneity of freezing degree days in detail. In this section, we examine the assumption of exogeneity in four other economic applications.

U.S. Income and Australian Exports

The United States is an important source of demand for Australian exports. Precisely how sensitive Australian exports are to fluctuations in U.S. aggregate income could be investigated by regressing Australian exports to the United States against a measure of U.S. income. Strictly speaking, because the world economy is integrated, there is simultaneous causality in this relationship: A decline in Australian exports reduces Australian income, which reduces demand for imports from the United States, which reduces U.S. income. As a practical matter, however, this effect is very small because the Australian economy is much smaller than the U.S. economy. Thus U.S. income plausibly can be treated as exogenous in this regression.

In contrast, in a regression of European Union exports to the United States against U.S. income, the argument for treating U.S. income as exogenous is less convincing because demand by residents of the European Union for U.S. exports constitutes a substantial fraction of the total demand for U.S. exports. Thus a decline in U.S. demand for EU exports would decrease EU income, which in turn would decrease demand for U.S. exports and thus decrease U.S. income. Because of these linkages through international trade, EU exports to the United States and U.S. income are simultaneously determined, so in this regression U.S. income arguably is not exogenous. This example illustrates a more general point that

NEWS FLASH: Commodity Traders Send Shivers Through Disney World

Although the weather at Disney World in Orlando, Florida, is usually pleasant, now and then a cold spell can settle in. If you are visiting Disney World on a winter evening, should you bring a warm coat? Some people might check the weather forecast on TV, but those in the know can do better: They can check that day's closing price on the New York orange juice futures market!

The financial economist Richard Roll undertook a detailed study of the relationship between orange juice prices and the weather. Roll (1984) examined the effect on prices of cold weather in Orlando, but he also studied the "effect" of changes in the price of an orange juice futures contract (a contract to buy frozen orange juice concentrate at a specified date in the future) on the weather. Roll used daily data from 1975 to 1981 on the prices of OJ futures contracts traded at the New York Cotton Exchange and on daily and overnight temperatures in Orlando. He found that a rise in the price of the futures contract during the trading day in New York predicted cold weather, in particular a freezing spell, in Orlando over the following night. In fact, the market was so effective in predicting cold weather in Florida that a price rise during the trading day actually predicted forecast errors in the official U.S. government weather forecasts for that night.

Roll's study is also interesting for what he did *not* find: Although his detailed weather data explained some of the variation in daily OJ futures prices, most of the daily movements in OJ prices remained unexplained. He therefore suggested that the OJ futures market exhibits "excess volatility," that is, more volatility than can be attributed to movements in fundamentals. Understanding why (and if) there is excess volatility in financial markets is now an important area of research in financial economics.

Roll's finding also illustrates the difference between forecasting and estimating dynamic causal effects. Price changes on the OJ futures market are a useful predictor of cold weather, but that does not mean that commodity traders are so powerful that they can *cause* the temperature to fall. Visitors to Disney World might shiver after an OJ futures contract price rise, but they are not shivering *because* of the price rise—unless, of course, they went short in the OJ futures market.

whether a variable is exogenous depends on the context: U.S. income is plausibly exogenous in a regression explaining Australian exports, but not in a regression explaining EU exports.

Oil Prices and Inflation

Ever since the oil price increases of the 1970s, macroeconomists have been interested in estimating the dynamic effect of an increase in the international price of crude oil on the U.S. rate of inflation. Because oil prices are set in world markets in large part by foreign oil-producing countries, initially one might think that oil prices are exogenous. But oil prices are not like the weather: Members of OPEC set oil production levels strategically, taking many factors, including the state of the world economy, into account. To the extent that oil prices (or quantities) are set based on an assessment of current and future world economic conditions, including inflation in the United States, oil prices are endogenous.

Monetary Policy and Inflation

The central bankers in charge of monetary policy need to know the effect on inflation of monetary policy. Because an important tool of monetary policy is the short-term interest rate (the "short rate"), they need to know the dynamic causal effect on inflation of a change in the short rate. Although the short rate is determined by the central bank, it is not set by the central bankers at random (as it would be in an ideal randomized experiment) but rather is set endogenously: The central bank determines the short rate based on an assessment of the current and future states of the economy, especially including the current and future rates of inflation. The rate of inflation in turn depends on the interest rate (higher interest rates reduce aggregate demand), but the interest rate depends on the rate of inflation, its past value, and its (expected) future value. Thus the short rate is endogenous, and the causal dynamic effect of a change in the short rate on future inflation cannot be consistently estimated by an OLS regression of the rate of inflation on current and past interest rates.

The Growth Rate of GDP and the Term Spread

In Chapter 14 lagged values of the term spread were used to forecast future values of the growth rate of GDP. Because lags of the term spread happened in the past, one might initially think that there cannot be feedback from current growth rates of GDP to past values of the term spread, so past values of the term spread can be treated as exogenous. But past values of the term spread were not randomly assigned in an experiment; instead, the past term spread was simultaneously determined with past values of the growth rate of GDP. Because GDP and the interest rates making up the term spread are simultaneously determined, the other factors that determine the growth rate of GDP contained in u_t are correlated with past values of the term spread; that is, the term spread is not exogenous. It follows that the term spread is not strictly exogenous, so the dynamic multipliers computed using an ADL model [for example, the ADL model in Equation (14.17)] are not consistent estimates of the dynamic causal effect on the growth rate of GDP of a change in the term spread.

15.8 Conclusion

Time series data provide the opportunity to estimate the time path of the effect on Y of a change in X, that is, the dynamic causal effect on Y of a change in X. To estimate dynamic causal effects using a distributed lag regression, however, X must be exogenous, as it would be if it were set randomly in an ideal randomized experiment. If X is not just exogenous but is *strictly* exogenous, then the dynamic causal effects can be estimated using an autoregressive distributed lag model or by GLS.

In some applications, such as estimating the dynamic causal effect on the price of orange juice of freezing weather in Florida, a convincing case can be made that the regressor (freezing degree days) is exogenous; thus the dynamic causal effect can be estimated by OLS estimation of the distributed lag coefficients. Even in this application, however, economic theory suggests that the weather is not strictly exogenous, so the ADL or GLS methods are inappropriate. Moreover, in many relations of interest to econometricians, there is simultaneous causality, so the regressor in these specifications are not exogenous, strictly or otherwise. Ascertaining whether the regressor is exogenous (or strictly exogenous) ultimately requires combining economic theory, institutional knowledge, and careful judgment.

Summary

1. Dynamic causal effects in time series are defined in the context of a randomized experiment, where the same subject (entity) receives different randomly assigned treatments at different times. The coefficients in a distributed lag regression of Y on X and its lags can be interpreted as the dynamic causal effects when the time path of X is determined randomly and independently of other factors that influence Y.
2. The variable X is (past and present) exogenous if the conditional mean of the error u_t in the distributed lag regression of Y on current and past values of X does not depend on current and past values of X. If in addition the conditional mean of u_t does not depend on future values of X, then X is strictly exogenous.
3. If X is exogenous, then the OLS estimators of the coefficients in a distributed lag regression of Y on current and past values of X are consistent estimators of the dynamic causal effects. In general, the error u_t in this regression is serially correlated, so conventional standard errors are misleading, and HAC standard errors must be used instead.

4. If X is strictly exogenous, then the dynamic multipliers can be estimated using OLS estimation of an ADL model or using GLS.
5. Exogeneity is a strong assumption that often fails to hold in economic time series data because of simultaneous causality, and the assumption of strict exogeneity is even stronger.

Key Terms

dynamic causal effect (589)
distributed lag model (595)
exogeneity (596)
strict exogeneity (596)
dynamic multiplier (600)
impact effect (600)
cumulative dynamic multiplier (600)
long-run cumulative dynamic multiplier (601)
heteroskedasticity- and autocorrelation-consistent (HAC) standard error (604)
truncation parameter (605)
Newey–West variance estimator (605)
generalized least squares (GLS) (606)
quasi-difference (608)
infeasible GLS estimator (612)
feasible GLS estimator (612)

Review the Concepts

15.1 In the 1970s a common practice was to estimate a distributed lag model relating changes in nominal gross domestic product (Y) to current and past changes in the money supply (X). Under what assumptions will this regression estimate the causal effects of money on nominal GDP? Are these assumptions likely to be satisfied in a modern economy like that of the United States?

15.2 Suppose that X is strictly exogenous. A researcher estimates an ADL(1,1) model, calculates the regression residual, and finds the residual to be highly serially correlated. Should the researcher estimate a new ADL model with additional lags or simply use HAC standard errors for the ADL(1,1) estimated coefficients?

15.3 Suppose that a distributed lag regression is estimated, where the dependent variable is ΔY_t instead of Y_t. Explain how you would compute the dynamic multipliers of X_t on Y_t.

15.4 Suppose that you added FDD_{t+1} as an additional regressor in Equation (15.2). If *FDD* is strictly exogenous, would you expect the coefficient on FDD_{t+1} to be zero or nonzero? Would your answer change if *FDD* is exogenous but not strictly exogenous?

Exercises

15.1 Increases in oil prices have been blamed for several recessions in developed countries. To quantify the effect of oil prices on real economic activity, researchers have run regressions like those discussed in this chapter. Let GDP_t denote the value of quarterly gross domestic product in the United States and let $Y_t = 100\ln(GDP_t/GDP_{t-1})$ be the quarterly percentage change in GDP. James Hamilton, an econometrician and macroeconomist, has suggested that oil prices adversely affect that economy only when they jump above their values in the recent past. Specifically, let O_t equal the greater of zero or the percentage point difference between oil prices at date *t* and their maximum value during the past 3 years. A distributed lag regression relating Y_t and O_t, estimated over 1960:Q1–2013:Q4, is

$$\begin{aligned}\hat{Y}_t = \underset{(0.1)}{1.0} - \underset{(0.013)}{0.007}O_t - \underset{(0.011)}{0.015}O_{t-1} - \underset{(0.011)}{0.019}O_{t-2} - \underset{(0.010)}{0.024}O_{t-3} - \underset{(0.012)}{0.037}O_{t-4}\\ -\underset{(0.008)}{0.012}O_{t-5} + \underset{(0.010)}{0.005}O_{t-6} - \underset{(0.008)}{0.008}O_{t-7} + \underset{(0.008)}{0.006}O_{t-8}.\end{aligned}$$

a. Suppose that oil prices jump 25% above their previous peak value and stay at this new higher level (so that $O_t = 25$ and $O_{t+1} = O_{t+2} = \cdots = 0$). What is the predicted effect on output growth for each quarter over the next 2 years?

b. Construct a 95% confidence interval for your answers in (a).

c. What is the predicted cumulative change in GDP growth over 8 quarters?

d. The HAC *F*-statistic testing whether the coefficients on O_t and its lags are zero is 5.79. Are the coefficients significantly different from zero?

15.2 Macroeconomists have also noticed that interest rates change following oil price jumps. Let R_t denote the interest rate on 3-month Treasury bills (in percentage points at an annual rate). The distributed lag regression relating the change in R_t (ΔR_t) to O_t estimated over 1960:Q1–2013:Q4 is

$$\begin{aligned}\widehat{\Delta R_t} = \underset{(0.05)}{0.03} + \underset{(0.010)}{0.013O_t} + \underset{(0.010)}{0.013O_{t-1}} - \underset{(0.008)}{0.004O_{t-2}} - \underset{(0.015)}{0.024O_{t-3}} - \underset{(0.010)}{0.000O_{t-4}} \\ + \underset{(0.015)}{0.006O_{t-5}} - \underset{(0.015)}{0.005O_{t-6}} - \underset{(0.010)}{0.018O_{t-7}} - \underset{(0.006)}{0.004O_{t-8}}.\end{aligned}$$

a. Suppose that oil prices jump 25% above their previous peak value and stay at this new higher level (so that $O_t = 25$ and $O_{t+1} = O_{t+2} = \cdots = 0$). What is the predicted change in interest rates for each quarter over the next 2 years?

b. Construct 95% confidence intervals for your answers to (a).

c. What is the effect of this change in oil prices on the level of interest rates in period $t + 8$? How is your answer related to the cumulative multiplier?

d. The HAC F-statistic testing whether the coefficients on O_t and its lags are zero is 1.93. Are the coefficients significantly different from zero?

15.3 Consider two different randomized experiments. In experiment A, oil prices are set randomly, and the central bank reacts according to its usual policy rules in response to economic conditions, including changes in the oil price. In experiment B, oil prices are set randomly, and the central bank holds interest rates constant and in particular does not respond to the oil price changes. In both experiments, GDP growth is observed. Now suppose that oil prices are exogenous in the regression in Exercise 15.1. To which experiment, A or B, does the dynamic causal effect estimated in Exercise 15.1 correspond?

15.4 Suppose that oil prices are strictly exogenous. Discuss how you could improve on the estimates of the dynamic multipliers in Exercise 15.1.

15.5 Derive Equation (15.7) from Equation (15.4) and show that $\delta_0 = \beta_0$, $\delta_1 = \beta_1$, $\delta_2 = \beta_1 + \beta_2$, $\delta_3 = \beta_1 + \beta_2 + \beta_3$ (etc.). (*Hint:* Note that $X_t = \Delta X_t + \Delta X_{t-1} + \cdots + \Delta X_{t-p+1} + X_{t-p}$.)

15.6 Consider the regression model $Y_t = \beta_0 + \beta_1 X_t + u_t$, where u_t follows the stationary AR(1) model $u_t = \phi_1 u_{t-1} + \widetilde{u}_t$ with $\widetilde{u}_t$ i.i.d. with mean 0 and variance $\sigma^2_{\widetilde{u}}$ and $|\phi_1| < 1$; the regressorX_t follows the stationary AR(1) model $X_t = \gamma_1 X_{t-1} + e_t$ with e_t i.i.d. with mean 0 and variance σ^2_e and $|\gamma| < 1$; and e_t is independent of $\widetilde{u}_i$ for all t and i.

a. Show that $\text{var}(u_t) = \dfrac{\sigma_{\tilde{u}}^2}{1 - \phi_1^2}$ and $\text{var}(X_t) = \dfrac{\sigma_e^2}{1 - \gamma_1^2}$.

b. Show that $\text{cov}(u_t, u_{t-j}) = \phi_1^j \text{var}(u_t)$ and $\text{cov}(X_t, X_{t-j}) = \gamma_1^j \text{var}(X_t)$.

c. Show that $\text{corr}(u_t, u_{t-j}) = \phi_1^j$ and $\text{corr}(X_t, X_{t-j}) = \gamma_1^j$.

d. Consider the terms σ_v^2 and f_T in Equation (15.14).

i. Show that $\sigma_v^2 = \sigma_X^2 \sigma_u^2$, where σ_X^2 is the variance of X and σ_u^2 is the variance of u.

ii. Derive an expression for f_∞.

15.7 Consider the regression model $Y_t = \beta_0 + \beta_1 X_t + u_t$, where u_t follows the stationary AR(1) model $u_t = \phi_1 u_{t-1} + \tilde{u}_t$ with $\tilde{u}_t$ i.i.d. with mean 0 and variance $\sigma_{\tilde{u}}^2$ and $|\phi_1| < 1$.

a. Suppose that X_t is independent of $\tilde{u}_j$ for all t and j. Is X_t exogenous (past and present)? Is X_t strictly exogenous (past, present, and future)?

b. Suppose that $X_t = \tilde{u}_{t+1}$. Is X_t exogenous? Is X_t strictly exogenous?

15.8 Consider the model in Exercise 15.7 with $X_t = \tilde{u}_{t+1}$.

a. Is the OLS estimator of β_1 consistent? Explain.

b. Explain why the GLS estimator of β_1 is not consistent.

c. Show that the infeasible GLS estimator $\hat{\beta}_1^{GLS} \xrightarrow{p} \beta_1 - \dfrac{\phi_1}{1 + \phi_1^2}$.

[*Hint:* Use the omitted variable formula (6.1) applied to the quasi-differenced regression in Equation (15.23)].

15.9 Consider the "constant-term-only" regression model $Y_t = \beta_0 + u_t$, where u_t follows the stationary AR(1) model $u_t = \phi_1 u_{t-1} + \tilde{u}_t$ with $\tilde{u}_t$ i.i.d. with mean 0 and variance $\sigma_{\tilde{u}}^2$ and $|\phi_1| < 1$.

a. Show that the OLS estimator is $\hat{\beta}_0 = T^{-1}\Sigma_{t=1}^T Y_t$.

b. Show that the (infeasible) GLS estimator is $\hat{\beta}_0^{GLS} = (1 - \phi_1)^{-1}(T - 1)^{-1}\Sigma_{t=2}^T (Y_t - \phi_1 Y_{t-1})$. [*Hint:* The GLS estimator of β_0 is $(1 - \phi_1)^{-1}$ multiplied by the OLS estimator of α_0 in Equation (15.23). Why?]

c. Show that $\hat{\beta}_0^{GLS}$ can be written as $\hat{\beta}_0^{GLS} = (T - 1)^{-1}\Sigma_{t=2}^{T-1} Y_t + (1 - \phi_1)^{-1}(T - 1)^{-1}(Y_T - \phi_1 Y_1)$. [*Hint:* Rearrange the formula in (b).]

d. Derive the difference $\hat{\beta}_0 - \hat{\beta}_0^{GLS}$ and discuss why it is likely to be small when T is large.

15.10 Consider the ADL model $Y_t = 3.1 + 0.4Y_{t-1} + 2.0X_t - 0.8X_{t-1} + \tilde{u}_t$, where X_t is strictly exogenous.

a. Derive the impact effect of X on Y.

b. Derive the first five dynamic multipliers.

c. Derive the first five cumulative multipliers.

d. Derive the long-run cumulative dynamic multiplier.

15.11 Suppose that $a(\text{L}) = (1 - \phi\text{L})$, with $|\phi_1| < 1$, and $b(\text{L}) = 1 + \phi\text{L} + \phi^2\text{L}^2 + \phi^3\text{L}^3 \ldots.$

a. Show that the product $b(\text{L})a(\text{L}) = 1$, so that $b(\text{L}) = a(\text{L})^{-1}$.

b. Why is the restriction $|\phi_1| < 1$ important?

Empirical Exercises

(Only two empirical exercises for this chapter are given in the text, but you can find more on the text website, **http://www.pearsonhighered.com/stock_watson/**.)

E15.1 In this exercise you will estimate the effect of oil prices on macroeconomic activity, using monthly data on the Index of Industrial Production (IP) and the monthly measure of O_t described in Exercise 15.1. The data can be found on the textbook website, **http://www.pearsonhighered.com/stock_watson**, in the file **USMacro_Monthly**.

a. Compute the monthly growth rate in IP, expressed in percentage points, $ip\_growth_t = 100 \times \ln(IP_t/IP_{t-1})$. What are the mean and standard deviation of *ip_growth* over the 1960:M1–2012:M12 sample period? What are the units for *ip_growth* (percent, percent per annum, percent per month, or something else)?

b. Plot the value of O_t. Why are so many values of O_t equal to zero? Why aren't some values of O_t negative?

c. Estimate a distributed lag model by regressing *ip_growth* onto the current value and 18 lagged values of O_t, including an intercept. What value of the HAC standard truncation parameter m did you choose? Why?

d. Taken as a group, are the coefficients on O_t statistically significantly different from zero?

e. Construct graphs like those in Figure 15.2, showing the estimated dynamic multipliers, cumulative multipliers, and 95% confidence intervals. Comment on the real-world size of the multipliers.

f. Suppose that high demand in the United States (evidenced by large values of *ip_growth*) leads to increases in oil prices. Is O_t exogenous? Are the estimated multipliers shown in the graphs in (e) reliable? Explain.

E15.2 In the data file **USMacro_Quarterly**, you will find data on two aggregate price series for the United States: the price index for personal consumption expenditures (PCEP) that you used in Empirical Exercise 14.1 and the Consumer Price Index (CPI). These series are alternative measures of consumer prices in the United States. The CPI prices a basket of goods whose composition is updated every 5–10 years. PCEP uses chain-weighting to price a basket of goods whose composition changes from month to month. Economists have argued that the CPI will overstate inflation because it does not take into account the substitution that occurs when relative prices change. If this substitution bias is important, then average CPI inflation should be systematically higher than PCEP inflation. Let $\pi_t^{CPI} = 400 \times [\ln(CPI_t) - \ln(CPI_{t-1})]$, and $\pi_t^{PCEP} = 400 \times [\ln(PCEP_t) - \ln(PCEP_{t-1})]$, and $Y_t = \pi_t^{CPI} - \pi_t^{PCEP}$, so π_t^{CPI} is the quarterly rate of price inflation (measured in percentage points at an annual rate) based on the CPI, π_t^{PCEP} is the quarterly rate of price inflation from the PCEP, and Y_t is their difference. Using data from 1963:Q1 through 2012:Q4, carry out the following exercises.

a. Compute the sample means of π_t^{CPI} and π_t^{PCED}. Are these point estimates consistent with the presence of economically significant substitution bias in the CPI?

b. Compute the sample mean of Y_t. Explain why it is numerically equal to the difference in the means computed in (a).

c. Show that the population mean of Y is equal to the difference of the population means of the two inflation rates.

d. Consider the "constant-term-only" regression: $Y_t = \beta_0 + u_t$. Show that $\beta_0 = E(Y)$. Do you think that u_t is serially correlated? Explain.

e. Construct a 95% confidence interval for β_0. What value of the HAC standard truncation parameter m did you choose? Why?

f. Is there statistically significant evidence that the mean inflation rate for the CPI is greater than the rate for the PCEP?

g. Is there evidence of instability in β_0? Carry out a QLR test. (*Hint:* Make sure you use HAC standard errors for the regressions in the QLR procedure.)

APPENDIX

15.1 The Orange Juice Data Set

The orange juice price data are the frozen orange juice component of processed foods and feeds group of the Producer Price Index (PPI), collected by the U.S. Bureau of Labor Statistics (BLS series wpu02420301). The orange juice price series was divided by the overall PPI for finished goods to adjust for general price inflation. The freezing degree days series was constructed from daily minimum temperatures recorded at Orlando-area airports, obtained from the National Oceanic and Atmospheric Administration (NOAA) of the U.S. Department of Commerce. The *FDD* series was constructed so that its timing and the timing of the orange juice price data were approximately aligned. Specifically, the frozen orange juice price data are collected by surveying a sample of producers in the middle of every month, although the exact date varies from month to month. Accordingly, the *FDD* series was constructed to be the number of freezing degree days from the 11th of one month to the 10th of the next month; that is, *FDD* is the maximum of zero and 32 minus the minimum daily temperature, summed over all days from the 11th to the 10th. Thus $\%ChgP_t$ for February is the percentage change in real orange juice prices from mid-January to mid-February, and FDD_t for February is the number of freezing degree days from January 11 to February 10.

APPENDIX

15.2 The ADL Model and Generalized Least Squares in Lag Operator Notation

This appendix presents the distributed lag model in lag operator notation, derives the ADL and quasi-differenced representations of the distributed lag model, and discusses the conditions under which the ADL model can have fewer parameters than the original distributed lag model.

The Distributed Lag, ADL, and Quasi-Difference Models, in Lag Operator Notation

As defined in Appendix 14.3, the lag operator, L, has the property that $\mathrm{L}^j X_t = X_{t-j}$, and the distributed lag $\beta_1 X_t + \beta_2 X_{t-1} + \cdots + \beta_{r+1} X_{t-r}$ can be expressed as $\beta(\mathrm{L})X_t$, where $\beta(\mathrm{L}) = \Sigma_{j=0}^{r} \beta_{j+1}\mathrm{L}^j$, where $\mathrm{L}^0 = 1$. Thus the distributed lag model in Key Concept 15.1 [Equation (15.4)] can be written in lag operator notation as

$$Y_t = \beta_0 + \beta(\mathrm{L})X_t + u_t. \tag{15.40}$$

In addition, if the error term u_t follows an AR(p), then it can be written as

$$\phi(\mathrm{L})u_t = \widetilde{u}_t, \tag{15.41}$$

where $\phi(\mathrm{L}) = \Sigma_{j=0}^{p}\phi_j L^j$, where $\phi_0 = 1$ and $\widetilde{u}_t$ is serially uncorrelated [note that $\phi_1, \ldots, \phi_p$ as defined here are the negatives of $\phi_1, \ldots, \phi_p$ in the notation of Equation (15.31)].

