CHAPTER 15

Introduction to Time Series Regression and Forecasting

Time series data—data collected for a single entity at multiple points in time—can be used to answer quantitative questions for which cross-sectional data are inadequate. One such question is, what is the causal effect on a variable of interest, Y, of a change in another variable, X, over time? In other words, what is the *dynamic* causal effect on Y of a change in X? For example, what is the effect on traffic fatalities of a law requiring passengers to wear seatbelts, both initially and subsequently, as drivers adjust to the law? Another such question is, what is your best forecast of the value of some variable at a future date? For example, what is your best forecast of next month's unemployment rate, interest rates, or stock prices? Both of these questions—one about dynamic causal effects, the other about economic forecasting—can be answered using time series data.

This chapter and Chapters 16 and 17 introduce techniques for econometric analysis of time series data and apply those techniques to the problems of forecasting and estimating dynamic causal effects. This chapter introduces the basic concepts and tools of regression using time series data and applies them to economic forecasting. Chapter 16 applies these tools to the estimation of dynamic causal effects. Chapter 17 takes up some more advanced topics in time series econometrics, including forecasting multiple time series, forecasting with many predictors, and modeling changes in volatility over time.

Economic forecasting is the prediction of future values of economic variables. Firms use economic forecasts when they plan production levels. Governments use revenue forecasts when they develop their budgets for the upcoming year. Economists at central banks, like the U.S. Federal Reserve System, forecast economic variables including the inflation rate and the growth of Gross Domestic Product (GDP) as part of setting monetary policy. Wall Street investors rely on forecasts of profits when deciding whether to invest in a company.

Forecasting is an application of the more general prediction problem in statistics, in which a given set of data is used to predict observations not in the data set. Forecasting refers to the prediction of *future* values of time series data. As with prediction more generally, forecasting models need not and generally do not have a causal interpretation.

Section 15.1 presents some examples of economic time series data and introduces basic concepts of time series analysis. Section 15.2 sets out the forecasting problem and introduces a measure of forecast accuracy, the mean squared forecast error. It also introduces the concept of stationarity, which implies that historical relationships among variables hold in the future, so that past data can reliably be used to make forecasts. Section 15.3 introduces autoregressions, time series regression models in

which the regressors are past values of the dependent variable, and Section 15.4 explains how to include additional regressors. For example, we find that including the term spread (the difference between long- and short-term interest rates) improves forecasts of the growth of U.S. GDP relative to using only lagged values of GDP growth. Section 15.5 discusses how to estimate the mean squared forecast error and how to compute forecast intervals—that is, ranges that are likely to contain the actual value of the variable being forecasted. Section 15.6 describes methods for choosing the number of lags in forecasting models. Sections 15.7 and 15.8 take up two common departures from the assumption of stationarity, trends and breaks, and show how to modify forecasting regressions if they are present.

15.1 Introduction to Time Series Data and Serial Correlation

A good place to start any empirical analysis is plotting the data, so that is where we begin.

Real GDP in the United States

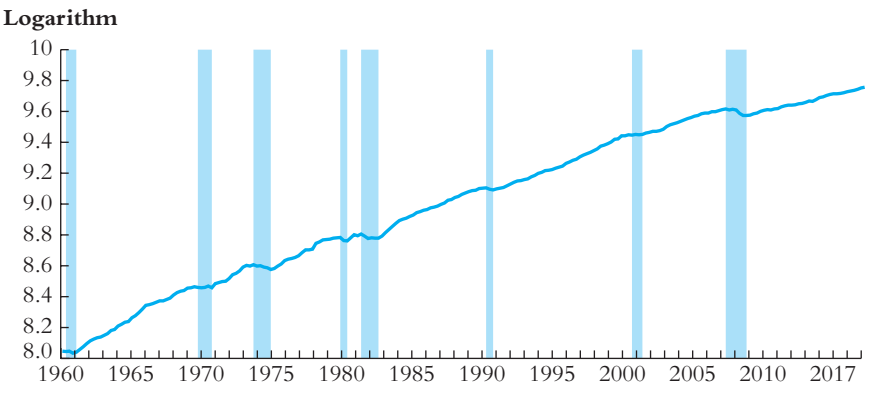
Gross Domestic Product (GDP) measures the value of goods and services produced in an economy over a given time period. Figure 15.1a plots the value of "real" GDP per year in the United States from 1960 through 2017, where "real" indicates that the values have been adjusted for inflation. The values of GDP are expressed in $2009, which means that the price level is held fixed at its 2009 value. Because U.S. GDP grows at approximately an exponential rate, Figure 15.1a plots GDP on a logarithmic scale. GDP increased dramatically over a recent 58-year period, from approximately $3 trillion in 1960 to over $17 trillion in 2017. Measured on a logarithmic scale, this greater-than-five-fold increase corresponds to an increase of 1.7 log points. The rate of growth was not constant, however, and the figure shows declines in GDP during the recessions of 1960–1961, 1970, 1974–1975, 1980, 1981–1982, 1990–1991, 2001, and 2007–2009, episodes denoted by shading in Figure 15.1.

Lags, First Differences, Logarithms, and Growth Rates

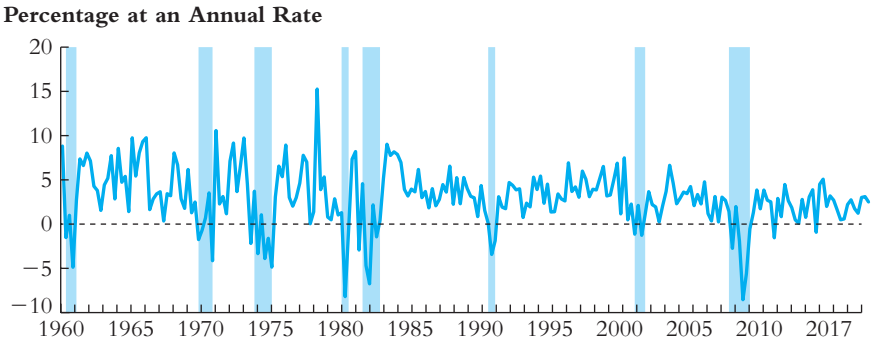
The observation on the time series variable Y made at date t is denoted Y_t, and the total number of observations is denoted T. The interval between observations—that is, the period of time between observation t and observation $t + 1$—is some unit of time such as weeks, months, quarters (three-month units), or years. A set of T observations on a time series variable Y is denoted $Y_1, \ldots, Y_T$, or $\{Y_t\}, t = 1, \ldots, T$. This notation parallels the notation for cross-sectional data, in which the observations are denoted by $i = 1, \ldots, n$. In a given data set, the date $t = 1$ corresponds to the first date in the data set, and $t = T$ corresponds to the final date in the data set. For example, the GDP data studied in this chapter are quarterly, so the unit of time

FIGURE 15.1 The Logarithm and the Growth Rate of Real GDP in the United States, 1960–2017

GDP increased from $3 trillion per year in 1960 to over $17 trillion per year in 2017 when measured in inflation-adjusted 2009 dollars. This greater-than-five-fold increase corresponds to an increase of 1.7 log points. The growth rate of GDP was not constant, and it varied considerably from quarter to quarter.



(a) U.S. GDP ($2009, billions)



(b) Growth rate in U.S. GDP

(a period) is a quarter of a year. The data plotted in Figure 15.1b are quarterly growth rates of GDP from the first quarter of 1960, or 1960:Q1, through the fourth quarter of 2017, or 2017:Q4, for a total of $T = 232$ observations.

The change in the value of Y between period $t - 1$ and period t is $Y_t - Y_{t-1}$; this change is called the **first difference** in the variable Y_t. In time series data, "Δ" is used to represent the first difference, so $\Delta Y_t = Y_t - Y_{t-1}$.

Special terminology and notation are used to indicate future and past values of Y. The value of Y in the previous period (relative to the current period, t) is called its *first lagged value* (or, more simply, its **first lag**) and is denoted Y_{t-1}. Its j^{th} *lagged value* (or, more simply, its j^{th} **lag**) is its value j periods ago, which is Y_{t-j}. Similarly, Y_{t+1} denotes the value of Y one period into the future.

Economic time series are often analyzed after computing their logarithms or the changes in their logarithms. One reason for this is that many economic series exhibit growth that is approximately exponential; that is, over the long run, the series tends to grow by a certain percentage per year on average. This implies that the logarithm of the series grows approximately linearly and is why Figure 15.1a plots the logarithm

Lags, First Differences, Logarithms, and Growth Rates

KEY CONCEPT 15.1

- The first lag of a time series Y_t is Y_{t-1}; its j^{th} lag is Y_{t-j}.
- The first difference of a series, ΔY_t, is its change between periods $t-1$ and t; that is, $\Delta Y_t = Y_t - Y_{t-1}$.
- The first difference of the logarithm of Y_t is $\Delta\ln(Y_t) = \ln(Y_t) - \ln(Y_{t-1})$.
- The percentage change of a time series Y_t between periods $t-1$ and t is approximately $100\Delta\ln(Y_t)$, where the approximation is most accurate when the percentage change is small.

of U.S. GDP. Another reason is that the standard deviation of many economic time series is approximately proportional to its level; that is, the standard deviation is well expressed as a percentage of the level of the series. This implies that the standard deviation of the logarithm of the series is approximately constant. In either case, it is useful to transform the series so that changes in the transformed series are proportional (or percentage) changes in the original series, and this is achieved by taking the logarithm of the series.[1]

Lags, first differences, and growth rates are summarized in Key Concept 15.1.

Lags, changes, and percentage changes are illustrated using the U.S. GDP data in Table 15.1. The first column shows the date, or period, where the fourth quarter of 2016 is denoted 2016:Q4, the first quarter of 2017 is denoted 2017:Q1, and so forth. The second column shows the value of GDP in that quarter, the third column shows the logarithm of GDP, and the fourth column shows the growth rate of GDP (in percent at an annual rate). For example, from the fourth quarter of 2016 to the first quarter of 2017, GDP increased from \$16,851 to \$16,903 billion, which is a percentage increase of $100 \times (16{,}903 - 16{,}851)/16{,}851 = 0.31\%$. This is the percentage increase from one quarter to the next. It is conventional to report rates of growth in quarterly macroeconomic time series on an annual basis, which is the percentage increase in GDP that would occur over a year if the series were to continue to increase at the same rate. Because there are four quarters in a year, the annualized rate of GDP growth in 2017:Q1 is $0.31 \times 4 = 1.24$, or 1.24%.

[1]The change of the logarithm of a variable is approximately equal to the proportional change of that variable; that is, $\ln(X + a) - \ln(X) \cong a/X$, where the approximation works best when a/X is small [see Equation (8.16) and the surrounding discussion]. Now, replace X with Y_{t-1} and a with ΔY_t, and note that $Y_t = Y_{t-1} + \Delta Y_t$. This means that the proportional change in the series Y_t between periods $t-1$ and t is approximately $\ln(Y_t) - \ln(Y_{t-1}) = \ln(Y_{t-1} + \Delta Y_t) - \ln(Y_{t-1}) \cong \Delta Y_t/Y_{t-1}$ (see Equation 18.16). The expression $\ln(Y_t) - \ln(Y_{t-1})$ is the first difference of $\ln(Y_t)$—that is, $\Delta\ln(Y_t)$. Thus $\Delta\ln(Y_t) \cong \Delta Y_t/Y_{t-1}$. The percentage change is 100 times the fractional change, so the percentage change in the series Y_t is approximately $100\Delta\ln(Y_t)$.

TABLE 15.1 GDP in the United States in the Last Quarter of 2016 and in 2017

Quarter	U.S. GDP (billions of \$2009), GDP_t	Logarithm of GDP, $\ln(GDP_t)$	Growth Rate of GDP at an Annual Rate, $GDPGR_t = 400 \times \Delta\ln(GDP_t)$	First Lag, $GDPGR_{t-1}$
2016:Q4	16,851	9.732	1.74	2.74
2017:Q1	16,903	9.735	1.23	1.74
2017:Q2	17,031	9.743	3.01	1.23
2017:Q3	17,164	9.751	3.11	3.01
2017:Q4	17,272	9.757	2.50	3.11

Note: The quarterly rate of GDP growth is the first difference of the logarithm. This is converted into percentages at an annual rate by multiplying by 400. The first lag is its value in the previous quarter. All entries are rounded to the nearest decimal.

In the table, this percentage change is computed using the differences-of-logarithms approximation in Key Concept 15.1. The difference in the logarithm of GDP from 2016:Q4 to 2017:Q1 is $\ln(16{,}903) - \ln(16{,}851) = 0.00308$, yielding the approximate quarterly percentage difference $100 \times 0.00308 = 0.308\%$. On an annualized basis, this is $0.308 \times 4 = 1.23$, or 1.23%, essentially the same as the change obtained by directly computing the percentage growth. These calculations can be summarized as

$$\begin{aligned}\text{Annualized rate of GDP growth} &= GDPGR_t \cong 400\,[\ln(GDP_t) - \ln(GDP_{t-1})] \\ &= 400\Delta\ln(GDP_t), \qquad (15.1)\end{aligned}$$

where GDP_t is the value of GDP at date t. The factor of 400 arises from converting the decimal change to a percentage (multiplying by 100) and then converting the quarterly percentage change to an equivalent annual rate (multiplying by 4).

The final column of Table 15.1 illustrates lags. The first lag of *GDPGR* in 2017:Q1 is 1.74%, the value of *GDPGR* in 2016:Q4.

Figure 15.1b plots $GDPGR_t$ from 1960:Q1 through 2017:Q4. It shows substantial variability in the growth rate of GDP. For example, GDP grew at an annual rate of over 15% in 1978:Q2 and fell at an annual rate of over 8% in 2008:Q4. Over the entire period, the growth rate averaged 3.0% (which is responsible for the increase of GDP from \$3.1 trillion in 1960 to \$17.3 trillion in 2017), and the sample standard deviation was 3.3%.

Autocorrelation

In time series data, the value of Y in one period typically is correlated with its value in the next period. The correlation of a series with its own lagged values is called **autocorrelation** or **serial correlation**. The first autocorrelation (or **autocorrelation coefficient**) is the correlation between Y_t and Y_{t-1}—that is, the correlation between

Autocorrelation (Serial Correlation) and Autocovariance

KEY CONCEPT 15.2

The j^{th} autocovariance of a series Y_t is the covariance between Y_t and its j^{th} lag, Y_{t-j}, and the j^{th} autocorrelation coefficient is the correlation between Y_t and Y_{t-j}. That is,

$$j^{\text{th}} \text{ autocovariance} = \text{cov}(Y_t, Y_{t-j}) \tag{15.2}$$

$$j^{\text{th}} \text{ autocorrelation} = \rho_j = \text{corr}(Y_t, Y_{t-j}) = \frac{\text{cov}(Y_t, Y_{t-j})}{\sqrt{\text{var}(Y_t)\text{var}(Y_{t-j})}}. \tag{15.3}$$

The j^{th} autocorrelation coefficient is sometimes called the j^{th} serial correlation coefficient.

values of Y at two adjacent dates. The second autocorrelation is the correlation between Y_t and Y_{t-2}, and the j^{th} autocorrelation is the correlation between Y_t and Y_{t-j}. Similarly, the ***j*th autocovariance** is the covariance between Y_t and Y_{t-j}. Autocorrelation and autocovariance are summarized in Key Concept 15.2.

The j^{th} population autocovariances and autocorrelations in Key Concept 15.2 can be estimated by the j^{th} sample autocovariances and autocorrelations, $\widehat{\text{cov}(Y_t, Y_{t-j})}$ and $\hat{\rho}_j$:

$$\widehat{\text{cov}(Y_t, Y_{t-j})} = \frac{1}{T}\sum_{t=j+1}^{T}(Y_t - \overline{Y}_{j+1:T})(Y_{t-j} - \overline{Y}_{1:T-j}) \tag{15.4}$$

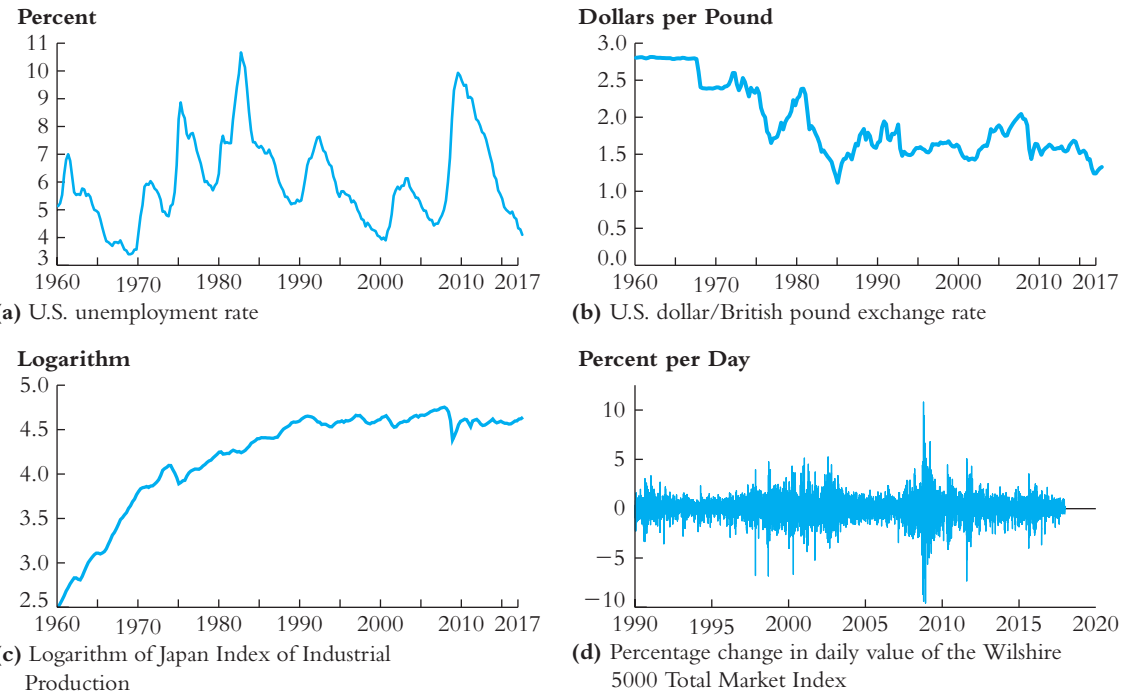
$$\hat{\rho}_j = \frac{\widehat{\text{cov}(Y_t, Y_{t-j})}}{\widehat{\text{var}(Y_t)}}, \tag{15.5}$$

where $\overline{Y}_{j+1:T}$ denotes the sample average of Y_t computed using the observations $t = j + 1, \ldots, T$ and where $\widehat{\text{var}(Y_t)}$ is the sample variance of Y.[2]

The first four sample autocorrelations of *GDPGR*, the growth rate of GDP, are $\hat{\rho}_1 = 0.33$, $\hat{\rho}_2 = 0.26$, $\hat{\rho}_3 = 0.10$, and $\hat{\rho}_4 = 0.11$. These values suggest that GDP growth rates are mildly positively autocorrelated: If GDP grows faster than average in one period, it tends to also grow faster than average in the following period.

