20

SERIAL CORRELATION

20.1 INTRODUCTION

Time-series data often display autocorrelation or serial correlation of the disturbances across periods. Consider, for example, the plot of the least squares residuals in the following example.

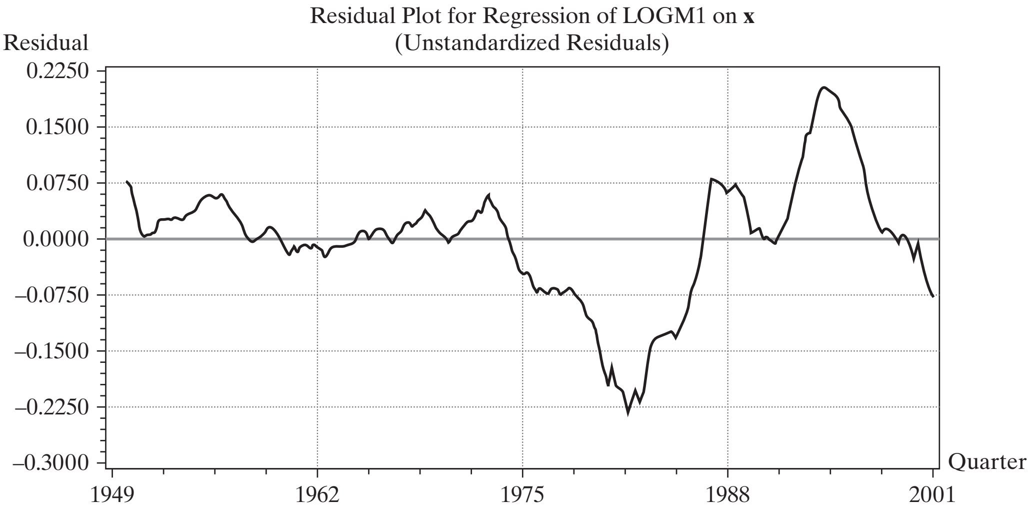
Example 20.1 Money Demand Equation

Appendix Table F5.2 contains quarterly data from 1950I to 2000IV on the U.S. money stock (M1), output (real GDP), and the price level (CPI_U). Consider a simple (extremely) model of money demand,[1]

$$\ln M1_t = \beta_1 + \beta_2 \ln GDP_t + \beta_3 \ln CPI_t + \varepsilon_t.$$

A plot of the least squares residuals is shown in Figure 20.1. The pattern in the residuals suggests that knowledge of the sign of a residual in one period is a good indicator of the sign of the residual in the next period. This knowledge suggests that the effect of a

FIGURE 20.1 Autocorrelated Least Squares Residuals.



[1] Because this chapter deals exclusively with time-series data, we shall use the index *t* for observations and *T* for the sample size throughout.

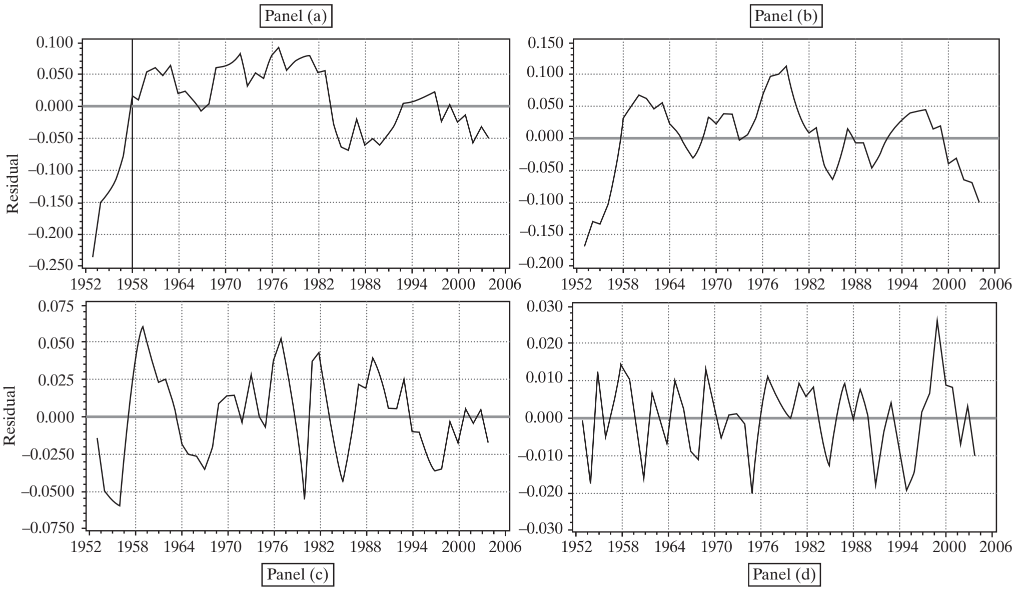
given disturbance is carried, at least in part, across periods. This sort of memory in the disturbances creates the long, slow swings from positive values to negative ones that are evident in Figure 20.1. One might argue that this pattern is the result of an obviously naïve model, but that is one of the important points in this discussion. Patterns such as this usually do not arise spontaneously; to a large extent, they are, indeed, a result of an incomplete or flawed model specification.