To derive the ADL model, premultiply each side of Equation (15.40) by $\phi(\mathrm{L})$ so that

$$\phi(\mathrm{L})Y_t = \phi(\mathrm{L})[\beta_0 + \beta(\mathrm{L})X_t + u_t] = \alpha_0 + \delta(\mathrm{L})X_t + \widetilde{u}_t, \tag{15.42}$$

where

$$\alpha_0 = \phi(1)\beta_0 \text{ and } \delta(\mathrm{L}) = \phi(\mathrm{L})\beta(\mathrm{L}), \text{ where } \phi(1) = \sum_{j=0}^{p}\phi_j. \tag{15.43}$$

To derive the quasi-differenced model, note that $\phi(\mathrm{L})\beta(\mathrm{L})X_t = \beta(\mathrm{L})\phi(\mathrm{L})X_t = \beta(\mathrm{L})\widetilde{X}_t$, where $\widetilde{X}_t = \phi(\mathrm{L})X_t$. Thus rearranging Equation (15.42) yields

$$\widetilde{Y}_t = \alpha_0 + \beta(\mathrm{L})\widetilde{X}_t + \widetilde{u}_t, \tag{15.44}$$

where $\widetilde{Y}_t$ is the quasi-difference of Y_t; that is, $\widetilde{Y}_t = \phi(\mathrm{L})Y_t$.

The Inverse of a Lag Polynomial

Let $a(x) = \Sigma_{j=0}^{p}a_j x^j$ denote a polynomial of order p. The inverse of $a(x)$, say $b(x)$, is a function that satisfies $b(x)a(x) = 1$. If the roots of the polynomial $a(x)$ are greater than 1 in absolute value, then $b(x)$ is a polynomial in nonnegative powers of x: $b(x) = \Sigma_{j=0}^{\infty}b_j x^j$. Because $b(x)$ is the inverse of $a(x)$, it is denoted as $a(x)^{-1}$ or as $1/a(x)$.

The inverse of a lag polynomial $a(\mathrm{L})$ is defined analogously: $a(\mathrm{L})^{-1} = 1/a(\mathrm{L}) = b(\mathrm{L}) = \Sigma_{j=0}^{\infty}b_j\mathrm{L}^j$, where $b(\mathrm{L})a(\mathrm{L}) = 1$. For example, if $a(\mathrm{L}) = (1 - \phi\mathrm{L})$, with $|\phi| < 1$, you can verify that $a(\mathrm{L})^{-1} = 1 + \phi\mathrm{L} + \phi^2\mathrm{L}^2 + \phi^3\mathrm{L}^3 \ldots = \Sigma_{j=0}^{\infty}\phi^j\mathrm{L}^j$. (See Exercise 15.11.)

The ADL and GLS Estimators

The OLS estimator of the ADL coefficients is obtained by OLS estimation of Equation (15.42). The original distributed lag coefficients are $\beta(\mathrm{L})$, which, in terms of the estimated coefficients, is $\beta(\mathrm{L}) = \phi(\mathrm{L})^{-1}\delta(\mathrm{L})$; that is, the coefficients in $\beta(\mathrm{L})$ satisfy the restrictions implied by $\phi(\mathrm{L})\beta(L) = \delta(\mathrm{L})$. Thus the estimator of the dynamic multipliers based on the OLS estimators of the coefficients of the ADL model, $\hat{\delta}(\mathrm{L})$ and $\hat{\phi}(\mathrm{L})$, is

$$\hat{\beta}^{ADL}(\mathrm{L}) = \hat{\phi}(\mathrm{L})^{-1}\hat{\delta}(\mathrm{L}). \tag{15.45}$$

The expressions for the coefficients in Equation (15.29) in the text are obtained as a special case of Equation (15.45) when $r = 1$ and $p = 1$.

The feasible GLS estimator is computed by obtaining a preliminary estimator of $\phi(\mathrm{L})$, computing estimated quasi-differences, estimating $\beta(\mathrm{L})$ in Equation (15.44) using these estimated quasi-differences, and (if desired) iterating until convergence. The iterated GLS estimator is the NLLS estimator computed by NLLS estimation of the ADL model in Equation (15.42), subject to the nonlinear restrictions on the parameters contained in Equation (15.43).

As stressed in the discussion surrounding Equation (15.36) in the text, it is not enough for X_t to be (past and present) exogenous to use either of these estimation methods, for exogeneity alone does not ensure that Equation (15.36) holds. If, however, X is strictly exogenous, then Equation (15.36) does hold, and assuming that Assumptions 2 through 4 of Key Concept 14.6 hold, these estimators are consistent and asymptotically normal. Moreover, the usual (cross-sectional heteroskedasticity-robust) OLS standard errors provide a valid basis for statistical inference.

Parameter reduction using the ADL model. Suppose that the distributed lag polynomial $\beta(\mathrm{L})$ can be written as a ratio of lag polynomials, $\theta_2(\mathrm{L})^{-1}\theta_1(\mathrm{L})$, where $\theta_1(\mathrm{L})$ and $\theta_2(\mathrm{L})$ are both lag polynomials of a low degree. Then $\phi(\mathrm{L})\beta(\mathrm{L})$ in Equation (15.43) is $\phi(\mathrm{L})\beta(\mathrm{L}) = \phi(\mathrm{L})[\theta_2(\mathrm{L})^{-1}\theta_1(\mathrm{L})] = [\phi(\mathrm{L})\theta_2(\mathrm{L})^{-1}]\theta_1(\mathrm{L})$. If it so happens that $\phi(\mathrm{L}) = \theta_2(\mathrm{L})$, then $\delta(\mathrm{L}) = \phi(\mathrm{L})\beta(\mathrm{L}) = \theta_1(\mathrm{L})$. If the degree of $\theta_1(\mathrm{L})$ is low, then q, the number of lags of X_t in the ADL model, can be much less than r. Thus, under these assumptions, estimation of the ADL model entails estimating potentially many fewer parameters than the original distributed lag model. It is in this sense that the ADL model can achieve more parsimonious parameterizations (that is, use fewer unknown parameters) than the distributed lag model.

As developed here, the assumption that $\phi(\mathrm{L})$ and $\theta_2(\mathrm{L})$ happen to be the same seems like a coincidence that would not occur in an application. However, the ADL model is able to capture a large number of shapes of dynamic multipliers with only a few coefficients.

ADL or GLS: Bias versus variance. A good way to think about whether to estimate dynamic multipliers by first estimating an ADL model and then computing the dynamic multipliers from the ADL coefficients or, alternatively, by estimating the distributed lag model directly using GLS is to view the decision in terms of a trade-off between bias and variance. Estimating the dynamic multipliers using an approximate ADL model introduces bias; however, because there are few coefficients, the variance of the estimator of the dynamic multipliers can be small. In contrast, estimating a long distributed lag model using GLS produces less bias in the multipliers; however, because there are so many coefficients, their variance can be large. If the ADL approximation to the dynamic multipliers is a good one, then the bias of the implied dynamic multipliers will be small, so the ADL approach will have a smaller variance than the GLS approach with only a small increase in the bias. For this reason, unrestricted estimation of an ADL model with small number of lags of Y and X is an attractive way to approximate a long distributed lag when X is strictly exogenous.

16 Additional Topics in Time Series Regression

This chapter takes up some further topics in time series regression, starting with forecasting. Chapter 14 considered forecasting a single variable. In practice, however, you might want to forecast two or more variables, such as the growth rate of GDP and the rate of inflation. Section 16.1 introduces a model for forecasting multiple variables, vector autoregressions (VARs), in which lagged values of two or more variables are used to forecast future values of those variables. Chapter 14 also focused on making forecasts one period (e.g., one quarter) into the future, but making forecasts two, three, or more periods into the future is important as well. Methods for making multiperiod forecasts are discussed in Section 16.2.

Sections 16.3 and 16.4 return to the topic of Section 14.6, stochastic trends. Section 16.3 introduces additional models of stochastic trends and an alternative test for a unit autoregressive root. Section 16.4 introduces the concept of cointegration, which arises when two variables share a common stochastic trend—that is, when each variable contains a stochastic trend, but a weighted difference of the two variables does not.

In some time series data, especially financial data, the variance changes over time: Sometimes the series exhibits high volatility, while at other times the volatility is low, so the data exhibit clusters of volatility. Section 16.5 discusses volatility clustering and introduces models in which the variance of the forecast error changes over time, that is, models in which the forecast error is conditionally heteroskedastic. Models of conditional heteroskedasticity have several applications. One application is computing forecast intervals, where the width of the interval changes over time to reflect periods of high or low uncertainty. Another application is forecasting the uncertainty of returns on an asset, such as a stock, which in turn can be useful in assessing the risk of owning that asset.

16.1 Vector Autoregressions

Chapter 14 focused on forecasting the growth rate of GDP, but in reality economic forecasters are in the business of forecasting other key macroeconomic variables as well, such as the rate of inflation, the unemployment rate, and interest rates. One approach is to develop a separate forecasting model for each variable,

Vector Autoregressions

KEY CONCEPT

16.1

A vector autoregression (VAR) is a set of k time series regressions, in which the regressors are lagged values of all k series. A VAR extends the univariate autoregression to a list, or "vector," of time series variables. When the number of lags in each of the equations is the same and is equal to p, the system of equations is called a VAR(p).

In the case of two time series variables, Y_t and X_t, the VAR(p) consists of the two equations

$$Y_t = \beta_{10} + \beta_{11}Y_{t-1} + \cdots + \beta_{1p}Y_{t-p} + \gamma_{11}X_{t-1} + \cdots + \gamma_{1p}X_{t-p} + u_{1t} \quad (16.1)$$

$$X_t = \beta_{20} + \beta_{21}Y_{t-1} + \cdots + \beta_{2p}Y_{t-p} + \gamma_{21}X_{t-1} + \cdots + \gamma_{2p}X_{t-p} + u_{2t}, \quad (16.2)$$

where the β's and the γ's are unknown coefficients and u_{1t} and u_{2t} are error terms.

The VAR assumptions are the time series regression assumptions of Key Concept 14.6, applied to each equation. The coefficients of a VAR are estimated by estimating each equation by OLS.

using the methods of Section 14.4. Another approach is to develop a single model that can forecast all the variables, which can help to make the forecasts mutually consistent. One way to forecast several variables with a single model is to use a vector autoregression (VAR). A VAR extends the univariate autoregression to multiple time series variables, that is, it extends the univariate autoregression to a "vector" of time series variables.

The VAR Model

A **vector autoregression (VAR)** with two time series variables, Y_t and X_t, consists of two equations: In one, the dependent variable is Y_t; in the other, the dependent variable is X_t. The regressors in both equations are lagged values of both variables. More generally, a VAR with k time series variables consists of k equations, one for each of the variables, where the regressors in all equations are lagged values of all the variables. The coefficients of the VAR are estimated by estimating each of the equations by OLS.

VARs are summarized in Key Concept 16.1.

Inference in VARs. Under the VAR assumptions, the OLS estimators are consistent and have a joint normal distribution in large samples. Accordingly, statistical inference proceeds in the usual manner; for example, 95% confidence intervals on coefficients can be constructed as the estimated coefficient ±1.96 standard errors.

One new aspect of hypothesis testing arises in VARs because a VAR with k variables is a collection, or system, of k equations. Thus it is possible to test joint hypotheses that involve restrictions across multiple equations.

For example, in the two-variable VAR(p) in Equations (16.1) and (16.2), you could ask whether the correct lag length is p or $p - 1$; that is, you could ask whether the coefficients on Y_{t-p} and X_{t-p} are zero in these two equations. The null hypothesis that these coefficients are zero is

$$H_0: \beta_{1p} = 0, \beta_{2p} = 0, \gamma_{1p} = 0, \text{ and } \gamma_{2p} = 0. \tag{16.3}$$

The alternative hypothesis is that at least one of these four coefficients is nonzero. Thus the null hypothesis involves coefficients from *both* of the equations, two from each equation.

Because the estimated coefficients have a jointly normal distribution in large samples, it is possible to test restrictions on these coefficients by computing an F-statistic. The precise formula for this statistic is complicated because the notation must handle multiple equations, so we omit it. In practice, most modern software packages have automated procedures for testing hypotheses on coefficients in systems of multiple equations.

How many variables should be included in a VAR? The number of coefficients in each equation of a VAR is proportional to the number of variables in the VAR. For example, a VAR with 5 variables and 4 lags will have 21 coefficients (4 lags each of 5 variables, plus the intercept) in each of the 5 equations, for a total of 105 coefficients! Estimating all these coefficients increases the amount of estimation error entering a forecast, which can result in deterioration of the accuracy of the forecast.

The practical implication is that one needs to keep the number of variables in a VAR small and, especially, to make sure the variables are plausibly related to each other so that they will be useful for forecasting one another. For example, we know from a combination of empirical evidence (such as that discussed in Chapter 14) and economic theory that the growth rate of GDP, the term spread, and the rate of inflation are related to one another, suggesting that these variables could help forecast one another in a VAR. Including an unrelated variable in a

VAR, however, introduces estimation error without adding predictive content, thereby reducing forecast accuracy.

Determining lag lengths in VARs. Lag lengths in a VAR can be determined using either F-tests or information criteria.

The information criterion for a system of equations extends the single-equation information criterion in Section 14.5. To define this information criterion, we need to adopt matrix notation. Let Σ_u be the $k \times k$ covariance matrix of the VAR errors and let $\hat{\Sigma}_u$ be the estimate of the covariance matrix where the i, j element of $\hat{\Sigma}_u$ is $\frac{1}{T}\sum_{t=1}^{T} \hat{u}_{it}\, \hat{u}_{jt}$, where $\hat{u}_{it}$ is the OLS residual from the i^{th} equation and $\hat{u}_{jt}$ is the OLS residual from the j^{th} equation. The BIC for the VAR is

$$\mathrm{BIC}(p) = \ln[\det(\hat{\Sigma}_u)] + k(kp + 1)\frac{\ln(T)}{T}, \tag{16.4}$$

where $\det(\hat{\Sigma}_u)$ is the determinant of the matrix $\hat{\Sigma}_u$. The AIC is computed using Equation (16.4), modified by replacing the term "$\ln(T)$" with "2."

The expression for the BIC for the k equations in the VAR in Equation (16.4) extends the expression for a single equation given in Section 14.5. When there is a single equation, the first term simplifies to $\ln[SSR(p)/T]$. The second term in Equation (16.4) is the penalty for adding additional regressors; $k(kp + 1)$ is the total number of regression coefficients in the VAR. (There are k equations, each of which has an intercept and p lags of each of the k time series variables.)

Lag length estimation in a VAR using the BIC proceeds analogously to the single equation case: Among a set of candidate values of p, the estimated lag length $\hat{p}$ is the value of p that minimizes $\mathrm{BIC}(p)$.

Using VARs for causal analysis. The discussion so far has focused on using VARs for forecasting. Another use of VAR models is for analyzing causal relationships among economic time series variables; indeed, it was for this purpose that VARs were first introduced to economics by the econometrician and macroeconomist Christopher Sims (1980). (See the box "Nobel Laureates in Time Series Econometrics" at the end of this chapter.) The use of VARs for causal inference is known as *structural VAR modeling*, "structural" because in this application VARs are used to model the underlying structure of the economy. Structural VAR analysis uses the techniques introduced in this section in the context of forecasting, plus some additional tools. The biggest conceptual difference between using VARs for forecasting and using them for structural modeling, however, is that structural modeling requires very specific assumptions, derived from economic theory and institutional knowledge, of what is exogenous and what is not. The discussion of structural VARs is best undertaken in the context of estimation of systems of simultaneous equations, which goes beyond the scope of this book. For an introduction to using VARs for forecasting and policy analysis, see Stock and Watson (2001). For additional mathematical detail on structural VAR modeling, see Hamilton (1994) or Watson (1994).

A VAR Model of the Growth Rate of GDP and the Term Spread

As an illustration, consider a two-variable VAR for the growth rate of GDP, $GDPGR_t$, and the term spread, $TSpread_t$. The VAR for $GDPGR_t$ and $TSpread_t$ consists of two equations: one in which $GDPGR_t$ is the dependent variable and one in which $TSpread_t$ is the dependent variable. The regressors in both equations are lagged values of $GDPGR_t$ and $TSpread_t$. Because of the apparent break in the relation in the early 1980s found in Section 14.7 using the QLR test, the VAR is estimated using data from 1981:Q1 to 2012:Q4.

The first equation of the VAR is the GDP growth rate equation:

$$\begin{aligned}\widehat{GDPGR_t} = \underset{(0.52)}{0.52} &+ \underset{(0.11)}{0.29}\,GDPGR_{t-1} + \underset{(0.09)}{0.22}\,GDPGR_{t-2}\\ &\underset{(0.36)}{-0.90}\,TSpread_{t-1} + \underset{(0.39)}{1.33}\,TSpread_{t-2}.\end{aligned} \tag{16.5}$$

The adjusted R^2 is $\overline{R}^2 = 0.29$.

The second equation of the VAR is the term spread equation, in which the regressors are the same as in the $GDPGR$ equation, but the dependent variable is the term spread:

$$\begin{aligned}\widehat{TSpread_t} = \underset{(0.12)}{0.46} &+ \underset{(0.02)}{0.01}\,GDPGR_{t-1} - \underset{(0.03)}{0.06}\,GDPGR_{t-2}\\ &+ \underset{(0.10)}{1.06}\,TSpread_{t-1} - \underset{(0.11)}{0.22}\,TSpread_{t-2}.\end{aligned} \tag{16.6}$$

The adjusted R^2 is $\overline{R}^2 = 0.83$.

Equations (16.5) and (16.6), taken together, are a VAR(2) model of the growth rate of GDP, $GDPGR_t$, and the term spread, $TSpread_t$.

These VAR equations can be used to perform Granger causality tests. The F-statistic testing the null hypothesis that the coefficients on $TSpread_{t-1}$ and $TSpread_{t-2}$ are zero in the GDP growth rate equation [Equation (16.5)] is 5.91, which has a p-value less than 0.001. Thus the null hypothesis is rejected, so we can conclude that the term spread is a useful predictor of the growth rate of GDP, given lags in the growth rate of GDP (that is, the term spread rate Granger-causes the growth rate of GDP). The F-statistic testing the hypothesis that the coefficients on the two lags of $GDPGR_t$ are zero in the term spread equation [Equation (16.6)] is 3.48, which has a p-value of 0.03. Thus the growth rate of GDP Granger-causes the term spread at the 5% significance level.

Forecasts of the growth rate of GDP and the term spread one period ahead are obtained exactly as discussed in Section 14.4. The forecast of the growth rate of GDP for 2013:Q1, based on Equation (16.5), is $\widehat{GDP}_{2013:Q1|2012:Q4} = 1.7$ percentage point. A similar calculation using Equation (16.6) gives a forecast of the term spread 2013:Q1, based on data through 2012:Q4 of $\widehat{TSpread}_{2013:Q1|2012:Q4} = 1.7\%$. The actual values for 2013:Q1 are $GDPGR_{2013:Q1} = 1.1\%$ and $TSpread_{2013:Q1} = 1.9\%$.

16.2 Multiperiod Forecasts

The discussion of forecasting so far has focused on making forecasts one period in advance. Often, however, forecasters are called upon to make forecasts further into the future. This section describes two methods for making multiperiod forecasts. The usual method is to construct "iterated" forecasts, in which a one-period-ahead model is iterated forward one period at a time, in a way that is made precise in this section. The second method is to make "direct" forecasts by using a regression in which the dependent variable is the multiperiod variable that one wants to forecast. For reasons discussed at the end of this section, in most applications, the iterated method is recommended over the direct method.

Iterated Multiperiod Forecasts

The essential idea of an iterated forecast is that a forecasting model is used to make a forecast one period ahead, for period $T + 1$, using data through period T. The model then is used to make a forecast for date $T + 2$, given the data through date T, where the forecasted value for date $T + 1$ is treated as data for the purpose of making the forecast for period $T + 2$. Thus the one-period-ahead forecast (which is also referred to as a one-step-ahead forecast) is used as an intermediate step to make the two-period-ahead forecast. This process repeats, or iterates, until the forecast is made for the desired forecast horizon h.

The iterated AR forecast method: AR(1). An iterated AR(1) forecast uses an AR(1) for the one-period-ahead model. For example, consider the first-order autoregression for $GDPGR$ [Equation (14.7)]:

$$\underset{}{\widehat{GDPGR_t}} = \underset{(0.35)}{1.99} + \underset{(0.08)}{0.34}\, GDPGR_{t-1}. \tag{16.7}$$

The first step in computing the two-quarter-ahead forecast of $GDPGR_{2013:Q2}$ based on Equation (16.7) using data through 2012:Q4 is to compute the one-quarter-ahead forecast of $GDPGR_{2013:Q1}$ based on data through 2012:Q4: $\widehat{GDPGR}_{2013:Q1|2012:Q4} = 1.99 + 0.34\, GDPGR_{2012:Q4} = 1.99 + 0.34 \times 0.15 = 2.0$. The second step is to substitute this forecast into Equation (16.7) so that $\widehat{GDPGR}_{2013:Q2|2012:Q4} = 1.99 + 0.34\, \widehat{GDPGR}_{2013:Q1|2012:Q4} = 1.99 + 0.34 \times 2.0 = 2.7$. Thus, based on information through the fourth quarter of 2012, this forecast states that the growth rate of GDP will be 2.7% in the second quarter of 2013.

The iterated AR forecast method: AR(p). The iterated AR(1) strategy is extended to an AR(p) by replacing Y_{T+1} with its forecast, $\hat{Y}_{T+1|T}$, and then treating that forecast as data for the AR(p) forecast of Y_{T+2}. For example, consider the iterated two-period-ahead forecast of the growth rate of GDP based on the AR(2) model from Section 14.3 [Equation (14.13)]:

$$\widehat{GDPGR_t} = \underset{(0.40)}{1.63} + \underset{(0.08)}{0.28}\, GDPGR_{t-1} + \underset{(0.08)}{0.18}\, GDPGR_{t-2}. \tag{16.8}$$

The forecast of $GDPGR_{2013:Q1}$ based on data through 2012:Q4 using this AR(2), computed in Section 14.3, is $\widehat{GDPGR}_{2013:Q1|2012:Q4} = 2.1$. Thus the two-quarter-ahead iterated forecast based on the AR(2) is $\widehat{GDPGR}_{2013:Q2|2012:Q4} = 1.63 + 0.28\, \widehat{GDPGR}_{2013:Q1|2012:Q4} + 0.18\, GDPGR_{2012:Q4} = 1.63 + 0.28 \times 2.1 + 0.18 \times 0.15 = 2.2$. According to this iterated AR(2) forecast, based on data through the fourth quarter of 2012, the growth rate of GDP is predicted to be 2.2 percentage points in the second quarter of 2013.

Iterated multivariate forecasts using an iterated VAR. Iterated multivariate forecasts can be computed using a VAR in much the same way as iterated univariate forecasts are computed using an autoregression. The main new feature of an iterated multivariate forecast is that the two-step-ahead (period $T + 2$) forecast of one variable depends on the forecasts of all variables in the VAR in period $T + 1$. For example, to compute the forecast of the growth rate of GDP in period $T + 2$ using a VAR with the variables $GDPGR_t$ and $TSpread_t$, one must forecast both $GDPGR_{T+1}$ and $TSpread_{T+1}$, using data through period T as an intermediate step in forecasting $GDPGR_{T+2}$. More generally, to compute multiperiod iterated VAR forecasts h periods ahead, it is necessary to compute forecasts of all variables for all intervening periods between T and $T + h$.

As an example, we will compute the iterated VAR forecast of $GDPGR_{2013:Q2}$ based on data through 2012:Q4, using the VAR(2) for $GDPGR_t$ and $TSpread_t$ in Section 16.1 [Equations (16.5) and (16.6)]. The first step is to compute the one-quarter-ahead forecasts $\widehat{GDPGR}_{2013:Q1|2012:Q4}$ and $\widehat{TSpread}_{2013:Q1|2012:Q4}$ from that VAR. These one-period-ahead forecasts were computed in Section 16.1 based on Equations (16.5) and (16.6). The forecasts were $\widehat{GDPGR}_{2013:Q1|2012:Q4} = 1.7$ and $\widehat{TSpread}_{2013:Q1|2012:Q4} = 1.7$. In the second step, these forecasts are substituted into Equations (16.5) and (16.6) to produce the two-quarter-ahead forecast:

$$\begin{aligned}\widehat{GDPGR}_{2013:Q2|2012:Q4} &= 0.52 + 0.29\,\widehat{GDPGR}_{2013:Q1|2012:Q4} + 0.22GDPGR_{2012:Q4}\\ &\quad - 0.90\,\widehat{TSpread}_{2013:Q1|2012:Q4} + 1.33TSpread_{2012:Q4}\\ &= 0.52 + 0.30 \times 1.7 + 0.22 \times 0.15\\ &\quad - 0.90 \times 1.7 + 1.33 \times 1.6 = 1.7. \end{aligned} \tag{16.9}$$

Thus the iterated VAR(2) forecast, based on data through the fourth quarter of 2012, is that the growth rate of GDP will be 1.7% in the second quarter of 2013.

Iterated multiperiod forecasts are summarized in Key Concept 16.2.