[2]The summation in Equation (15.4) is divided by T, whereas in the usual formula for the sample covariance [see Equation (3.24)], the summation is divided by the number of observations in the summation minus a degrees-of-freedom adjustment. The formula in Equation (15.4) is conventional for the purpose of computing the autocovariance. Equation (15.5) uses the assumption that $\text{var}(Y_t)$ and $\text{var}(Y_{t-j})$ are the same—an implication of the assumption that Y is stationary, a concept introduced in Section 15.3.

FIGURE 15.2 Four Economic Time Series



(a) U.S. unemployment rate

(b) U.S. dollar/British pound exchange rate

(c) Logarithm of Japan Index of Industrial Production

(d) Percentage change in daily value of the Wilshire 5000 Total Market Index

The four time series have markedly different patterns. The unemployment rate (Figure 15.2a) increases during recessions and declines during expansions. The exchange rate between the U.S. dollar and the British pound (Figure 15.2b) shows a discrete change after the 1972 collapse of the Bretton Woods system of fixed exchange rates. The logarithm of the Japan Index of Industrial Production (Figure 15.2c) shows decreasing growth. The daily percentage changes in the Wilshire 5000 Total Market Index, a stock price index (Figure 15.2d), are essentially unpredictable, but the variance changes: This series shows *volatility clustering*.

Other Examples of Economic Time Series

Economic time series differ greatly. Four examples of economic time series are plotted in Figure 15.2: the U.S. unemployment rate; the rate of exchange between the U.S. dollar and the British pound; the logarithm of the Japan Index of Industrial Production; and the percentage change in daily values of the Wilshire 5000 Total Market Index, a stock price index.

The U.S. unemployment rate (Figure 15.2a) is the fraction of the labor force out of work, as measured in the Current Population Survey (see Appendix 3.1). Figure 15.2a shows that the unemployment rate increases by large amounts during recessions (the shaded areas in Figure 15.1) and falls during expansions.

The dollar/pound exchange rate (Figure 15.2b) is the price of a British pound (£) in U.S. dollars. Before 1972, the developed economies ran a system of fixed exchange rates—called the Bretton Woods system—under which governments kept exchange

rates from fluctuating. In 1972, inflationary pressures led to the breakdown of this system; thereafter, the major currencies were allowed to "float"; that is, their values were determined by the supply and demand for currencies in the market for foreign exchange. Prior to 1972, the exchange rate was approximately constant, with the exception of a single devaluation in 1968, in which the official value of the pound relative to the dollar was decreased to $2.40. Since 1972, the exchange rate has fluctuated over a very wide range.

The Japan Index of Industrial Production (Figure 15.2c) measures Japan's output of industrial commodities. The logarithm of the series is plotted in Figure 15.2c, and changes in this series can be interpreted as (fractional) growth rates. During the 1960s and early 1970s, Japanese industrial production grew quickly, but this growth slowed in the late 1970s and 1980s, and industrial production has grown little since the early 1990s.

The Wilshire 5000 Total Market Index is an index of the share prices of all firms traded on exchanges in the United States. Figure 15.2d plots the daily percentage change in this index for trading days from January 2, 1990, to December 29, 2017 (a total of 7305 observations). Unlike the other series in Figure 15.2, there is very little serial correlation in these daily percentage changes; if there were, then you could predict them using past daily changes and make money by buying when you expect the market to rise and selling when you expect it to fall. Although the changes are essentially unpredictable, inspection of Figure 15.2d reveals patterns in their volatility. For example, the standard deviation of daily percentage changes was relatively large in 1998–2003 and 2007–2012, and it was relatively small in 1994, 2004, and 2017. This *volatility clustering* is found in many financial time series, and econometric models for modeling this special type of heteroskedasticity are taken up in Section 17.5.

15.2 Stationarity and the Mean Squared Forecast Error

Stationarity

Time series forecasts use data on the past to forecast the future. Doing so presumes that the future is similar to the past in the sense that the correlations, and more generally the distributions, of the data in the future will be like they were in the past. If the future differs fundamentally from the past, then historical relationships might not be reliable guides to the future.

In the context of regression with time series data, the idea that historical relationships can be generalized to the future is formalized by the concept of **stationarity**. The precise definition of stationarity, given in Key Concept 15.3, is that the probability distribution of the time series variable does not change over time. Under the assumption of stationarity, regression models estimated using past data can be used to forecast future values.

KEY CONCEPT 15.3

Stationarity

A time series Y_t is *stationary* if its probability distribution does not change over time—that is, if the joint distribution of $(Y_{s+1}, Y_{s+2}, \ldots, Y_{s+T})$ does not depend on s, regardless of the value of T; otherwise, Y_t is said to be *nonstationary*. A pair of time series, X_t and Y_t, are said to be *jointly stationary* if the joint distribution of $(X_{s+1}, Y_{s+1}, X_{s+2}, Y_{s+2}, \ldots, X_{s+T}, Y_{s+T})$ does not depend on s, regardless of the value of T. Stationarity requires the future to be like the past, at least in a probabilistic sense.

Stationarity can fail to hold for multiple reasons, in which case the time series is said to be **nonstationary**. One reason is that the unconditional mean might have a trend. For example, the logarithm of U.S. GDP plotted in Figure 15.1a has a persistent upward trend, reflecting long-term economic growth. Another type of nonstationarity arises when the population regression coefficients change at a given point in time. Ways to detect and to address these two types of nonstationarity are taken up in Sections 15.6 and 15.7. Until then, we assume that the time series is stationary.

Forecasts and Forecast Errors

This chapter considers the problem of forecasting the value of a time series variable Y in the period immediately following the end of the available data—that is, of forecasting Y_{T+1} using data through date T. This forecast answers questions such as, Given data through the current quarter, what is my forecast of GDP growth for the next quarter? Because the forecast is for the next time period, this forecast is called a **one-step ahead forecast**. A more ambitious question is, Given data through the current quarter, what is my forecast of GDP growth for *each* of the next eight quarters? Answering that question entails making a forecast over a longer horizon, called a **multi-step ahead forecast**. Multi-step ahead forecasts are taken up in Chapter 17.

We let $\hat{Y}_{T+1|T}$ denote a candidate one-step ahead forecast of Y_{T+1}. In this notation, the subscript $T + 1|T$ indicates that the forecast is of the value of Y at time $T + 1$, made using data through time T, and the caret (^) indicates that the forecast is based on an estimated model. For example, suppose you have quarterly observations on GDP growth (Y) from 1960:Q1 to 2017:Q3. The one-step ahead forecasting problem is to use these data to forecast GDP growth in 2017:Q4, and the forecast is denoted $\hat{Y}_{2017:Q4|2017:Q3}$.

Because the future is unknown, errors in forecasting are inevitable. The **forecast error** is the difference between the actual value of Y_{T+1} and its forecast:

$$\text{Forecast error} = Y_{T+1} - \hat{Y}_{T+1|T}. \tag{15.6}$$

A forecast refers to a prediction made for a future date that is not in the data set used to make the forecast—that is, the forecast is for an out-of-sample future observation. The forecast error is the mistake made by the forecast, which is realized only after time has elapsed and the actual value of Y_{T+1} is observed.

The Mean Squared Forecast Error

Because forecast errors are inevitable, the aim of the forecaster is not to eliminate errors but rather to make them as small as possible—that is, to make the forecasts as accurate as possible. To make this goal precise, we need a quantitative measure of what it means for a forecast error to be small. The most commonly used measure, which we adopt in this text, is the **mean squared forecast error (MSFE)**, which is the expected value of the square of the forecast error:

$$\text{MSFE} = E[(Y_{T+1} - \hat{Y}_{T+1|T})^2]. \tag{15.7}$$

The MSFE is the time series counterpart of the mean squared prediction error introduced in Section 14.2 for out-of-sample prediction with cross-sectional data.

In practice, large forecast errors can be much more costly than small ones. A series of small forecast errors often causes only minor problems for the user, but a single very large forecast error can call the entire forecasting activity into question. The MSFE captures this idea by using the square of the forecast error, so that large errors receive a much greater penalty than small ones.

The **root mean squared forecast error (RMSFE)** is the square root of the MSFE. The RMSFE is easily interpreted because it has the same units as Y. If the forecast is unbiased, forecast errors have mean zero and the RMSFE is the standard deviation of the out-of-sample forecast made using a given model.

The MSFE incorporates two sources of randomness. The first is the randomness of the future value, Y_{T+1}. The second is the randomness arising from estimating a forecasting model. For example, suppose a forecaster uses a very simple model, in which the value of Y_{T+1} is forecasted to be its historical mean value, μ_Y. (This simple model is a plausible starting point for forecasting stock returns, as discussed in the box "Can You Beat the Market?" later in this section.) Because the mean is unknown, it must be estimated—say, by $\hat{\mu}_Y$. In this example, the forecast is $\hat{Y}_{T+1|T} = \hat{\mu}_Y$, the forecast error is $Y_{T+1} - \hat{Y}_{T+1|T} = Y_{T+1} - \hat{\mu}_Y$, and the MSFE is $\text{MSFE} = E[(Y_{T+1} - \hat{\mu}_Y)^2]$. By adding and subtracting μ_Y, if Y_{T+1} is uncorrelated with $\hat{\mu}_Y$, the MSFE can be written as $\text{MSFE} = E[(Y_{T+1} - \mu_Y)^2] + E[(\hat{\mu}_Y - \mu_Y)^2]$. The first term in this expression is the error the forecaster would make if the population mean were known: This term captures the random future (out-of-sample) fluctuations in Y_{T+1} around the population mean. The second term in this expression is the additional error made because the population mean is unknown, so the forecaster must estimate it.

From the perspective of the MSFE, the best-possible prediction is the conditional mean given the in-sample observations on Y—that is, $E(Y_{T+1}|Y_1, \ldots, Y_T)$

Can You Beat the Market?

Have you ever dreamed of getting rich quickly by beating the stock market? If you think that the market will be going up, you should buy stocks today and sell them later, before the market turns down. If you are good at forecasting swings in stock prices, then this active trading strategy will produce better returns than a passive "buy and hold" strategy, in which you purchase stocks and just hang onto them. The trick, of course, is having a reliable forecast of future stock returns.

Forecasts based on past values of stock returns are sometimes called momentum forecasts: If the value of a stock rose this month, perhaps it has momentum and will also rise next month. If so, then returns will be autocorrelated, and the autoregressive model will provide useful forecasts. You can implement a momentum-based strategy for a specific stock or for a stock index that measures the overall value of the market.

Table 15.2 presents autoregressive models of the excess return on a broad-based index of stock prices, called the CRSP value-weighted index, using monthly data from 1960:M1 to 2002:M12, where M1 denotes the first month of the year (January), M2 denotes the second month, and so forth. The monthly excess return is what you earn, in percentage terms, by purchasing a stock at the end of the previous month and selling it at the end of this month minus

TABLE 15.2 Autoregressive Models of Monthly Excess Stock Returns, 1960:M1–2002:M12

Dependent variable: excess returns on the CRSP value-weighted index

	(1)	(2)	(3)
Specification	AR(1)	AR(2)	AR(4)
Regressors			
excess return$_{t-1}$	0.050 (0.051)	0.053 (0.051)	0.054 (0.051)
excess return$_{t-2}$		–0.053 (0.048)	–0.054 (0.048)
excess return$_{t-3}$			0.009 (0.050)
excess return$_{t-4}$			–0.016 (0.047)
Intercept	0.312 (0.197)	0.328 (0.199)	0.331 (0.202)
F-statistic for coefficients on lags of *excess return* (*p*-value)	0.968 (0.325)	1.342 (0.261)	0.707 (0.587)
$\overline{R}^2$	0.0006	0.0014	–0.0022

Note: Excess returns are measured in percentage points per month. The data are described in Appendix 15.1. All regressions are estimated over 1960:M1–2002:M12 ($T = 516$ observations), with earlier observations used for initial values of lagged variables. Entries in the regressor rows are coefficients, with standard errors in parentheses. The final two rows report the *F*-statistic testing the hypothesis that the coefficients on lags of *excess return* in the regression are 0, with its *p*-value in parentheses, and the adjusted R^2, or $\overline{R}^2$.

what you would have earned had you purchased a safe asset (a U.S. Treasury bill). The return on the stock includes the capital gain (or loss) from the change in price plus any dividends you receive during the month. The data are described further in Appendix 15.1.

Sadly, the results in Table 15.2 are negative. The coefficient on lagged returns in the AR(1) model is not statistically significant, and we cannot reject the null hypothesis that the coefficients on lagged returns are all 0 in the AR(2) or AR(4) model. In fact, the adjusted R^2, or $\overline{R}^2$, of one of the models is negative, and those of the other two are only slightly positive, suggesting that none of these models is useful for forecasting.

These negative results are consistent with the theory of efficient capital markets, which holds that excess returns should be unpredictable because stock prices already embody all currently available information. The reasoning is simple: If market participants think that a stock will have a positive excess return next month, then they will buy that stock now, but doing so will drive up the price of the stock to exactly the point at which there is no expected excess return. As a result, we should not be able to forecast future excess returns by using past publicly available information, and we cannot do it, at least using the regressions in Table 15.2.

(Appendix 2.2). This best-possible forecast, $E(Y_{T+1}|Y_1, \ldots, Y_T)$, is called the **oracle forecast**. The oracle forecast is infeasible because the conditional mean is unknown in practice. Because it minimizes the MSFE, the oracle forecast is a conceptual benchmark against which to assess an actual forecast.

The MSFE is an unknown population expectation, so to use it in practice it must be estimated using data. We discuss estimation of the RMSFE in Section 15.4.

15.3 Autoregressions

If you want to predict the future, a good place to start is the immediate past. For example, if you want to forecast the rate of GDP growth in the next quarter, you might use data on how fast GDP grew in the current quarter or perhaps over the past several quarters as well. To do so, a forecaster would fit an autoregression.

The First-Order Autoregressive Model

An **autoregression** expresses the conditional mean of a time series variable Y_t as a linear function of its own lagged values. A **first-order autoregression** uses only one lag of Y in this conditional expectation. That is, in a first-order autoregression, $E(Y_t|Y_{t-1}, Y_{t-2}, \ldots) = \beta_0 + \beta_1 Y_{t-1}$. The first-order autoregression [AR(1)] model can be written in the familiar form of a regression model as

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + u_t, \tag{15.8}$$

where u_t is the error term. The first-order autoregression in Equation (15.8) is a population autoregression with two unknown coefficients, β_0 and β_1.

The unknown population coefficients β_0 and β_1 in Equation (15.8) can be estimated by ordinary least square (OLS). How to estimate β_0 and β_1 might initially seem puzzling: Unlike a cross-sectional regression with X on the right-hand side, Equation (15.8) has Y on both the right- *and* the left-hand sides! The solution to this puzzle is to realize that the variable Y_{t-1} on the right-hand side differs from the dependent variable Y_t because the regressor is the first lag of Y. That is, Equation (15.8) has the form of a standard regression model, with X being the first lag of Y. Thus, to estimate β_0 and β_1, you must create a new variable—the first lag of Y—and then use that as the regressor. Doing so yields the OLS estimators, $\hat{\beta}_0$ and $\hat{\beta}_1$.

To make this concrete, consider estimating a first-order autoregression for GDP growth. Observations on the dependent variable, $Y_t = GDPGR_t$, are given in the fourth column of Table 15.1 for 2016:Q4-2017:Q4. Data on the regressor, $Y_{t-1} = GDPGR_{t-1}$ for those dates are given in the final column of Table 15.1. Thus the OLS estimator is obtained by regressing the data in the fourth column of Table 15.1 (extended back to the start of the sample) against the data in the final column, including an intercept. To estimate this AR(1) model, we use data starting in 1962:Q1 and reserve the final observation, 2017:Q4, to illustrate computing the forecast and forecast error. The resulting first-order autoregression, estimated using data from 1962:Q1–2017:Q3, is

$$\widehat{GDPGR_t} = \underset{(0.322)}{1.950} + \underset{(0.073)}{0.341}\, GDPGR_{t-1}. \tag{15.9}$$

As usual, standard errors are given in parentheses under the estimated coefficients, and $\widehat{GDPGR}$ is the predicted value of $GDPGR$ based on the estimated regression line.

Forecasts and forecast errors. If the population coefficients in Equation (15.8) were known, then the one-step ahead forecast of Y_{T+1}, made using data through date T, would be $\beta_0 + \beta_1 Y_T$. Although β_0 and β_1 are unknown, the forecaster can use their OLS estimates instead. Accordingly, the forecast based on the AR(1) model in Equation (15.8) is

$$\hat{Y}_{T+1|T} = \hat{\beta}_0 + \hat{\beta}_1 Y_T, \tag{15.10}$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are estimated using historical data through time T. The forecast error is $Y_{T+1} - \hat{Y}_{T+1|T}$.

Application to GDP growth. What is the forecast of the growth rate of GDP in the fourth quarter of 2017 (2017:Q4) that a forecaster would have made in 2017:Q3, based on the estimated AR(1) model in Equation (15.9) (which was estimated using data through 2017:Q3)? According to Table 15.1, the growth rate of GDP in 2017:Q3 was 3.11% (so $GDPGR_{2017:Q3} = 3.11$). Plugging this value

into Equation (15.8), the forecast of the growth rate of GDP in 2017:Q4 is $\widehat{GDPGR}_{2017:Q4|2017:Q3} = 1.950 + 0.341 \times GDPGR_{2017:Q3} = 1.950 + 0.341 \times 3.11 = 3.0$ (rounded to the nearest tenth).Thus, the AR(1) model forecasts that the growth rate of GDP will be 3.0% in 2017:Q4. Because data for 2017:Q4 are available, we can evaluate the forecast error for this forecast. Table 15.1 shows that the actual growth rate of GDP in 2017:Q4 was 2.5%, so the AR(1) forecast is high by 0.5 percentage points; that is, the forecast error is –0.5.[3]

The $\overline{R}^2$ of the AR(1) model in Equation (15.9) is only 0.11, so the lagged value of GDP growth explains only a small fraction of the variation in GDP growth in the sample used to fit the autoregression. It is therefore of interest to see whether including additional variables, beyond the first lag, could improve the fit of the forecasting model.