One explanation for autocorrelation is that relevant factors omitted from the time-series regression, like those included, are correlated across periods. This fact may be due to serial correlation in factors that should be in the regression model. It is easy to see why this situation would arise. Example 20.2 shows an obvious case.

Example 20.2 Autocorrelation Induced by Misspecification of the Model

In Examples 2.3, 4.2, 4.7, and 4.8, we examined yearly time-series data on the U.S. gasoline market from 1953 to 2004. The evidence in the examples was convincing that a regression model of variation in ln *G/Pop* should include, at a minimum, a constant, ln P_G and ln *Income/Pop* price variables and a time trend also provide significant explanatory power, but these two are a bare minimum. Moreover, we also found on the basis of a Chow test of structural change that apparently this market changed structurally after 1974. Figure 20.2 displays plots of four sets of least squares residuals. Parts (a) through (c) show clearly that as the specification of the regression is expanded, the autocorrelation in the "residuals" diminishes. Part (c) shows the effect of forcing the coefficients in the equation to be the same both before and after the structural shift. In part (d), the residuals in the two subperiods 1953 to 1974 and 1975 to 2004 are produced by separate unrestricted regressions. This latter set of residuals is almost nonautocorrelated. (*Note:* The range of variation of the residuals falls as the model is improved, i.e., as its fit improves.) The full equation is

FIGURE 20.2 Regression Residuals.



$$\ln\frac{G_t}{Pop_t} = \beta_1 + \beta_2 \ln P_{Gt} + \beta_3 \ln\frac{I_t}{Pop_t} + \beta_4 \ln P_{NCt} + \beta_5 \ln P_{UCt}$$
$$+ \beta_6 \ln P_{PTt} + \beta_7 \ln P_{Nt} + \beta_8 \ln P_{Dt} + \beta_9 \ln P_{St} + \beta_{10} t + \varepsilon_t.$$

Finally, we consider an example in which serial correlation is an anticipated part of the model.

Example 20.3 Negative Autocorrelation in the Phillips Curve

The Phillips curve [Phillips (1957)] has been one of the most intensively studied relationships in the macroeconomics literature. As originally proposed, the model specifies a negative relationship between wage inflation and unemployment in the United Kingdom over a period of 100 years. Recent research has documented a similar relationship between unemployment and price inflation. It is difficult to justify the model when cast in simple levels; labor market theories of the relationship rely on an uncomfortable proposition that markets persistently fall victim to money illusion, even when the inflation can be anticipated. Recent research[2] has reformulated a short-run (disequilibrium) “expectations augmented Phillips curve” in terms of unexpected inflation and unemployment that deviates from a long-run equilibrium or “natural rate.” The **expectations-augmented Phillips curve** can be written as

$$\Delta p_t - E[\Delta p_t \mid \Psi_{t-1}] = \beta[u_t - u^*] + \varepsilon_t,$$

where Δp_t is the rate of inflation in year t, $E[\Delta p_t \mid \Psi_{t-1}]$ is the forecast of Δp_t made in period $t - 1$ based on information available at time $t - 1$, Ψ_{t-1}, u_t is the unemployment rate, and u^* is the natural, or equilibrium rate. (Whether u^* can be treated as an unchanging parameter, as we are about to do, is controversial.) By construction, $[u_t - u^*]$ is disequilibrium, or cyclical unemployment. In this formulation, ε_t would be the supply shock (i.e., the stimulus that produces the disequilibrium situation). To complete the model, we require a model for the expected inflation. For the present, we’ll assume that economic agents are rank empiricists. The forecast of next year’s inflation is simply this year’s value. This produces the estimating equation,