Direct Multiperiod Forecasts

Direct multiperiod forecasts are computed without iterating by using a single regression in which the dependent variable is the multiperiod-ahead variable to be forecasted and the regressors are the predictor variables. Forecasts computed this way are called direct forecasts because the regression coefficients can be used directly to make the multiperiod forecast.

The direct multiperiod forecasting method. Suppose that you want to make a forecast of Y_{T+2} using data through time T. The direct multivariate method takes the ADL model as its starting point but lags the predictor variables by an additional time period. For example, if two lags of the predictors are used, then the

KEY CONCEPT 16.2

Iterated Multiperiod Forecasts

The **iterated multiperiod AR forecast** is computed in steps: First compute the one-period-ahead forecast, then use that to compute the two-period-ahead forecast, and so forth. The two- and three-period-ahead iterated forecasts based on an AR(p) are

$$\hat{Y}_{T+2|T} = \hat{\beta}_0 + \hat{\beta}_1\hat{Y}_{T+1|T} + \hat{\beta}_2 Y_T + \hat{\beta}_3 Y_{T-1} + \cdots + \hat{\beta}_p Y_{T-p+2} \quad (16.10)$$

$$\hat{Y}_{T+3|T} = \hat{\beta}_0 + \hat{\beta}_1\hat{Y}_{T+2|T} + \hat{\beta}_2\hat{Y}_{T+1|T} + \hat{\beta}_3 Y_T + \cdots + \hat{\beta}_p Y_{T-p+3}, \quad (16.11)$$

where the $\hat{\beta}$'s are the OLS estimates of the AR(p) coefficients. Continuing this process ("iterating") produces forecasts further into the future.

The **iterated multiperiod VAR forecast** is also computed in steps: First compute the one-period-ahead forecast of all the variables in the VAR, then use those forecasts to compute the two-period-ahead forecasts, and continue this process iteratively to the desired forecast horizon. The two-period-ahead iterated forecast of Y_{T+2}, based on the two-variable VAR(p) in Key Concept 16.1, is

$$\hat{Y}_{T+2|T} = \hat{\beta}_{10} + \hat{\beta}_{11}\hat{Y}_{T+1|T} + \hat{\beta}_{12}Y_T + \hat{\beta}_{13}Y_{T-1} + \cdots + \hat{\beta}_{1p}Y_{T-p+2}$$
$$+ \hat{\gamma}_{11}\hat{X}_{T+1|T} + \hat{\gamma}_{12}X_T + \hat{\gamma}_{13}X_{T-1} + \cdots + \hat{\gamma}_{1p}X_{T-p+2}, \quad (16.12)$$

where the coefficients in Equation (16.12) are the OLS estimates of the VAR coefficients. Iterating produces forecasts further into the future.

dependent variable is Y_t and the regressors are Y_{t-2}, Y_{t-3}, X_{t-2}, and X_{t-3}. The coefficients from this regression can be used directly to compute the forecast of Y_{T+2} using data on Y_T, Y_{T-1}, X_T, and X_{T-1}, without the need for any iteration. More generally, in a direct h-period-ahead forecasting regression, all predictors are lagged h periods to produce the h-period-ahead forecast.

For example, the forecast of $GDPGR_t$ two quarters ahead using two lags each of $GDPGR_{t-2}$ and $TSpread_{t-2}$ is computed by first estimating the regression:

$$\begin{aligned}\widehat{GDPGR}_{t|t-2} = \underset{(0.67)}{0.57} + \underset{(0.07)}{0.34}GDPGR_{t-2} + \underset{(0.10)}{0.03}GDPGR_{t-3}\\ + \underset{(0.47)}{0.62}TSpread_{t-2} - \underset{(0.46)}{0.01}TSpread_{t-3}.\end{aligned} \quad (16.13)$$

The two-quarter-ahead forecast of the growth rate of GDP in 2013:Q2 based on data through 2012:Q4 is computed by substituting the values of $GDPGR_{2012:Q4}$, $GDPGR_{2012:Q3}$, $TSpread_{2012:Q4}$, and $TSpread_{2012:Q3}$ into Equation (16.13); this yields

$$\widehat{GDPGR}_{2013:Q2|2012:Q4} = 0.57 + 0.34GDPGR_{2012:Q4} + 0.03GDPGR_{2012:Q3} + 0.62TSpread_{2012:Q4} - 0.01TSpread_{2012:Q3} = 1.68. \tag{16.14}$$

The three-quarter-ahead direct forecast of$GDPGR_{T+3}$ is computed by lagging all the regressors in Equation (16.13) by one additional quarter, estimating that regression, and then computing the forecast. The h-quarter-ahead direct forecast of $GDPGR_{T+h}$ is computed by using $GPDGR_t$ as the dependent variable and the regressors $GPDGR_{t-h}$ and $TSpread_{t-h}$, plus additional lags of $GPDGR_{t-h}$ and $TSpread_{t-h}$, as desired.

Standard errors in direct multiperiod regressions. Because the dependent variable in a multiperiod regression occurs two or more periods into the future, the error term in a multiperiod regression is serially correlated. To see this, consider the two-period-ahead forecast of the growth rate of GDP and suppose that a surprise jump in oil prices occurs in the next quarter. Today's two-period-ahead forecast of the growth rate of GDP will be too low because it does not incorporate this unexpected event. Because the oil price rise was also unknown in the previous quarter, the two-period-ahead forecast made last quarter will also be too low. Thus the surprise oil price jump next quarter means that *both* last quarter's and this quarter's two-period-ahead forecasts are too low. Because of such intervening events, the error term in a multiperiod regression is serially correlated.

As discussed in Section 15.4, if the error term is serially correlated, the usual OLS standard errors are incorrect or, more precisely, they are not a reliable basis for inference. Therefore, heteroskedasticity- and autocorrelation-consistent (HAC) standard errors must be used with direct multiperiod regressions. The standard errors reported in Equation (16.13) for direct multiperiod regressions therefore are Newey–West HAC standard errors, where the truncation parameter m is set according to Equation (15.17); for these data (for which $T = 128$), Equation (15.17) yields $m = 4$. For longer forecast horizons, the amount of overlap—and thus the degree of serial correlation in the error—increases: In general, the first $h - 1$ autocorrelation coefficients of the errors in an h-period-ahead regression are nonzero. Thus larger values of m than indicated by Equation (15.17) are appropriate for multiperiod regressions with long forecast horizons.

Direct multiperiod forecasts are summarized in Key Concept 16.3.

KEY CONCEPT

16.3

Direct Multiperiod Forecasts

The **direct multiperiod forecast** h periods into the future based on p lags each of Y_t and an additional predictor X_t is computed by first estimating the regression

$$\begin{aligned} Y_t = \delta_0 + \delta_1 Y_{t-h} + \cdots + \delta_p Y_{t-p-h+1} + \delta_{p+1} X_{t-h} \\ + \cdots + \delta_{2p} X_{t-p-h+1} + u_t, \end{aligned} \tag{16.15}$$

and then using the estimated coefficients directly to make the forecast of Y_{T+h} using data through period T.

Which Method Should You Use?

In most applications, the iterated method is the recommended procedure for multiperiod forecasting, for two reasons. First, from a theoretical perspective, if the underlying one-period-ahead model (the AR or VAR that is used to compute the iterated forecast) is specified correctly, then the coefficients are estimated more efficiently if they are estimated by a one-period-ahead regression (and then iterated) than by a multiperiod-ahead regression. Second, from a practical perspective, forecasters are usually interested in forecasts not just at a single horizon but at multiple horizons. Because they are produced using the same model, iterated forecasts tend to have time paths that are less erratic across horizons than do direct forecasts. Because a different model is used at every horizon for direct forecasts, sampling error in the estimated coefficients can add random fluctuations to the time paths of a sequence of direct multiperiod forecasts.

Under some circumstances, however, direct forecasts are preferable to iterated forecasts. One such circumstance is when you have reason to believe that the one-period-ahead model (the AR or VAR) is not specified correctly. For example, you might believe that the equation for the variable you are trying to forecast in a VAR is specified correctly, but that one or more of the other equations in the VAR is specified incorrectly, perhaps because of neglected nonlinear terms. If the one-step-ahead model is specified incorrectly, then in general the iterated multiperiod forecast will be biased, and the MSFE of the iterated forecast can exceed the MSFE of the direct forecast, even though the direct forecast has a larger variance. A second circumstance in which a direct forecast might be desirable arises in multivariate forecasting models with many predictors, in which case a VAR specified in terms of all the variables could be unreliable because it would have very many estimated coefficients.

16.3 Orders of Integration and the DF-GLS Unit Root Test

This section extends the treatment of stochastic trends in Section 14.6 by addressing two further topics. First, the trends of some time series are not well described by the random walk model, so we introduce an extension of that model and discuss its implications for regression modeling of such series. Second, we continue the discussion of testing for a unit root in time series data and, among other things, introduce a second test for a unit root, the DF-GLS test.

Other Models of Trends and Orders of Integration

Recall that the random walk model for a trend, introduced in Section 14.6, specifies that the trend at date t equals the trend at date $t - 1$, plus a random error term. If Y_t follows a random walk with drift β_0, then

$$Y_t = \beta_0 + Y_{t-1} + u_t, \tag{16.16}$$

where u_t is serially uncorrelated. Also recall from Section 14.6 that, if a series has a random walk trend, then it has an autoregressive root that equals 1.

Although the random walk model of a trend describes the long-run movements of many economic time series, some economic time series have trends that are smoother—that is, vary less from one period to the next—than is implied by Equation (16.16). A different model is needed to describe the trends of such series.

One model of a smooth trend makes the first difference of the trend follow a random walk—that is,

$$\Delta Y_t = \beta_0 + \Delta Y_{t-1} + u_t, \tag{16.17}$$

where u_t is serially uncorrelated. Thus, if Y_t follows Equation (16.17), ΔY_t follows a random walk, so $\Delta Y_t - \Delta Y_{t-1}$ is stationary. The difference of the first differences, $\Delta Y_t - \Delta Y_{t-1}$, is called the **second difference** of Y_t and is denoted $\Delta^2 Y_t = \Delta Y_t - \Delta Y_{t-1}$. In this terminology, if Y_t follows Equation (16.17), then its second

KEY CONCEPT 16.4

Orders of Integration, Differencing, and Stationarity

- If Y_t is integrated of order one—that is, if Y_t is $I(1)$—then Y_t has a unit autoregressive root and its first difference, ΔY_t, is stationary.
- If Y_t is integrated of order two—that is, if Y_t is $I(2)$—then ΔY_t has a unit autoregressive root and its second difference, $\Delta^2 Y_t$, is stationary.
- If Y_t is **integrated of order** $\boldsymbol{d}$—that is, if Y_t is $\boldsymbol{I(d)}$—then Y_t must be differenced d times to eliminate its stochastic trend; that is, $\Delta^d Y_t$ is stationary.

difference is stationary. If a series has a trend of the form in Equation (16.17), then the first difference of the series has an autoregressive root that equals 1.

"Orders of integration" terminology. Some additional terminology is useful for distinguishing between these two models of trends. A series that has a random walk trend is said to be **integrated of order one**, or $\boldsymbol{I(1)}$. A series that has a trend of the form in Equation (16.17) is said to be **integrated of order two**, or $\boldsymbol{I(2)}$. A series that does not have a stochastic trend and is stationary is said to be **integrated of order zero**, or $\boldsymbol{I(0)}$.

The **order of integration** in the $I(1)$ and $I(2)$ terminology is the number of times that the series needs to be differenced for it to be stationary: If Y_t is $I(1)$, then the first difference of Y_t, ΔY_t, is stationary, and if Y_t is $I(2)$, then the second difference of Y_t, $\Delta^2 Y_t$, is stationary. If Y_t is $I(0)$, then Y_t is stationary.

Orders of integration are summarized in Key Concept 16.4.

How to test whether a series is I(2) or I(1). If Y_t is $I(2)$, then ΔY_t is $I(1)$, so ΔY_t has an autoregressive root that equals 1. If, however, Y_t is $I(1)$, then ΔY_t is stationary. Thus the null hypothesis that Y_t is $I(2)$ can be tested against the alternative hypothesis that Y_t is $I(1)$ by testing whether ΔY_t has a unit autoregressive root. If the hypothesis that ΔY_t has a unit autoregressive root is rejected, then the hypothesis that Y_t is $I(2)$ is rejected in favor of the alternative that Y_t is $I(1)$.

Examples of I(2) and I(1) series: The price level and the rate of inflation. The rate of inflation is the growth rate of the price level. Recall from Section 14.2 that the growth rate of a time series X_t can be computed as the first difference of the logarithm of X_t; that is $\Delta\ln(X_t)$ is the growth rate of X_t (expressed as fraction). If P_t is a time series for the price level measured quarterly, then $\Delta\ln(P_t)$ is its growth rate, and $Infl_t = 400 \times \Delta\ln(P_t)$ is the quarterly rate of inflation, measured in percentage points at an annual rate. As in the expression for the growth of GDP, *GDPGR* in Equation (14.2), the factor 400 arises from converting fractional changes to percentage changes (multiplying by 100) and converting quarterly percentages to an annual rate (multiplying by 4).

In Empirical Exercise 14.1, you analyzed the rate of inflation, $Infl_t$, computed using the price index for personal consumption expenditures in the United States as P_t. In that exercise you concluded that the rate of inflation in the United States plausibly has a random walk stochastic trend—that is, that the rate of inflation is $I(1)$. If inflation is $I(1)$, then its stochastic trend is removed by first differencing, so ΔInf_t is stationary. But treating inflation as $I(1)$ is equivalent to treating $\Delta\ln(P_t)$ as $I(1)$, but this in turn is equivalent to treating the logarithm of the price level, $\ln(P_t)$, as $I(2)$.

The logarithm of the price level and the rate of inflation are plotted in Figure 16.1. The long-run trend of the logarithm of the price level (Figure 16.1a) varies more smoothly than the long-run trend in the rate of inflation (Figure 16.1b). The smoothly varying trend in the logarithm of the price level is typical of $I(2)$ series.

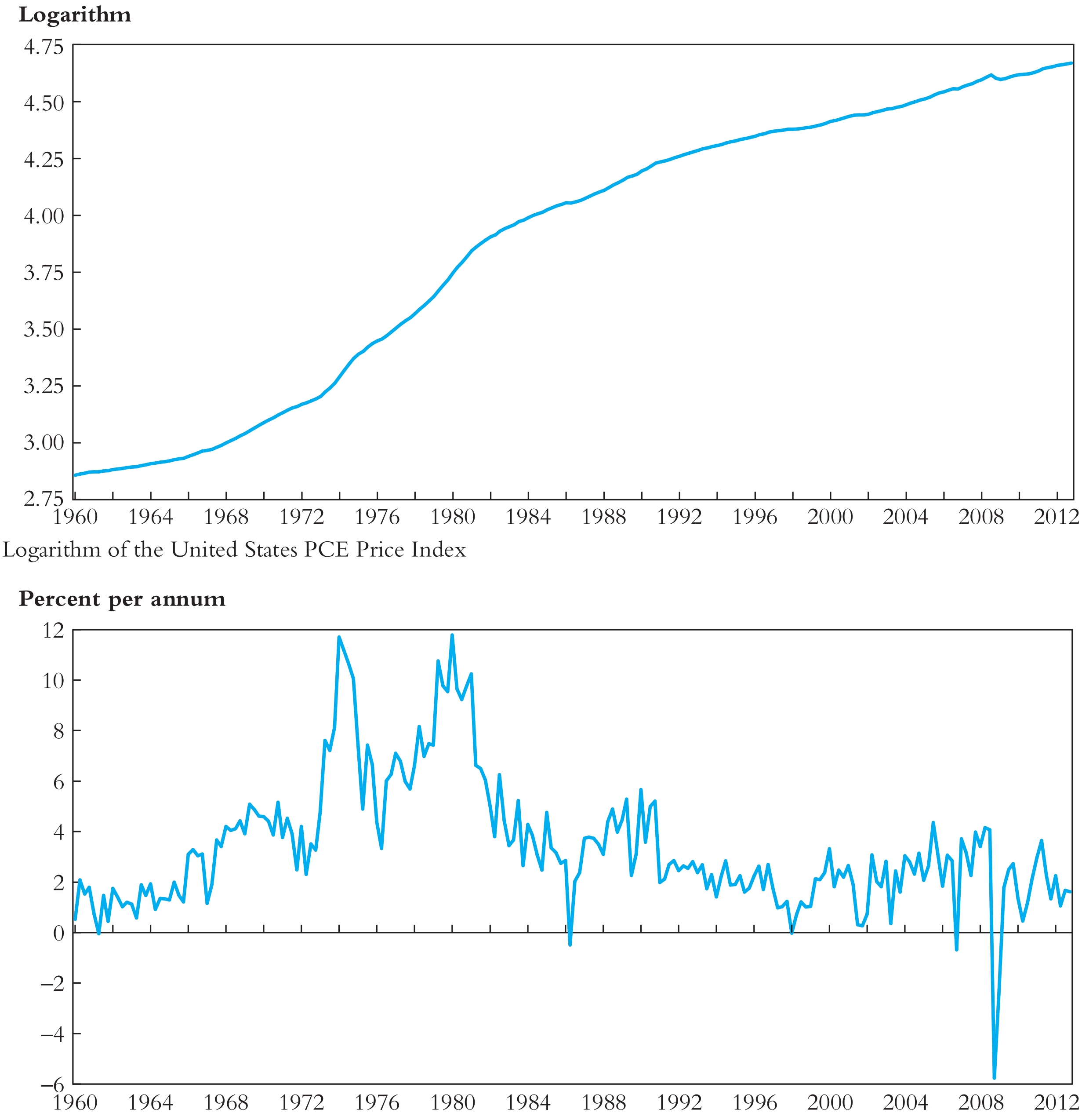
The DF-GLS Test for a Unit Root

This section continues the discussion of Section 14.6 regarding testing for a unit autoregressive root. We first describe another test for a unit autoregressive root, the so-called DF-GLS test. Next, in an optional mathematical section, we discuss why unit root test statistics do not have normal distributions, even in large samples.

The DF-GLS test. The ADF test was the first test developed for testing the null hypothesis of a unit root and is the most commonly used test in practice. Other tests subsequently have been proposed, however, many of which have higher power (Key Concept 3.5) than the ADF test. A test with higher power than the ADF test is more likely to reject the null hypothesis of a unit root against the stationary alternative when the alternative is true; thus a more powerful test is better able to distinguish between a unit AR root and a root that is large but less than 1.

This section discusses one such test, the **DF-GLS test** developed by Elliott, Rothenberg, and Stock (1996). The test is introduced for the case that, under the null hypothesis, Y_t has a random walk trend, possibly with drift, and under the alternative Y_t is stationary around a linear time trend.

FIGURE 16.1 The Logarithm of the Price Level and the Inflation Rate in the United States, 1960–2012



(a) Logarithm of the United States PCE Price Index

(b) United States PCE price inflation

The trend in the logarithm of prices (Figure 16.1a) is much smoother than the trend in inflation (Figure 16.1b).

The DF-GLS test is computed in two steps. In the first step, the intercept and trend are estimated by generalized least squares (GLS; see Section 15.5). The GLS estimation is performed by computing three new variables, V_t, X_{1t}, and X_{2t}, where $V_1 = Y_1$ and $V_t = Y_t - \alpha^* Y_{t-1}, t = 2, \ldots, T, X_{11} = 1$ and $X_{1t} = 1 - \alpha^*$, $t = 2, \ldots, T$, and $X_{21} = 1$ and $X_{2t} = t - \alpha^*(t-1)$, where α^* is computed using the formula $\alpha^* = 1 - 13.5/T$. Then V_t is regressed against X_{1t} and X_{2t}; that is, OLS is used to estimate the coefficients of the population regression equation

$$V_t = \delta_0 X_{1t} + \delta_1 X_{2t} + e_t, \tag{16.18}$$

using the observations $t = 1, \ldots, T$, where e_t is the error term. Note that there is no intercept in the regression in Equation (16.18). The OLS estimators $\hat{\delta}_0$ and $\hat{\delta}_1$ are then used to compute a "detrended" version of Y_t, $Y_t^d = Y_t - (\hat{\delta}_0 + \hat{\delta}_1 t)$.

In the second step, the Dickey–Fuller test is used to test for a unit autoregressive root in Y_t^d, where the Dickey–Fuller regression does not include an intercept or a time trend. That is, ΔY_t^d is regressed against Y_{t-1}^d and $\Delta Y_{t-1}^d, \ldots, \Delta Y_{t-p}^d$, where the number of lags p is determined, as usual, either by expert knowledge or by using a data-based method such as the AIC or BIC, as discussed in Section 14.5.

If the alternative hypothesis is that Y_t is stationary with a mean that might be nonzero but without a time trend, the preceding steps are modified. Specifically, α^* is computed using the formula $\alpha^* = 1 - 7/T$, X_{2t} is omitted from the regression in Equation (16.18), and the series Y_t^d is computed as $Y_t^d = Y_t - \hat{\delta}_0$.

The GLS regression in the first step of the DF-GLS test makes this test more complicated than the conventional ADF test, but it is also what improves its ability to discriminate between the null hypothesis of a unit autoregressive root and the alternative that Y_t is stationary. This improvement can be substantial. For example, suppose that Y_t is in fact a stationary AR(1) with autoregressive coefficient $\beta_1 = 0.95$, that there are $T = 200$ observations, and that the unit root tests are computed without a time trend [that is, t is excluded from the Dickey–Fuller regression, and X_{2t} is omitted from Equation (16.18)]. Then the probability that the ADF test correctly rejects the null hypothesis at the 5% significance level is approximately 31% compared to 75% for the DF-GLS test.

Critical values for DF-GLS test. Because the coefficients on the deterministic terms are estimated differently in the ADF and DF-GLS tests, the tests have different critical values. The critical values for the DF-GLS test are given in Table 16.1. If the DF-GLS test statistic (the t-statistic on Y_{t-1}^d in the regression in the second

TABLE 16.1 Critical Values of the DF-GLS Test

Deterministic Regressors [Regressors in Equation (16.18)]	10%	5%	1%
Intercept only (X_{1t} only)	−1.62	−1.95	−2.58
Intercept and time trend (X_{1t} and X_{2t})	−2.57	−2.89	−3.48

Source: Fuller (1976) and Elliott, Rothenberg, and Stock (1996, Table 1).

step) is less than the critical value (that is, it is more negative than the critical value), then the null hypothesis that Y_t has a unit root is rejected. Like the critical values for the Dickey–Fuller test, the appropriate critical value depends on which version of the test is used—that is, on whether or not a time trend is included [whether or not X_{2t} is included in Equation (16.18)].

Application to the logarithm of GDP. The DF-GLS statistic, computed for the logarithm of GDP, $\ln(GDP_t)$, over the period 1962:Q1 to 2012:Q4 with an intercept and time trend, is −2.85 when two lags of ΔY_t^d are included in the Dickey–Fuller regression in the second stage, where the choice of two lags was based on the AIC (out of a maximum of six lags). This value is greater than the 5% critical value in Table 16.1, −2.89, so the DF-GLS test does not reject the null hypothesis of a unit root at the 5% significance level.

Why Do Unit Root Tests Have Nonnormal Distributions?

In Section 14.6, it was stressed that the large-sample normal distribution on which regression analysis relies so heavily does not apply if the regressors are nonstationary. Under the null hypothesis that the regression contains a unit root, the regressor Y_{t-1} in the Dickey–Fuller regression (and the regressor Y_{t-1}^d in the modified Dickey–Fuller regression in the second step of the DF-GLS test) is nonstationary. The nonnormal distribution of the unit root test statistics is a consequence of this nonstationarity.