The p^{th}-Order Autoregressive Model

The AR(1) model uses Y_{t-1} to forecast Y_t, but doing so ignores potentially useful information in the more distant past. One way to incorporate this information is to include additional lags in the AR(1) model; this yields the p^{th}-order autoregressive model.

The ***p*th-order autoregressive [AR(*p*)] model** represents Y_t as a linear function of p of its lagged values; that is, in the AR(p) model, the regressors are Y_{t-1}, $Y_{t-2}, \ldots, Y_{t-p}$, plus an intercept. The number of lags, p, included in an AR(p) model is called the order, or lag length, of the autoregression.

For example, an AR(2) model of GDP growth uses two lags of GDP growth as regressors. Estimated by OLS over the period 1962:Q1–2017:Q3, the AR(2) model is

$$\underset{}{\widehat{GDPGR_t}} = \underset{}{1.60} + \underset{(0.37)}{0.28}\,GDPGR_{t-1} + \underset{(0.08)}{0.18}\,GDPGR_{t-2}. \qquad (15.11)$$

The coefficient on the additional lag in (Equation (15.11)) is significantly different from 0 at the 5% significance level: The t-statistic is 2.30 (p-value = 0.02). This is reflected in an improvement in the $\overline{R}^2$ from 0.11 for the AR(1) model in Equation (15.8) to 0.14 for the AR(2) model.

The AR(p) model is summarized in Key Concept 15.4.

Properties of the forecast and error term in the AR(p) model. The assumption that the conditional expectation of u_t is 0 given past values of Y_t—that is, $E(u_t \mid Y_{t-1}, Y_{t-2}, \ldots) = 0$]—has two important implications.

The first implication is that the best forecast of Y_{T+1} based on its entire history depends on only the most recent p past values. Specifically, let $Y_{T+1|T} = E(Y_{T+1} \mid Y_T, Y_{T-1}, \ldots)$ denote the conditional mean of Y_{T+1} given its

[3]The units of the arithmetic difference between two percentages is percentage points. For example, if an interest rate is 3.5% at an annual rate and it rises to 3.8%, then it has risen by 0.3 percentage points.

KEY CONCEPT 15.4

Autoregressions

The p^{th}-order autoregressive [AR(p)] model represents the conditional expectation of Y_t as a linear function of p of its lagged values:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p} + u_t, \tag{15.12}$$

where $E(u_t \mid Y_{t-1}, Y_{t-2}, \ldots) = 0$. The number of lags p is called the order, or the lag length, of the autoregression.

entire history. Then $Y_{T+1|T}$ is the oracle forecast and has the smallest MSFE of any forecast, based on the history of Y (Exercise 15.5). That is, if Y_t follows an AR(p), then the oracle forecast of Y_{T+1} based on $Y_T, Y_{T-1}, \ldots$ is

$$Y_{T+1|T} = \beta_0 + \beta_1 Y_T + \beta_2 Y_{T-1} + \cdots + \beta_p Y_{T-p+1}. \tag{15.13}$$

In practice, the coefficients $\beta_0, \beta_1, \ldots, \beta_p$ are unknown, so actual forecasts from an AR(p) use Equation (15.13) with estimated coefficients.

The second implication is that the errors u_t are serially uncorrelated. This result follows from Equation (2.28) (Exercise 15.5).

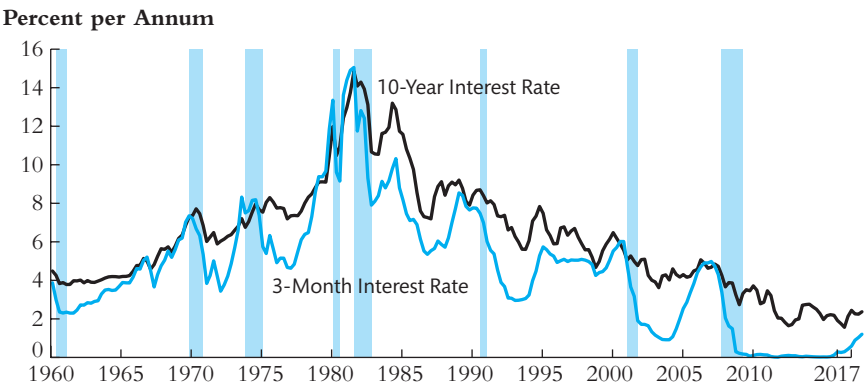
Application to GDP growth. What is the forecast of the growth rate of GDP in 2017:Q4, using data through 2017:Q3, based on the AR(2) model of GDP growth in Equation (15.11)? To compute this forecast, substitute the values of GDP growth in 2017:Q2 and 2017:Q3 into Equation (15.11): $GDPGR_{2017:Q4|2017:Q3} = 160 + 0.28\, GDPGR_{2017:Q3} + 0.18\, GDPGR_{2017:Q2} = 1.60 + 0.28 \times 3.11 + 0.18 \times 3.01 \approx 3.0$, where the 2017:Q3 and 2017:Q2 values for $GDPGR$ are taken from the fourth column of Table 15.1. The forecast error is the actual value, 2.5%, minus the forecast, or 2.5% − 3.0% = −0.5 percentage points, essentially the same as the AR(1) forecast error.

15.4 Time Series Regression with Additional Predictors and the Autoregressive Distributed Lag Model

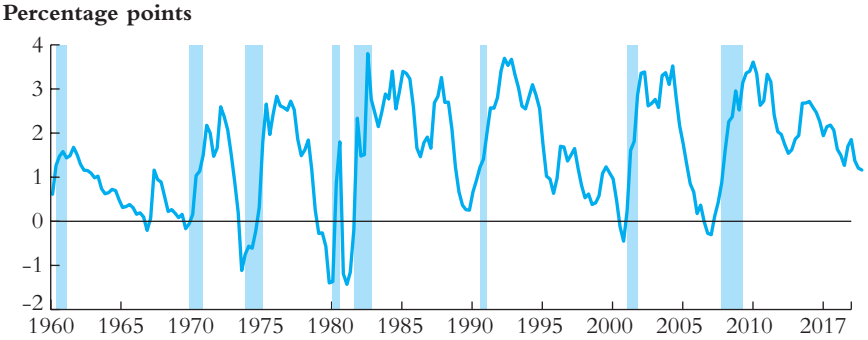
Economic theory often suggests other variables that could help forecast a variable of interest. These other variables, or predictors, can be added to an autoregression to produce a time series regression model with multiple predictors. When other

FIGURE 15.3 Interest Rates and the Term Spread, 1960–2017

Long-term and short-term interest rates on bonds move together but not one-for-one. The difference between long-term rates and short-term rates is called the term spread. The term spread has fallen sharply before U.S. recessions, which are shown as shaded regions in the figures.



(a) 10-year interest rate and 3-month interest rate



(b) Term spread

variables and their lags are added to an autoregression, the result is an autoregressive distributed lag model.

Forecasting GDP Growth Using the Term Spread

Interest rates on long-term and short-term bonds move together but not one for one. Figure 15.3a plots interest rates on 10-year U.S. Treasury bonds and 3-month Treasury bills from 1960 through 2017. These interest rates show the same long-run tendencies: Both were low in the 1960s, both rose through the 1970s and peaked in the early 1980s, and both fell subsequently. But the gap, or difference, between the two interest rates has not been constant: While short-term rates are generally below long-term rates, the gap between them narrows and even disappears shortly before the start of a recession; recessions are shown as the shaded bars in the figure. This difference between long-term and short-term interest rates is called the **term spread** and is plotted in Figure 15.3b. The term spread is generally positive, but it falls toward or below 0 before recessions.

Figure 15.3 suggests that the term spread might contain information about the future GDP growth that is not already contained in past values of GDP growth. This conjecture is readily checked by augmenting the AR(2) model in Equation (15.11) to include the first lag of the term spread:

$$\begin{aligned}\widehat{GDPGR_t} &= 0.94 + 0.27\,GDPGR_{t-1} + 0.19\,GDPGR_{t-2} + 0.42\,TSpread_{t-1}. \\ &\quad (0.47)\ (0.08) \qquad\quad (0.08) \qquad\quad (0.18)\end{aligned} \tag{15.14}$$

The t-statistic on $TSpread_{t-1}$ is -2.34, so this coefficient is significant at the 1% level. The $\overline{R}^2$ of this regression is 0.16, an improvement over the AR(2) $\overline{R}^2$ of 0.14.

The forecast of the rate of GDP growth in 2017:Q4 is obtained by substituting the 2017:Q2 and 2017:Q3 values of GDP growth into Equation (15.14), along with the value of the term spread in 2017:Q3 (which is 1.21); the resulting forecast is $\widehat{GDPGR}_{2017:Q4|2017:Q3} = 2.9\%$, and the forecast error is –0.4%.

If one lag of the term spread is helpful for forecasting GDP growth, more lags might be even more helpful; adding an additional lag of the term spread yields

$$\begin{aligned}\widehat{GDPGR_t} &= 0.94 + 0.25\,GDPGR_{t-1} + 0.18\,GDPGR_{t-2} \\ &\quad (0.46)\ (0.08) \qquad\quad (0.08) \\ &\quad - 0.13\,TSpread_{t-1} + 0.62\,TSpread_{t-2}. \\ &\quad (0.42) \qquad\quad (0.43)\end{aligned} \tag{15.15}$$

The t-statistic testing the significance of the second lag of the term spread is 1.46 (p-value $= 0.14$), so it falls just short of statistical significance at the 10% level. The $\overline{R}^2$ of the regression in Equation (15.15) is 0.16, essentially the same as that in Equation (15.14).

The forecasted rate of GDP growth in 2017:Q4 is computed by substituting the values of the variables into Equation (15.15). The term spread was 1.37 in 2017:Q2 and 1.21 in 2017:Q3. The forecasted value of the rate of GDP growth in 2017:Q4, based on Equation (15.15), is

$$\begin{aligned}\widehat{GDPGR}_{2017:Q4|2017:Q3} &= 0.94 + 0.25 \times 3.11 + 0.18 \times 3.01 \\ &\quad - 0.13 \times 1.21 + 0.62 \times 1.37 \approx 2.9.\end{aligned} \tag{15.16}$$

The forecast error is –0.4 percentage points.

The Autoregressive Distributed Lag Model

Each model in Equations (15.14) and (15.15) is an **autoregressive distributed lag (ADL) model**: *autoregressive* because lagged values of the dependent variable are included as regressors, as in an autoregression, and *distributed lag* because the regression also includes multiple lags (a "distributed lag") of an additional predictor. In general, an ADL model with p lags of the dependent variable Y_t and q lags of an

The Autoregressive Distributed Lag Model

KEY CONCEPT 15.5

The autoregressive distributed lag model with p lags of Y_t and q lags of X_t, denoted ADL(p, q), is

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p} + \delta_1 X_{t-1} + \delta_2 X_{t-2} + \cdots + \delta_q X_{t-q} + u_t, \quad (15.17)$$

where $\beta_0, \beta_1, \ldots, \beta_p, \delta_1, \ldots, \delta_q$, are unknown coefficients and u_t is the error term with $E(u_t \mid Y_{t-1}, Y_{t-2}, \ldots, X_{t-1}, X_{t-2}, \ldots) = 0$.

additional predictor X_t is called an **ADL(*p*, *q*)** model. In this notation, the model in Equation (15.14) is an ADL(2, 1) model, and the model in Equation (15.15) is an ADL(2, 2) model.

The ADL model is summarized in Key Concept 15.5. The notation in Equation (15.17) is somewhat cumbersome, and alternative optional notation, based on the so-called lag operator, is presented in Appendix 15.3.

The assumption that the errors in the ADL model have a conditional mean of 0 given all past values of Y and X—that is, that $E(u_t \mid Y_{t-1}, Y_{t-2}, \ldots, X_{t-1}, X_{t-2}, \ldots) = 0$—implies that no additional lags of either Y or X belong in the ADL model. In other words, the lag lengths p and q are the true lag lengths, and the coefficients on additional lags are 0.

The Least Squares Assumptions for Forecasting with Multiple Predictors

The general time series regression model with multiple predictors extends the ADL model to include multiple predictors and their lags. The model is summarized in Key Concept 15.6. The presence of multiple predictors and their lags leads to double subscripting of the regression coefficients and regressors.

The assumptions in Key Concept 15.6 are the time series counterparts of the four least squares assumptions for prediction with multiple regression using cross-sectional data (Appendix 6.4).

The first assumption is that u_t has conditional mean 0 given the history of all the regressors. This assumption extends the assumption used in the AR and ADL models and implies that the oracle forecast of Y_t using all past values of Y and the X's is given by the regression in Equation (15.18).

The second least squares assumption for cross-sectional data is that $(X_{1i}, \ldots, X_{ki}, Y_i)$, $i = 1, \ldots, n$, are independently and identically distributed (i.i.d.). The second assumption for time series regression replaces the i.i.d. assumption by a more appropriate one with two parts. Part (a) is that the data are drawn

KEY CONCEPT 15.6

The Least Squares Assumptions for Forecasting with Time Series Data

The general time series regression model allows for k additional predictors, where q_1 lags of the first predictor are included, q_2 lags of the second predictor are included, and so forth:

$$\begin{aligned} Y_t = {} & \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p} \\ & + \delta_{11} X_{1t-1} + \delta_{12} X_{1t-2} + \cdots + \delta_{1q_1} X_{1t-q_1} \\ & + \cdots + \delta_{k1} X_{kt-1} + \delta_{k2} X_{kt-2} + \cdots + \delta_{kq_k} X_{kt-q_k} + u_t, \end{aligned} \tag{15.18}$$

where

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

from a stationary distribution, so that the distribution of the time series today is the same as its distribution in the past. This assumption is a time series version of the *identically distributed* part of the i.i.d. assumption: The cross-sectional requirement of each draw being identically distributed is replaced by the time series requirement that the joint distribution of the variables, *including lags*, not change over time. If the time series variables are nonstationary, then one or more problems can arise in time series regression, including biased forecasts.

The assumption of stationarity implies that the conditional mean for the data used to estimate the model is also the conditional mean for the out-of-sample observation of interest. Thus the assumption of stationarity is also an assumption about external validity, and it plays the role of the first least squares assumption for prediction in Appendix 6.4.

Part (b) of the second assumption requires that the random variables become independently distributed when the amount of time separating them becomes large. This replaces the cross-sectional requirement that the variables be independently distributed from one observation to the next with the time series requirement that they be independently distributed when they are separated by long periods of time. This assumption is sometimes referred to as **weak dependence**, and it ensures that in large samples there is sufficient randomness in the data for the law of large numbers and the central limit theorem to hold. For a precise mathematical statement of the weak dependence condition, see Hayashi (2000, Chapter 2).

The third assumption (no outliers) and fourth assumption (no perfect multicollinearity) are the same as for cross-sectional data.

Under the assumptions of Key Concept 15.6, inference on the regression coefficients using OLS proceeds in the same way as it usually does using cross-sectional data.

15.5 Estimation of the MSFE and Forecast Intervals

An estimate of the MSFE can be used to summarize forecast uncertainty and to construct forecast intervals.

Estimation of the MSFE

The MSFE, defined in Equation (15.7), is an expected value that depends on the distribution of Y and on the forecasting model. Because it is an expectation, its value is not known and must be estimated from the data.

A natural instinct would be to estimate the MSFE by replacing the expectation in Equation (15.7) with an average over out-of-sample observations. The out-of-sample data, however, are not observed, so this approach is not feasible. Instead, there are three commonly used methods, with increasing complexity, for estimation of the MSFE. All three methods necessarily rely on the in-sample data. The simplest estimator focuses only on future uncertainty and ignores uncertainty associated with estimation of the regression coefficients. The second estimator incorporates future uncertainty and estimation error, under the assumption of stationarity so that the conditional expectation estimated by the model applies to the out-of-sample forecast. The third incorporates uncertainty and estimation error and in addition allows for the possibility that the conditional expectation might change over the course of the sample.

The first two methods are based on an expression for the MSFE derived from Equation (15.7) and the assumption of stationarity. We provide this expression here for an AR(p); it extends directly to the models with additional predictors in Key Concept 15.6. Under the assumption of stationarity,

$$\text{MSFE} = \sigma_u^2 + \text{var}(\hat{\beta}_0 + \hat{\beta}_1 Y_T + \cdots + \hat{\beta}_p Y_{T-p+1}). \tag{15.19}$$

This result is shown for an AR(1) in Exercise 15.12.

The first term in Equation (15.19) is the variance of Y_{T+1} around its conditional mean. This is the variance of the oracle forecast. The second term in Equation (15.19) arises because the coefficients of the autoregression are unknown and must be estimated.

Method 1: Estimating the MSFE by the standard error of the regression. Because the variance of the OLS estimator is proportional to $1/T$, the second term in

Equation (15.19) is proportional to $1/T$. Consequently, if the number of observations T is large relative to the number of autoregressive lags p, then the contribution of the second term is small relative to the first term. That is, if T is large relative to p, Equation (15.19) simplifies to the approximation $\text{MSFE} \approx \sigma_u^2$. This simplification in turn suggests estimating the MSFE by

$$\widehat{\text{MSFE}}_{SER} = s_{\hat{u}}^2, \text{ where } s_{\hat{u}}^2 = \frac{SSR}{T - p - 1}, \tag{15.20}$$

where SSR is the sum of squared residuals of the autoregression. The statistic $s_{\hat{u}}^2$ is the square of the standard error of the regression (SER), originally defined in Equation (6.13) and restated in Equation (15.20) using the notation of autoregressions.

Method 2: Estimating the MSFE by the final prediction error. If T is not large relative to p, the sampling error of the estimated autoregression coefficients can be sufficiently large that the second term in Equation (15.19) should not be ignored. The **final prediction error (FPE)** is an estimate of the MSFE that incorporates both terms in Equation (15.19), under the additional assumption that the errors are homoskedastic. With homoskedastic errors, $\text{var}(\hat{\beta}_0 + \hat{\beta}_1 Y_T + \cdots + \hat{\beta}_p Y_{T-p+1}) \approx \sigma_u^2[(p+1)/T]$ (shown in Appendix 19.7); substitution of this expression into Equation (15.19) yields, $\text{MSFE} = \sigma_u^2 + \sigma_u^2 \frac{p+1}{T} = \sigma_u^2(1 + \frac{p+1}{T})$. The FPE uses this final expression, along with the estimator s_u^2, to estimate the MSFE:

$$\widehat{MSFE}_{FPE} = \left(\frac{T + p + 1}{T}\right) s_{\hat{u}}^2 = \left(\frac{T + p + 1}{T - p - 1}\right) \frac{SSR}{T}. \tag{15.21}$$

The FPE estimator improves upon the squared SER in Equation (15.20) by adjusting for the sampling uncertainty in estimating the autoregression coefficients.