$$\Delta p_t - \Delta p_{t-1} = \beta_1 + \beta_2 u_t + \varepsilon_t,$$

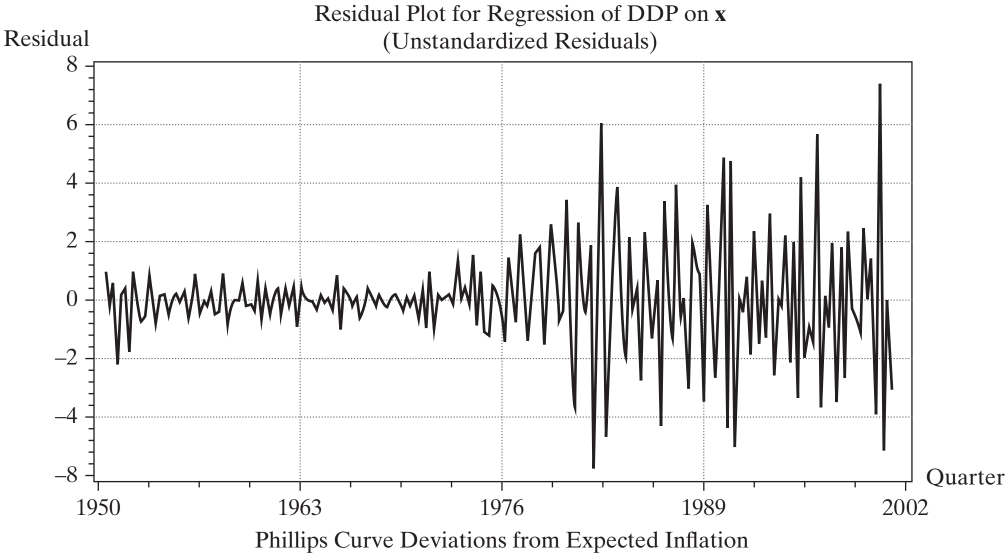
where $\beta_2 = \beta$ and $\beta_1 = -\beta u^*$. Note that there is an implied estimate of the natural rate of unemployment embedded in the equation. After estimation, u^* can be estimated by $-b_1/b_2$. The equation was estimated with the 1950.1 to 2000.4 data in Appendix Table F5.2 that were used in Example 20.1 (minus two quarters for the change in the rate of inflation). Least squares estimates (with standard errors in parentheses) are as follows:

$$\Delta p_t - \Delta p_{t-1} = 2.23567\,(0.49213) - 0.04155\,(0.08360)\,u_t + e_t,\ R^2 = 0.00123,\ T = 202.$$

The implied estimate of the natural rate of unemployment is 5.67 percent, which is in line with other estimates. The estimated asymptotic covariance of b_1 and b_2 is -0.03964. Using the delta method, we obtain a standard error of 3.17524 for this estimate, so a confidence interval for the natural rate is $5.67\% \pm 1.96(3.17\%) = (-0.55\%, 11.89\%)$. (This seems fairly wide, but, again, whether it is reasonable to treat this as a parameter is at least questionable). The regression of the least squares residuals on their past values gives a slope of -0.51843 with a highly significant t ratio of -8.48. We thus conclude that the residuals (and, apparently, the disturbances) in this model are highly negatively autocorrelated. This is consistent with the striking pattern in Figure 20.3.

[2] For example, Staiger et al. (1996).

FIGURE 20.3 Negatively Autocorrelated Residuals.



The problems for estimation and inference caused by autocorrelation are similar to (although, unfortunately, more involved than) those caused by heteroscedasticity. As before, least squares is inefficient, and inference based on the least squares estimates is adversely affected. Depending on the underlying process, however, GLS and FGLS estimators can be devised that circumvent these problems. There is one qualitative difference to be noted. In Section 20.10, we will examine models in which the generalized regression model can be viewed as an extension of the regression model to the conditional second moment of the dependent variable. In the case of autocorrelation, the phenomenon arises in almost all cases from a misspecification of the model. Views differ on how one should react to this failure of the classical assumptions, from a pragmatic one that treats it as another problem in the data to an orthodox methodological view that it represents a major specification issue.[3]

We should emphasize that the models we shall examine here are quite far removed from the classical regression. The exact or small-sample properties of the estimators are rarely known, and only their asymptotic properties have been derived.