To gain some mathematical insight into this nonnormality, consider the simplest possible Dickey–Fuller regression, in which ΔY_t is regressed against the single regressor Y_{t-1} and the intercept is excluded. In the notation of Key Concept 14.8, the OLS estimator in this regression is $\hat{\delta} = \sum_{t=1}^{T} Y_{t-1}\Delta Y_t / \sum_{t=1}^{T} Y_{t-1}^2$, so

$$T\hat{\delta} = \frac{\frac{1}{T}\sum_{t=1}^{T} Y_{t-1}\Delta Y_t}{\frac{1}{T^2}\sum_{t=1}^{T} Y_{t-1}^2}. \tag{16.19}$$

Consider the numerator in Equation (16.19). Under the additional assumption that $Y_0 = 0$, a bit of algebra (Exercise 16.5) shows that

$$\frac{1}{T}\sum_{t=1}^{T} Y_{t-1}\Delta Y_t = \frac{1}{2}\left[\left(\frac{Y_T}{\sqrt{T}}\right)^2 - \frac{1}{T}\sum_{t=1}^{T}(\Delta Y_t)^2\right]. \tag{16.20}$$

Under the null hypothesis, $\Delta Y_t = u_t$, which is serially uncorrelated and has a finite variance, so the second term in Equation (16.20) has the probability limit $\frac{1}{T}\Sigma_{t=1}^{T}(\Delta Y_t)^2 \xrightarrow{p} \sigma_u^2$. Under the assumption that $Y_0 = 0$, the first term in Equation (16.20) can be written $Y_T/\sqrt{T} = \sqrt{\frac{1}{T}}\Sigma_{t=1}^{T}\,\Delta Y_t = \sqrt{\frac{1}{T}}\Sigma_{t=1}^{T} u_t$, which in turn obeys the central limit theorem; that is, $Y_T/\sqrt{T} \xrightarrow{d} N(0, \sigma_u^2)$. Thus $(Y_T/\sqrt{T})^2 - \frac{1}{T}\Sigma_{t=1}^{T}\,(\Delta Y_t)^2 \xrightarrow{d} \sigma_u^2(Z^2 - 1)$, where Z is a standard normal random variable. Recall, however, that the square of a standard normal distribution has a chi-squared distribution with 1 degree of freedom. It therefore follows from Equation (16.20) that, under the null hypothesis, the numerator in Equation (16.19) has the limiting distribution

$$\frac{1}{T}\sum_{t=1}^{T} Y_{t-1}\Delta Y_t \xrightarrow{d} \frac{\sigma_u^2}{2}(\chi_1^2 - 1). \tag{16.21}$$

The large-sample distribution in Equation (16.21) is different than the usual large-sample normal distribution when the regressor is stationary. Instead, the numerator of the OLS estimator of the coefficient on Y_t in this Dickey–Fuller regression has a distribution that is proportional to a chi-squared distribution with 1 degree of freedom minus 1.

This discussion has considered only the numerator of $T\hat{\delta}$. The denominator also behaves unusually under the null hypothesis: Because Y_t follows a random walk under the null hypothesis, $\frac{1}{T}\Sigma_{t=1}^{T} Y_{t-1}^2$ does not converge in probability to a constant. Instead, the denominator in Equation (16.19) is a random variable, even in large samples: Under the null hypothesis, $\frac{1}{T^2}\Sigma_{t=1}^{T} Y_{t-1}^2$ converges in distribution jointly with the numerator. The unusual distributions of the numerator and denominator in Equation (16.19) are the source of the nonstandard distribution of the Dickey–Fuller test statistic and the reason that the ADF statistic has its own special table of critical values.

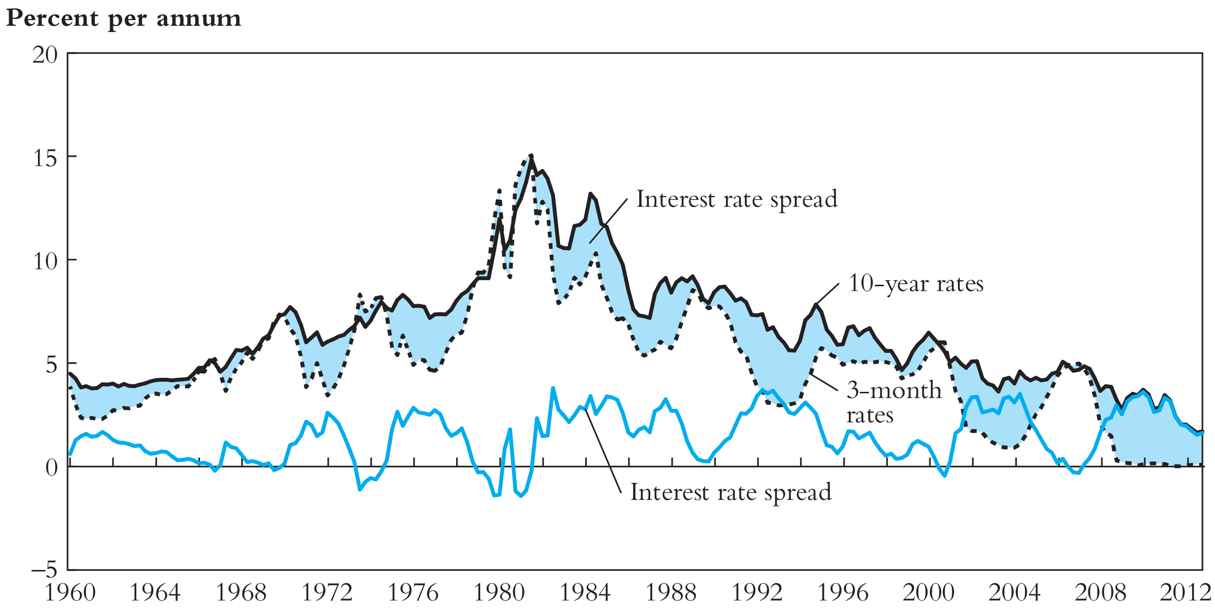
16.4 Cointegration

Sometimes two or more series have the same stochastic trend in common. In this special case, referred to as cointegration, regression analysis can reveal long-run relationships among time series variables, but some new methods are needed.

Cointegration and Error Correction

Two or more time series with stochastic trends can move together so closely over the long run that they appear to have the same trend component; that is, they appear to have a **common trend**. For example, Figure 16.2 reproduces the plot of the 10-year and 3-month interest rates from Figure 14.3. The interest rates exhibit the same long-run tendencies or trends: Both were low in the 1960s, both rose through the 1970s to peaks in the early 1980s, then both fell through the 1990s. However, the difference between the long-term and short-term interest rates, the term spread, does not appear to have a trend. That is, subtracting the short-term rate from the long-term rate appears to eliminate the trends in both of the

FIGURE 16.2 10-Year Interest Rate, 3-Month Interest Rate, and the Term Spread



10-year and 3-month interest rates share a common stochastic trend. The term spread, or the difference, between the two rates does not exhibit a trend. These two interest rates appear to be cointegrated.

Cointegration

KEY CONCEPT 16.5

Suppose that X_t and Y_t are integrated of order one. If, for some coefficient θ, $Y_t - \theta X_t$ is integrated of order zero, then X_t and Y_t are said to be *cointegrated*. The coefficient θ is called the **cointegrating coefficient**.

If X_t and Y_t are cointegrated, then they have the same, or common, stochastic trend. Computing the difference $Y_t - \theta X_t$ eliminates this common stochastic trend.

individual rates. Said differently, although the two interest rates differ, they appear to share a common stochastic trend: Because the trend in each individual series is eliminated by subtracting one series from the other, the two series must have the same trend; that is, they must have a common stochastic trend.

Two or more series that have a common stochastic trend are said to be cointegrated. The formal definition of **cointegration** (due to the econometrician Clive Granger, 1983; see the box "Nobel Laureates in Time Series Econometrics") is given in Key Concept 16.5. In this section, we introduce a test for whether cointegration is present, discuss estimation of the coefficients of regressions relating cointegrated variables, and illustrate the use of the cointegrating relationship for forecasting. The discussion initially focuses on the case that there are only two variables, X_t and Y_t.

Vector error correction model. Until now, we have eliminated the stochastic trend in an $I(1)$ variable Y_t by computing its first difference, ΔY_t; the problems created by stochastic trends were then avoided by using ΔY_t instead of Y_t in time series regressions. If X_t and Y_t are cointegrated, however, another way to eliminate the trend is to compute $Y_t - \theta X_t$, where θ is chosen to eliminate the common trend from the difference. Because the term $Y_t - \theta X_t$ is stationary, it too can be used in regression analysis.

In fact, if X_t and Y_t are cointegrated, the first differences of X_t and Y_t can be modeled using a VAR, augmented by including $Y_{t-1} - \theta X_{t-1}$ as an additional regressor:

$$\Delta Y_t = \beta_{10} + \beta_{11}\Delta Y_{t-1} + \cdots + \beta_{1p}\Delta Y_{t-p} + \gamma_{11}\Delta X_{t-1} + \cdots + \gamma_{1p}\Delta X_{t-p} + \alpha_1(Y_{t-1} - \theta X_{t-1}) + u_{1t} \quad (16.22)$$

$$\Delta X_t = \beta_{20} + \beta_{21}\Delta Y_{t-1} + \cdots + \beta_{2p}\Delta Y_{t-p} + \gamma_{21}\Delta X_{t-1} + \cdots + \gamma_{2p}\Delta X_{t-p} + \alpha_2(Y_{t-1} - \theta X_{t-1}) + u_{2t}. \quad (16.23)$$

The term $Y_t - \theta X_t$ is called the **error correction term**. The combined model in Equations (16.22) and (16.23) is called a **vector error correction model (VECM)**. In a VECM, past values of $Y_t - \theta X_t$ help to predict future values of ΔY_t and/or ΔX_t.

How Can You Tell Whether Two Variables Are Cointegrated?

There are three ways to determine whether two variables can plausibly be modeled as cointegrated: Use expert knowledge and economic theory, graph the series and see whether they appear to have a common stochastic trend, and perform statistical tests for cointegration. All three methods should be used in practice.

First, you must use your expert knowledge of these variables to decide whether cointegration is in fact plausible. For example, the two interest rates in Figure 16.2 are linked together by the so-called expectations theory of the term structure of interest rates. According to this theory, the interest rate on January 1 on the 10-year Treasury bond is the average of the interest rate on a 3-month Treasury bill for the first quarter of the year and the expected interest rates on future 3-month Treasury bills issued in the subsequent 39 quarters, for total of 40 quarters, or 10 years. If this was not the case, then investors could expect to make money by holding either the 10-year Treasury note or a sequence of forty 3-month Treasury bills, and they would bid up prices until the expected returns were equalized. If the 3-month interest rate has a random walk stochastic trend, this theory implies that this stochastic trend is inherited by the 10-year interest rate and that the difference between the two rates—that is, the term spread—is stationary. Thus the expectations theory of the term structure implies that if the interest rates are $I(1)$, then they will be cointegrated with a cointegrating coefficient of $\theta = 1$ (Exercise 16.2).

Second, visual inspection of the series helps to identify cases in which cointegration is plausible. For example, the graph of the two interest rates in Figure 16.2 shows that each of the series appears to be $I(1)$ but that the term spread appears to be $I(0)$, so the two series appear to be cointegrated.

Third, the unit root testing procedures introduced so far can be extended to tests for cointegration. The insight on which these tests are based is that if Y_t and X_t are cointegrated with cointegrating coefficient θ, then $Y_t - \theta X_t$ is stationary; otherwise, $Y_t - \theta X_t$ is nonstationary [is $I(1)$]. The hypothesis that Y_t and X_t are not cointegrated [that is, that $Y_t - \theta X_t$ is $I(1)$] therefore can be tested by testing the null hypothesis that $Y_t - \theta X_t$ has a unit root; if this hypothesis is rejected, then Y_t and X_t can be modeled as cointegrated. The details of this test depend on whether the cointegrating coefficient θ is known.

TABLE 16.2 Critical Values for the Engle–Granger ADF Statistic

Number of *X*'s in Equation (16.24)	10%	5%	1%
1	−3.12	−3.41	−3.96
2	−3.52	−3.80	−4.36
3	−3.84	−4.16	−4.73
4	−4.20	−4.49	−5.07

Testing for cointegration when θ is known. In many cases expert knowledge or economic theory suggests a value for θ. When θ is known, the Dickey–Fuller and DF-GLS unit root tests can be used to test for cointegration by first constructing the series $z_t = Y_t - \theta X_t$ and then testing the null hypothesis that z_t has a unit autoregressive root.

Testing for cointegration when θ is unknown. If the cointegrating coefficient θ is unknown, then it must be estimated prior to testing for a unit root in the error correction term. This preliminary step makes it necessary to use different critical values for the subsequent unit root test.

Specifically, in the first step the cointegrating coefficient θ is estimated by OLS estimation of the regression

$$Y_t = \alpha + \theta X_t + z_t. \tag{16.24}$$

In the second step, a Dickey–Fuller t-test (with an intercept but no time trend) is used to test for a unit root in the residual from this regression, $\hat{z}_t$. This two-step procedure is called the Engle–Granger Augmented Dickey–Fuller test for cointegration, or **EG-ADF test** (Engle and Granger, 1987).

Critical values of the EG-ADF statistic are given in Table 16.2.[1] The critical values in the first row apply when there is a single regressor in Equation (16.26), so there are two cointegrated variables (X_t and Y_t). The subsequent rows apply to the case of multiple cointegrated variables, which is discussed at the end of this section.

Estimation of Cointegrating Coefficients

If X_t and Y_t are cointegrated, then the OLS estimator of the coefficient in the cointegrating regression in Equation (16.24) is consistent. However, in general the OLS

[1] The critical values in Table 16.2 are taken from Fuller (1976) and Phillips and Ouliaris (1990). Following a suggestion by Hansen (1992), the critical values in Table 16.2 are chosen so that they apply whether or not X_t and Y_t have drift components.

estimator has a nonnormal distribution, and inferences based on its t-statistics can be misleading whether or not those t-statistics are computed using HAC standard errors. Because of these drawbacks of the OLS estimator of θ, econometricians have developed a number of other estimators of the cointegrating coefficient.

One such estimator of θ that is simple to use in practice is the **dynamic OLS (DOLS) estimator** (Stock and Watson, 1993). The DOLS estimator is based on a modified version of Equation (16.24) that includes past, present, and future values of the change in X_t:

$$Y_t = \beta_0 + \theta X_t + \sum_{j=-p}^{p} \delta_j \Delta X_{t-j} + u_t. \tag{16.25}$$

Thus, in Equation (16.25), the regressors are X_t, $\Delta X_{t+p}, \ldots, \Delta X_{t-p}$. The DOLS estimator of θ is the OLS estimator of θ in the regression of Equation (16.25).

If X_t and Y_t are cointegrated, then the DOLS estimator is efficient in large samples. Moreover, statistical inferences about θ and the δ's in Equation (16.25) based on HAC standard errors are valid. For example, the t-statistic constructed using the DOLS estimator with HAC standard errors has a standard normal distribution in large samples.

One way to interpret Equation (16.25) is to recall from Section 15.3 that cumulative dynamic multipliers can be computed by modifying the distributed lag regression of Y_t on X_t and its lags. Specifically, in Equation (15.7), the cumulative dynamic multipliers were computed by regressing Y_t on ΔX_t, lags of ΔX_t, and X_{t-r}; the coefficient on X_{t-r} in that specification is the long-run cumulative dynamic multiplier. Similarly, if X_t were strictly exogenous, then in Equation (16.25) the coefficient on X_t, θ would be the long-run cumulative multiplier—that is, the long-run effect on Y of a change in X. If X_t is not strictly exogenous, then the coefficients do not have this interpretation. Nevertheless, because X_t and Y_t have a common stochastic trend if they are cointegrated, the DOLS estimator is consistent even if X_t is endogenous.

The DOLS estimator is not the only efficient estimator of the cointegrating coefficient. The first such estimator was developed by Søren Johansen (Johansen, 1988). For a discussion of Johansen's method and of other ways to estimate the cointegrating coefficient, see Hamilton (1994, Chapter 20).

Even if economic theory does not suggest a specific value of the cointegrating coefficient, it is important to check whether the estimated cointegrating relationship makes sense in practice. Because cointegration tests can be misleading (they can improperly reject the null hypothesis of no cointegration more frequently than they should, and frequently they improperly fail to reject the null hypothesis), it is especially important to rely on economic theory, institutional knowledge, and common sense when estimating and using cointegrating relationships.

Extension to Multiple Cointegrated Variables

The concepts, tests, and estimators discussed here extend to more than two variables. For example, if there are three variables, Y_t, X_{1t}, and X_{2t}, each of which is $I(1)$, then they are cointegrated with cointegrating coefficients θ_1 and θ_2 if $Y_t - \theta_1 X_{1t} - \theta_2 X_{2t}$ is stationary. When there are three or more variables, there can be multiple cointegrating relationships. For example, consider modeling the relationship among three interest rates: the 3-month rate ($R3m$), the 1-year ($R1y$) rate, and the 10-year rate ($R10y$). If they are $I(1)$, then the expectations theory of the term structure of interest rates suggests that they will all be cointegrated. One cointegrating relationship suggested by the theory is $R10y_t - R3m_t$, and a second relationship is $R1y_t - R3m_t$. (The relationship $R10y_t - R1y_t$ is also a cointegrating relationship, but it contains no additional information beyond that in the other relationships because it is perfectly multicollinear with the other two cointegrating relationships.)

The EG-ADF procedure for testing for a single cointegrating relationship among multiple variables is the same as for the case of two variables, except that the regression in Equation (16.24) is modified so that both X_{1t} and X_{2t} are regressors; the critical values for the EG-ADF test are given in Table 16.2, where the appropriate row depends on the number of regressors in the first-stage OLS cointegrating regression. The DOLS estimator of a single cointegrating relationship among multiple X's involves including the level of each X along with leads and lags of the first difference of each X. Tests for multiple cointegrating relationships can be performed using system methods, such as Johansen's (1988) method, and the DOLS estimator can be extended to multiple cointegrating relationships by estimating multiple equations, one for each cointegrating relationship. For additional discussion of cointegration methods for multiple variables, see Hamilton (1994).

A cautionary note. If two or more variables are cointegrated, then the error correction term can help to forecast these variables and, possibly, other related variables. However, cointegration requires the variables to have the same stochastic trends. Trends in economic variables typically arise from complex interactions of disparate forces, and closely related series can have different trends for subtle reasons. If variables that are not cointegrated are incorrectly modeled using a VECM, then the error correction term will be $I(1)$; this introduces a trend into the forecast that can result in poor out-of-sample forecast performance. Thus forecasting using a VECM must be based on a combination of compelling theoretical arguments in favor of cointegration and careful empirical analysis.

Application to Interest Rates

As discussed earlier, the expectations theory of the term structure of interest rates implies that if two interest rates of different maturities are $I(1)$, then they will be cointegrated with a cointegrating coefficient of $\theta = 1$; that is, the spread between the two rates will be stationary. Inspection of Figure 16.2 provides qualitative support for the hypothesis that the 10-year and 3-month interest rates are cointegrated. We first use unit root and cointegration test statistics to provide more formal evidence on this hypothesis, then estimate a vector error correction model for these two interest rates.

Unit root and cointegration tests. Various unit root and cointegration test statistics for these two series are reported in Table 16.3. The unit root test statistics in the first two rows examine the hypothesis that the two interest rates, the 3-month rate ($R3m$) and the 10-year rate ($R10y$), individually have a unit root. The ADF and DF-GLS test statistics are larger than the 10% critical values, so the null hypothesis of a unit root is not rejected for either series at the 10% significance level. Thus, these results suggest that the interest rates are plausibly modeled as $I(1)$.

The unit root statistics for the term spread, $R10y_t - R3m_t$, test the further hypothesis that these variables are not cointegrated against the alternative hypothesis that they are. The null hypothesis that the term spread contains a unit root is rejected at the 1% level, using both unit root tests. Thus we reject the hypothesis that the series are not cointegrated against the alternative that they are, with a cointegrating coefficient $\theta = 1$. Taken together, the evidence in the first three rows of Table 16.3 suggests that these variables plausibly can be modeled as cointegrated with $\theta = 1$.

TABLE 16.3 Unit Root and Cointegration Test Statistics for Two Interest Rates

Series	ADF Statistic	DF-GLS Statistic
$R3m$	−2.17	−1.84
$R10y$	−1.03	−0.96
$R10y - R3m$	−3.97**	−3.92**
$R10y - 0.814 \times R3m$	−3.15	—

R3m is the interest rate on 3-month U.S. Treasury bills, and *R10y* is the interest rate on 10-year U.S. Treasury bonds. Regressions were estimated using quarterly data over the period 1962:Q1–2012:Q4. The number of lags in the unit root test statistic regressions were chosen by AIC (six lags maximum). Unit root test statistics are significant at the *5% or **1% significance level.

Because in this application economic theory suggests a value for θ (the expectations theory of the term structure suggests that $\theta = 1$) and because the error correction term is $I(0)$ when this value is imposed (the spread is stationary), in principle it is not necessary to use the EG-ADF test, in which θ is estimated. Nevertheless, we compute the test as an illustration. The first step in the EG-ADF test is to estimate θ by the OLS regression of one variable on the other; the result is

$$\widehat{R10y_t} = 2.46 + 0.81R3m_t, \overline{R}^2 = 0.83. \tag{16.26}$$

The second step is to compute the ADF statistic for the residual from this regression, $\hat{z}_t$. The result, given in the final row of Table 16.3, is –3.15. This value is smaller than the 10% critical value (which is -3.12) but not smaller than the 5% critical value (-3.41), so the null hypothesis of no cointegration is rejected at the 10% significance level but not the 5% significance level. An interpretation of this result is that the EG-ADF test, which uses an estimated value of θ, is less powerful than the test that uses what is arguably the correct value of $\theta = 1$.

A vector error correction model of the two interest rates. If Y_t and X_t are cointegrated, then forecasts of ΔY_t and ΔX_t can be improved by augmenting a VAR of ΔY_t and ΔX_t by the lagged value of the error correction term—that is, by computing forecasts using the VECM in Equations (16.22) and (16.23). If θ is known, then the unknown coefficients of the VECM can be estimated by OLS, including $z_{t-1} = Y_{t-1} - \theta X_{t-1}$ as an additional regressor. If θ is unknown, then the VECM can be estimated using $\hat{z}_{t-1}$ as a regressor, where $\hat{z}_t = Y_t - \hat{\theta}X_t$, and where $\hat{\theta}$ is an estimator of θ.

In the application to the two interest rates, theory suggests that $\theta = 1$, and the unit root tests support modeling the two interest rates as cointegrated with a cointegrating coefficient of 1. We therefore specify the VECM using the theoretically suggested value of $\theta = 1$—that is, by adding the lagged value of the term spread, $R10y$ –$R3m$, to a VAR in $\Delta R10y_t$ and $\Delta R3m_t$. Specified with two lags of first differences, the resulting VECM is

$$\begin{aligned}\widehat{\Delta R3m_t} = \underset{(0.12)}{-0.06} + \underset{(0.13)}{0.24}\Delta R3m_{t-1} - \underset{(0.18)}{0.16}\Delta R3m_{t-2} + \underset{(0.20)}{0.11}\Delta R10y_{t-1} \\ -\underset{(0.15)}{0.15}\Delta R10y_{t-2} + \underset{(0.05)}{0.03}(R10y_{t-1} - R3m_{t-1})\end{aligned} \tag{16.27}$$

$$\begin{aligned}\widehat{\Delta R10y_t} = \underset{(0.06)}{0.12} - \underset{(0.09)}{0.00}\Delta R3m_{t-1} - \underset{(0.07)}{0.07}\Delta R3m_{t-2} + \underset{(0.11)}{0.22}\Delta R10y_{t-1} \\ -\underset{(0.09)}{0.07}\Delta R10y_{t-2} - \underset{(0.03)}{0.09}(R10y_{t-1} - R3m_{t-1}). \qquad (16.28)\end{aligned}$$

In Equation (16.27), none of the coefficients is individually significant at the 5% level, and the coefficients on the lagged first differences of the interest rates are not jointly significant at the 5% level. In Equation (16.28), the coefficients on the lagged first differences are not jointly significant, but the coefficient on the lagged spread (the error correction term), which is estimated to be −0.09, has a *t*-statistic of −2.74, so it is statistically significant at the 1% level. Although lagged values of the first difference of the interest rates are not useful for predicting future interest rates, the lagged spread does help predict the change in the 10-year Treasury bond rate. When the 10-year rate exceeds the 3-month rate, the 10-year rate is forecasted to fall in the future.