Method 3: Estimating the MSFE by pseudo out-of-sample forecasting. The third estimate of the MSFE uses the data to simulate out-of-sample forecasting. This method proceeds by first dividing the data set into two parts: an initial estimation sample (the first T-P observations) and a reserved sample (the final P observations). The initial estimation sample is used to estimate the forecasting model, which is then used to forecast the first observation in the reserved sample. Next the estimation sample is augmented by the first observation in the reserved sample, and the model is reestimated and is used to forecast the second observation in the reserved sample. This procedure is repeated until the forecast is made for the final observation in the reserved sample and produces P forecasts and thus P forecast errors. Those P forecast errors can then be used to estimate the MSFE.[4]

[4]Readers of Chapter 14 will recognize that this method for estimating the MSFE is related to estimation of the mean squared prediction error by cross validation.

Pseudo Out-of-Sample Forecasts

KEY CONCEPT 15.7

Pseudo out-of-sample forecasts are computed using the following steps:

1. Choose a number of observations, P, for which you will generate pseudo out-of-sample forecasts; for example, P might be 10% or 20% of the sample size. Let $s = T - P$.
2. Estimate the forecasting regression using the estimation sample—that is, using observations $t = 1, \ldots, s$.
3. Compute the forecast for the first period beyond this shortened sample, $s + 1$; call this $\widetilde{Y}_{s+1|s}$.
4. Compute the forecast error, $\widetilde{u}_{s+1} = Y_{s+1} - \widetilde{Y}_{s+1|s}$.
5. Repeat steps 2 through 4 for the remaining periods, $s = T - P + 1$ to $T - 1$ (reestimate the regression for each period). The pseudo out-of-sample forecasts are $\widetilde{Y}_{s+1|s}, s = T - P, \ldots, T - 1$, and the pseudo out-of-sample forecast errors are $\widetilde{u}_{s+1}, s = T - P, \ldots, T - 1$.

This method of estimating a model on a subsample of the data and then using that model to forecast on a reserved sample is called **pseudo out-of-sample forecasting**: *out-of-sample* because the observations being forecasted were not used for model estimation but *pseudo* because the reserved data are not truly out-of-sample observations. The construction of pseudo out-of-sample forecasts is summarized in Key Concept 15.7.

With the resulting pseudo out-of-sample forecast errors $\widetilde{u}_s, s = T - P + 1, \ldots, T$ in hand, the pseudo out-of-sample estimate of the MSFE is

$$\widehat{MSFE}_{POOS} = \frac{1}{P}\sum_{s=T-P+1}^{T} \widetilde{u}_s^2. \tag{15.22}$$

Compared to the squared *SER* estimate in Equation (15.20) and the final prediction error estimate in Equation (15.21), the pseudo out-of-sample estimate in Equation (15.22) has both advantages and disadvantages. The main advantage is that it does not rely on the assumption of stationarity, so that the conditional mean might differ between the estimation and the reserved samples. For example, the coefficients of the autoregression need not be the same in the two samples, and indeed the pseudo out-of-sample forecast error need not have mean 0. Thus any bias in the forecast arising because of a change in coefficients will be captured by $\widehat{MSFE}_{POOS}$ but not by the other two estimators [which rely on Equation (15.19), which was derived under the assumption of stationarity]. Three disadvantages of the pseudo out-of-sample estimate are that it is more difficult to compute, that the estimate of the MSFE will have greater sampling variability than the other two estimates if Y is, in fact,

stationary (because $\widehat{MSFE}_{POOS}$ uses only P forecast errors), and that it requires choosing P.

The choice of P entails a trade-off between the precision of the coefficient estimates and the number of observations available for estimating the MSFE. In practice, choosing P to be 10% or 20% of the total number of observations can provide a reasonable balance between these two considerations.

Application to GDP growth. For the AR(1) in Equation (15.9), $\widehat{\text{RMSFE}}_{SER} = 3.05$, $\widehat{\text{RMSFE}}_{FPE} = 3.07$, and $\widehat{\text{RMSFE}}_{POOS} = 2.60$ (computed over the final 44 quarters or 20% of the sample). For the AR(2) in Equation (15.11), $\widehat{\text{RMSFE}}_{SER} = 3.01$, $\widehat{\text{RMSFE}}_{FPE} = 3.03$, and $\widehat{\text{RMSFE}}_{POOS} = 2.52$. The FPE estimates are larger than the SER estimates because of the additional factor that estimates the variance from estimating the coefficients. The pseudo out-of-sample estimates of the RMSFE are smaller than the in-sample estimates. In part, this reflects the reduction in the variability of GDP growth that occurred in the early 1980s that is evident in Figure 15.1b, a phenomenon known as the Great Moderation.

Forecast Uncertainty and Forecast Intervals

In any estimation problem, it is good practice to report a measure of the uncertainty of that estimate, and forecasting is no exception. One measure of the uncertainty of a forecast is its root mean squared forecast error (RMSFE). Under the additional assumption that the errors u_t are normally distributed, the estimates of the RMSFE introduced in Section 15.3 can be used to construct a forecast interval—that is, an interval that contains the future value of the variable with a certain probability.

Forecast intervals. A forecast interval is like a confidence interval except that it pertains to a forecast. For example, a 95% **forecast interval** is an interval that contains the future value of the variable being forecasted in 95% of repeated applications.

One important difference between a forecast interval and a confidence interval is that the usual formula for a 95% confidence interval (the estimator ± 1.96 standard errors) is justified by the central limit theorem and therefore holds for a wide range of distributions of the error term. In contrast, because the forecast error in Equation (15.15) includes the future value of the error u_{T+1}, computing a forecast interval requires either estimating the distribution of the error term or making some assumption about that distribution.

In practice, it is convenient to assume that u_{T+1} is normally distributed. Under the assumption of stationarity, the forecast error is the sum of u_{T+1} and a term reflecting the estimation error of the regression coefficients. In large samples, this second term is approximately normally distributed (by the central limit theorem) and is uncorrelated with u_{T+1}. Thus, if u_{T+1} is normally distributed, the forecast error is approximately normally distributed and has a variance equal to the MSFE (Exercise 15.12).

The River of Blood

As part of its efforts to inform the public about monetary policy decisions, the Bank of England regularly publishes forecasts of inflation. These forecasts combine output from econometric models maintained by professional econometricians at the bank with the expert judgment of the members of the bank's senior staff and Monetary Policy Committee. The forecasts are presented as a set of forecast intervals designed to reflect what these economists consider to be the range of probable paths that inflation might take. In its *Inflation Report*, the bank prints these ranges in red, with the darkest red reserved for the central band. Although the bank prosaically refers to this as the "fan chart," the press has called these spreading shades of red the "river of blood."

The river of blood for February 2017 is shown in Figure 15.4. (In this figure, the blood is blue, not red, so you will need to use your imagination.) This chart shows that, as of February 2017, the bank's economists expected the rate of inflation to rise from below its 2.0% target in early 2017 to 2.7% in the first quarter of 2018. The economists cited an expected strengthening of demand and a depreciation in the British pound as reasons for the increase in the inflation rate. As it turned out, inflation rose over the next year by more than they had forecasted, to 3.0% in early 2018.

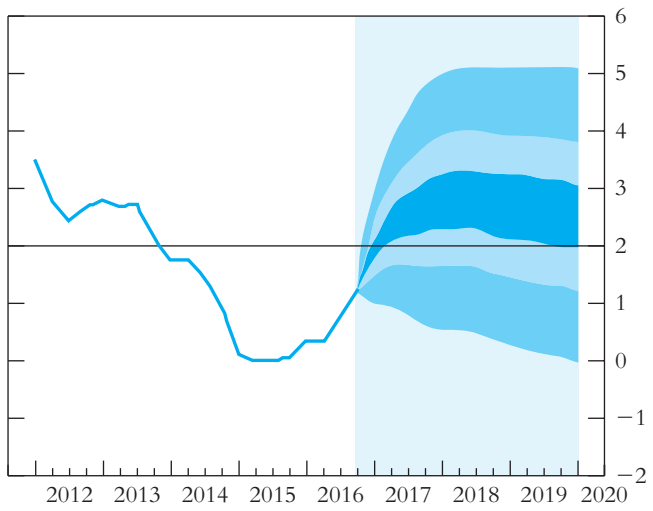
The Bank of England has been a pioneer in the movement toward greater openness by central banks, and other central banks now also publish inflation forecasts. The decisions made by monetary policy makers are difficult ones and affect the lives—and wallets—of many of their fellow citizens. In a democracy in the information age, reasoned the economists at the Bank of England, it is particularly important for citizens to understand the bank's economic outlook and the reasoning behind its difficult decisions.

To see the river of blood in its original red hue, visit the Bank of England's website, at **http://www.bankofengland.co.uk.** To learn more about the performance of the Bank of England inflation forecasts, see Clements (2004).

FIGURE 15.4 The River of Blood

The Bank of England's fan chart for February 2017 shows forecast ranges for inflation.

Source: Reprinted with permission from the Bank of England.



The second two of the estimators of the MSFE, $\widehat{MSE}_{FPE}$ and $\widehat{MSE}_{POOS}$, incorporate estimation error, and either one can be used to construct forecast intervals. That is, if u_{T+1} is normally distributed, a 95% forecast interval is given by $\hat{Y}_{T+1|T} \pm 1.96\,\widehat{RMSE}$, where $\widehat{RMSE}$ is either $\widehat{RMSE}_{FPE}$ in Equation (15.21) or $\widehat{RMSE}_{POOS}$ in Equation (15.22).

This discussion has focused on the case that u_t is homoskedastic. If instead it is heteroskedastic, then one needs to develop a model of the heteroskedasticity so that the term σ_u^2 in Equation (15.19) can be estimated given the most recent values of Y and X. Methods for modeling this conditional heteroskedasticity are presented in Chapter 17.

Fan charts. To convey the full range of uncertainty about future values of a variable, professional forecasters sometimes report multiple forecast intervals. Taken together, multiple forecast intervals summarize the full distribution of future values of the variable. A forecast of the distribution of future values of a variable provides a great deal more information to consumers of forecasts than does a forecast of just its mean.

Forecast distributions are frequently conveyed graphically in what is known as a **fan chart**. Fan charts portray the distribution at a future date by shaded overlaid forecast intervals, connected over an expanding forecast horizon. The Bank of England was one of the early users of fan charts as a way to convey forecast paths and uncertainty to the public and to monetary policy makers (see the box "The River of Blood").

15.6 Estimating the Lag Length Using Information Criteria

The estimated GDP growth regressions in Sections 15.3 and 15.4 have either one or two lags of the predictors. Why not more lags? How many lags should you include in a time series regression? This section discusses statistical methods for choosing the number of lags, first in an autoregression and then in a time series regression model with multiple predictors.

Determining the Order of an Autoregression

In practice, choosing the order p of an autoregression requires balancing the marginal benefit of including more lags against the marginal cost of additional estimation error. On the one hand, if the order of an estimated autoregression is too low, you will omit potentially valuable information contained in the more distant lagged values. On the other hand, if it is too high, you will be estimating more coefficients than necessary, which in turn introduces additional estimation error into your forecasts.

The F-statistic approach. One approach to choosing p is to start with a model with many lags and to perform hypothesis tests on the final lag. For example, you might

start by estimating an AR(6) and test whether the coefficient on the sixth lag is significant at the 5% level; if not, drop it and estimate an AR(5), test the coefficient on the fifth lag, and so forth. The drawback to this method is that it will tend to produce large models: Even if the true AR order is five, so the sixth coefficient is 0, a 5% test using the *t*-statistic will incorrectly reject this null hypothesis 5% of the time just by chance. Thus, if the true value of p is five, this method will estimate p to be six 5% of the time.

The BIC. One way around this problem is to estimate p by minimizing an information criterion. One such information criterion is the **Bayes information criterion (BIC)**, also called the *Schwarz information criterion (SIC)*, which is

$$\text{BIC}(p) = \ln\left[\frac{SSR(p)}{T}\right] + (p+1)\frac{\ln(T)}{T}, \tag{15.23}$$

where $SSR(p)$ is the sum of squared residuals of the estimated AR(p). The BIC estimator of p, $\hat{p}$, is the value that minimizes BIC(p) among the possible choices $p = 0, 1, \ldots, p_{max}$, where p_{max} is the largest value of p considered and $p = 0$ corresponds to the model that contains only an intercept.

The formula for the BIC might look a bit mysterious at first, but it has an intuitive interpretation. Consider the first term in Equation (15.23). Because the regression coefficients are estimated by OLS, the sum of squared residuals necessarily decreases (or at least does not increase) when you add a lag. In contrast, the second term is the number of estimated regression coefficients (the number of lags, p, plus one for the intercept) times the factor $\ln(T)/T$. This second term increases when you add a lag and thus provides a penalty for including another lag. The BIC trades off these two forces so that the number of lags that minimizes the BIC is a consistent estimator of the true lag length. Appendix 15.5 provides the mathematics of this argument.

As an example, consider estimating the AR order for an autoregression of the growth rate of GDP. The various steps in the calculation of the BIC are carried out in Table 15.3 for autoregressions of maximum order six ($p_{max} = 6$). For example, for the AR(1) model in Equation (15.8), $[SSR(1)/T] = 9.247$, so $\ln[SSR(1)/T] = 2.224$. Because $T = 223$ (1962:Q1–2017:Q3), $\ln(T)/T = 0.024$, and $(p+1)\ln(T)/T = 2 \times 0.024 = 0.048$. Thus $\text{BIC}(1) = 2.224 + 0.048 = 2.273$.

The BIC is smallest when $p = 2$ in Table 15.3. Thus the BIC estimate of the lag length is 2. As can be seen in Table 15.3, as the number of lags increases, the R^2 increases, and the SSR decreases. The increase in the R^2 is large from zero to one lag, smaller for one to two lags, and smaller yet for other lags. The BIC helps decide precisely how large the increase in the R^2 must be to justify including the additional lag.

The AIC. Another information criterion is the **Akaike information criterion (AIC)**:

$$\text{AIC}(p) = \ln\left[\frac{SSR(p)}{T}\right] + (p+1)\frac{2}{T}. \tag{15.24}$$

TABLE 15.3 The Bayes Information Criterion (BIC) and the R^2 for Autoregressive Models of U.S. GDP Growth Rates, 1962:Q1–2017:Q3

p	$SSR(p)/T$	$\ln[SSR(p)/T]$	$(p+1)\ln(T)/T$	BIC(p)	R^2
0	10.477	2.349	0.024	2.373	0.000
1	9.247	2.224	0.048	2.273	0.117
2	8.954	2.192	0.073	2.265	0.145
3	8.954	2.192	0.097	2.289	0.145
4	8.920	2.188	0.121	2.310	0.149
5	8.788	2.173	0.145	2.319	0.161
6	8.779	2.172	0.170	2.342	0.162

The difference between the AIC and the BIC is that the term $\ln(T)$ in the BIC is replaced by 2 in the AIC, so the second term in the AIC is smaller. For example, for the 223 observations used to estimate the GDP autoregressions, $\ln(T) = \ln(223) = 5.41$, so the second term for the BIC is more than twice as large as the term in the AIC. Thus a smaller decrease in the SSR is needed in the AIC to justify including another lag.

The AIC has an appealing motivation: In large samples, it corresponds to choosing p to minimize the MSFE as estimated by the final prediction error; that is, it minimizes $\widehat{MSFE}_{FPE}$ in Equation (15.21).[5] However, as a matter of theory, the second term in the AIC is not large enough to ensure that the correct lag length is chosen, even in large samples, so the AIC estimator of p is not consistent. As is discussed in Appendix 15.5, in large samples the AIC will overestimate p with nonzero probability.

Both the AIC and the BIC are widely used in practice. If you are concerned that the BIC might yield a model with too few lags, the AIC provides a reasonable alternative.[6]

---

[5]Start with Equation (15.21) to write $\widehat{MSFE}_{FPE} = \left[\frac{T+p+1}{T-(p+1)}\right]\frac{SSR}{T} = \left[\frac{1+(p+1)/T}{1-(p+1)/T}\right]\frac{SSR}{T}$. Taking logarithms of the final expression yields $\ln(\widehat{MSFE}_{FPE}) = \ln\left(1+\frac{p+1}{T}\right) - \ln\left(1-\frac{p+1}{T}\right) + \ln\left(\frac{SSR}{T}\right) \cong 2\left(\frac{p+1}{T}\right) + \ln\left(\frac{SSR}{T}\right)$, where the final expression uses the approximation that $\ln(1+x) \cong x$ when x is small [Equation (8.16)]. The final expression is the AIC in Equation (15.24). The approximation $\widehat{MSFE}_{FPE} \approx$ AIC holds when $(p+1)/T$ is small.

[6]The BIC and the AIC tackle the same problem—restricting the number of parameters to estimate—as the penalized least squares methods of ridge regression and the LASSO discussed in Sections 14.3 and 14.4. One difference between the variable selection problem discussed in Chapter 14 and the lag selection problem discussed here is that, in the general prediction problem with cross-sectional data, there is no natural ordering of the potential regressors. In contrast, in the lag selection problem, it is natural to think that the first lag will be the most useful predictor, followed by the second lag, and so forth, so the predictors have a natural ordering. The AIC and the BIC exploit that natural ordering.

A note on calculating information criteria. For the AIC and BIC to decide between competing regressions with different numbers of lags, those regressions must be estimated using the same observations. For example, in Table 15.3 all the regressions were estimated using data from 1962:Q1 to 2017:Q3, for a total of 223 observations. Because the autoregressions involve lags of the GDP growth rate, this means that the regression uses earlier values of GDP growth (values before 1962:Q1) for initial observations. Said differently, each of the regressions examined in Table 15.3 includes observations on $GDPGR_t$, $GDPGR_{t-1}, \ldots, GDPGR_{t-p}$ for $t =$ 1962:Q1, . . . , 2017:Q3 corresponding to 223 observations on the dependent variable and regressors, so $T = 223$ in Equations (15.23) and (15.24).

Lag Length Selection in Time Series Regression with Multiple Predictors

The trade-off involved with lag length choice in the general time series regression model with multiple predictors [Equation (15.18)] is similar to that in an autoregression: Using too few lags can decrease forecast accuracy because valuable information is lost, but adding lags increases estimation error. The choice of lags must balance the benefit of using additional information against the cost of estimating the additional coefficients.

The F-statistic approach. As in the univariate autoregression, one way to determine the number of lags is to use F-statistics to test joint hypotheses that sets of coefficients are equal to 0. For example, in the discussion of Equation (15.15), we tested the hypothesis that the coefficient on the second lag of the term spread was equal to 0 against the alternative that it was nonzero; this hypothesis was not rejected at the 10% significance level, suggesting that the second lag of the term spread could be dropped from the regression. If the number of models being compared is small, then this F-statistic method is easy to use. In general, however, the F-statistic method can produce models that are large and thus have considerable estimation error.

Information criteria. As in an autoregression, the BIC and the AIC can be used to estimate the number of lags and variables in the time series regression model with multiple predictors. If the regression model has K coefficients (including the intercept), the BIC is

$$\text{BIC}(K) = \ln\left[\frac{SSR(K)}{T}\right] + K\frac{\ln(T)}{T}. \tag{15.25}$$

The AIC is defined in the same way, but with 2 replacing $\ln(T)$ in Equation (15.25). For each candidate model, the BIC (or the AIC) can be evaluated, and the model with the lowest value of the BIC (or the AIC) is the preferred model, based on the information criterion.

There are two important practical considerations when using an information criterion to estimate the lag lengths. First, as is the case for the autoregression, all the candidate models must be estimated over the same sample; in the notation of Equation (15.25), the number of observations used to estimate the model, T, must be the same for all models. Second, when there are multiple predictors, this approach is computationally demanding because it requires computing many different models (many combinations of the lag parameters). In practice, a convenient shortcut is to require all the regressors to have the same number of lags—that is, to require that $p = q_1 = \cdots = q_k$, so that only $p_{max} + 1$ models need to be compared (corresponding to $p = 0, 1, \ldots, p_{max}$. Applying this lag-length selection method to the ADL for GDP growth and the term spread results in the ADL(2, 2) model in Equation (15.15).

15.7 Nonstationarity I: Trends

In Key Concept 15.6, it was assumed that the dependent variable and the regressors are stationary. If this is not the case—that is, if the dependent variable and/or the regressors are nonstationary—then conventional hypothesis tests, confidence intervals, and forecasts can be unreliable. The precise problem created by nonstationarity, and the solution to that problem, depends on the nature of that nonstationarity.

In this and the next section, we examine two types of nonstationarity that are frequently encountered in economic time series: trends and breaks. In each section, we first describe the nature of the nonstationarity and then discuss the consequences for time series regression if this type of nonstationarity is present but ignored. We next present tests for nonstationarity and discuss remedies for, or solutions to, the problems caused by that particular type of nonstationarity. We begin by discussing trends.

What Is a Trend?

A **trend** is a persistent long-term movement of a variable over time. A time series variable fluctuates around its trend.

Inspection of Figure 15.1a suggests that the logarithm of U.S. GDP has a clear upwardly increasing trend. The series in Figures 15.2a, 15.2b, and 15.2c also have trends, but their trends are quite different. The trend in the unemployment rate is increasing from the late 1960s through the early 1980s, then decreasing until the early 2000s, and then increasing again. The $/£ exchange rate clearly had a prolonged downward trend after the collapse of the fixed exchange rate system in 1972. The logarithm of the Japan Industrial Production Index has a complicated trend: fast growth at first, then moderate growth, and finally no growth.

Deterministic and stochastic trends. There are two types of trends in time series data: deterministic and stochastic. A **deterministic trend** is a nonrandom function of

time. For example, a deterministic trend might be linear in time; if the logarithm of U.S. GDP had a deterministic linear trend, so that it increased by 0.75 percentage points per quarter, this trend could be written as $0.75t$, where t is measured in quarters. In contrast, a **stochastic trend** is random and varies over time. For example, a stochastic trend might exhibit a prolonged period of increase followed by a prolonged period of decrease, like the unemployment rate trend in Figure 15.2a. But stochastic trends can be more subtle. For example, if you look carefully at Figure 15.1a, you will notice that the trend growth rate of GDP is not constant; for example, GDP grew faster in the 1960s than in the 1970s (the plot is steeper in the 1960s than in the 1970s), and it grew faster in the 1990s than in the 2000s.

Like many econometricians, we think it is more appropriate to model economic time series as having stochastic rather than deterministic trends. It is hard to reconcile the predictability implied by a deterministic trend with the complications and surprises faced year after year by workers, businesses, and governments. For example, although the U.S. unemployment rate rose through the 1970s, it was neither destined to rise forever nor destined to fall again. Rather, the slow rise of unemployment rates is now understood to have occurred because of a combination of demographic changes (including an influx of younger workers), bad luck (such as oil price shocks and a productivity slowdown), and monetary policy mistakes. Similarly, the \$/£ exchange rate trended down from 1972 to 1985 and subsequently drifted up, but these movements, too, were the consequences of complex economic forces; because these forces change unpredictably, these trends are usefully thought of as having a large unpredictable, or random, component.

For these reasons, our treatment of trends in economic time series focuses on stochastic rather than deterministic trends, and when we refer to "trends" in time series data, we mean stochastic trends unless we explicitly say otherwise.

The random walk model of a trend. The simplest model of a variable with a stochastic trend is the random walk. A time series Y_t is said to follow a **random walk** if the change in Y_t is i.i.d.—that is, if

$$Y_t = Y_{t-1} + u_t, \tag{15.26}$$

where u_t is i.i.d. We will, however, use the term *random walk* more generally to refer to a time series that follows Equation (15.26), where u_t has conditional mean 0; that is, $E(u_t \mid Y_{t-1}, Y_{t-2}, \ldots) = 0$. Another term for a time series for which $E(\Delta Y_t \mid Y_{t-1}, Y_{t-2}, \ldots) = 0$ is a *martingale*.

The basic idea of a random walk is that the value of the series tomorrow is its value today plus an unpredictable change: Because the path followed by Y_t consists of random "steps" u_t, that path is a "random walk." The conditional mean of Y_t based on data through time $t - 1$ is Y_{t-1}; that is, because $E(u_t \mid Y_{t-1}, Y_{t-2}, \ldots) = 0$, $E(Y_t \mid Y_{t-1}, Y_{t-2}, \ldots) = Y_{t-1}$. In other words, if Y_t follows a random walk, then the best forecast of tomorrow's value is its value today.

If Y_t follows a random walk, its variance increases over time. Because it does not have a constant variance, a random walk is nonstationary (Exercise 15.13).

Some series, such as the logarithm of U.S. GDP in Figure 15.1a, have an obvious upward tendency, in which case the best forecast of the series must include an adjustment for the tendency of the series to increase. This adjustment leads to an extension of the random walk model to include a tendency to move, or drift, in one direction or the other. This extension is referred to as a **random walk with drift**:

$$Y_t = \beta_0 + Y_{t-1} + u_t, \tag{15.27}$$

where $E(u_t | Y_{t-1}, Y_{t-2}, \ldots) = 0$ and β_0 is the drift in the random walk. If β_0 is positive, then Y_t increases on average. In the random walk with drift model, the best forecast of the series tomorrow is the value of the series today plus the drift β_0.

The random walk model (with drift, as appropriate) is simple yet versatile, and it is the primary model for trends used in this book.

Stochastic trends, autoregressive models, and a unit root. The random walk model is a special case of the AR(1) model [Equation (15.8)] in which $\beta_1 = 1$. In other words, if Y_t follows an AR(1) with $\beta_1 = 1$, then Y_t contains a stochastic trend and is nonstationary. If, however, $|\beta_1| < 1$ and u_t is stationary, then the joint distribution of Y_t and its lags does not depend on t (a result shown in Appendix 15.2), so Y_t is stationary.

The analogous condition for an AR(p) to be stationary is more complicated than the condition $|\beta_1| < 1$ for an AR(1). Its formal statement involves the roots of the polynomial, $1 - \beta_1 z - \beta_2 z^2 - \beta_3 z^3 - \cdots - \beta_p z^p$. (The roots of this polynomial are the values of z that satisfy $1 - \beta_1 z - \beta_2 z^2 - \beta_3 z^3 - \cdots - \beta_p z^p = 0$.) For an AR($p$) to be stationary, the roots of this polynomial must all be greater than 1 in absolute value. In the special case of an AR(1), the root is the value of z that solves $1 - \beta_1 z = 0$, so its root is $z = 1/\beta_1$. Thus the statement that the root must be greater than 1 in absolute value is equivalent to $|\beta_1| < 1$.

If an AR(p) has a root that equals 1, the series is said to have a *unit autoregressive root* or, more simply, a **unit root**. If Y_t has a unit root, then it contains a stochastic trend. If Y_t is stationary (and thus does not have a unit root), it does not contain a stochastic trend. For this reason, we will use the terms *stochastic trend* and *unit root* interchangeably.

Problems Caused by Stochastic Trends

If a regressor has a stochastic trend (that is, has a unit root), then inferences made using the OLS estimator of the autoregressive coefficient can be misleading. Moreover, two series that are independent but have stochastic trends will, with high probability, misleadingly appear to be related, a situation known as spurious regression.

Downward bias and nonnormal distributions of the OLS estimator and t-statistic. If a regressor has a stochastic trend, then its usual OLS t-statistic can have a nonnormal distribution under the null hypothesis, even in large samples, and the estimate of the autoregressive coefficient is biased toward 0. This nonnormal distribution means that conventional confidence intervals are not valid and hypothesis tests cannot be conducted as usual.

The downward bias of the OLS estimator poses a problem for forecasts. Recall that the oracle forecast is the conditional mean. If the coefficient in an AR(1) model of the conditional mean is 1 (a unit root), then the OLS estimator will tend to take on a value less than 1, and its sampling distribution has a mean that is less than 1. In a forecasting application, this can lead to systematic bias in the forecast. Moreover, because the distribution of the t-statistic testing that coefficient is not normal, even in large samples, standard inferences based on that t-statistic will not detect this mistake of downward-biased forecasts. Fortunately, as is discussed later in this section, there are ways to detect whether a series has a unit root and thus to avoid these problems.

Spurious regression. Stochastic trends can lead two time series to appear related when they are not, a problem called **spurious regression**.

For example, the U.S. unemployment rate was steadily rising from the mid-1960s through the early 1980s, and at the same time, Japanese industrial production (plotted in logarithms in Figure 15.2c) was steadily rising. These two trends conspire to produce a regression that appears to be "significant" using conventional measures. Estimated by OLS using data from 1962 through 1985, this regression is

$$\begin{aligned}\widehat{U.\,S.\,Unemployment\,Rate}_t &= -2.37 + 2.22 \times \ln(Japanese\,IP_t),\ \overline{R}^2 = 0.34.\\ &\qquad\quad (1.19)\ \ (0.32)\end{aligned} \tag{15.28}$$

The t-statistic on the slope coefficient is 7, which by usual standards indicates a strong positive relationship between the two series, and the $\overline{R}^2$ is moderately high. However, running this regression using data from 1986 through 2017 yields

$$\begin{aligned}\widehat{U.\,S.\,Unemployment\,Rate}_t &= 42.37 - 7.92 \times \ln(Japanese\,IP_t),\ \overline{R}^2 = 0.14.\\ &\qquad\quad (7.74)\ \ (1.69)\end{aligned} \tag{15.29}$$

The regressions in Equations (15.28) and (15.29) could hardly be more different. Interpreted literally, Equation (15.28) indicates a strong positive relationship, while Equation (15.29) indicates a negative relationship.

The source of these conflicting results is that both series have stochastic trends. These trends happened to align from 1962 through 1985 but were reversed from 1986 through 2017. There is, in fact, no compelling economic or political reason to think that the trends in these two series are related. In short, these regressions are spurious.

The regressions in Equations (15.28) and (15.29) illustrate empirically the theoretical point that OLS can be misleading when the series contain stochastic trends. (See Exercise 15.6 for a computer simulation that demonstrates this result.)

One special case in which certain regression-based methods *are* reliable is when the trend component of the two series is the same—that is, when the series contain a *common* stochastic trend; in such a case, the series are said to be cointegrated. Econometric methods for detecting and analyzing cointegrated economic time series are discussed in Chapter 17.

Detecting Stochastic Trends: Testing for a Unit AR Root

The starting point for detecting a trend in a time series is inspecting its time series plot. If the series looks like it might have a trend, the hypothesis that it has a stochastic trend can be tested using a Dickey–Fuller test.

The Dickey–Fuller test in the AR(1) model. The random walk in Equation (15.27) is a special case of the AR(1) model with $\beta_1 = 1$. Thus, when Y_t follows an AR(1), the hypothesis that Y_t has a stochastic trend corresponds to

$$H_0: \beta_1 = 1 \text{ vs. } H_1: \beta_1 < 1, \text{ where } Y_t = \beta_0 + \beta_1 Y_{t-1} + u_t. \tag{15.30}$$

The null hypothesis in Equation (15.30) is that the AR(1) has a unit root, and the one-sided alternative is that it is stationary.

This test is most easily implemented by estimating a modified version of Equation (15.30), obtained by subtracting Y_{t-1} from both sides. Let $\delta = \beta_1 - 1$; then Equation (15.30) becomes

$$H_0: \delta = 0 \text{ vs. } H_1: \delta < 0, \text{ where } \Delta Y_t = \beta_0 + \delta Y_{t-1} + u_t. \tag{15.31}$$

The OLS t-statistic testing $\delta = 0$ in Equation (15.31) is called the **Dickey–Fuller statistic** [Dickey and Fuller (1979)]. The Dickey–Fuller statistic is computed using nonrobust standard errors—that is, the homoskedasticity-only standard errors, presented in Appendix 5.1.[7]

Critical values for the ADF statistic. Under the null hypothesis of a unit root, the Dickey–Fuller statistic does not have a normal distribution, even in large samples. Because its distribution is nonnormal, a different set of critical values is required.

The critical values for the ADF test of the null and alternative hypotheses in Equation (15.31) are given in the first row of Table 15.4. Because the alternative hypothesis of stationarity implies that $\delta < 0$ in Equation (15.31), the ADF test is one-sided. For example, if the regression does not include a time trend, then the hypothesis of a unit root is rejected at the 5% significance level if the ADF statistic is less than -2.86.

[7]Under the null hypothesis of a unit root, the usual nonrobust standard errors produce a t-statistic that is, in fact, robust to heteroskedasticity, a surprising and special result.

TABLE 15.4 Large-Sample Critical Values of the Augmented Dickey–Fuller Statistic

Deterministic Regressors	10%	5%	1%
Intercept only	−2.57	−2.86	−3.43
Intercept and time trend	−3.12	−3.41	−3.96

The critical values in Table 15.4 are substantially larger (more negative) than the one-sided critical values of −1.28 (at the 10% level) and –1.64 (at the 5% level) from the standard normal distribution. The nonstandard distribution of the ADF statistic is an example of how OLS t-statistics for regressors with stochastic trends can have nonnormal distributions.

The Dickey–Fuller test in the AR(p) model. The Dickey–Fuller statistic in Equation (15.31) applies to first-order autoregression. The extension of the Dickey–Fuller test to the AR(p) model entails including $p - 1$ lags of ΔY_t as additional regressors, so that Equation (15.31) becomes

$$\Delta Y_t = \beta_0 + \delta Y_{t-1} + \gamma_1 \Delta Y_{t-1} + \gamma_2 \Delta Y_{t-2} + \cdots + \gamma_{p-1} \Delta Y_{t-p+1} + u_t. \quad (15.32)$$

Under the null hypothesis that $\delta = 0$, Y_t has a stochastic trend; under the alternative hypothesis that $\delta < 0$, Y_t is stationary. The t-statistic testing the hypothesis that $\Delta = 0$ in Equation (15.32) is called the **augmented Dickey–Fuller (ADF) statistic**. In general, the lag length p is unknown, but it can be estimated using an information criterion applied to regressions of the form in Equation (15.32) for various values of p. Studies of the ADF statistic suggest that it is better to have too many lags than too few, so it is recommended to use the AIC instead of the BIC to estimate p for the ADF statistic.[8]

Testing against the alternative of stationarity around a linear deterministic time trend. The discussion so far has considered the null hypothesis that a series has a unit root and the alternative hypothesis that it is stationary. This alternative hypothesis of stationarity is appropriate for series such as the unemployment rate that do not exhibit growth over the long run. But for series such as U.S. GDP, the alternative of stationarity around a constant mean is inappropriate, and it makes more sense to test for stationarity around a deterministic trend. One specific formulation of this alternative hypothesis is that the trend is a linear function of t. Thus the null hypothesis is that the series has a unit root, and the alternative is that it does not have a unit root but does have a deterministic time trend.

[8]See Stock (1994) and Haldrup and Jansson (2006) for reviews of simulation studies of the finite-sample properties of the Dickey–Fuller and other unit root test statistics.

If the alternative hypothesis is that Y_t is stationary around a deterministic linear time trend, then this trend, t (the observation number), must be added as an additional regressor, in which case the Dickey–Fuller regression becomes

$$\Delta Y_t = \beta_0 + \alpha t + \delta Y_{t-1} + \gamma_1 \Delta Y_{t-1} + \gamma_2 \Delta Y_{t-2} + \cdots + \gamma_{p-1} \Delta Y_{t-p+1} + u_t, \quad (15.33)$$

where α is an unknown coefficient. The ADF statistic now is the OLS t-statistic testing $\delta = 0$ in Equation (15.33), and the one-sided critical values are given in the second row of Table 15.4.[9]

Does U.S. GDP have a stochastic trend? The null hypothesis that the logarithm of U.S. GDP has a stochastic trend can be tested against the alternative that it is stationary by performing the ADF test for a unit autoregressive root. The ADF regression with two lags of $\Delta \ln(GDP_t)$ is

$$\begin{aligned} \widehat{\Delta \ln(GDP_t)} &= \underset{(0.080)}{0.162} + \underset{(0.0001)}{0.0001}t - \underset{(0.010)}{0.019} \ln(GDP_{t-1}) \\ &\quad + \underset{(0.066)}{0.261} \Delta \ln(GDP_{t-1}) + \underset{(0.066)}{0.165} \Delta \ln(GDP_{t-2}). \end{aligned} \quad (15.34)$$

The ADF t-statistic is the t-statistic testing the hypothesis that the coefficient on $\ln(GDP_{t-1})$ is 0; this is, $t = -1.95$. From Table 15.4, the 10% critical value is -3.12. Because the ADF statistic of -1.95 is less negative than -3.12, the test does not reject the null hypothesis at the 10% significance level. Based on the regression in Equation (15.34), we therefore cannot reject (at the 10% significance level) the null hypothesis that the logarithm of GDP has a unit autoregressive root—that is, that $\ln(GDP)$ has a stochastic trend—against the alternative that it is stationary around a linear trend.