16.5 Volatility Clustering and Autoregressive Conditional Heteroskedasticity

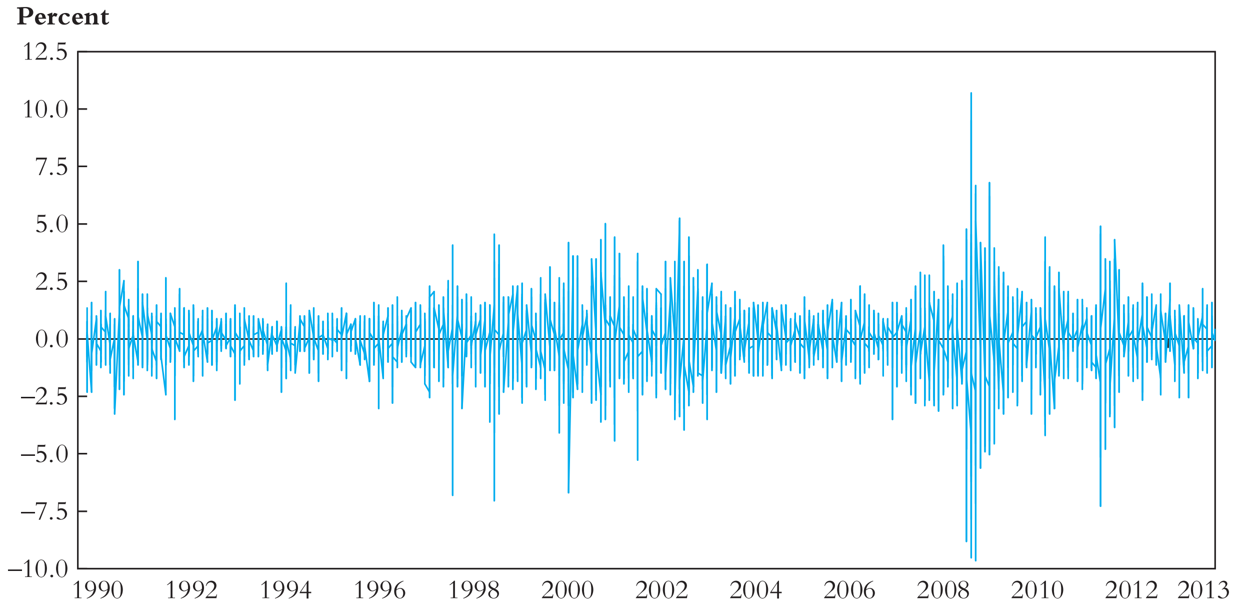
The phenomenon that some times are tranquil while others are not—that is, that volatility comes in clusters—shows up in many economic time series. This section presents a pair of models for quantifying volatility clustering or, as it is also known, conditional heteroskedasticity.

Volatility Clustering

The volatility of many financial and macroeconomic variables changes over time. For example, daily percentage changes in the Wilshire 5000 stock price index, shown in Figure 16.3, exhibit periods of high volatility, such as in 2001 and 2008, and other periods of low volatility, such as in 2004. A series with some periods of low volatility and some periods of high volatility is said to exhibit **volatility clustering**. Because the volatility appears in clusters, the variance of the daily percentage price change in the Wilshire 5000 index can be forecasted, even though the daily price change itself is very difficult to forecast.

Forecasting the variance of a series is of interest for several reasons. First, the variance of an asset price is a measure of the risk of owning that asset: The larger

FIGURE 16.3 Daily Percentage Changes in the Wilshire Index, 1990–2013



Daily percentage price changes in the Wilshire 5000 index exhibit volatility clustering, in which there are some periods of high volatility, such as in 2008, and other periods of relative tranquility, such as in 2004.

the variance of daily stock price changes, the more a stock market participant stands to gain—or lose—on a typical day. An investor who is worried about risk would be less tolerant of participating in the stock market during a period of high—rather than low—volatility.

Second, the value of some financial derivatives, such as options, depends on the variance of the underlying asset. An options trader wants the best available forecasts of future volatility to help him or her know the price at which to buy or sell options.

Third, forecasting variances makes it possible to have accurate forecast intervals. Suppose that you are forecasting the rate of inflation. If the variance of the forecast error is constant, then an approximate forecast confidence interval can be constructed along the lines discussed in Section 14.4—that is, as the forecast plus or minus a multiple of the *SER*. If, however, the variance of the forecast error changes over time, then the width of the forecast interval should change over time: At periods when inflation is subject to particularly large disturbances or shocks, the interval should be wide; during periods of relative tranquility, the interval should be tighter.

Volatility clustering can be thought of as clustering of the variance of the error term over time: If the regression error has a small variance in one period, its variance tends to be small in the next period, too. In other words, volatility clustering implies that the error exhibits time-varying heteroskedasticity.

Autoregressive Conditional Heteroskedasticity

Two models of volatility clustering are the **autoregressive conditional heteroskedasticity (ARCH)** model and its extension, the **generalized ARCH (GARCH)** model.

ARCH. Consider the ADL(1,1) regression

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \gamma_1 X_{t-1} + u_t. \tag{16.29}$$

In the ARCH model, which was developed by the econometrician Robert Engle (Engle, 1982; see the box "Nobel Laureates in Time Series Econometrics"), the error u_t is modeled as being normally distributed with mean zero and variance σ_t^2, where σ_t^2 depends on past squared values u_t. Specifically, the ARCH model of order p, denoted ARCH(p), is

$$\sigma_t^2 = \alpha_0 + \alpha_1 u_{t-1}^2 + \alpha_2 u_{t-2}^2 + \cdots + \alpha_p u_{t-p}^2, \tag{16.30}$$

where $\alpha_0, \alpha_1, \ldots, \alpha_p$ are unknown coefficients. If these coefficients are positive, then if recent squared errors are large, the ARCH model predicts that the current squared error will be large in magnitude, in the sense that its variance, σ_t^2, is large.

Although it is described here for the ADL(1,1) model in Equation (16.29), the ARCH model can be applied to the error variance of any time series regression model with an error that has a conditional mean of zero, including higher-order ADL models, autoregressions, and time series regressions with multiple predictors.

GARCH. The generalized ARCH (GARCH) model, developed by the econometrician Tim Bollerslev (Bollerslev, 1986), extends the ARCH model to let σ_t^2 depend on its own lags as well as lags of the squared error. The GARCH(p,q) model is

$$\sigma_t^2 = \alpha_0 + \alpha_1 u_{t-1}^2 + \cdots + \alpha_p u_{t-p}^2 + \phi_1 \sigma_{t-1}^2 + \cdots + \phi_q \sigma_{t-q}^2, \tag{16.31}$$

where $\alpha_0, \alpha_1, \ldots, \alpha_p, \phi_1, \ldots, \phi_q$ are unknown coefficients.

The ARCH model is analogous to a distributed lag model, and the GARCH model is analogous to an ADL model. As discussed in Appendix 15.2, the ADL model (when appropriate) can provide a more parsimonious model of dynamic multipliers than can the distributed lag model. Similarly, by incorporating lags of σ_t^2, the GARCH model can capture slowly changing variances with fewer parameters than the ARCH model.

An important application of ARCH and GARCH models is to measuring and forecasting the time-varying volatility of returns on financial assets, particularly assets observed at high sampling frequencies such as the daily stock returns in Figure 16.3. In such applications, the return itself is often modeled as unpredictable, so the regression in Equation (16.29) only includes the intercept.

Estimation and inference. ARCH and GARCH models are estimated by the method of maximum likelihood (Appendix 11.2). The estimators of the ARCH and GARCH coefficients are normally distributed in large samples, so in large samples, t-statistics have standard normal distributions, and confidence intervals can be constructed as the maximum likelihood estimate ±1.96 standard errors.

Application to Stock Price Volatility

A GARCH(1,1) model of the Wilshire daily percentage stock price changes, R_t, estimated using data on all trading days from January 2, 1990, through December 31, 2013, is

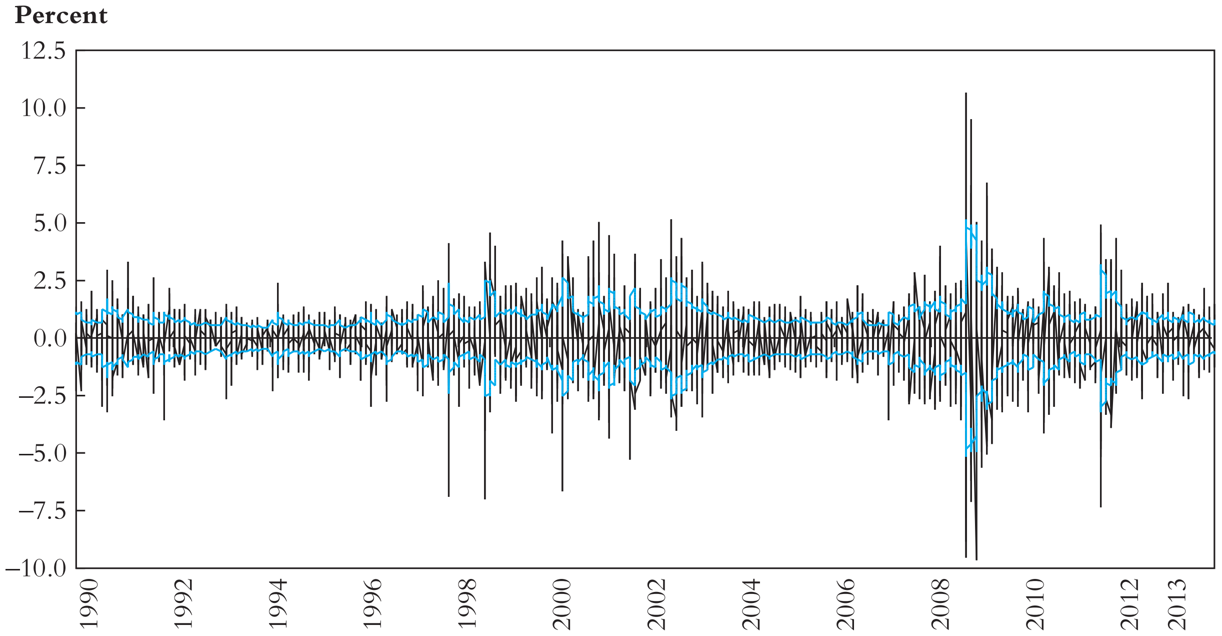
$$\underset{(0.010)}{\hat{R}_t = 0.057} \tag{16.32}$$

$$\hat{\sigma}_t^2 = \underset{(0.002)}{0.011} + \underset{(0.007)}{0.082}\, u_{t-1}^2 + \underset{(0.008)}{0.908}\sigma_{t-1}^2. \tag{16.33}$$

No lagged predictors appear in Equation (16.32) because daily Wilshire 5000 percentage price changes are essentially unpredictable.

The two coefficients in the GARCH model (the coefficients on u_{t-1}^2 and σ_{t-1}^2) are both individually statistically significant at the 5% significance level. One measure of the persistence of movements in the variance is the sum of the coefficients on u_{t-1}^2 and σ_{t-1}^2 in the GARCH model (Exercise 16.9). This sum (0.99) is large, indicating that changes in the conditional variance are persistent. Said

FIGURE 16.4 Daily Percentage Changes in the Wilshire 5000 Index and GARCH(1,1) Bands



The GARCH(1,1) bands, which are $\pm\hat{\sigma}_t$, where $\hat{\sigma}_t$ is computed using Equation (16.33), are narrow when the conditional variance is small and wide when it is large. The conditional volatility of stock price changes varies considerably over the 1990–2013 period.

differently, the estimated GARCH model implies that periods of high volatility in stock prices will be long-lasting. This implication is consistent with the long periods of volatility clustering seen in Figure 16.3.

The estimated conditional variance at date t, $\hat{\sigma}_t^2$, can be computed using the residuals from Equation (16.32) and the coefficients in Equation (16.33). Figure 16.4 plots bands of plus or minus one conditional standard deviation (that is, $\pm\hat{\sigma}_t$), based on the GARCH(1,1) model, along with deviations of the percentage price change series from its mean. The conditional standard deviation bands quantify the time-varying volatility of the daily price changes. During the mid-1990s, the conditional standard deviation bands are tight, indicating lower levels of risk for investors holding a portfolio of stocks making up the Wilshire index. In contrast, during 2008, these conditional standard deviation bands are wide, indicating a period of greater daily stock price volatility.

Nobel Laureates in Time Series Econometrics

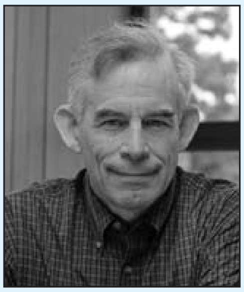
In 2003 Robert Engle and Clive Granger won the Nobel Prize in economics for fundamental theoretical research in time series econometrics. Engle's work was motivated by the volatility clustering evident in plots like Figure 16.3. Engle wondered whether series like these could be stationary and whether econometric models could be developed to explain and predict their time-varying volatility. Engle's answer was to develop the autoregressive conditional heteroskedasticity (ARCH) model, described in Section 16.5. The ARCH model and its extensions proved especially useful for modeling the volatility of asset returns, and the resulting volatility forecasts are used to price financial derivatives and to assess changes over time in the risk of holding financial assets. Today, measures and forecasts of volatility are a core component of financial econometrics, and the ARCH model and its descendants are the workhorse tools for modeling volatility.

Clive W. J. Granger

Robert F. Engle

Granger's work focused on how to handle stochastic trends in economic time series data. From his earlier work, he knew that two unrelated series with stochastic trends could, by the usual statistical measures of t-statistics and regression R^2's, falsely appear to be meaningfully related; this is the "spurious regression" problem exemplified by the regressions in Equations (14.28) and (14.29). But are all regressions involving stochastic trending variables spurious? Granger discovered that when variables shared common trends—in his terminology, were "co-integrated"—meaningful relationships could be uncovered by regression analysis using a vector error correction model. The methods of cointegration analysis are now a staple in modern macroeconometrics.

In 2011, Thomas Sargent and Christopher Sims won the Nobel Prize for their empirical research on cause and effect in the macroeconomy. Sargent was recognized for developing models that featured the important role that expectations about the future play in disentangling cause and effect. Sims was recognized for developing structural VAR (SVAR) models. Sims's key insight concerned the forecast errors in a VAR model—the u_t errors in Equations (16.1) and (16.2). These errors, he realized, arose because of unforeseen "shocks" that buffeted the macroeconomy, and in many cases, these shocks had well defined sources like OPEC (oil price shocks), the Fed (interest rate shocks), or Congress (tax shocks). By disentangling the various sources of shocks that comprise the VAR errors, Sims was able to estimate the dynamic causal effect of these shocks on the variables appearing in the VAR. This disentangling of shocks is never without controversy, but SVARs are now a standard tool for estimating dynamic causal effects in macroeconomics.


Christopher A. Sims

Lars Peter Hansen

continued on next page

In 2013, Eugene Fama, Lars Peter Hansen, and Robert Shiller won the Nobel Prize for their empirical analysis of asset prices. The work in the two "Can You Beat the Market" boxes in Chapter 14 and the box "Commodity Traders Send Shivers Through Disney World" in Chapter 15 was motivated in part by the "efficient markets" (unpredictability) work of Fama and the "irrational exuberance" (unexplained volatility) work of Shiller. Hansen was honored for developing "Generalized Method of Moments" (GMM) methods to investigate whether asset returns are consistent with expected utility theory. Microeconomics says that investors should equate the marginal cost of an investment (today's foregone utility from investing rather than consuming) with its marginal benefit (tomorrow's boost in utility from consumption financed by the investment's return). But a simple test of this proposition is complicated because marginal utility is difficult to measure, asset returns are uncertain, and the argument should hold across all asset returns. Hansen developed GMM methods to test asset-pricing models. As it turned out, Hansen's GMM methods had applications well beyond finance and are now widely used in econometrics. Section 18.7 introduces GMM.

For more information on these and other Nobel laureates in economics, visit the Nobel Foundation website, **http://www.nobel.se/economics**.

16.6 Conclusion

This part of the book has covered some of the most frequently used tools and concepts of time series regression. Many other tools for analyzing economic time series have been developed for specific applications. If you are interested in learning more about economic forecasting, see the introductory textbooks by Diebold (2007) and Enders (2009). For an advanced treatment of econometrics with time series data, see Hamilton (1994) and Hayashi (2000).

Summary

1. Vector autoregressions model a "vector" of k time series variables as each depends on its own lags and the lags of the $k - 1$ other series. The forecasts of each of the time series produced by a VAR are mutually consistent, in the sense that they are based on the same information.
2. Forecasts two or more periods ahead can be computed either by iterating forward a one-step-ahead model (an AR or a VAR) or by estimating a multiperiod-ahead regression.
3. Two series that share a common stochastic trend are cointegrated; that is, Y_t and X_t are cointegrated if Y_t and X_t are $I(1)$ but $Y_t - \theta X_t$ is $I(0)$. If Y_t and

X_t are cointegrated, the error correction term $Y_t - \theta X_t$ can help predict ΔY_t and/or ΔX_t. A vector error correction model is a VAR model of ΔY_t and ΔX_t, augmented to include the lagged error correction term.

4. Volatility clustering—in which the variance of a series is high in some periods and low in others—is common in economic time series, especially financial time series.
5. The ARCH model of volatility clustering expresses the conditional variance of the regression error as a function of recent squared regression errors. The GARCH model augments the ARCH model to include lagged conditional variances as well. Estimated ARCH and GARCH models produce forecast intervals with widths that depend on the volatility of the most recent regression residuals.

Key Terms

vector autoregression (VAR) (639)
iterated multiperiod AR forecast (646)
iterated multiperiod VAR forecast (646)
direct multiperiod forecast (648)
integrated of order $d[I(d)]$ (650)
second difference (649)
integrated of order zero $[I(0)]$, one $[I(1)]$, or two $[I(2)]$ (650)
order of integration (650)
DF-GLS test (651)
common trend (656)
cointegration (657)
cointegrating coefficient (657)
error correction term (658)
vector error correction model (VECM) (658)
EG-ADF test (659)
dynamic OLS (DOLS) estimator (660)
volatility clustering (664)
autoregressive conditional heteroskedasticity (ARCH) (666)
generalized ARCH (GARCH) (666)

Review the Concepts

16.1 A macroeconomist wants to construct forecasts for the following macroeconomic variables: GDP, consumption, investment, government purchases, exports, imports, short-term interest rates, long-term interest rates, and the rate of price inflation. He has quarterly time series for each of these variables from 1970 to 2014. Should he estimate a VAR for these variables and use this for forecasting? Why or why not? Can you suggest an alternative approach?

16.2 Suppose that Y_t follows a stationary AR(1) model with $\beta_0 = 0$ and $\beta_1 = 0.7$. If $Y_t = 5$, what is your forecast of Y_{t+2} (that is, what is $Y_{t+2|t}$)? What is $Y_{t+h|t}$ for $h = 30$? Does this forecast for $h = 30$ seem reasonable to you?

16.3 A version of the permanent income theory of consumption implies that the logarithm of real GDP (Y) and the logarithm of real consumption (C) are cointegrated with a cointegrating coefficient equal to 1. Explain how you would investigate this implication by (a) plotting the data and (b) using a statistical test.

16.4 Consider the ARCH model, $\sigma_t^2 = 1.0 + 0.8\,u_{t-1}^2$. Explain why this will lead to volatility clustering. (*Hint:* What happens when u_{t-1}^2 is unusually large?)

16.5 The DF-GLS test for a unit root has higher power than the Dickey–Fuller test. Why should you use a more powerful test?

Exercises

16.1 Suppose that Y_t follows a stationary AR(1) model, $Y_t = \beta_0 + \beta_1 Y_{t-1} + u_t$.

a. Show that the h-period-ahead forecast of Y_t is given by $Y_{t+h|t} = \mu_Y + \beta_1^h(Y_t - \mu_Y)$, where $\mu_Y = \beta_0/(1 - \beta_1)$.

b. Suppose that X_t is related to Y_t by $X_t = \sum_{i=0}^{\infty} \delta^i Y_{t+i|t}$, where $|\delta| < 1$. Show that $X_t = [\mu_Y/(1 - \delta)] + [(Y_t - \mu_Y)/(1 - \beta_1\delta)]$.

16.2 One version of the expectations theory of the term structure of interest rates holds that a long-term rate equals the average of the expected values of short-term interest rates into the future, plus a term premium that is $I(0)$. Specifically, let Rk_t denote a k-period interest rate, let $R1_t$ denote a one-period interest rate, and let e_t denote an $I(0)$ term premium. Then $Rk_t = \frac{1}{k}\sum_{i=0}^{k-1} R1_{t+i|t} + e_t$, where $R1_{t+i|t}$ is the forecast made at date t of the value of $R1$ at date $t + i$. Suppose that $R1_t$ follows a random walk so that $R1_t = R1_{t-1} + u_t$.

a. Show that $Rk_t = R1_t + e_t$.

b. Show that Rk_t and $R1_t$ are cointegrated. What is the cointegrating coefficient?

c. Now suppose that $\Delta R1_t = 0.5\Delta R1_{t-1} + u_t$. How does your answer to (b) change?

d. Now suppose that $R1_t = 0.5R1_{t-1} + u_t$. How does your answer to (b) change?

16.3 Suppose that u_t follows the ARCH process, $\sigma_t^2 = 1.0 + 0.5u_{t-1}^2$.

a. Let $E(u_t^2) = \text{var}(u_t)$ be the unconditional variance of u_t. Show that $\text{var}(u_t) = 2$. (*Hint:* Use the law of iterated expectations, $E(u_t^2) = E[E(u_t^2|u_{t-1})]$.)

b. Suppose that the distribution of u_t conditional on lagged values of u_t is $N(0, \sigma_t^2)$. If $u_{t-1} = 0.2$, what is $\Pr(-3 \leq u_t \leq 3)$? If $u_{t-1} = 2.0$, what is $\Pr(-3 \leq u_t \leq 3)$?

16.4 Suppose that Y_t follows the AR(p) model $Y_t = \beta_0 + \beta_1 Y_{t-1} + \cdots + \beta_p Y_{t-p} + u_t$, where $E(u_t|Y_{t-1}, Y_{t-2}, \ldots) = 0$ Let $Y_{t+h|t} = E(Y_{t+h}|Y_t, Y_{t-1}, \ldots)$. Show that $Y_{t+h|t} = \beta_0 + \beta_1 Y_{t-1+h|t} + \cdots + \beta_p Y_{t-p+h|t}$ for $h > p$.

16.5 Verify Equation (16.20). [*Hint:* Use $\Sigma_{t=1}^T Y_t^2 = \Sigma_{t=1}^T (Y_{t-1} + \Delta Y_t)^2$ to show that $\Sigma_{t=1}^T Y_t^2 = \Sigma_{t=1}^T Y_{t-1}^2 + 2\Sigma_{t=1}^T Y_{t-1}\Delta Y_t + \Sigma_{t=1}^T \Delta Y_t^2$ and solve for $\Sigma_{t=1}^T Y_{t-1}\Delta Y_t$.]

16.6 A regression of Y_t onto current, past, and future values of X_t yields

$$Y_t = 3.0 + 1.7X_{t+1} + 0.8X_t - 0.2X_{t-1} + u_t.$$

a. Rearrange the regression so that it has the form shown in Equation (16.25). What are the values of θ, δ_{-1}, δ_0, and δ_1?

b. i. Suppose that X_t is $I(1)$ and u_t is $I(1)$. Are Y and X cointegrated?

ii. Suppose that X_t is $I(0)$ and u_t is $I(1)$. Are Y and X cointegrated?

iii. Suppose that X_t is $I(1)$ and u_t is $I(0)$. Are Y and X cointegrated?

16.7 Suppose that $\Delta Y_t = u_t$, where u_t is i.i.d. $N(0, 1)$, and consider the regression $Y_t = \beta X_t + error$, where $X_t = \Delta Y_{t+1}$ and *error* is the regression error.

Show that $\hat{\beta} \xrightarrow{d} \frac{1}{2}(\chi_1^2 - 1)$. [*Hint:* Analyze the numerator of $\hat{\beta}$ using analysis like that in Equation (16.21). Analyze the denominator using the law of large numbers.]

16.8 Consider the following two-variable VAR model with one lag and no intercept:

$$Y_t = \beta_{11}Y_{t-1} + \gamma_{11}X_{t-1} + u_{1t}$$
$$X_t = \beta_{21}Y_{t-1} + \gamma_{21}X_{t-1} + u_{2t}.$$

a. Show that the iterated two-period-ahead forecast for Y can be written as $Y_{t|t-2} = \delta_1 Y_{t-2} + \delta_2 X_{t-2}$ and derive values for δ_1 and δ_2 in terms of the coefficients in the VAR.

b. In light of your answer to (a), do iterated multiperiod forecasts differ from direct multiperiod forecasts? Explain.

16.9 **a.** Suppose that $E(u_t|u_{t-1}, u_{t-2}, \ldots) = 0$, that $\text{var}(u_t|u_{t-1}, u_{t-2}, \ldots)$ follows the ARCH(1) model $\sigma_t^2 = \alpha_0 + \alpha_1 u_{t-1}^2$, and that the process for u_t is stationary. Show that $\text{var}(u_t) = \alpha_0/(1 - \alpha_1)$. (*Hint:* Use the law of iterated expectations $E(u_t^2) = E[E(u_t^2|u_{t-1})]$.)

b. Extend the result in (a) to the ARCH(p) model.

c. Show that $\sum_{i=1}^{p}\alpha_i < 1$ for a stationary ARCH(p) model.

d. Extend the result in (a) to the GARCH(1,1) model.

e. Show that $\alpha_1 + \phi_1 < 1$ for a stationary GARCH(1,1) model.

16.10 Consider the cointegrated model $Y_t = \theta X_t + v_{1t}$ and $X_t = X_{t-1} + v_{2t}$, where v_{1t} and v_{2t} are mean zero serially uncorrelated random variables with $E(v_{1t}v_{2j}) = 0$ for all t and j. Derive the vector error correction model [Equations (16.22) and (16.23)] for X and Y.

Empirical Exercises

(Only two empirical exercises for this chapter are given in the text, but you can find more on the text website, **http://www.pearsonhighered.com/stock_watson/**.)

E16.1 This exercise is an extension of Empirical Exercise 14.1. On the text website, **http://www.pearsonhighered.com/stock_watson**, you will find the data file **USMacro_Quarterly**, which contains quarterly data on several macroeconomic series for the United States; the data are described in the file **USMacro_Description**. Compute inflation, *Infl*, using the price index for personal consumption expenditures. For all regressions use the sample period 1963:Q1–2012:Q4 (where data before 1963 may be used as initial values for lags in regressions).

a. Using the data on inflation through 2012:Q4 and an estimated AR(2) model:

i. Forecast $\Delta Infl_{2013:Q1}$, the change in inflation from 2012:Q4 to 2013:Q1.

ii. Forecast $\Delta Infl_{2013:Q2}$, the change in inflation from 2013:Q1 to 2013:Q2. (Use an iterated forecast.)

iii. Forecast $Infl_{2013:Q2} - Infl_{2012:Q4}$, the change in inflation from 2012:Q4 to 2013:Q2.

iv. Forecast $Infl_{2013:Q2}$, the level of inflation in 2013:Q2.

b. Repeat (a) using the direct forecasting method.

c. In Exercise 14.1 you carried out an ADF test for a unit root in the autoregression for *Infl*. Now carry out the unit root test using the DF-GLS test. Are the conclusions based on the DF-GLS test the same as you reached using the ADF test? Explain.

E16.2 On the text website, **http://www.pearsonhighered.com/stock_watson**, you will find the data file **USMacro_Quarterly**, which contains quarterly data on real GDP, measured in \$1996. Compute $GDPGR_t = 400 \times [\ln(GDP_t) - \ln(GDP_{t-1})]$, the growth rate of GDP.

a. Using data on $GDPGR_t$ from 1960:1 to 2012:4, estimate an AR(2) model with GARCH(1,1) errors.

b. Plot the residuals from the AR(2) model along with $\pm\hat{\sigma}_t$ bands as in Figure 16.4.

c. Some macroeconomists have claimed that there was a sharp drop in the variability of the growth rate of GDP around 1983, which they call the "Great Moderation." Is this Great Moderation evident in the plot that you formed in (b)?

17 The Theory of Linear Regression with One Regressor

Why should an applied econometrician bother learning any econometric theory? There are several reasons. Learning econometric theory turns your statistical software from a "black box" into a flexible tool kit from which you are able to select the right tool for the job at hand. Understanding econometric theory helps you appreciate why these tools work and what assumptions are required for each tool to work properly. Perhaps most importantly, knowing econometric theory helps you recognize when a tool will *not* work well in an application and when you should look for a different econometric approach.

This chapter provides an introduction to the econometric theory of linear regression with a single regressor. This introduction is intended to supplement—not replace—the material in Chapters 4 and 5, which should be read first.

This chapter extends Chapters 4 and 5 in two ways.

First, it provides a mathematical treatment of the sampling distribution of the OLS estimator and *t*-statistic, both in large samples under the three least squares assumptions of Key Concept 4.3 and in finite samples under the two additional assumptions of homoskedasticity and normal errors. These five extended least squares assumptions are laid out in Section 17.1. Sections 17.2 and 17.3, augmented by Appendix 17.2, mathematically develop the large-sample normal distributions of the OLS estimator and *t*-statistic under the first three assumptions (the least squares assumptions of Key Concept 4.3). Section 17.4 derives the exact distributions of the OLS estimator and *t*-statistic under the two additional assumptions of homoskedasticity and normally distributed errors.

Second, this chapter extends Chapters 4 and 5 by providing an alternative method for handling heteroskedasticity. The approach of Chapters 4 and 5 is to use heteroskedasticity-robust standard errors to ensure that statistical inference is valid even if the errors are heteroskedastic. This method comes with a cost, however: If the errors are heteroskedastic, then in theory a more efficient estimator than OLS is available. This estimator, called weighted least squares, is presented in Section 17.5. Weighted least squares requires a great deal of prior knowledge about the precise nature of the heteroskedasticity—that is, about the conditional variance of *u* given *X*. When such knowledge is available, weighted least squares improves upon OLS. In most applications, however, such knowledge is unavailable; in those cases, using OLS with heteroskedasticity-robust standard errors is the preferred method.

17.1 The Extended Least Squares Assumptions and the OLS Estimator

This section introduces a set of assumptions that extend and strengthen the three least squares assumptions of Chapter 4. These stronger assumptions are used in subsequent sections to derive stronger theoretical results about the OLS estimator than are possible under the weaker (but more realistic) assumptions of Chapter 4.

The Extended Least Squares Assumptions

Extended least squares Assumptions #1, #2, and #3. The first three extended least squares assumptions are the three assumptions given in Key Concept 4.3: that the conditional mean of u_i, given X_i, is zero; that $(X_{i,}Y_i)$, $i = 1, \dots, n$, are i.i.d. draws from their joint distribution; and that X_i and u_i have four moments.

Under these three assumptions, the OLS estimator is unbiased, is consistent, and has an asymptotically normal sampling distribution. If these three assumptions hold, then the methods for inference introduced in Chapter 4—hypothesis testing using the t-statistic and construction of 95% confidence intervals as ± 1.96 standard errors—are justified when the sample size is large. To develop a theory of efficient estimation using OLS or to characterize the exact sampling distribution of the OLS estimator, however, requires stronger assumptions.

Extended least squares Assumption #4. The fourth extended least squares assumption is that u_i is homoskedastic; that is, $\text{var}(u_i|X_i) = \sigma_u^2$, where σ_u^2 is a constant. As seen in Section 5.5, if this additional assumption holds, then the OLS estimator is efficient among all linear estimators that are unbiased, conditional on $X_1, \dots, X_n$.

Extended least squares Assumption #5. The fifth extended least squares assumption is that the conditional distribution of u_i, given X_i, is normal.

Under least squares Assumptions #1 and #2 and the extended least squares Assumptions #4 and #5, u_i is i.i.d. $N(0, \sigma_u^2)$, and u_i and X_i are independently distributed. To see this, note that the fifth extended least squares assumption states that the conditional distribution of $u_i|X_i$ is $N(0, \text{var}(u_i|X_i))$, where the distribution has mean zero by the first extended least squares assumption. By the fourth least

KEY CONCEPT 17.1

The Extended Least Squares Assumptions for Regression with a Single Regressor

The linear regression model with a single regressor is

$$Y_i = \beta_0 + \beta_1 X_i + u_i, i = 1, \dots, n. \tag{17.1}$$

The extended least squares assumptions are

1. $E(u_i|X_i) = 0$ (conditional mean zero);
2. $(X_i, Y_i), i = 1, \dots, n,$ are independent and identically distributed (i.i.d.) draws from their joint distribution;
3. (X_i, u_i) have nonzero finite fourth moments;
4. $\text{var}(u_i|X_i) = \sigma_u^2$ (homoskedasticity); and
5. The conditional distribution of u_i given X_i is normal (normal errors).

squares assumption, however, $\text{var}(u_i|X_i) = \sigma_u^2$, so the conditional distribution of $u_i|X_i$ is $N(0, \sigma_u^2)$. Because this conditional distribution does not depend on X_i, u_i and X_i are independently distributed. By the second least squares assumption, u_i is distributed independently of u_j for all $j \neq i$. It follows that, under the extended least squares Assumptions #1, #2, #4, and #5, u_i and X_i are independently distributed and u_i is i.i.d. $N(0, \sigma_u^2)$.

It is shown in Section 17.4 that, if all five extended least squares assumptions hold, the OLS estimator has an exact normal sampling distribution and the homoskedasticity-only t-statistic has an exact Student t distribution.

The fourth and fifth extended least squares assumptions are much more restrictive than the first three. Although it might be reasonable to assume that the first three assumptions hold in an application, the final two assumptions are less realistic. Even though these final two assumptions might not hold in practice, they are of theoretical interest because if one or both of them hold, then the OLS estimator has additional properties beyond those discussed in Chapters 4 and 5. Thus we can enhance our understanding of the OLS estimator and the theory of estimation in the linear regression model by exploring estimation under these stronger assumptions.

The five extended least squares assumptions for the single-regressor model are summarized in Key Concept 17.1.

The OLS Estimator

For easy reference, we restate the OLS estimators of β_0 and β_1 here:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \tag{17.2}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{X}. \tag{17.3}$$

Equations (17.2) and (17.3) are derived in Appendix 4.2.

17.2 Fundamentals of Asymptotic Distribution Theory

Asymptotic distribution theory is the theory of the distribution of statistics—estimators, test statistics, and confidence intervals—when the sample size is large. Formally, this theory involves characterizing the behavior of the sampling distribution of a statistic along a sequence of ever-larger samples. The theory is asymptotic in the sense that it characterizes the behavior of the statistic in the limit as $n \rightarrow \infty$.

Even though sample sizes are, of course, never infinite, asymptotic distribution theory plays a central role in econometrics and statistics for two reasons. First, if the number of observations used in an empirical application is large, then the asymptotic limit can provide a high-quality approximation to the finite sample distribution. Second, asymptotic sampling distributions typically are much simpler, and thus easier to use in practice, than exact finite-sample distributions. Taken together, these two reasons mean that reliable and straightforward methods for statistical inference—tests using t-statistics and 95% confidence intervals calculated as ±1.96 standard errors—can be based on approximate sampling distributions derived from asymptotic theory.

The two cornerstones of asymptotic distribution theory are the law of large numbers and the central limit theorem, both introduced in Section 2.6. We begin this section by continuing the discussion of the law of large numbers and the central limit theorem, including a proof of the law of large numbers. We then introduce two more tools, Slutsky's theorem and the continuous mapping theorem, that extend the usefulness of the law of large numbers and the central limit theorem. As an illustration, these tools are then used to prove that the distribution of the t-statistic based on $\overline{Y}$ testing the hypothesis $E(Y) = \mu_0$ has a standard normal distribution under the null hypothesis.

Convergence in Probability and the Law of Large Numbers

The concepts of convergence in probability and the law of large numbers were introduced in Section 2.6. Here we provide a precise mathematical definition of convergence in probability, followed by a statement and proof of the law of large numbers.

Consistency and convergence in probability. Let $S_1, S_2, \ldots, S_n, \ldots$ be a sequence of random variables. For example, S_n could be the sample average $\overline{Y}$ of a sample of n observations of the random variable Y. The sequence of random variables $\{S_n\}$ is said to **converge in probability** to a limit, μ (that is, $S_n \xrightarrow{p} \mu$), if the probability that S_n is within $\pm\delta$ of μ tends to 1 as $n \rightarrow \infty$, as long as the constant δ is positive. That is,

$$S_n \xrightarrow{p} \mu \text{ if and only if } \Pr(|S_n - \mu| \geq \delta) \longrightarrow 0 \qquad (17.4)$$

as $n \rightarrow \infty$ for every $\delta > 0$. If $S_n \xrightarrow{p} \mu$ then S_n is said to be a **consistent estimator** of μ.

The law of large numbers. The law of large numbers says that, under certain conditions on $Y_1, \ldots, Y_n$, the sample average $\overline{Y}$ converges in probability to the population mean. Probability theorists have developed many versions of the law of large numbers, corresponding to various conditions on $Y_1, \ldots, Y_n$. The version of the law of large numbers used in this book is that $Y_1, \ldots, Y_n$ are i.i.d. draws from a distribution with finite variance. This law of large numbers (also stated in Key Concept 2.6) is

$$\text{if } Y_1, \ldots, Y_n \text{ are i.i.d., } E(Y_i) = \mu_Y, \text{ and } \operatorname{var}(Y_i) < \infty, \text{ then } \overline{Y} \xrightarrow{p} \mu_Y. \qquad (17.5)$$

The idea of the law of large numbers can be seen in Figure 2.8: As the sample size increases, the sampling distribution of $\overline{Y}$ concentrates around the population mean, μ_Y. One feature of the sampling distribution is that the variance of $\overline{Y}$ decreases as the sample size increases; another feature is that the probability that $\overline{Y}$ falls outside $\pm\delta$ of μ_Y vanishes as n increases. These two features of the sampling distribution are in fact linked, and the proof of the law of large numbers exploits this link.

Proof of the law of large numbers. The link between the variance of $\overline{Y}$ and the probability that $\overline{Y}$ is within $\pm\delta$ of μ_Y is provided by Chebychev's inequality, which is stated and proven in Appendix 17.2 [see Equation (17.42)]. Written in terms of $\overline{Y}$, Chebychev's inequality is

$$\Pr(|\overline{Y} - \mu_Y| \geq \delta) \leq \frac{\text{var}(\overline{Y})}{\delta^2}, \tag{17.6}$$

for any positive constant δ. Because $Y_1, \ldots, Y_n$ are i.i.d. with variance σ_Y^2, $\text{var}(\overline{Y}) = \sigma_Y^2/n$; thus, for any $\delta > 0$, $\text{var}(\overline{Y})/\delta^2 = \sigma_Y^2/(\delta^2 n) \longrightarrow 0$. It follows from Equation (17.6) that $\Pr(|\overline{Y} - \mu_Y| \geq \delta) \longrightarrow 0$ for every $\delta > 0$, proving the law of large numbers.

Some examples. Consistency is a fundamental concept in asymptotic distribution theory, so we present some examples of consistent and inconsistent estimators of the population mean, μ_Y. Suppose that $Y_i, i = 1, \ldots, n$ are i.i.d. with variance σ_Y^2 that is positive and finite. Consider the following three estimators of μ_Y: (1) $m_a = Y_1$; (2) $m_b = \left(\frac{1-a^n}{1-a}\right)^{-1}\sum_{i=1}^{n} a^{i-1}Y_i$, where $0 < a < 1$; and (3) $m_c = \overline{Y} + 1/n$. Are these estimators consistent?

The first estimator, m_a, is just the first observation, so $E(m_a) = E(Y_1) = \mu_Y$ and m_a is unbiased. However, m_a is not consistent: $\Pr(|m_a - \mu_Y| \geq \delta) = \Pr(|Y_1 - \mu_Y| \geq \delta)$, which must be positive for sufficiently small δ (because $\sigma_Y^2 > 0$), so $\Pr(|m_a - \mu_Y| \geq \delta)$ does not tend to zero as $n \rightarrow \infty$, so m_a is not consistent. This inconsistency should not be surprising: Because m_a uses the information in only one observation, its distribution cannot concentrate around μ_Y as the sample size increases.

The second estimator, m_b, is unbiased but is not consistent. It is unbiased because

$$E(m_b) = E\left[\left(\frac{1-a^n}{1-a}\right)^{-1}\sum_{i=1}^{n} a^{i-1}Y_i\right] = \left(\frac{1-a^n}{1-a}\right)^{-1}\sum_{i=1}^{n} a^{i-1}\mu_Y = \mu_Y,$$

$$\text{since } \sum_{i=1}^{n} a^{i-1} = \left(1 - a^n\right)\sum_{i=0}^{\infty} a^i = \frac{1-a^n}{1-a}.$$

The variance of m_b is

$$\text{var}(m_b) = \left(\frac{1-a^n}{1-a}\right)^{-2}\sum_{i-1}^{n} a^{2(i-1)}\sigma_Y^2 = \sigma_Y^2\frac{(1-a^{2n})(1-a)^2}{(1-a^2)(1-a^n)^2} = \sigma_Y^2\frac{(1+a^n)(1-a)}{(1-a^n)(1+a)},$$

which has the limit $\text{var}(m_b) \rightarrow \sigma_Y^2(1-a)/(1+a)$ as $n \rightarrow \infty$. Thus the variance of this estimator does not tend to zero, the distribution does not concentrate around μ_Y, and the estimator, although unbiased, is not consistent. This is perhaps surprising, because all the observations enter this estimator. But most of the observations receive very small weight (the weight of the i^{th} observation is proportional to a^{i-1}, a very small number when i is large), and for this reason there is an insufficient amount of cancellation of sampling errors for the estimator to be consistent.

The third estimator, m_c, is biased but consistent. Its bias is $1/n$: $E(m_c) = E(\overline{Y} + 1/n) = \mu_Y + 1/n$, so the bias tends to zero as the sample size increases. To see why m_c is consistent: $\Pr(|m_c - \mu_Y| \geq \delta) = \Pr(|\overline{Y} + 1/n - \mu_Y| \geq \delta)$. Now, from Equation (17.43) in Appendix 17.2, a generalization of Chebychev's inequality implies that for any random variable W, $\Pr(|W| \geq \delta) \leq E(W^2)/\delta^2$ for any positive constant δ. Thus. $\Pr(|\overline{Y} + 1/n - \mu_Y| \geq \delta) \leq E[(\overline{Y} + 1/n - \mu_Y)^2]/\delta^2$. But $E[(\overline{Y} + 1/n - \mu_Y)^2] = \text{var}(\overline{Y}) + 1/n^2 = \sigma^2/n + 1/n^2 \longrightarrow 0$ as n grows large. It follows that $\Pr(|\overline{Y} + 1/n - \mu_Y| \geq \delta) \longrightarrow 0$, and m_c is consistent. This example illustrates the general point that an estimator can be biased in finite samples but, if that bias vanishes as the sample size gets large, the estimator can still be consistent (Exercise 17.10).

The Central Limit Theorem and Convergence in Distribution

If the distributions of a sequence of random variables converge to a limit as $n \longrightarrow \infty$, then the sequence of random variables is said to converge in distribution. The central limit theorem says that, under general conditions, the standardized sample average converges in distribution to a normal random variable.

Convergence in distribution. Let $F_1, F_2, \ldots, F_n, \ldots$ be a sequence of cumulative distribution functions corresponding to a sequence of random variables, S_1, $S_2, \ldots, S_n, \ldots$. For example, S_n might be the standardized sample average, $(\overline{Y} - \mu_Y)/\sigma_{\overline{Y}}$. Then the sequence of random variables S_n is said to **converge in distribution** to S (denoted $S_n \xrightarrow{d} S$) if the distribution functions $\{F_n\}$ converge to F, the distribution of S. That is,

$$S_n \xrightarrow{d} S \text{ if and only if } \lim_{n\to\infty} F_n(t) = F(t), \tag{17.7}$$

where the limit holds at all points t at which the limiting distribution F is continuous. The distribution F is called the **asymptotic distribution** of S_n.

It is useful to contrast the concepts of convergence in probability ($\xrightarrow{p}$) and convergence in distribution ($\xrightarrow{d}$). If $S_n \xrightarrow{p} \mu$, then S_n becomes close to μ with high probability as n increases. In contrast, if $S_n \xrightarrow{d} S$, then the *distribution* of S_n becomes close to the *distribution* of S as n increases.

The central limit theorem. We now restate the central limit theorem using the concept of convergence in distribution. The central limit theorem in Key Concept 2.7 states that if $Y_1, \ldots, Y_n$ are i.i.d. and $0 < \sigma_Y^2 < \infty$, then the asymptotic distribution of $(\overline{Y} - \mu_Y)/\sigma_{\overline{Y}}$ is $N(0, 1)$. Because $\sigma_{\overline{Y}} = \sigma_Y/\sqrt{n}$, $(\overline{Y} - \mu_Y)/\sigma_{\overline{Y}} = \sqrt{n}(\overline{Y} - \mu_Y)/\sigma_Y$. Thus the central limit theorem can be restated as $\sqrt{n}(\overline{Y} - \mu_Y) \xrightarrow{d} \sigma_Y Z$, where Z is a standard normal random variable. This means that the distribution of $\sqrt{n}(\overline{Y} - \mu_Y)$ converges to $N(0, \sigma_Y^2)$ as $n \longrightarrow \infty$. Conventional shorthand for this limit is

$$\sqrt{n}(\overline{Y} - \mu_Y) \xrightarrow{d} N(0, \sigma_Y^2). \tag{17.8}$$

That is, if $Y_1, \ldots, Y_n$ are i.i.d. and $0 < \sigma_Y^2 < \infty$, then the distribution of $\sqrt{n}(\overline{Y} - \mu_Y)$ converges to a normal distribution with mean zero and variance σ_Y^2.

Extensions to time series data. The law of large numbers and central limit theorem stated in Section 2.6 apply to i.i.d. observations. As discussed in Chapter 14, the i.i.d. assumption is inappropriate for time series data, and these theorems need to be extended before they can be applied to time series observations. Those extensions are technical in nature, in the sense that the conclusion is the same—versions of the law of large numbers and the central limit theorem apply to time series data—but the conditions under which they apply are different. This is discussed briefly in Section 16.4, but a mathematical treatment of asymptotic distribution theory for time series variables is beyond the scope of this book and interested readers are referred to Hayashi (2000, Chapter 2).

Slutsky's Theorem and the Continuous Mapping Theorem

Slutsky's theorem combines consistency and convergence in distribution. Suppose that $a_n \xrightarrow{p} a$, where a is a constant, and $S_n \xrightarrow{d} S$. Then

$$a_n + S_n \xrightarrow{d} a + S,\ a_nS_n \xrightarrow{d} aS, \text{ and, if } a \neq 0,\ S_n/a_n \xrightarrow{d} S/a. \tag{17.9}$$

These three results are together called Slutsky's theorem.

The **continuous mapping theorem** concerns the asymptotic properties of a continuous function, g, of a sequence of random variables, S_n. The theorem has two parts. The first is that if S_n converges in probability to the constant a, then $g(S_n)$

converges in probability to $g(a)$; the second is that if S_n converges in distribution to S, then $g(S_n)$ converges in distribution to $g(S)$. That is, if g is a continuous function, then

$$\begin{aligned}&\text{(i) if } S_n \xrightarrow{p} a, \text{ then } g(S_n) \xrightarrow{p} g(a), \text{ and}\\ &\text{(ii) if } S_n \xrightarrow{d} S, \text{ then } g(S_n) \xrightarrow{d} g(S).\end{aligned} \tag{17.10}$$

As an example of (i), if $s_Y^2 \xrightarrow{p} \sigma_Y^2$, then $\sqrt{s_Y^2} = s_Y \xrightarrow{p} \sigma_Y$. As an example of (ii), suppose that $S_n \xrightarrow{d} Z$, where Z is a standard normal random variable, and let $g(S_n) = S_n^2$. Because g is continuous, the continuous mapping theorem applies and $g(S_n) \xrightarrow{d} g(Z)$; that is, $S_n^2 \xrightarrow{d} Z^2$. In other words, the distribution of S_n^2 converges to the distribution of a squared standard normal random variable, which in turn has a χ_1^2 distribution; that is, $S_n^2 \xrightarrow{d} \chi_1^2$.

Application to the *t*-Statistic Based on the Sample Mean

We now use the central limit theorem, the law of large numbers, and Slutsky's theorem to prove that, under the null hypothesis, the t-statistic based on $\overline{Y}$ has a standard normal distribution when $Y_1, \ldots, Y_n$ are i.i.d. and $0 < E(Y_i^4) < \infty$.

The t-statistic for testing the null hypothesis that $E(Y_i) = \mu_0$ based on the sample average $\overline{Y}$ is given in Equations (3.8) and (3.11), and can be written

$$t = \frac{\overline{Y} - \mu_0}{s_Y/\sqrt{n}} = \frac{\sqrt{n}\,(\overline{Y} - \mu_0)}{\sigma_Y} \div \frac{s_Y}{\sigma_Y}, \tag{17.11}$$

where the second equality uses the trick of dividing both the numerator and the denominator by σ_Y.

Because $Y_1, \ldots, Y_n$ have two moments (which is implied by their having four moments; see Exercise 17.5), and because $Y_1, \ldots, Y_n$ are i.i.d., the first term after the final equality in Equation (17.11) obeys the central limit theorem: Under the null hypothesis, $\sqrt{n}(\overline{Y} - \mu_0)/\sigma_Y \xrightarrow{d} N(0, 1)$. In addition, $s_Y^2 \xrightarrow{p} \sigma_Y^2$ (as proven in Appendix 3.3), so $s_Y^2/\sigma_Y^2 \xrightarrow{p} 1$ and the ratio in the second term in Equation (17.11) tends to 1 (Exercise 17.4). Thus the expression after the final equality in Equation (17.11) has the form of the final expression in Equation (17.9), where [in the notation of Equation (17.9)] $S_n = \sqrt{n}(\overline{Y} - \mu_0)/\sigma_Y \xrightarrow{d} N(0, 1)$ and $a_n = s_Y/\sigma_Y \xrightarrow{p} 1$. It follows by applying Slutsky's theorem that $t \xrightarrow{d} N(0, 1)$.