Avoiding the Problems Caused by Stochastic Trends

The most reliable way to handle a trend in a series is to transform the series so that it does not have the trend. If the series has a stochastic trend, then its difference does not. For example, if Y_t follows a random walk, so that $Y_t = \beta_0 + Y_{t-1} + u_t$, then $\Delta Y_t = \beta_0 + u_t$ is stationary. Thus using first differences eliminates random walk trends in a series.

In practice, you can rarely be sure whether a series has a stochastic trend. Recall that, as a general point, failure to reject the null hypothesis does not necessarily mean that the null hypothesis is true; rather, it simply means that you have insufficient evidence to conclude that it is false. Thus failure to reject the null hypothesis of a unit root using the ADF statistic does not mean that the series actually *has* a unit root. Even though failure to reject the null hypothesis of a unit root does not mean the series has

[9]For extensions of the Dickey–Fuller test to nonlinear time trends, see Maddala and Kim (1998).

a unit root, it still can be reasonable to approximate the true autoregressive root as equaling 1 and therefore to use differences of the series rather than its levels.[10]

15.8 Nonstationarity II: Breaks

A second type of nonstationarity arises when the population regression function changes over the course of the sample. In economics, this can occur for a variety of reasons, such as changes in economic policy, changes in the structure of the economy, or changes in a specific industry due to an invention. If such changes, or **breaks**, occur, then a regression model that neglects those changes can provide a misleading basis for inference and forecasting. It is therefore important to check a forecasting model for breaks and to adjust the model if one is found.

What Is a Break?

Breaks can arise either from a discrete change in the population regression coefficients at a distinct date or from a gradual evolution of the coefficients over a longer period of time.

One source of discrete breaks in macroeconomic data is a major change in macroeconomic policy. For example, the breakdown of the Bretton Woods system of fixed exchange rates in 1972 produced the break in the time series behavior of the \$/£ exchange rate that is evident in Figure 15.2b. Prior to 1972, the exchange rate was essentially constant, with the exception of a single devaluation in 1968, when the official value of the pound relative to the dollar was decreased. In contrast, since 1972 the exchange rate has fluctuated over a very wide range.

Breaks also can occur more slowly, as the population regression evolves over time. For example, such changes can arise because of slow evolution of economic policy and ongoing changes in the structure of the economy. The methods for detecting breaks described in this section can detect both types of breaks: distinct changes and slow evolution.

Problems caused by breaks. If a break occurs in the population regression function during the sample, then the OLS regression estimates over the full sample will estimate a relationship that holds on average in the sense that the estimate combines the two different periods. Depending on the location and the size of the break, the "average" regression function can be quite different from the true regression function at the end of the sample, and this leads to poor forecasts.

Testing for Breaks

One way to detect breaks is to test for discrete changes, or breaks, in the regression coefficients. How this is done depends on whether the **break date** (the date of the suspected break) is known.

[10]For additional discussion of stochastic trends in economic time series variables and of the problems they pose for regression analysis, see Stock and Watson (1988).

Testing for a break at a known date. In some applications, you might suspect that there is a break at a known date. For example, if you are studying international trade relationships using data from the 1970s, you might hypothesize that there is a break in the population regression function of interest in 1972, when the Bretton Woods system of fixed exchange rates was abandoned in favor of floating exchange rates.

If the date of the hypothesized break in the coefficients is known, then the null hypothesis of no break can be tested using a binary variable interaction regression (Key Concept 8.4). To keep things simple, consider an ADL(1, 1) model, so there is an intercept, a single lag of Y_t, and a single lag of X_t. Let τ denote the hypothesized break date, and let $D_t(\tau)$ be a binary variable that equals 0 before the break date and 1 after, so $D_t(\tau) = 0$ if $t \leq \tau$ and $D_t(\tau) = 1$ if $t > \tau$. Then the regression including the binary break indicator and all interaction terms is

$$\begin{aligned} Y_t = \beta_0 + \beta_1 Y_{t-1} + \delta_1 X_{t-1} + \gamma_0 D_t(\tau) + \gamma_1[D_t(\tau) \times Y_{t-1}] \\ + \gamma_2[D_t(\tau) \times X_{t-1}] + u_t. \end{aligned} \tag{15.35}$$

If there is not a break, then the population regression function is the same over both parts of the sample, so the terms involving the break binary variable $D_t(\tau)$ do not enter Equation (15.35). That is, under the null hypothesis of no break, $\gamma_0 = \gamma_1 = \gamma_2 = 0$. Under the alternative hypothesis that there is a break, the population regression function is different before and after the break date τ, in which case at least one of the γ's is nonzero. Thus the hypothesis of a break can be tested using the F-statistic that tests the hypothesis that $\gamma_0 = \gamma_1 = \gamma_2 = 0$ against the hypothesis that at least one of the γ's is nonzero. This is often called a Chow test for a break at a known break date, named for its inventor, Gregory Chow (1960).

If there are multiple predictors or more lags, then this test can be extended by constructing binary variable interaction variables for all the regressors and testing the hypothesis that all the coefficients on terms involving $D_t(\tau)$ are 0.

This approach can be modified to check for a break in a subset of the coefficients by including only the binary variable interactions for that subset of regressors of interest.

Testing for a break at an unknown date. Often the date of a possible break is unknown or known only within a range. Suppose, for example, you suspect that a break occurred sometime between two dates, τ_0 and τ_1. The Chow test can be extended to handle this situation by testing for breaks at all possible dates τ between τ_0 and τ_1 and then using the largest of the resulting F-statistics to test for a break at an unknown date. This modified Chow test is variously called the **Quandt likelihood ratio (QLR) statistic** (Quandt 1960) (the term we shall use) or, more obscurely, the sup-Wald statistic.

Because the QLR statistic is the largest of many F-statistics, its distribution is not the same as an individual F-statistic. Instead, the critical values for the QLR statistic must be obtained from a special distribution. Like the F-statistic, this distribution depends on the number of restrictions being tested, q—that is, the number of coefficients (including the intercept) that are being allowed to break, or change, under the alternative hypothesis. The distribution of the QLR statistic also depends on

τ_0/T and τ_1/T—that is, on the endpoints, τ_0 and τ_1, of the subsample over which the F-statistics are computed, expressed as a fraction of the total sample size.

For the large-sample approximation to the distribution of the QLR statistic to be a good one, the subsample endpoints, τ_0 and τ_1, cannot be too close to the beginning or the end of the sample. For this reason, in practice the QLR statistic is computed over a "trimmed" range, or subset, of the sample. A common choice is to use 15% trimming—that is, to set $\tau_0 = 0.15T$ and $\tau_1 = 0.85T$ (rounded to the nearest integer). With 15% trimming, the F-statistic is computed for break dates in the central 70% of the sample.

The critical values for the QLR statistic, computed with 15% trimming, are given in Table 15.5. Comparing these critical values with those of the $F_{q,\infty}$ distribution (Appendix Table 4) shows that the critical values for the QLR statistics are larger.

TABLE 15.5 Critical Values of the QLR Statistic with 15% Trimming

Number of Restrictions (q)	10%	5%	1%
1	7.12	8.68	12.16
2	5.00	5.86	7.78
3	4.09	4.71	6.02
4	3.59	4.09	5.12
5	3.26	3.66	4.53
6	3.02	3.37	4.12
7	2.84	3.15	3.82
8	2.69	2.98	3.57
9	2.58	2.84	3.38
10	2.48	2.71	3.23
11	2.40	2.62	3.09
12	2.33	2.54	2.97
13	2.27	2.46	2.87
14	2.21	2.40	2.78
15	2.16	2.34	2.71
16	2.12	2.29	2.64
17	2.08	2.25	2.58
18	2.05	2.20	2.53
19	2.01	2.17	2.48
20	1.99	2.13	2.43

Note: These critical values apply when $\tau_0 = 0.15T$ and $\tau_1 = 0.85T$ (rounded to the nearest integer), so the F-statistic is computed for all potential break dates in the central 70% of the sample. The number of restrictions q is the number of restrictions tested by each individual F-statistic. Critical values for other trimming percentages are given in Andrews (2003).

KEY CONCEPT 15.8

The QLR Test for Coefficient Stability

Let $F(\tau)$ denote the F-statistic testing the hypothesis of a break in the regression coefficients at date τ; in the regression in Equation (15.35), for example, this is the F-statistic testing the null hypothesis that $\gamma_0 = \gamma_1 = \gamma_2 = 0$. The QLR (or sup-Wald) test statistic is the largest of the F-statistics in the range $\tau_0 \leq \tau \leq \tau_1$:

$$\text{QLR} = \max[F(\tau_0), F(\tau_0 + 1), \ldots, F(\tau_1)]. \tag{15.36}$$

1. Like the F-statistic, the QLR statistic can be used to test for a break in all or just some of the regression coefficients.
2. In large samples, the distribution of the QLR statistic under the null hypothesis depends on the number of restrictions being tested, q, and on the endpoints τ_0 and τ_1 as a fraction of T. Critical values are given in Table 15.5 for 15% trimming ($\tau_0 = 0.15T$ and $\tau_1 = 0.85T$, rounded to the nearest integer).
3. The QLR test can detect a single discrete break, multiple discrete breaks, and/or slow evolution of the regression function.
4. If there is a distinct break in the regression function, the date at which the largest Chow statistic occurs is an estimator of the break date.

This reflects the fact that the QLR statistic looks at the largest of many individual F-statistics. By examining F-statistics at many possible break dates, the QLR statistic has many opportunities to reject the null hypothesis, leading to QLR critical values that are larger than the individual F-statistic critical values.

The QLR test can be used to test for a break in only some of the regression coefficients by using interactions between the date binary indicators and only the variables in question, and then computing the largest of the resulting F-statistics. The critical values for this version of the QLR test are also taken from Table 15.5, where the number of restrictions (q) is the number of restrictions tested.

If there is a discrete break at a date within the range tested, the date at which the constituent F-statistic is at its maximum, $\hat{\tau}$, is an estimate of the break date τ.

The QLR statistic also rejects the null hypothesis with high probability in large samples when there are multiple discrete breaks or when the break comes in the form of a slow evolution of the regression function. This means that the QLR statistic detects forms of instability other than a single discrete break. As a result, if the QLR statistic rejects the null hypothesis, it can mean that there is a single discrete break, that there are multiple discrete breaks, or that there is slow evolution of the regression function.

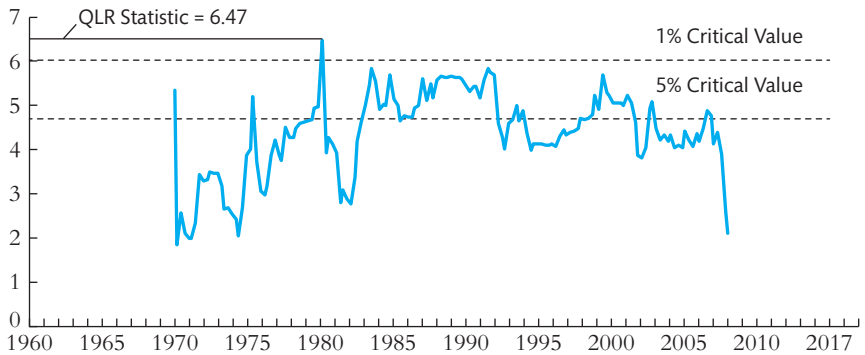
The QLR statistic is summarized in Key Concept 15.8.

Warning: You probably don't know the break date even if you think you do. Sometimes an expert might believe that he or she knows the date of a possible break, so that the Chow test can be used instead of the QLR test. But if this knowledge is based on the expert's knowledge of the series being analyzed, then, in fact, this date was estimated using the data, albeit in an informal way. Preliminary estimation of the break date means that the usual F critical values cannot be used for the Chow test for a break at that date. Thus it remains appropriate to use the QLR statistic in this circumstance.

Application: Has the predictive power of the term spread been stable? The QLR test provides a way to check whether the GDP–term spread relation has been stable from 1962 to 2017. Specifically, we focus on whether there have been changes in the coefficients on the lagged values of the term spread and the intercept in the ADL(2,2) specification in Equation (15.15), containing two lags each of $GDPGR_t$ and $TSpread_t$.

The Chow F-statistics testing the hypothesis that the intercept and the coefficients on $TSpread_{t-1}$, $TSpread_{t-2}$, and the intercept in Equation (15.15) are constant against the alternative that they break at a given date are plotted in Figure 15.5 for breaks in the central 70% of the sample. For example, the F-statistic testing for a break in 1975:Q1 is 2.07, the value plotted at that date in the figure. Each F-statistic tests three restrictions (no change in the intercept and in the two coefficients on lags of the term spread), so $q = 3$. The largest of these F-statistics is 6.47, which occurs in 1980:Q4; this is the QLR statistic. Comparing 6.47 to the critical values for $q = 3$ in Table 15.5 indicates that the hypothesis that these coefficients are stable is rejected at the 1% significance level. (The 1% critical value is 6.02.) Thus, there is statistically significant evidence that at least one of these coefficients changed over the sample.

FIGURE 15.5 *F*-Statistics Testing for a Break in Equation (15.15) at Different Dates



At a given break date, the F-statistic plotted here tests the null hypothesis of a break in at least one of the coefficients on $TSpread_{t-1}$, $TSpread_{t-2}$, or the intercept in Equation (15.15). For example, the F-statistic testing for a break in 1975:Q1 is 2.07. The QLR statistic, 6.47, is the largest of these F-statistics and exceeds the 1% critical value of 6.02.

Detecting Breaks Using Pseudo Out-of-Sample Forecasts

The ultimate test of a forecasting model is its out-of-sample performance—that is, its forecasting performance in "real time," after the model has been estimated. Pseudo out-of-sample forecasting, introduced in Key Concept 15.7 for the purpose of estimating the MSFE, simulates the real-time performance of a forecasting model and can be used to detect breaks near the end of the sample.

The most direct and often most useful way to do so is via a time series plot of the in-sample predicted values, the pseudo out-of-sample forecasts, and the actual values of the series. A visible deterioration of the forecasts in the pseudo out-of-sample period is a red flag warning of a possible breakdown of the forecasting model. Another check is to compare $\widehat{MSFE}_{POOS}$ with $\widehat{MSFE}_{FPE}$, where $\widehat{MSFE}_{FPE}$ is computed on the same estimation sample as used for $\widehat{MSFE}_{POOS}$ (the first $T - P$ observations). If the series is stationary, these two estimates of the MSFE should be numerically close. A value of $\widehat{MSFE}_{POOS}$ that is much larger than $\widehat{MSFE}_{FPE}$ suggests some violation of stationarity, possibly a breakdown of the forecasting equation.

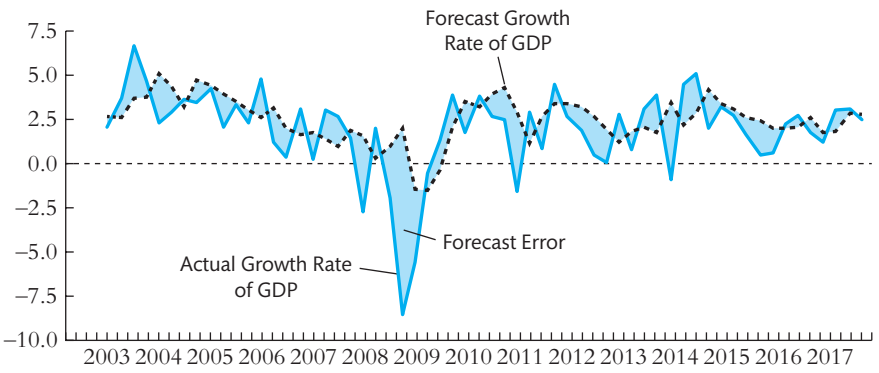
Application: Did the predictive power of the term spread change during the 2000s? Using the QLR statistic, we rejected the null hypothesis that the predictive power of the term spread has been stable against the alternative of a break at the 1% significance level, with a break occurring in the early 1980s. Does the ADL(2, 2) model provide a stable forecasting model subsequent to the 1980:Q4 break?

If the coefficients of the ADL(2, 2) model changed toward the end of the 1981:Q1–2017:Q3 period, then pseudo out-of-sample forecasts computed using an estimation sample starting in 1981:Q1 should deteriorate. The pseudo out-of-sample forecasts of the growth rate of GDP for the period 2003:Q1–2017:Q3, computed using the estimation sample of 1981:Q1–2002:Q4 and the method of Key Concept 5.7, are plotted in Figure 15.6, along with the actual values of the growth rate of GDP. The pseudo out-of-sample forecast errors are the differences between the actual growth rate of GDP and its pseudo out-of-sample forecast—that is, the differences between the two lines in Figure 15.6. For example, in 2006:Q4, the growth rate of GDP was 3.1 percentage points (at an annual rate), but the pseudo out-of-sample forecast of $GDPGR_{2006:Q4}$ was 1.6 percentage points, so the pseudo out-of-sample forecast error was $GDPGR_{2006:Q4} - \widehat{GDPGR}_{2006:Q4|2006:Q3} = 1.5$ percentage points. In other words, a forecaster using the ADL(2, 2) estimated through 2006:Q3 would have forecasted GDP growth of 1.6 percentage points in 2006:Q4, whereas in reality GDP grew by 3.1 percentage points.

How do the mean and standard deviation of the pseudo out-of-sample forecast errors compare with the in-sample fit of the model? If the forecasting model is stable, the pseudo out-of-sample forecast errors should have mean 0. However, over the 2003:Q1–2017:Q4 pseudo out-of-sample forecast period, the average forecast error is –0.57, and the t-statistic testing the hypothesis that the mean forecast error equals 0 is –2.00; thus the hypothesis that the forecasts have mean 0 is rejected

FIGURE 15.6 U.S. GDP Growth Rates and Pseudo Out-of-Sample Forecasts

The pseudo out-of-sample forecasts made using the ADL(2, 2) model of the form in Equation (15.15) generally track the actual growth rate of GDP from 2003 to 2017 but fail to forecast the sharp decline in GDP following the financial crisis of 2008.



at the 5% significance level. That said, $\widehat{RMSFE}_{FPE} = 2.45$ (1981:Q1–2002:Q4) and $\widehat{RMSFE}_{POOS} = 2.29$ (2003:Q1–2017:Q4), indicating a slight improvement of the forecast in the out-of-sample period. Figure 15.6 shows that the pseudo out-of-sample forecasts track actual GDP growth reasonably well except during late 2008 and early 2009, the period of steepest decline of GDP during the financial crisis and its immediate aftermath. Excluding the single quarter 2008:Q4 lowers $\widehat{RMSFE}_{POOS}$ from 2.29 to 1.85.

According to the pseudo out-of-sample forecasting exercise, the performance of the ADL(2, 2) forecasting model during the pseudo out-of-sample period 2003:Q1–2017:Q4 was, with the exception of the sharp decline in GDP in late 2008, better than its performance during the in-sample period of 1981:Q1–2002:Q4.[11]

Avoiding the Problems Caused by Breaks

How best to adjust for a break in the population regression function depends on the source of that break. If a distinct break occurs at a specific date, that break will be detected with high probability by the QLR statistic, and the break date can be estimated. The regression function can then be reestimated using a binary variable indicating the two subsamples associated with this break and including interactions with the other regressors as appropriate. If all the coefficients break, then this simplifies to reestimating the regression using the post-break data. If there is, in fact, a distinct break, then subsequent inference on the regression coefficients can proceed as usual—for example, using normal critical values for hypothesis tests based on

[11]The ADL(2, 2) was not alone in failing to forecast GDP growth in 2008:Q4. Researchers at the Federal Reserve Bank of Philadelphia surveyed 47 professional forecasters in the third quarter of 2008 and asked for their forecasts of the growth rate of GDP in the fourth quarter. The median of the 47 forecasts was 0.7%, lower than the ADL(2, 2) forecast of 2.0%. The actual growth rate of GDP in 2008:Q4 was –8.5%.

t-statistics. In addition, forecasts can be produced using the regression function estimated using the post-break model.

If the break is not distinct but rather arises from a slow, ongoing change in the parameters, the remedy is more difficult and goes beyond the scope of this book.[12]

15.9 Conclusion

In time series data, a variable generally is correlated from one observation, or date, to the next. A consequence of this correlation is that linear regression can be used to forecast future values of a time series based on its current and past values. The starting point for time series regression is an autoregression, in which the regressors are lagged values of the dependent variable. If additional predictors are available, then their lags can be added to the regression. This chapter has described methods for specifying and estimating forecasting regressions, for selecting among competing forecasting regressions, for handling trends in the data, and for assessing the stability of forecasting models.

The time series regressions in this chapter were developed for forecasting, and in general, the coefficients do not have a causal interpretation. In some applications, however, the task is not to develop a forecasting model but rather to estimate causal relationships among time series variables—that is, to estimate the *dynamic* causal effect on Y *over time* of a change in X. Under the right conditions, the methods of this chapter, or closely related methods, can be used to estimate dynamic causal effects, and that is the topic of the next chapter.

Summary

1. Regression models used for forecasting need not have a causal interpretation.
2. A time series variable generally is correlated with one or more of its lagged values; that is, it is serially correlated.
3. The accuracy of a forecast is measured by its mean squared forecast error.
4. An autoregression of order p is a linear multiple regression model in which the regressors are the first p lags of the dependent variable. The coefficients of an AR(p) can be estimated by OLS, and the estimated regression function can be used for forecasting. The lag order p can be estimated using an information criterion such as the BIC or the AIC.
5. Adding other variables and their lags to an autoregression can improve forecasting performance. Under the least squares assumptions for prediction with time series regression (Key Concept 15.6), the OLS estimators have normal distributions in large samples, and statistical inference proceeds the same way as for cross-sectional data.

[12]For additional discussion of estimation and testing in the presence of discrete breaks, see Hansen (2001). For an advanced discussion of estimation and forecasting when there are slowly evolving coefficients, see Hamilton (1994, Chapter 13).

6. Forecast intervals quantify forecast uncertainty. If the errors are normally distributed, an approximate 68% forecast interval can be constructed as the forecast plus or minus an estimate of the root mean squared forecast error.
7. A series that contains a stochastic trend is nonstationary. A random walk stochastic trend can be detected using the ADF statistic and can be eliminated by using the first difference of the series.
8. If the population regression function changes over time, then OLS estimates neglecting this instability produce unreliable forecasts. The QLR statistic can be used to test for a break, and if a discrete break is found, the regression function can be reestimated allowing for the break.
9. Pseudo out-of-sample forecasts can be used to estimate the root mean squared forecast error, to compare different forecasting models, and to assess model stability toward the end of the sample.

Key Terms

gross domestic product (GDP) (555)
first difference (556)
first lag (556)
j^{th} lag (556)
autocorrelation (558)
serial correlation (558)
autocorrelation coefficient (558)
j^{th} autocovariance (559)
stationarity (561)
nonstationarity (562)
one-step ahead forecast (562)
multi-step ahead forecast (562)
forecast error (562)
mean squared forecast error (MSFE) (563)
root mean squared forecast error (RMSFE) (563)
oracle forecast (565)
autoregression (565)
first-order autoregression (565)
p^{th}-order autoregressive [AR(p)] model (567)
term spread (569)
autoregressive distributed lag (ADL) model (570)
ADL(p, q) (571)
weak dependence (572)
final prediction error (FPE) (574)
pseudo out-of-sample forecasting (575)
forecast interval (576)
fan chart (578)
Bayes information criterion (BIC) (579)
Akaike information criterion (AIC) (579)
trend (582)
deterministic trend (582)
stochastic trend (583)
random walk (583)
random walk with drift (584)
unit root (584)
spurious regression (585)
Dickey–Fuller statistic (586)
augmented Dickey–Fuller (ADF) statistic (587)
break (589)
break date (589)
Quandt likelihood ratio (QLR) statistic (590)
lag operator (606)
lag polynomial (606)
autoregressive–moving average (ARMA) model (607)

Review the Concepts

15.1 Look at the four plots in Figure 15.2—U.S. unemployment rate, U.S. dollar/British pound exchange rate, logarithm of Japan index of industrial production, and the percentage change in daily values. Which of these series appears to be nonstationary? Which of them appears to resemble a random walk?

15.2 Many financial economists believe that the random walk model is a good description of the logarithm of stock prices. It implies that the percentage changes in stock prices are unforecastable. A financial analyst claims to have a new model that makes better predictions than the random walk model. Explain how you would examine the analyst's claim that his model is superior.

15.3 A researcher estimates an AR(1) with an intercept and finds that the OLS estimate of β_1 is 0.88, with a standard error of 0.03. Does a 95% confidence interval include $\beta_1 = 1$? Explain.

15.4 Suppose you suspected that the intercept in Equation (15.15) changed in 1992:Q1. How would you modify the equation to incorporate this change? How would you test for a change in the intercept? How would you test for a change in the intercept if you did not know the date of the change?

Exercises

15.1 Consider the AR(1) model $Y_t = \beta_0 + \beta_1 Y_{t-1} + u_t$. Suppose the process is stationary.

a. Show that $E(Y_t) = E(Y_{t-1})$. (*Hint:* Read Key Concept 15.3.)

b. Show that $E(Y_t) = \beta_0/(1 - \beta_1)$.

15.2 The Index of Industrial Production (IP_t) is a monthly time series that measures the quantity of industrial commodities produced in a given month. This problem uses data on this index for the United States. All regressions are estimated over the sample period 1986:M1–2017:M12 (that is, January 1986 through December 2017). Let $Y_t = 1200 \times \ln(IP_t/IP_{t-1})$.

a. A forecaster states that Y_t shows the monthly percentage change in *IP*, measured in percentage points per annum. Is this correct? Why?

b. Suppose she estimates the following AR(4) model for Y_t:

$$\underset{}{\hat{Y}_t} = \underset{(0.488)}{0.749} + \underset{(0.088)}{0.071}Y_{t-1} + \underset{(0.053)}{0.170}Y_{t-2} + \underset{(0.078)}{0.216}Y_{t-3} + \underset{(0.064)}{0.167}Y_{t-4}.$$

Use this AR(4) to forecast the value of Y_t in January 2018, using the following values of *IP* for July 2017 through December 2017:

Date	2017:M7	2017:M8	2017:M9	2017:M10	2017:M11	2017:M12
IP	105.01	104.56	104.82	106.58	106.86	107.30

c. Worried about potential seasonal fluctuations in production, she adds Y_{t-12} to the autoregression. The estimated coefficient on Y_{t-12} is –0.061, with a standard error of 0.043. Is this coefficient statistically significant?

d. Worried about a potential break, she computes a QLR test (with 15% trimming) on the constant and AR coefficients in the AR(4) model. The resulting QLR statistic is 1.80. Is there evidence of a break? Explain.

e. Worried that she might have included too few or too many lags in the model, the forecaster estimates AR(p) models for $p = 0, 1, \ldots, 6$ over the same sample period. The sum of squared residuals from each of these estimated models is shown in the table. Use the BIC to estimate the number of lags that should be included in the autoregression. Do the results differ if you use the AIC?

AR Order	0	1	2	3	4	5	6
SSR	21,045	20,043	18,870	17,838	17,344	17,337	17,306

15.3. Using the same data as in Exercise 15.2, a researcher tests for a stochastic trend in $\ln(IP_t)$, using the following regression:

$$\begin{aligned}\widehat{\Delta\ln(IP_t)} = \underset{(0.013)}{0.026} + \underset{(0.000067)}{0.000097}t - \underset{(0.0037)}{0.0070}\ln(IP_{t-1}) + \underset{(0.050)}{0.068}\Delta\ln(IP_{t-1}) \\ + \underset{(0.049)}{0.169}\Delta\ln(IP_{t-2}) + \underset{(0.050)}{0.219}\Delta\ln(IP_{t-3}) + \underset{(0.051)}{0.173}\Delta\ln(IP_{t-4}),\end{aligned}$$

where the standard errors shown in parentheses are computed using the homoskedasticity-only formula and the regressor t is a linear time trend.

a. Use the ADF statistic to test for a stochastic trend (unit root) in $\ln(IP)$.

b. Do these results support the specification used in Exercise 15.2? Explain.

15.4 The forecaster in Exercise 15.2 augments her AR(4) model for *IP* growth to include four lagged values of ΔR_t, where R_t is the interest rate on three-month U.S. Treasury bills (measured in percentage points at an annual rate).

a. The F-statistic on the four lags of ΔR_t is 3.91. Do interest rates help predict *IP* growth? Explain.

b. The researcher also regresses ΔR_t on a constant, four lags of ΔR_t, and four lags of *IP* growth. The resulting F-statistic on the four lags of *IP* growth is 1.48. Does *IP* growth help to predict interest rates? Explain.

15.5 Prove the following results about conditional means, forecasts, and forecast errors:

a. Let W be a random variable with mean μ_W and variance σ_w^2, and let c be a constant. Show that $E[(W - c)^2] = \sigma_w^2 + (\mu_W - c)^2$.

b. Consider the problem of forecasting Y_t, using data on $Y_{t-1}, Y_{t-2}, \ldots$. Let f_{t-1} denote some forecast of Y_t, where the subscript $t - 1$ on f_{t-1} indicates that the forecast is a function of data through date $t - 1$. Let $E[(Y_t - f_{t-1})^2 \mid Y_{t-1}, Y_{t-2}, \ldots]$ be the conditional mean squared error of the forecast f_{t-1}, conditional on values of Y observed through date $t - 1$. Show that the conditional mean squared forecast error is minimized when $f_{t-1} = Y_{t|t-1}$, where $Y_{t|t-1} = E(Y_t \mid Y_{t-1}, Y_{t-2}, \ldots)$. (*Hint:* Review Appendix 2.2.)

c. Let u_t denote the error in Equation (15.12). Show that $\text{cov}(u_t, u_{t-j}) = 0$ for $j \neq 0$. [*Hint:* Use Equation (2.28).]

15.6 In this exercise, you will conduct a Monte Carlo experiment to study the phenomenon of spurious regression discussed in Section 15.7. In a Monte Carlo study, artificial data are generated using a computer, and then those artificial data are used to calculate the statistics being studied. This makes it possible to compute the distribution of statistics for known models when mathematical expressions for those distributions are complicated (as they are here) or even unknown. In this exercise, you will generate data so that two series, Y_t and X_t, are independently distributed random walks. The specific steps are as follows:

i. Use your computer to generate a sequence of $T = 100$ i.i.d. standard normal random variables. Call these variables $e_1, e_2, \ldots, e_{100}$. Set $Y_1 = e_1$ and $Y_t = Y_{t-1} + e_t$ for $t = 2, 3, \ldots, 100$.

ii. Use your computer to generate a new sequence, $a_1, a_2, \ldots, a_{100}$, of $T = 100$ i.i.d. standard normal random variables. Set $X_1 = a_1$ and $X_t = X_{t-1} + a_t$ for $t = 2, 3, \ldots, 100$.

iii. Regress Y_t onto a constant and X_t. Compute the OLS estimator, the regression R^2, and the (homoskedasticity-only) t-statistic testing the null hypothesis that β_1 (the coefficient on X_t) is 0.

Use this algorithm to answer the following questions:

a. Run the algorithm (i) through (iii) once. Use the t-statistic from (iii) to test the null hypothesis that $\beta_1 = 0$, using the usual 5% critical value of 1.96. What is the R^2 of your regression?

b. Repeat (a) 1000 times, saving each value of R^2 and the t-statistic. Construct a histogram of the R^2 and t-statistic. What are the 5%, 50%, and 95% percentiles of the distributions of the R^2 and the t-statistic? In what fraction of your 1000 simulated data sets does the t-statistic exceed 1.96 in absolute value?

c. Repeat (b) for different numbers of observations, such as $T = 50$ and $T = 200$. As the sample size increases, does the fraction of times that you reject the null hypothesis approach 5%, as it should because you have generated Y and X to be independently distributed? Does this fraction seem to approach some other limit as T gets large? What is that limit?

15.7 Suppose Y_t follows the stationary AR(1) model $Y_t = 2.5 + 0.7Y_{t-1} + u_t$, where u_t is i.i.d. with $E(u_t) = 0$ and $\text{var}(u_t) = 9$.

a. Compute the mean and variance of Y_t. (*Hint:* See Exercise 15.1.)

b. Compute the first two autocovariances of Y_t. (*Hint:* Read Appendix 15.2.)

c. Compute the first two autocorrelations of Y_t.

d. Suppose $Y_T = 102.3$. Compute $Y_{T+1|T} = E(Y_{T+1} \mid Y_T, Y_{t-1}, \ldots)$.

15.8 Suppose Y_t is the monthly value of the number of new home construction projects started in the United States. Because of the weather, Y_t has a pronounced seasonal pattern; for example, housing starts are low in January and high in June. Let μ_{Jan} denote the average value of housing starts in January, and let $\mu_{Feb}, \mu_{Mar}, \ldots, \mu_{Dec}$ denote the average values in the other months. Show that the values of $\mu_{Jan}, \mu_{Feb}, \ldots, \mu_{Dec}$ can be estimated from the OLS regression $Y_t = \beta_0 + \beta_1 Feb_t + \beta_2 Mar_t + \cdots + \beta_{11} Dec_t + u_t$, where Feb_t is a binary variable equal to 1 if t is February, Mar_t is a binary variable equal to 1 if t is March, and so forth. (*Hint:* Show that $\beta_0 + \beta_2 = \mu_{Mar}$ and so forth.)

15.9 The moving average model of order q has the form

$$Y_t = \beta_0 + e_t + b_1 e_{t-1} + b_2 e_{t-2} + \cdots + b_q e_{t-q},$$

where e_t is a serially uncorrelated random variable with mean 0 and variance σ_e^2.

a. Show that $E(Y_t) = \beta_0$.

b. Show that the variance of Y_t is $\text{var}(Y_t) = \sigma_e^2(1 + b_1^2 + b_2^2 + \cdots + b_q^2)$.

c. Show that $\rho_j = 0$ for $j > q$.

d. Suppose $q = 1$. Derive the autocovariances for Y.

15.10 A researcher carries out a QLR test using 30% trimming, and there are $q = 5$ restrictions. Answer the following questions, using the values in Table 15.5 ("Critical Values of the QLR Statistic with 15% Trimming") and Appendix Table 4 ("Critical Values for the $F_{m,\infty}$ Distribution").

a. The QLR F-statistic is 3.9. Should the researcher reject the null hypothesis at the 5% level?

b. The QLR F-statistic is 1.1. Should the researcher reject the null hypothesis at the 5% level?

c. The QLR F-statistic is 3.6. Should the researcher reject the null hypothesis at the 5% level?

15.11 Suppose ΔY_t follows the AR(1) model $\Delta Y_t = \beta_0 + \beta_1 \Delta Y_{t-1} + u_t$.

a. Show that Y_t follows an AR(2) model.

b. Derive the AR(2) coefficients for Y_t as a function of β_0 and β_1.

15.12 Consider the stationary AR(1) model $Y_t = \beta_0 + \beta_1 Y_{t-1} + u_t$, where u_t is i.i.d. with mean 0 and variance σ_u^2. The model is estimated using data from time periods $t = 1$ through $t = T$, yielding the OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$. You are interested in forecasting the value of Y at time $T + 1$—that is, Y_{T+1}. Denote the forecast by $\hat{Y}_{T+1|T} = \hat{\beta}_0 + \hat{\beta}_1 Y_T$.

a. Show that the forecast error is $Y_{T+1} - \hat{Y}_{T+1|T} = u_{T+1} - [(\hat{\beta}_0 - \beta_0) + (\hat{\beta}_1 - \beta_1)Y_T]$.

b. Show that u_{T+1} is independent of Y_T.

c. Show that u_{T+1} is independent of $\hat{\beta}_0$ and $\hat{\beta}_1$

d. Show that $\text{var}(Y_{T+1|T} - \hat{Y}_{T+1|T}) = \sigma_u^2 + \text{var}[(\hat{\beta}_0 - \beta_0) + (\hat{\beta}_1 - \beta_1)Y_T]$.

15.13 Suppose Y_t follows a random walk, $Y_t = Y_{t-1} + u_t$, for $t = 1, \ldots, T$, where $Y_0 = 0$ and u_t is i.i.d. with mean 0 and variance σ_u^2.

a. Compute the mean and variance of Y_t.

b. Compute the covariance between Y_t and Y_{t-k}.

c. Use the results in (a) and (b) to show that Y_t is nonstationary.