17.3 Asymptotic Distribution of the OLS Estimator and *t*-Statistic

Recall from Chapter 4 that, under the assumptions of Key Concept 4.3 (the first three assumptions of Key Concept 17.1), the OLS estimator $\hat{\beta}_1$ is consistent and $\sqrt{n}(\hat{\beta}_1 - \beta_1)$ has an asymptotic normal distribution. Moreover, the *t*-statistic testing the null hypothesis $\beta_1 = \beta_{1,0}$ has an asymptotic standard normal distribution under the null hypothesis. This section summarizes these results and provides additional details of their proofs.

Consistency and Asymptotic Normality of the OLS Estimators

The large-sample distribution of $\hat{\beta}_1$, originally stated in Key Concept 4.4, is

$$\sqrt{n}(\hat{\beta}_1 - \beta_1) \xrightarrow{d} N\left(0, \frac{\text{var}(v_i)}{[\text{var}(X_i)]^2}\right), \tag{17.12}$$

where $v_i = (X_i - \mu_X)u_i$. The proof of this result was sketched in Appendix 4.3, but that proof omitted some details and involved an approximation that was not formally shown. The missing steps in that proof are left as Exercise 17.3.

An implication of Equation (17.12) is that $\hat{\beta}_1$ is consistent (Exercise 17.4).

Consistency of Heteroskedasticity-Robust Standard Errors

Under the first three least squares assumptions, the heteroskedasticity-robust standard error for $\hat{\beta}_1$ forms the basis for valid statistical inferences. Specifically,

$$\frac{\hat{\sigma}^2_{\hat{\beta}_1}}{\sigma^2_{\hat{\beta}_1}} \xrightarrow{p} 1, \tag{17.13}$$

where $\sigma^2_{\hat{\beta}_1} = \text{var}(v_i)/\{n[\text{var}(X_i)]^2\}$ and $\hat{\sigma}^2_{\hat{\beta}_1}$ is square of the heteroskedasticity-robust standard error defined in Equation (5.4); that is,

$$\hat{\sigma}^2_{\hat{\beta}_1} = \frac{1}{n} \frac{\frac{1}{n-2}\sum_{i=1}^{n}(X_i - \overline{X})^2\hat{u}_i^2}{\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right]^2}. \tag{17.14}$$

To show the result in Equation (17.13), first use the definitions of $\sigma^2_{\hat{\beta}_1}$ and $\hat{\sigma}^2_{\hat{\beta}_1}$ to rewrite the ratio in Equation (17.13) as

$$\frac{\hat{\sigma}^2_{\hat{\beta}_1}}{\sigma^2_{\hat{\beta}_1}} = \left[\frac{n}{n-2}\right]\left[\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\hat{u}_i^2}{\operatorname{var}(v_i)}\right] \div \left[\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}{\operatorname{var}(X_i)}\right]^2. \tag{17.15}$$

We need to show that each of the three terms in brackets on the right-hand side of Equation (17.15) converge in probability to 1. Clearly the first term converges to 1, and by the consistency of the sample variance (Appendix 3.3) the final term converges in probability to 1. Thus all that remains is to show that the second term converges in probability to 1, that is, that $\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\hat{u}_i^2 \xrightarrow{p} \operatorname{var}(v_i)$.

The proof that $\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\hat{u}_i^2 \xrightarrow{p} \operatorname{var}(v_i)$ proceeds in two steps. The first shows that $\frac{1}{n}\sum_{i=1}^{n}v_i^2 \xrightarrow{p} \operatorname{var}(v_i)$; the second shows that $\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\hat{u}_i^2 - \frac{1}{n}\sum_{i=1}^{n}v_i^2 \xrightarrow{p} 0$.

For the moment, suppose that X_i and u_i have eight moments [that is, $E(X_i^8) < \infty$ and $E(u_i^8) < \infty$], which is a stronger assumption than the four moments required by the third least squares assumption. To show the first step, we must show that $\frac{1}{n}\sum_{i=1}^{n}v_i^2$ obeys the law of large numbers in Equation (17.5). To do so, v_i^2 must be i.i.d. (which it is by the second least squares assumption) and $\operatorname{var}(v_i^2)$ must be finite. To show that $\operatorname{var}(v_i^2) < \infty$, apply the Cauchy–Schwarz inequality (Appendix 17.2): $\operatorname{var}(v_i^2) \le E(v_i^4) = E[(X_i - \mu_X)^4u_i^4] \le \{E[(X_i - \mu_X)^8]E(u_i^8)\}^{1/2}$. Thus, if X_i and u_i have eight moments, then v_i^2 has a finite variance and thus satisfies the law of large numbers in Equation (17.5).

The second step is to prove that $\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\hat{u}_i^2 - \frac{1}{n}\sum_{i=1}^{n}v_i^2 \xrightarrow{p} 0$. Because $v_i = (X_i - \mu_X)u_i$, this second step is the same as showing that

$$\frac{1}{n}\sum_{i=1}^{n}[(X_i - \overline{X})^2\hat{u}_i^2 - (X_i - \mu_X)^2u_i^2] \xrightarrow{p} 0. \tag{17.16}$$

Showing this result entails setting $\hat{u}_i = u_i - (\hat{\beta}_0 - \beta_0) - (\hat{\beta}_1 - \beta_1)X_i$, expanding the term in Equation (17.16) in brackets, repeatedly applying the Cauchy–Schwarz inequality, and using the consistency of $\hat{\beta}_0$ and $\hat{\beta}_1$. The details of the algebra are left as Exercise 17.9.

The preceding argument supposes that X_i and u_i have eight moments. This is not necessary, however, and the result $\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\hat{u}_i^2 \xrightarrow{p} \operatorname{var}(v_i)$ can be proven under the weaker assumption that X_i and u_i have four moments, as stated in the third least squares assumption. That proof, however, is beyond the scope of this textbook; see Hayashi (2000, Section 2.5) for details.

Asymptotic Normality of the Heteroskedasticity-Robust *t*-Statistic

We now show that, under the null hypothesis, the heteroskedasticity-robust OLS *t*-statistic testing the hypothesis $\beta_1 = \beta_{1,0}$ has an asymptotic standard normal distribution if least squares Assumptions #1, #2, and #3 hold.

The *t*-statistic constructed using the heteroskedasticity-robust standard error $SE(\hat{\beta}_1) = \hat{\sigma}_{\hat{\beta}_1}$ [defined in Equation (17.14)] is

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{\sqrt{n}(\hat{\beta}_1 - \beta_{1,0})}{\sqrt{n\sigma^2_{\hat{\beta}_1}}} \div \sqrt{\frac{\hat{\sigma}^2_{\hat{\beta}_1}}{\sigma^2_{\hat{\beta}_1}}}. \tag{17.17}$$

It follows from Equation (17.12) and the definition of $\sigma^2_{\hat{\beta}_1}$ that first term after the second equality in Equation (17.17) converges in distribution to a standard normal random variable. In addition, because the heteroskedasticity-robust standard error is consistent [Equation (17.13)], $\sqrt{\hat{\sigma}^2_{\hat{\beta}_1}/\sigma^2_{\hat{\beta}_1}} \xrightarrow{p} 1$ (Exercise 17.4). It follows from Slutsky's theorem that $t \xrightarrow{d} N(0, 1)$.

17.4 Exact Sampling Distributions When the Errors Are Normally Distributed

In small samples, the distribution of the OLS estimator and *t*-statistic depends on the distribution of the regression error and typically is complicated. As discussed in Section 5.6, however, if the regression errors are homoskedastic and normally distributed, then these distributions are simple. Specifically, if all five extended least squares assumptions in Key Concept 17.1 hold, then the OLS estimator has a normal sampling distribution, conditional on $X_1, \ldots, X_n$. Moreover, the *t*-statistic has a Student *t* distribution. We present these results here for $\hat{\beta}_1$.

Distribution of $\hat{\beta}_1$ with Normal Errors

If the errors are i.i.d. normally distributed and independent of the regressors, then the distribution of $\hat{\beta}_1$, conditional on $X_1, \ldots, X_n$, is $N(\beta_1, \sigma^2_{\hat{\beta}_{1|X}})$, where

$$\sigma^2_{\hat{\beta}_{1|X}} = \frac{\sigma^2_u}{\sum_{i=1}^{n}(X_i - \overline{X})^2}. \tag{17.18}$$

The derivation of the normal distribution $N(\beta_1, \sigma^2_{\hat{\beta}_1|X})$, conditional on $X_1, \ldots, X_n$, entails (i) establishing that the distribution is normal; (ii) showing that $E(\hat{\beta}_1 | X_1, \ldots, X_n) = \beta_1$; and (iii) verifying Equation (17.18).

To show (i), note that, conditional on $X_1, \ldots, X_n$, $\hat{\beta}_1 - \beta_1$ is a weighted average of $u_1, \ldots, u_n$:

$$\hat{\beta}_1 = \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})u_i}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}. \tag{17.19}$$

This equation was derived in Appendix 4.3 [Equation (4.30) and is restated here for convenience]. By extended least squares Assumptions #1, #2, #4, and #5, u_i is i.i.d. $N(0, \sigma_u^2)$, and u_i and X_i are independently distributed. Because weighted averages of normally distributed variables are themselves normally distributed, it follows that $\hat{\beta}_1$ is normally distributed, conditional on $X_1, \ldots, X_n$.

To show (ii), take conditional expectations of both sides of Equation (17.19): $E[(\hat{\beta}_1 - \beta_1)|X_1, \ldots, X_n)] = E[\sum_{i=1}^{n}(X_i - \overline{X})u_i / \sum_{i=1}^{n}(X_i - \overline{X})^2 | X_1, \ldots, X_n] = [\sum_{i=1}^{n}(X_i - \overline{X})\ E(u_i|X_1, \ldots, X_n)]/[\sum_{i=1}^{n}(X_i - \overline{X})^2] = 0$, where the final equality follows because $E(u_i|X_1, X_2, \ldots, X_n) = E(u_i|X_i) = 0$. Thus $\hat{\beta}_1$ is conditionally unbiased; that is,

$$E(\hat{\beta}_1 | X_1, \ldots, X_n) = \beta_1. \tag{17.20}$$

To show (iii), use that the errors are independently distributed, conditional on $X_1, \ldots, X_n$, to calculate the conditional variance of $\hat{\beta}_1$ using Equation (17.19):

$$\begin{aligned}\operatorname{var}(\hat{\beta}_1 | X_1, \ldots, X_n) &= \operatorname{var}\left[\frac{\sum_{i=1}^{n}(X_i - \overline{X})u_i}{\sum_{i=1}^{n}(X_i - \overline{X})^2} | X_1, \ldots, X_n\right] \\ &= \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2 \operatorname{var}(u_i | X_1, \ldots, X_n)}{\left[\sum_{i=1}^{n}(X_i - \overline{X})^2\right]^2} \\ &= \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sigma_u^2}{\left[\sum_{i=1}^{n}(X_i - \overline{X})^2\right]^2}.\end{aligned} \tag{17.21}$$

Canceling the term in the numerator in the final expression in Equation (17.21) yields the formula for the conditional variance in Equation (17.18).

Distribution of the Homoskedasticity-Only *t*-Statistic

The homoskedasticity-only t-statistic testing the null hypothesis $\beta_1 = \beta_{1,0}$ is

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}, \tag{17.22}$$

where $SE(\hat{\beta}_1)$ is computed using the homoskedasticity-only standard error of $\hat{\beta}_1$. Substituting the formula for $SE(\hat{\beta}_1)$ [Equation (5.29) of Appendix 5.1] into Equation (17.22) and rearranging yields

$$\begin{aligned} t &= \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{s_{\hat{u}}^2 \Big/ \sum_{i=1}^{n}(X_i - \overline{X})^2}} = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\sigma_u^2 \Big/ \sum_{i=1}^{n}(X_i - \overline{X})^2}} \div \sqrt{\frac{s_{\hat{u}}^2}{\sigma_u^2}} \\ &= \frac{(\hat{\beta}_1 - \beta_{1,0})/\sigma_{\hat{\beta}_1|X}}{\sqrt{W/(n-2)}}, \end{aligned} \tag{17.23}$$

where $s_{\hat{u}}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2$ and $W = \sum_{i=1}^{n} \hat{u}_i^2/\sigma_u^2$. Under the null hypothesis, $\hat{\beta}_1$ has an $N(\beta_{1,0}, \sigma_{\hat{\beta}_1|X}^2)$ distribution conditional on $X_1, \dots, X_n$, so the distribution of the numerator in the final expression in Equation (17.23) is $N(0, 1)$. It is shown in Section 18.4 that W has a chi-squared distribution with $n - 2$ degrees of freedom and moreover that W is distributed independently of the standardized OLS estimator in the numerator of Equation (17.23). It follows from the definition of the Student t distribution (Appendix 17.1) that, under the five extended least squares assumptions, the homoskedasticity-only t-statistic has a Student t distribution with $n - 2$ degrees of freedom.

Where does the degrees of freedom adjustment fit in? The degrees of freedom adjustment in $s_{\hat{u}}^2$ ensures that $s_{\hat{u}}^2$ is an unbiased estimator of σ_u^2 and that the t-statistic has a Student t distribution when the errors are normally distributed.

Because $W = \sum_{i=1}^{n} \hat{u}_i^2/\sigma_u^2$ is a chi-squared random variable with $n - 2$ degrees of freedom, its mean is $E(W) = n - 2$. Thus $E[W/(n - 2)] = (n - 2)/(n - 2) = 1$. Rearranging the definition of W, we have that $E(\frac{1}{n-2}\sum_{i=1}^{n} \hat{u}_i^2) = \sigma_u^2$. Thus the degrees of freedom correction makes $s_{\hat{u}}^2$ an unbiased estimator of σ_u^2. Also, by dividing by $n - 2$ rather than n, the term in the denominator of the final expression of

Equation (17.23) matches the definition of a random variable with a Student *t* distribution given in Appendix 17.1. That is, by using the degrees of freedom adjustment to calculate the standard error, the *t*-statistic has the Student *t* distribution when the errors are normally distributed.

17.5 Weighted Least Squares

Under the first four extended least squares assumptions, the OLS estimator is efficient among the class of linear (in $Y_1, \ldots, Y_n$), conditionally (on $X_1, \ldots, X_n$) unbiased estimators; that is, the OLS estimator is BLUE. This result is the Gauss–Markov theorem, which was discussed in Section 5.5 and proven in Appendix 5.2. The Gauss–Markov theorem provides a theoretical justification for using the OLS estimator. A major limitation of the Gauss–Markov theorem is that it requires homoskedastic errors. If, as is often encountered in practice, the errors are heteroskedastic, the Gauss–Markov theorem does not hold and the OLS estimator is not BLUE.

This section presents a modification of the OLS estimator, called **weighted least squares (WLS)**, which is more efficient than OLS when the errors are heteroskedastic.

WLS requires knowing quite a bit about the conditional variance function, $\text{var}(u_i|X_i)$. We consider two cases. In the first case, $\text{var}(u_i|X_i)$ is known up to a factor of proportionality, and WLS is BLUE. In the second case, the functional form of $\text{var}(u_i|X_i)$ is known, but this functional form has some unknown parameters that can be estimated. Under some additional conditions, the asymptotic distribution of WLS in the second case is the same as if the parameters of the conditional variance function were in fact known, and in this sense the WLS estimator is asymptotically BLUE. The section concludes with a discussion of the practical advantages and disadvantages of handling heteroskedasticity using WLS or, alternatively, heteroskedasticity-robust standard errors.

WLS with Known Heteroskedasticity

Suppose that the conditional variance $\text{var}(u_i|X_i)$ is known up to a factor of proportionality; that is,

$$\text{var}(u_i|X_i) = \lambda h(X_i), \tag{17.24}$$

where λ is a constant and h is a known function. In this case, the WLS estimator is the estimator obtained by first dividing the dependent variable and regressor by the square root of h and then regressing this modified dependent variable on the modified regressor using OLS. Specifically, divide both sides of the single-variable regressor model by $\sqrt{h(X_i)}$ to obtain

$$\widetilde{Y}_i = \beta_0\widetilde{X}_{0i} + \beta_1\widetilde{X}_{1i} + \widetilde{u}_i, \tag{17.25}$$

where $\widetilde{Y}_i = Y_i/\sqrt{h(X_i)}$, $\widetilde{X}_{0i} = 1/\sqrt{h(X_i)}$, $\widetilde{X}_{1i} = X_i/\sqrt{h(X_i)}$, and $\widetilde{u}_i = u_i/\sqrt{h(X_i)}$.

The **WLS estimator** is the OLS estimator of β_1 in Equation (17.25); that is, it is the estimator obtained by the OLS regression of $\widetilde{Y}_i$ on $\widetilde{X}_{0i}$ and $\widetilde{X}_{1i}$, where the coefficient on $\widetilde{X}_{0i}$ takes the place of the intercept in the unweighted regression.

Under the first three least squares assumptions in Key Concept 17.1 plus the known heteroskedasticity assumption in Equation (17.24), WLS is BLUE. The reason that the WLS estimator is BLUE is that weighting the variables has made the error term $\widetilde{u}_i$ in the weighted regression homoskedastic. That is,

$$\text{var}(\widetilde{u}_i|X_i) = \text{var}\left[\frac{u_i}{\sqrt{h(X_i)}}|X_i\right] = \frac{\text{var}(u_i|X_i)}{h(X_i)} = \frac{\lambda h(X_i)}{h(X_i)} = \lambda, \tag{17.26}$$

so the conditional variance of $\widetilde{u}_i$, $\text{var}(\widetilde{u}_i|X_i)$, is constant. Thus the first four least squares assumptions apply to Equation (17.25). Strictly speaking, the Gauss–Markov theorem was proven in Appendix 5.2 for Equation (17.1), which includes the intercept β_0, so it does not apply to Equation (17.25), in which the intercept is replaced by $\beta_0\widetilde{X}_{0i}$. However, the extension of the Gauss–Markov theorem for multiple regression (Section 18.5) does apply to estimation of β_1 in the weighted population regression, Equation (17.25). Accordingly, the OLS estimator of β_1 in Equation (17.25)—that is, the WLS estimators of β_1—is BLUE.

In practice, the function h typically is unknown, so neither the weighted variables in Equation (17.25) nor the WLS estimator can be computed. For this reason, the WLS estimator described here is sometimes called the **infeasible WLS** estimator. To implement WLS in practice, the function h must be estimated, the topic to which we now turn.

WLS with Heteroskedasticity of Known Functional Form

If the heteroskedasticity has a known functional form, then the heteroskedasticity function h can be estimated and the WLS estimator can be calculated using this estimated function.

Example #1: The variance of u is quadratic in X. Suppose that the conditional variance is known to be the quadratic function

$$\mathrm{var}(u_i|X_i) = \theta_0 + \theta_1 X_i^2, \tag{17.27}$$

where θ_0 and θ_1 are unknown parameters, $\theta_0 > 0$, and $\theta_1 \geq 0$.

Because θ_0 and θ_1 are unknown, it is not possible to construct the weighted variables $\tilde{Y}_i$, $\tilde{X}_{0i}$, and $\tilde{X}_{1i}$. It is, however, possible to estimate θ_0 and θ_1, and to use those estimates to compute estimates of $\mathrm{var}(u_i|X_i)$. Let $\hat{\theta}_0$ and $\hat{\theta}_1$ be estimators of θ_0 and θ_1, and let $\widehat{\mathrm{var}}(u_i|X_i) = \hat{\theta}_0 + \hat{\theta}_1 X_i^2$. Define the weighted regressors $\hat{\tilde{Y}}_i = Y_i/\sqrt{\widehat{\mathrm{var}}(u_i|X_i)}$, $\hat{\tilde{X}}_{0i} = 1/\sqrt{\widehat{\mathrm{var}}(u_i|X_i)}$, and $\hat{\tilde{X}}_{1i} = X_{1i}/\sqrt{\widehat{\mathrm{var}}(u_i|X_i)}$. The WLS estimator is the OLS estimator of the coefficients in the regression of $\hat{\tilde{Y}}_i$ on $\hat{\tilde{X}}_{0i}$ and $\hat{\tilde{X}}_{1i}$ (where $\beta_0\hat{\tilde{X}}_{0i}$ takes the place of the intercept β_0).

Implementation of this estimator requires estimating the conditional variance function, that is, estimating θ_0 and θ_1 in Equation (17.27). One way to estimate θ_0 and θ_1 consistently is to regress $\hat{u}_i^2$ on X_i^2 using OLS, where $\hat{u}_i^2$ is the square of the i^{th} OLS residual.

Suppose that the conditional variance has the form in Equation (17.27) and that $\hat{\theta}_0$ and $\hat{\theta}_1$ are consistent estimators of θ_0 and θ_1. Under Assumptions #1 through #3 of Key Concept 17.1, plus additional moment conditions that arise because θ_0 and θ_1 are estimated, the asymptotic distribution of the WLS estimator is the same as if θ_0 and θ_1 were known. Thus the WLS estimator with θ_0 and θ_1 estimated has the same asymptotic distribution as the infeasible WLS estimator and is in this sense asymptotically BLUE.

Because this method of WLS can be implemented by estimating unknown parameters of the conditional variance function, this method is sometimes called **feasible WLS** or *estimated WLS*.

Example #2: The variance depends on a third variable. WLS also can be used when the conditional variance depends on a third variable, W_i, which does not appear in the regression function. Specifically, suppose that data are collected on three variables, Y_i, X_i, and W_i, $i = 1, \ldots, n$; the population regression function depends on X_i but not W_i; and the conditional variance depends on W_i but not X_i. That is, the population regression function is $E(Y_i|X_i, W_i) = \beta_0 + \beta_1 X_i$ and the conditional variance is $\mathrm{var}(u_i|X_i, W_i) = \lambda h(W_i)$, where λ is a constant and h is a function that must be estimated.

For example, suppose that a researcher is interested in modeling the relationship between the unemployment rate in a state and a state economic policy variable (X_i). The measured unemployment rate (Y_i), however, is a survey-based estimate of the true unemployment rate (Y_i^*). Thus Y_i measures Y_i^* with error, where the source of the error is random survey error, so $Y_i = Y_i^* + v_i$, where v_i is the measurement error arising from the survey. In this example, it is plausible that the survey sample size, W_i, is not itself a determinant of the true state unemployment rate. Thus the population regression function does not depend on W_i; that is, $E(Y_i^* | X_i, W_i) = \beta_0 + \beta_1 X_i$. We therefore have the two equations

$$Y_i^* = \beta_0 + \beta_1 X_i + u_i^* \text{ and} \tag{17.28}$$

$$Y_i = Y_i^* + v_i, \tag{17.29}$$

where Equation (17.28) models the relationship between the state economic policy variable and the true state unemployment rate and Equation (17.29) represents the relationship between the measured unemployment rate Y_i and the true unemployment rate Y_i^*.

The model in Equations (17.28) and (17.29) can lead to a population regression in which the conditional variance of the error depends on W_i but not on X_i. The error term u_i^* in Equation (17.28) represents other factors omitted from this regression, while the error term v_i in Equation (17.29) represents measurement error arising from the unemployment rate survey. If u_i^* is homoskedastic, then $\text{var}(u_i^* | X_i, W_i) = \sigma_{u^*}^2$ is constant. The survey error variance, however, depends inversely on the survey sample size W_i; that is, $\text{var}(v_i | X_i, W_i) = a/W_i$ where a is a constant. Because v_i is random survey error, it is safely assumed to be uncorrelated with u_i^*, so $\text{var}(u_i^* + v_i | X_i, W_i) = \sigma_{u^*}^2 + a/W_i$ Thus, substituting Equation (17.28) into Equation (17.29) leads to the regression model with heteroskedasticity

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \tag{17.30}$$

$$\text{var}(u_i | X_i, W_i) = \theta_0 + \theta_1 \left(\frac{1}{W_i} \right), \tag{17.31}$$

where $u_i = u_i^* + v_i$, $\theta_0 = \sigma_{u^*}^2$, $\theta_1 = a$, and $E(u_i | X_i, W_i) = 0$.

If θ_0 and θ_1 were known, then the conditional variance function in Equation (17.31) could be used to estimate β_0 and β_1 by WLS. In this example, θ_0 and θ_1 are unknown, but they can be estimated by regressing the squared OLS residual [from OLS estimation of Equation (17.30)] on $1/W_i$. Then the estimated conditional variance function can be used to construct the weights in feasible WLS.

It should be stressed that it is critical that $E(u_i | X_i, W_i) = 0$; if not, the weighted errors will have nonzero conditional mean and WLS will be inconsistent. Said differently, if W_i is in fact a determinant of Y_i, then Equation (17.30) should be a multiple regression equation that includes both X_i and W_i.