Empirical Exercises

E15.1 On the text website, **http://www.pearsonglobaleditions.com**, 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**. The variable *PCEP* is the price index for personal consumption expenditures from the U.S. National Income and Product Accounts. In this exercise, you will construct forecasting models for the rate of inflation based on *PCEP*. For this analysis, use the sample period 1963:Q1–2017:Q4

(where data before 1963 may be used, as necessary, as initial values for lags in regressions).

a. i. Compute the inflation rate, $Infl = 400 \times [\ln(PCEP_t) - \ln(PCEP_{t-1})]$. What are the units of *Infl*? (Is *Infl* measured in dollars, percentage points, percentage per quarter, percentage per year, or something else? Explain.)

ii. Plot the value of *Infl* from 1963:Q1 through 2017:Q4. Based on the plot, do you think that *Infl* has a stochastic trend? Explain.

b. i. Compute the first four autocorrelations of $\Delta Infl$.

ii. Plot the value of $\Delta Infl$ from 1963:Q1 through 2017:Q4. The plot should look choppy or jagged. Explain why this behavior is consistent with the first autocorrelation that you computed in (i).

c. i. Run an OLS regression of $\Delta Infl_t$ on $\Delta Infl_{t-1}$. Does knowing the change in inflation over the current quarter help predict the change in inflation over the next quarter? Explain.

ii. Estimate an AR(2) model for $\Delta Infl$. Is the AR(2) model better than an AR(1) model? Explain.

iii. Estimate an AR(p) model for $p = 0, \ldots, 8$. What lag length is chosen by the BIC? What lag length is chosen by the AIC?

iv. Use the AR(2) model to predict the change in inflation from 2017:Q4 to 2018:Q1—that is, to predict the value of $\Delta Infl_{2018:Q1}$.

v. Use the AR(2) model to predict the level of the inflation rate in 2018:Q1—that is, $Infl_{2018:Q1}$.

d. i. Use the ADF test for the regression in Equation (15.32) with two lags of $\Delta Infl$ (so that $p = 3$ in Equation (15.32)) to test for a stochastic trend in *Infl*.

ii. Is the ADF test based on Equation (15.32) preferred to the test based on Equation (15.33) for testing for a stochastic trend in *Infl*? Explain.

iii. In (i), you used two lags of $\Delta Infl$. Should you use more lags? Fewer lags? Explain.

iv. Based on the test you carried out in (i), does the AR model for *Infl* contain a unit root? Explain carefully. (*Hint:* Does the failure to reject a null hypothesis mean that the null hypothesis is true?)

e. Use the QLR test with 15% trimming to test the stability of the coefficients in the AR(2) model for $\Delta Infl$. Is the AR(2) model stable? Explain.

f. i. Using the AR(2) model for $\Delta Infl$ with a sample period that begins in 1963:Q1, compute pseudo out-of-sample forecasts for the change in inflation beginning in 2003:Q1 and going through 2017:Q4. (That is, compute $\widehat{\Delta Infl}_{2003:Q1|2002:Q4}$, $\widehat{\Delta Infl}_{2003:Q2|2003:Q1}$, $\ldots$, $\widehat{\Delta Infl}_{2017:Q4|2017:Q3}$.)

ii. Are the pseudo out-of-sample forecasts biased? That is, do the forecast errors have a nonzero mean?

iii. How large is the RMSFE of the pseudo out-of-sample forecasts? Is this consistent with the AR(2) model for $\Delta Infl$ estimated over the 1963:Q1–2002:Q4 sample period?

iv. There is a large outlier in 2008:Q4. Why did inflation fall so much in 2008:Q4? (*Hint:* Collect some data on oil prices. What happened to oil prices during 2008?)

E15.2 Read the box "Can You Beat the Market?" Next go to the course website, where you will find an extended version of the data set described in the box; the data are in the file **Stock_Returns_1931_2002** and are described in the file **Stock_Returns_1931_2002_Description**.

a. Repeat the calculations reported in Table 15.2 using regressions estimated over the 1932:M1–2002:M12 sample period.

b. Construct pseudo out-of-sample forecasts of excess returns over the 1983:M1–2002:M12 period using regressions that begin in 1932:M1.

c. Do the results in (a)–(b) suggest any important changes to the conclusions reached in the box? Explain.

APPENDIX

15.1 Time Series Data Used in Chapter 15

Macroeconomic time series data for the United States are collected and published by various government agencies. The Bureau of Economic Analysis in the Department of Commerce publishes the National Income and Product Accounts, which include the GDP data used in this chapter. The unemployment rate is computed from the Bureau of Labor Statistics' Current Population Survey (see Appendix 3.1). The quarterly data used here were computed by averaging the monthly values. The 10-year Treasury bond rate, 3-month Treasury bill rate, and the \$/£ exchange rate data are quarterly averages of daily rates, as reported by the Federal Reserve System. The Japan Index of Industrial Production is published by the Organisation for Economic Co-operation and Development (OECD). The daily percentage change in the Wilshire 5000 Total Market Index, a stock price index, was computed as $100\Delta\ln(W5000_t)$, where $W5000_t$ is the daily value of the index; because the stock exchange is not open on weekends and holidays, the time period of analysis is a business day. We obtained all these data series from the Federal Reserve Economic Data (FRED) website at the Federal Reserve Bank of St. Louis. There you can find times series data on thousands of macroeconomic variables.

The regressions in Table 15.2 use monthly financial data for the United States. Stock prices (P_t) are measured by the broad-based (NYSE and AMEX), value-weighted index of stock prices constructed by the Center for Research in Security Prices (CRSP). The monthly

percentage excess return is $100 \times \{\ln[(P_t + Div_t)/P_{t-1}] - \ln(TBill_t)\}$, where Div_t is the dividends paid on the stocks in the CRSP index and $TBill_t$ is the gross return (1 plus the interest rate) on a 30-day Treasury bill during month t. We thank Motohiro Yogo for providing both his help and these data.

APPENDIX

15.2 Stationarity in the AR(1) Model

This appendix shows that if $|\beta_1| < 1$ and u_t is stationary, then Y_t is stationary. Recall from Key Concept 15.3 that the time series variable Y_t is stationary if the joint distribution of $(Y_{s+1}, \ldots, Y_{s+T})$ does not depend on s, regardless of the value of T. To streamline the argument, we show this for $T = 2$ under the simplifying assumptions that $\beta_0 = 0$ and $\{u_t\}$ are i.i.d. $N(0, \sigma_u^2)$.

The first step is deriving an expression for Y_t in terms of the u_t's. Because $\beta_0 = 0$, Equation (15.8) implies that $Y_t = \beta_1 Y_{t-1} + u_t$. Substituting $Y_{t-1} = \beta_1 Y_{t-2} + u_{t-1}$ into this expression yields $Y_t = \beta_1(\beta_1 Y_{t-2} + u_{t-1}) + u_t = \beta_1^2 Y_{t-2} + \beta_1 u_{t-1} + u_t$. Continuing this substitution another step yields $Y_t = \beta_1^3 Y_{t-3} + \beta_1^2 u_{t-2} + \beta_1 u_{t-1} + u_t$, and continuing indefinitely yields

$$Y_t = u_t + \beta_1 u_{t-1} + \beta_1^2 u_{t-2} + \beta_1^3 u_{t-3} + \cdots = \sum_{i=0}^{\infty} \beta_1^i u_{t-i}. \tag{15.37}$$

Thus Y_t is a weighted average of current and past u_t's. Because the u_t's are normally distributed and because the weighted average of normal random variables is normal (Section 2.4), Y_{s+1} and Y_{s+2} have a bivariate normal distribution. Recall from Section 2.4 that the bivariate normal distribution is completely determined by the means of the two variables, their variances, and their covariance. Thus, to show that Y_t is stationary, we need to show that the means, variances, and covariance of (Y_{s+1}, Y_{s+2}) do not depend on s. An extension of the argument used below can be used to show that the distribution of $(Y_{s+1}, Y_{s+2}, \ldots, Y_{s+T})$ does not depend on s.

The means and variances of Y_{s+1} and Y_{s+2} can be computed using Equation (15.37), with the subscript $s + 1$ or $s + 2$ replacing t. First, because $E(u_t) = 0$ for all t, $E(Y_t) = E(\sum_{i=0}^{\infty} \beta_1^i u_{t-i}) = \sum_{i=0}^{\infty} \beta_1^i E(u_{t-i}) = 0$, so the means of Y_{s+1} and Y_{s+2} are both 0 and in particular do not depend on s. Second, $\text{var}(Y_t) = \text{var}(\sum_{i=0}^{\infty} \beta_1^i u_{t-i}) = \sum_{i=0}^{\infty} (\beta_1^i)^2 \text{var}(u_{t-i}) = \sigma_u^2 \sum_{i=0}^{\infty} (\beta_1^2)^i = \sigma_u^2/(1 - \beta_1^2)$, where the final equality follows from the fact that if $|a| < 1$, $\sum_{i=0}^{\infty} a^i = 1/(1 - a)$; thus $\text{var}(Y_{s+1}) = \text{var}(Y_{s+2}) = \sigma_u^2/(1 - \beta_1^2)$. Finally, because $Y_{s+2} = \beta_1 Y_{s+1} + u_{s+2}$, $\text{cov}(Y_{s+1}, Y_{s+2}) = E(Y_{s+1}, Y_{s+2}) = E[Y_{s+1}(\beta_1 Y_{s+1} + u_{s+2})] = \beta_1 \text{var}(Y_{s+1}) + \text{cov}(Y_{s+1}, u_{s+2}) = \beta_1 \text{var}(Y_{s+1}) = \beta_1 \sigma_u^2/(1 - \beta_1^2)$.

The covariance does not depend on s, so Y_{s+1} and Y_{s+2} have a joint probability distribution that does not depend on s; that is, their joint distribution is stationary. If $|\beta_1| \geq 1$, this calculation breaks down because the infinite sum in Equation (15.37) does not converge, and the variance of Y_t is infinite. Thus Y_t is stationary if $|\beta_1| < 1$ but not if $|\beta_1| \geq 1$.

The preceding argument was made under the assumptions that $\beta_0 = 0$ and u_t is normally distributed. If $\beta_0 \neq 0$, the argument is similar except that the means of Y_{s+1} and Y_{s+2} are $\beta_0/(1 - \beta_1)$ and Equation (15.37) must be modified for this nonzero mean. The assumption that u_t is i.i.d. normal can be replaced with the assumption that u_t is stationary with a finite variance because, by Equation (15.37), Y_t can still be expressed as a function of current and past u_t's, so the distribution of Y_t is stationary as long as the distribution of u_t is stationary and the infinite sum expression in Equation (15.37) is meaningful in the sense that it converges, which requires that $|\beta_1| < 1$.

APPENDIX

15.3 Lag Operator Notation

The notation in this and the next two chapters is streamlined considerably by adopting what is known as lag operator notation. Let L denote the **lag operator**, which has the property that it transforms a variable into its lag. That is, the lag operator L has the property $LY_t = Y_{t-1}$. By applying the lag operator twice, one obtains the second lag: $L^2Y_t = L(LY_t) = LY_{t-1} = Y_{t-2}$. More generally, by applying the lag operator j times, one obtains the j^{th} lag. In summary, the lag operator has the property that

$$LY_t = Y_{t-1}, L^2Y_t = Y_{t-2}, \text{ and } L^jY_t = Y_{t-j}. \tag{15.38}$$

The lag operator notation permits us to define the **lag polynomial**, which is a polynomial in the lag operator:

$$a(L) = a_0 + a_1L + a_2L^2 + \cdots + a_pL^p = \sum_{j=0}^{p} a_jL^j, \tag{15.39}$$

where $a_0, \ldots, a_p$ are the coefficients of the lag polynomial and $L^0 = 1$. The degree of the lag polynomial $a(L)$ in Equation (15.39) is p. Multiplying Y_t by $a(L)$ yields

$$a(L)Y_t = \left(\sum_{j=0}^{p} a_jL^j\right)Y_t = \sum_{j=0}^{p} a_j(L^jY_t) = \sum_{j=0}^{p} a_jY_{t-j} = a_0Y_t + a_1Y_{t-1} + \cdots + a_pY_{t-p}. \tag{15.40}$$

The expression in Equation (15.40) implies that the AR(p) model in Equation (15.12) can be written compactly as

$$a(L)Y_t = \beta_0 + u_t, \tag{15.41}$$

where $a_0 = 1$ and $a_j = -\beta_j$, for $j = 1, \ldots, p$. Similarly, an ADL(p, q) model can be written

$$a(L)Y_t = \beta_0 + c(L)X_{t-1} + u_t, \tag{15.42}$$

where $a(L)$ is a lag polynomial of degree p with $a_0 = 1$ and $c(L)$ is a lag polynomial of degree $q - 1$.

APPENDIX

15.4 ARMA Models

The **autoregressive–moving average (ARMA) model** extends the autoregressive model by modeling u_t as serially correlated—specifically, as being a distributed lag (or moving average) of another unobserved error term. In the lag operator notation of Appendix 15.3, let $u_t = b(\mathrm{L})e_t$, where $b(\mathrm{L})$ is a lag polynomial of degree q with $b_0 = 1$ and e_t is a serially uncorrelated, unobserved random variable. Then the ARMA(p, q) model is

$$a(\mathrm{L})Y_t = \beta_0 + b(\mathrm{L})e_t, \tag{15.43}$$

where $a(\mathrm{L})$ is a lag polynomial of degree p with $a_0 = 1$.

Both the AR and ARMA models can be thought of as ways to approximate the autocovariances of Y_t. The reason for this is that any stationary time series Y_t with a finite variance can be written either as an AR or as a MA with a serially uncorrelated error term, although the AR or MA model might need to have an infinite order. The second of these results, that a stationary process can be written in moving average form, is known as the Wold decomposition theorem and is one of the fundamental results underlying the theory of stationary time series analysis.

The families of AR, MA, and ARMA models are equally rich as long as the lag polynomials have a sufficiently high degree. In some cases, the autocovariances can be better approximated by an ARMA(p, q) model with small p and q than by a pure AR model with only a few lags. That said, ARMA models are more difficult to extend to additional regressors than are AR models.

APPENDIX

15.5 Consistency of the BIC Lag Length Estimator

This appendix summarizes the argument that the BIC estimator of the lag length, $\hat{p}$, in an autoregression is correct in large samples; that is, $\Pr(\hat{p} = p) \to 1$. This is not true for the AIC estimator, which can overestimate p even in large samples.

BIC

First consider the special case in which the BIC is used to choose among autoregressions with zero, one, or two lags, when the true lag length is one. It is shown below that (i) $\Pr(\hat{p} = 0) \to 0$ and (ii) $\Pr(\hat{p} = 2) \to 0$, from which it follows that $\Pr(\hat{p} = 1) \to 1$. The extension of this argument to the general case of searching over $0 \leq p \leq p_{max}$ entails showing that $\Pr(\hat{p} < p) \to 0$ and $\Pr(\hat{p} > p) \to 0$; the strategy for showing these is the same as used in (i) and (ii) below.

Proof of (i) and (ii)

Proof of (i). To choose $\hat{p} = 0$, it must be the case that $\text{BIC}(0) < \text{BIC}(1)$; that is, $\text{BIC}(0) - \text{BIC}(1) < 0$. Now $\text{BIC}(0) - \text{BIC}(1) = [\ln(SSR(0)/T) + (\ln T)/T] - [\ln(SSR(1)/T) + 2(\ln T)/T] = \ln(SSR(0)/T) - \ln(SSR(1)/T) - (\ln T)/T$. Now $SSR(0)/T = [(T-1)/T]s_Y^2 \xrightarrow{p} \sigma_Y^2$, $SSR(1)/T \xrightarrow{p} \sigma_u^2$, and $(\ln T)/T \longrightarrow 0$; putting these pieces together, $\text{BIC}(0) - \text{BIC}(1) \xrightarrow{p} \ln\sigma_Y^2 - \ln\sigma_u^2 > 0$ because $\sigma_Y^2 > \sigma_u^2$. It follows that $\Pr[\text{BIC}(0) < \text{BIC}(1)] \to 0$, so $\Pr(\hat{p} = 0) \longrightarrow 0$.

Proof of (ii). To choose $\hat{p} = 2$, it must be the case that $\text{BIC}(2) < \text{BIC}(1)$ or $\text{BIC}(2) - \text{BIC}(1) < 0$. Now $T[\text{BIC}(2) - \text{BIC}(1)] = T\{[\ln(SSR(2)/T) + 3(\ln T)/T] - [\ln(SSR(1)/T) + 2(\ln T)/T]\} = T\ln[SSR(2)/SSR(1)] + \ln T = -T\ln[1 + F/(T-2)] + \ln T$, where $F = [SSR(1) - SSR(2)]/[SSR(2)/(T-2)]$ is the homoskedasticity-only F-statistic [Equation (7.13)] testing the null hypothesis that $\beta_2 = 0$ in the AR(2). If u_t is homoskedastic, then F has a χ_1^2 asymptotic distribution; if not, it has some other asymptotic distribution. Thus $\Pr[\text{BIC}(2) - \text{BIC}(1) < 0] = \Pr\{T[\text{BIC}(2) - \text{BIC}(1)] < 0\} = \Pr\{-T\ln[1 + F/(T-2)] + (\ln T) < 0\} = \Pr\{T\ln[1 + F/(T-2)] > \ln T\}$. As T increases, $T\ln[1 + F/(T-2)] - F \xrightarrow{p} 0$ [a consequence of the logarithmic approximation $\ln(1 + a) \cong a$, which becomes exact as $a \longrightarrow 0$]. Thus $\Pr[\text{BIC}(2) - \text{BIC}(1) < 0] \longrightarrow \Pr(F > \ln T) \longrightarrow 0$, so $\Pr(\hat{p} = 2) \longrightarrow 0$.

AIC

In the special case of an AR(1) when zero, one, or two lags are considered, the proof of (i) for the BIC applies to the AIC where the term $\ln T$ is replaced by 2, so $\Pr(\hat{p} = 0) \longrightarrow 0$. All the steps in the proof of (ii) for the BIC also apply to the AIC, with the modification that $\ln T$ is replaced by 2; thus $\Pr[\text{AIC}(2) - \text{AIC}(1) < 0] \longrightarrow \Pr(F > 2) > 0$. If u_t is homoskedastic, then $\Pr(F > 2) \longrightarrow Pr(\chi_1^2 > 2) = 0.16$, so $\Pr(\hat{p} = 2) \longrightarrow 0.16$. In general, when $\hat{p}$ is chosen using the AIC, $\Pr(\hat{p} < p) \longrightarrow 0$, but $\Pr(\hat{p} > p)$ tends to a positive number, so $\Pr(\hat{p} = p)$ does not tend to 1.