General method of feasible WLS. In general, feasible WLS proceeds in five steps:

1. Regress Y_i on X_i by OLS and obtain the OLS residuals $\hat{u}_i$, $i = 1, \ldots, n$.
2. Estimate a model of the conditional variance function $\text{var}(u_i|X_i)$. For example, if the conditional variance function has the form in Equation (17.27), this entails regressing $\hat{u}_i^2$ on X_i^2. In general, this step entails estimating a function for the conditional variance, $\text{var}(u_i|X_i)$.
3. Use the estimated function to compute predicted values of the conditional variance function, $\widehat{\text{var}}(u_i|X_i)$.
4. Weight the dependent variable and regressor (including the intercept) by the inverse of the square root of the estimated conditional variance function.
5. Estimate the coefficients of the weighted regression by OLS; the resulting estimators are the WLS estimators.

Regression software packages typically include optional weighted least squares commands that automate the fourth and fifth of these steps.

Heteroskedasticity-Robust Standard Errors or WLS?

There are two ways to handle heteroskedasticity: estimating β_0 and β_1 by WLS or estimating β_0 and β_1 by OLS and using heteroskedasticity-robust standard errors. Deciding which approach to use in practice requires weighing the advantages and disadvantages of each.

The advantage of WLS is that it is more efficient than the OLS estimator of the coefficients in the original regressors, at least asymptotically. The disadvantage of WLS is that it requires knowing the conditional variance function and estimating its parameters. If the conditional variance function has the quadratic form in Equation (17.27), this is easily done. In practice, however, the functional form of the conditional variance function is rarely known. Moreover, if the functional form is incorrect, then the standard errors computed by WLS regression routines are invalid in the sense that they lead to incorrect statistical inferences (tests have the wrong size).

The advantage of using heteroskedasticity-robust standard errors is that they produce asymptotically valid inferences even if you do not know the form of the conditional variance function. An additional advantage is that heteroskedasticity-robust standard errors are readily computed as an option in modern regression packages, so no additional effort is needed to safeguard against this threat. The disadvantage of heteroskedasticity-robust standard errors is that the OLS estimator will have a larger variance than the WLS estimator (based on the true conditional variance function).

In practice, the functional form of $\text{var}(u_i|X_i)$ is rarely if ever known, which poses a problem for using WLS in real-world applications. This problem is difficult enough with a single regressor, but in applications with multiple regressors it is even more difficult to know the functional form of the conditional variance. For this reason, practical use of WLS confronts imposing challenges. In contrast, in modern statistical packages it is simple to use heteroskedasticity-robust standard errors, and the resulting inferences are reliable under very general conditions; in particular, heteroskedasticity-robust standard errors can be used without needing to specify a functional form for the conditional variance. For these reasons, it is our opinion that, despite the theoretical appeal of WLS, heteroskedasticity-robust standard errors provide a better way to handle potential heteroskedasticity in most applications.

Summary

1. The asymptotic normality of the OLS estimator, combined with the consistency of heteroskedasticity-robust standard errors, implies that, if the first three least squares assumptions in Key Concept 17.1 hold, then the heteroskedasticity-robust *t*-statistic has an asymptotic standard normal distribution under the null hypothesis.
2. If the regression errors are i.i.d. and normally distributed, conditional on the regressors, then $\hat{\beta}_1$ has an exact normal sampling distribution, conditional on the regressors. In addition, the homoskedasticity-only *t*-statistic has an exact Student t_{n-2} sampling distribution under the null hypothesis.
3. The weighted least squares (WLS) estimator is OLS applied to a weighted regression, where all variables are weighted by the square root of the inverse of the conditional variance, $\text{var}(u_i|X_i)$, or its estimate. Although the WLS estimator is asymptotically more efficient than OLS, to implement WLS you must know the functional form of the conditional variance function, which usually is a tall order.

Key Terms

convergence in probability (680)
consistent estimator (680)
convergence in distribution (682)
asymptotic distribution (682)
Slutsky's theorem (683)
continuous mapping theorem (683)
weighted least squares (WLS) (690)
WLS estimator (691)
infeasible WLS (691)
feasible WLS (692)
normal p.d.f. (701)
bivariate normal p.d.f. (702)

Review the Concepts

17.1 Suppose that Assumption #4 in Key Concept 17.1 is true, but you construct a 95% confidence interval for β_1 using the heteroskedastic-robust standard error in a large sample. Would this confidence interval be valid asymptotically in the sense that it contained the true value of β_1 in 95% of all repeated samples for large n? Suppose instead that Assumption #4 in Key Concept 17.1 is false, but you construct a 95% confidence interval for β_1 using the homoskedasticity-only standard error formula in a large sample. Would this confidence interval be valid asymptotically?

17.2 Suppose that A_n is a sequence of random variables that converges in probability to 3. Suppose that B_n is a sequence of random variables that converges in distribution to a standard normal. What is the asymptotic distribution of A_nB_n? Use this asymptotic distribution to compute an approximate value of $\Pr(A_nB_n < 2)$.

17.3 Suppose that Y and X are related by the regression $Y = 1.0 + 2.0X + u$. A researcher has observations on Y and X, where $0 \leq X \leq 20$, where the conditional variance is $\text{var}(u_i|X_i = x) = 1$ for $0 \leq x \leq 10$ and $\text{var}(u_i|X_i = x) = 16$ for $10 < x \leq 20$. Draw a hypothetical scatterplot of the observations (X_i, Y_i), $i = 1, \ldots, n$. Does WLS put more weight on observations with $x \leq 10$ or $x > 10$? Why?

17.4 Instead of using WLS, the researcher in the previous problem decides to compute the OLS estimator using only the observations for which $x \leq 10$, then using only the observations for which $x > 10$, and then using the average the two OLS of estimators. Is this estimator more efficient than WLS?

Exercises

17.1 Consider the regression model without an intercept term, $Y_i = \beta_1 X_i + u_i$ (so the true value of the intercept, β_0, is zero).

a. Derive the least squares estimator of β_1 for the restricted regression model $Y_i = \beta_1 X_i + u_i$. This is called the restricted least squares estimator ($\hat{\beta}_1^{RLS}$) of β_1 because it is estimated under a restriction, which in this case is $\beta_0 = 0$.

b. Derive the asymptotic distribution of $\hat{\beta}_1^{RLS}$ under Assumptions #1 through #3 of Key Concept 17.1.

c. Show that $\hat{\beta}_1^{RLS}$ is linear [Equation (5.24)] and, under Assumptions #1 and #2 of Key Concept 17.1, conditionally unbiased [Equation (5.25)].

d. Derive the conditional variance of $\hat{\beta}_1^{RLS}$ under the Gauss–Markov conditions (Assumptions #1 through #4 of Key Concept 17.1).

e. Compare the conditional variance of $\hat{\beta}_1^{RLS}$ in (d) to the conditional variance of the OLS estimator $\hat{\beta}_1$ (from the regression including an intercept) under the Gauss–Markov conditions. Which estimator is more efficient? Use the formulas for the variances to explain why.

f. Derive the exact sampling distribution of $\hat{\beta}_1^{RLS}$ under Assumptions #1 through #5 of Key Concept 17.1.

g. Now consider the estimator $\widetilde{\beta}_1 = \Sigma_{i=1}^{n} Y_i / \Sigma_{i=1}^{n} X_i$. Derive an expression for $\text{var}(\widetilde{\beta}_1 | X_1, \ldots, X_n) - \text{var}(\hat{\beta}_1^{RLS} | X_1, \ldots, X_n)$ under the Gauss–Markov conditions and use this expression to show that $\text{var}(\widetilde{\beta}_1 | X_1, \ldots, X_n) \geq \text{var}(\hat{\beta}_1^{RLS} | X_1, \ldots, X_n)$.

17.2 Suppose that (X_i, Y_i) are i.i.d. with finite fourth moments. Prove that the sample covariance is a consistent estimator of the population covariance—that is, $s_{XY} \xrightarrow{p} \sigma_{XY}$, where s_{XY} is defined in Equation (3.24). (*Hint:* Use the strategy outlined in Appendix 3.3 and the Cauchy–Schwarz inequality.)

17.3. This exercise fills in the details of the derivation of the asymptotic distribution of $\hat{\beta}_1$ given in Appendix 4.3.

a. Use Equation (17.19) to derive the expression

$$\sqrt{n}(\hat{\beta}_1 - \beta_1) = \frac{\sqrt{\frac{1}{n}}\sum_{i=1}^{n} v_i}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2} - \frac{(\overline{X} - \mu_X)\sqrt{\frac{1}{n}}\sum_{i=1}^{n} u_i}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2},$$

where $v_i = (X_i - \mu_X)u_i$.

b. Use the central limit theorem, the law of large numbers, and Slutsky's theorem to show that the final term in the equation converges in probability to zero.

c. Use the Cauchy–Schwarz inequality and the third least squares assumption in Key Concept 17.1 to prove that $\text{var}(v_i) < \infty$. Does the term $\sqrt{\frac{1}{n}}\sum_{i=1}^{n} v_i/\sigma_v$ satisfy the central limit theorem?

d. Apply the central limit theorem and Slutsky's theorem to obtain the result in Equation (17.12).

17.4 Show the following results:

a. Show that $\sqrt{n}(\hat{\beta}_1 - \beta_1) \xrightarrow{d} N(0, a^2)$, where a^2 is a constant, implies that $\hat{\beta}_1$ is consistent. (*Hint:* Use Slutsky's theorem.)

b. Show that $s_u^2/\sigma_u^2 \xrightarrow{p} 1$ implies that $s_u/\sigma_u \xrightarrow{p} 1$.

17.5 Suppose that W is a random variable with $E(W^4) < \infty$. Show that $E(W^2) < \infty$.

17.6 Show that if $\hat{\beta}_1$ is conditionally unbiased, then it is unbiased; that is, show that if $E(\hat{\beta}_1|X_1, \ldots, X_n) = \beta_1$, then $E(\hat{\beta}_1) = \beta_1$.

17.7 Suppose that X and u are continuous random variables and $(X_i, u_i), i = 1, \ldots, n$, are i.i.d.

a. Show that the joint probability density function (p.d.f.) of (u_i, u_j, X_i, X_j) can be written as $f(u_i, X_i)f(u_j, X_j)$ for $i \neq j$, where $f(u_i, X_i)$ is the joint p.d.f. of u_i and X_i.

b. Show that $E(u_iu_j|X_i, X_j) = E(u_i|X_i)\,E(u_j|X_j)$ for $i \neq j$.

c. Show that $E(u_i|X_1, \ldots, X_n) = E(u_i|X_i)$.

d. Show that $E(u_iu_j|X_1, X_2, \ldots, X_n) = E(u_i|X_i)\,E(u_j|X_j)$ for $i \neq j$.

17.8 Consider the regression model in Key Concept 17.1 and suppose that Assumptions #1, #2, #3, and #5 hold. Suppose that Assumption #4 is replaced by the assumption that $\text{var}(u_i|X_i) = \theta_0 + \theta_1|X_i|$, where $|X_i|$ is the absolute value of X_i, $\theta_0 > 0$, and $\theta_1 \geq 0$.

a. Is the OLS estimator of β_1 BLUE?

b. Suppose that θ_0 and θ_1 are known. What is the BLUE estimator of β_1?

c. Derive the exact sampling distribution of the OLS estimator, $\hat{\beta}_1$, conditional on $X_1, \ldots, X_n$.

d. Derive the exact sampling distribution of the WLS estimator (treating θ_0 and θ_1 as known) of β_1, conditional on $X_1, \ldots, X_n$.

17.9 Prove Equation (17.16) under Assumptions #1 and #2 of Key Concept 17.1 plus the assumption that X_i and u_i have eight moments.

17.10 Let $\hat{\theta}$ be an estimator of the parameter θ, where $\hat{\theta}$ might be biased. Show that if $E[(\hat{\theta} - \theta)^2] \longrightarrow 0$ as $n \longrightarrow \infty$ (that is, the mean squared error of $\hat{\theta}$ tends to zero), then $\hat{\theta} \xrightarrow{p} \theta$. [*Hint*: Use Equation (17.43) with $W = \hat{\theta} - \theta$.]

17.11 Suppose that X and Y are distributed bivariate normal with density given in Equation (17.38).

a. Show that the density of Y given $X = x$ can be written as

$$f_{Y|X=x}(y) = \frac{1}{\sigma_{Y|X}\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{y - \mu_{Y|X}}{\sigma_{Y|X}}\right)^2\right]$$

where $\sigma_{YX} = \sqrt{\sigma_Y^2(1 - \rho_{XY}^2)}$ and $\mu_{Y|X} = \mu_Y - (\sigma_{XY}/\sigma_X^2)(x - \mu_X)$. [*Hint:* Use the definition of the conditional probability density $f_{Y|X=x}(y) = [g_{X,Y}(x, y)]/[f_X(x)]$, where $g_{X,Y}$ is the joint density of X and Y, and f_X is the marginal density of X.]

b. Use the result in (a) to show that $Y|X = x \sim N(\mu_{Y|X}, \sigma_{Y|X}^2)$.

c. Use the result in (b) to show that $E(Y|X = x) = a + bx$ for suitably chosen constants a and b.

17.12 **a.** Suppose that $u \sim N(0, \sigma_u^2)$. Show that $E(e^u) = e^{\frac{1}{2}\sigma_u^2}$

b. Suppose that the conditional distribution of u given $X = x$ is $N(0, a + bx^2)$, where a and b are positive constants. Show that $E(e^u|X = x) = e^{\frac{1}{2}(a+bx^2)}$.

17.13 Consider the heterogeneous regression model $Y_i = \beta_{0i} + \beta_{1i}X_i + u_i$, where β_{0i} and β_{1i} are random variables that differ from one observation to the next. Suppose that $E(u_i|X_i) = 0$ and (β_{0i}, β_{1i}) are distributed independently of X_i.

a. Let $\hat{\beta}_1^{OLS}$ denote the OLS estimator of β_1 given in Equation (17.2). Show that $\hat{\beta}_1^{OLS} \xrightarrow{p} E(\beta_1)$, where $E(\beta_1)$ is the average value of β_{1i} in the population. [*Hint:* See Equation (13.10).]

b. Suppose that $\text{var}(u_i|X_i) = \theta_0 + \theta_1 X_i^2$, where θ_0 and θ_1 are known positive constants. Let $\hat{\beta}_1^{WLS}$ denote the weighted least squares estimator. Does $\hat{\beta}_1^{WLS} \xrightarrow{p} E(\beta_1)$? Explain.

17.14 Suppose that $Y_i, i = 1, 2, \ldots, n$, are i.i.d. with $E(Y_i) = \mu$, $\text{var}(Y_i) = \sigma^2$, and finite fourth moment. Show the following:

a. $E(Y_i^2) = \mu^2 + \sigma^2$

b. $\overline{Y} \xrightarrow{p} \mu$

c. $\frac{1}{n}\sum_{i=1}^{n} Y_i^2 \xrightarrow{p} \mu^2 + \sigma^2$

d. $\frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \overline{Y}^2$

e. $\frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 \xrightarrow{p} \sigma^2$

f. $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 \xrightarrow{p} \sigma^2$

17.15 Z is distributed $N(0,1)$, W is distributed χ_n^2, and V is distributed χ_m^2. Show, as $n \longrightarrow \infty$ and m is fixed, that:

a. $W/n \xrightarrow{p} 1$.

b. $\frac{Z}{\sqrt{W/n}} \xrightarrow{d} N(0,1)$. Use the result to explain why the t_∞ distribution is the same as the standard normal distribution.

c. $\frac{V/m}{W/n} \xrightarrow{d} \chi_m^2/m$. Use the result to explain why the $F_{m,\infty}$ distribution is the same as the χ_m^2/m distribution.

APPENDIX

17.1 The Normal and Related Distributions and Moments of Continuous Random Variables

This appendix defines and discusses the normal and related distributions. The definitions of the chi-squared, F, and Student t distributions, given in Section 2.4, are restated here for convenient reference. We begin by presenting definitions of probabilities and moments involving continuous random variables.

Probabilities and Moments of Continuous Random Variables

As discussed in Section 2.1, if Y is a continuous random variable, then its probability is summarized by its probability density function (p.d.f.). The probability that Y falls between two values is the area under its p.d.f. between those two values. Because Y is continuous, however, the mathematical expressions for its probabilities involve integrals rather than the summations that are appropriate for discrete random variables.

Let f_Y denote the probability density function of Y. Because probabilities cannot be negative, $f_Y(y) \geq 0$ for all y. The probability that Y falls between a and b (where $a < b$) is

$$\Pr(a \leq Y \leq b) = \int_a^b f_Y(y)dy. \tag{17.32}$$

Because Y must take on some value on the real line, $\Pr(-\infty \leq Y \leq \infty) = 1$, which implies that $\int_{-\infty}^{\infty} f_Y(y)dy = 1$.

Expected values and moments of continuous random variables, like those of discrete random variables, are probability-weighted averages of their values, except that summations [for example, the summation in Equation (2.3)] are replaced by integrals. Accordingly, the expected value of Y is

$$E(Y) = \mu_Y = \int y f_Y(y)dy, \tag{17.33}$$

where the range of integration is the set of values for which f_Y is nonzero. The variance is the expected value of $(Y - \mu_Y)^2$, the r^{th} moment of a random variable is the expected value of Y^r, and the r^{th} central moment is the expected value of $(Y - \mu_Y)^r$. Thus

$$\text{var}(Y) = E(Y - \mu_Y)^2 = \int (y - \mu_Y)^2 f_Y(y)dy, \tag{17.34}$$

$$E(Y^r) = \int y^r f_Y(y)dy, \tag{17.35}$$

and similarly for the r^{th} central moment, $E(Y - \mu_Y)^r$.

The Normal Distribution

The normal distribution for a single variable. The probability density function of a normally distributed random variable (the **normal p.d.f.**) is

$$f_Y(y) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right], \tag{17.36}$$

where $\exp(x)$ is the exponential function of x. The factor $1/(\sigma\sqrt{2\pi})$ in Equation (17.36) ensures that $\Pr(-\infty \leq Y \leq \infty) = \int_{-\infty}^{\infty} f_Y(y)dy = 1$.

The mean of the normal distribution is μ, and its variance is σ^2. The normal distribution is symmetric, so all odd central moments of order three and greater are zero. The fourth central moment is $3\sigma^4$. In general, if Y is distributed $N(\mu, \sigma^2)$, then its even central moments are given by

$$E(Y - \mu)^k = \frac{k!}{2^{k/2}(k/2)!}\sigma^k \text{ } (k \text{ even}). \tag{17.37}$$

When $\mu = 0$ and $\sigma^2 = 1$, the normal distribution is called the standard normal distribution. The standard normal p.d.f. is denoted ϕ, and the standard normal c.d.f. is denoted Φ. Thus the standard normal density is $\phi(y) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right)$ and $\Phi(y) = \int_{-\infty}^{y} \phi(s)ds$.

The bivariate normal distribution. The **bivariate normal p.d.f.** for the two random variables X and Y is

$$g_{X,Y}(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1 - \rho_{XY}^2}} \times \exp\left\{\frac{1}{-2(1 - \rho_{XY}^2)}\left[\left(\frac{x - \mu_X}{\sigma_X}\right)^2 - 2\rho_{XY}\left(\frac{x - \mu_X}{\sigma_X}\right)\left(\frac{y - \mu_Y}{\sigma_Y}\right) + \left(\frac{y - \mu_Y}{\sigma_Y}\right)^2\right]\right\}, \tag{17.38}$$

where ρ_{XY} is the correlation between X and Y.

When X and Y are uncorrelated ($\rho_{XY} = 0$), $g_{X,Y}(x, y) = f_X(x)f_Y(y)$, where f is the normal density given in Equation (17.36). This proves that if X and Y are jointly normally distributed and are uncorrelated, then they are independently distributed. This is a special feature of the normal distribution that is typically not true for other distributions.

The multivariate normal distribution extends the bivariate normal distribution to handle more than two random variables. This distribution is most conveniently stated using matrices and is presented in Appendix 18.1.

The conditional normal distribution. Suppose that X and Y are jointly normally distributed. Then the conditional distribution of Y given X is $N(\mu_{Y|X}, \sigma_{Y|X}^2)$, with mean $\mu_{Y|X} = \mu_Y + (\sigma_{XY}/\sigma_X^2)(X - \mu_X)$ and variance $\sigma_{Y|X}^2 = (1 - \rho_{XY}^2)\sigma_Y^2$. The mean of this conditional distribution, conditional on $X = x$, is a linear function of x, and the variance does not depend on x (Exercise 17.11).

Related Distributions

The chi-squared distribution. Let $Z_1, Z_2, \ldots, Z_n$ be n i.i.d. standard normal random variables. The random variable

$$W = \sum_{i=1}^{n} Z_i^2 \tag{17.39}$$

has a chi-squared distribution with n degrees of freedom. This distribution is denoted χ_n^2. Because $E(Z_i^2) = 1$ and $E(Z_i^4) = 3$, $E(W) = n$ and $\text{var}(W) = 2n$.

The Student t distribution. Let Z have a standard normal distribution, let W have a χ_m^2 distribution, and let Z and W be independently distributed. Then the random variable

$$t = \frac{Z}{\sqrt{W/m}} \tag{17.40}$$

has a Student t distribution with m degrees of freedom, denoted t_m. The t_∞ distribution is the standard normal distribution. (See Exercise 17.15.)

The F distribution. Let W_1 and W_2 be independent random variables with chi-squared distributions with respective degrees of freedom n_1 and n_2. Then the random variable

$$F = \frac{W_1/n_1}{W_2/n_2} \tag{17.41}$$

has an F distribution with (n_1, n_2) degrees of freedom. This distribution is denoted F_{n_1,n_2}.

The F distribution depends on the numerator degrees of freedom n_1 and the denominator degrees of freedom n_2. As number of degrees of freedom in the denominator gets large, the F_{n_1,n_2} distribution is well approximated by a $\chi_{n_1}^2$ distribution, divided by n_1. In the limit, the $F_{n_1,\infty}$ distribution is the same as the $\chi_{n_1}^2$ distribution, divided by n_1; that is, it is the same as the $\chi_{n_1}^2/n_1$ distribution. (See Exercise 17.15.)

APPENDIX

17.2 Two Inequalities

This appendix states and proves Chebychev's inequality and the Cauchy–Schwarz inequality.

Chebychev's Inequality

Chebychev's inequality uses the variance of the random variable V to bound the probability that V is farther than $\pm\delta$ from its mean, where δ is a positive constant:

$$\Pr(|V - \mu_V| \geq \delta) \leq \frac{\text{var}(V)}{\delta^2} \quad \text{(Chebychev's inequality)}. \tag{17.42}$$

To prove Equation (17.42), let $W = V - \mu_V$, let f be the p.d.f. of W, and let δ be any positive number. Now

$$\begin{aligned} E(W^2) &= \int_{-\infty}^{\infty} w^2 f(w)dw \\ &= \int_{-\infty}^{-\delta} w^2 f(w)dw + \int_{-\delta}^{\delta} w^2 f(w)dw + \int_{\delta}^{\infty} w^2 f(w)dw \\ &\geq \int_{-\infty}^{-\delta} w^2 f(w)dw + \int_{\delta}^{\infty} w^2 f(w)dw \\ &\geq \delta^2 \left[\int_{-\infty}^{-\delta} f(w)dw + \int_{\delta}^{\infty} f(w)dw \right] \\ &= \delta^2 \Pr(|W| \geq \delta), \end{aligned} \tag{17.43}$$

where the first equality is the definition of $E(W^2)$, the second equality holds because the ranges of integration divides up the real line, the first inequality holds because the term that was dropped is nonnegative, the second inequality holds because $w^2 \geq \delta^2$ over the range of integration, and the final equality holds by the definition of $\Pr(|W| \geq \delta)$. Substituting $W = V - \mu_v$ into the final expression, noting that $E(W^2) = E[(V - \mu_V)^2] = \text{var}(V)$, and rearranging yields the inequality given in Equation (17.42). If V is discrete, this proof applies with summations replacing integrals.

The Cauchy–Schwarz Inequality

The Cauchy–Schwarz inequality is an extension of the correlation inequality, $|\rho_{XY}| \leq 1$, to incorporate nonzero means. The Cauchy–Schwarz inequality is

$$|E(XY)| \leq \sqrt{E(X^2)E(Y^2)} \quad \text{(Cauchy–Schwarz inequality)}. \tag{17.44}$$

The proof of Equation (17.44) is similar to the proof of the correlation inequality in Appendix 2.1. Let $W = Y + bX$, where b is a constant. Then $E(W^2) = E(Y^2) + 2bE(XY) + b^2E(X^2)$. Now let $b = -E(XY)/E(X^2)$ so that (after simplification) the expression becomes $E(W^2) = E(Y^2) - [E(XY)]^2/E(X^2)$. Because $E(W^2) \geq 0$ (since $W^2 \geq 0$), it must be the case that $[E(XY)]^2 \leq E(X^2)E(Y^2)$, and the Cauchy–Schwarz inequality follows by taking the square root.