20.2 THE ANALYSIS OF TIME-SERIES DATA

The treatment in this chapter will be the first structured analysis of time-series data in the text. Time-series analysis requires some revision of the interpretation of both data generation and sampling that we have maintained thus far.

[3] See, for example, "A Simple Message to Autocorrelation Correctors: Don't" [Mizon (1995)].

A time-series model will typically describe the path of a variable y_t in terms of contemporaneous (and perhaps lagged) factors $\mathbf{x}_t$, disturbances (**innovations**), ε_t, and its own past, $y_{t-1}, \ldots$. For example,

$$y_t = \beta_1 + \beta_2 x_t + \beta_3 y_{t-1} + \varepsilon_t.$$

The time series is a single occurrence of a random event. For example, the quarterly series on real output in the United States from 1950 to 2000 that we examined in Example 20.1 is a single realization of a process, GDP_t. The entire history over this period constitutes a realization of the process. At least in economics, the process could not be repeated. There is no counterpart to repeated sampling in a cross section or replication of an experiment involving a time-series process in physics or engineering. Nonetheless, were circumstances different at the end of World War II, the observed history *could* have been different. In principle, a completely different realization of the entire series might have occurred. The sequence of observations, $\{y_t\}_{t=-\infty}^{t=\infty}$, is a **time-series process**, which is characterized by its time ordering and its systematic correlation between observations in the sequence. The signature characteristic of a time-series process is that empirically, the data-generating mechanism produces exactly one realization of the sequence. Statistical results based on sampling characteristics concern not random sampling from a population, but from distributions of statistics constructed from sets of observations taken from this realization in a **time window**, $t = 1, \ldots, T$. Asymptotic distribution theory in this context concerns behavior of statistics constructed from an increasingly long window in this sequence.

The properties of y_t as a random variable in a cross section are straightforward and are conveniently summarized in a statement about its mean and variance or the probability distribution generating y_t. The statement is less obvious here. It is common to assume that innovations are generated independently from one period to the next, with the familiar assumptions

$$E[\varepsilon_t] = 0,$$
$$\text{Var}[\varepsilon_t] = \sigma_\varepsilon^2,$$

and

$$\text{Cov}[\varepsilon_t, \varepsilon_s] = 0 \quad \text{for } t \neq s.$$

In the current context, this distribution of ε_t is said to be **covariance stationary** or **weakly stationary**. Thus, although the substantive notion of random sampling must be extended for the time series ε_t, the mathematical results based on that notion apply here. It can be said, for example, that ε_t is generated by a time-series process whose mean and variance are not changing over time. As such, by the method we will discuss in this chapter, we could, at least in principle, obtain sample information and use it to characterize the distribution of ε_t. Could the same be said of y_t? There is an obvious difference between the series ε_t and y_t; observations on y_t at different points in time are necessarily correlated. Suppose that the y_t series *is* weakly stationary and that, for the moment, $\beta_2 = 0$. Then we could say that

$$E\,[y_t] = \beta_1 + \beta_3 E[y_{t-1}] + E[\varepsilon_t] = \beta_1/(1 - \beta_3)$$

and

$$\text{Var}[y_t] = \beta_3^2\,\text{Var}[y_{t-1}] + \text{Var}[\varepsilon_t],$$

or

$$\gamma_0 = \beta_3^2\gamma_0 + \sigma_\varepsilon^2,$$

so that

$$\gamma_0 = \frac{\sigma_\varepsilon^2}{1 - \beta_3^2}.$$

Thus, γ_0, the variance of y_t, is a fixed characteristic of the process generating y_t. Note how the stationarity assumption, which apparently includes $|\beta_3| < 1$, has been used. The assumption that $|\beta_3| < 1$ is needed to ensure a finite and positive variance.[4] Finally, the same results can be obtained for nonzero β_2 if it is further assumed that x_t is a weakly stationary series.[5]

Alternatively, consider simply repeated substitution of lagged values into the expression for y_t,

$$y_t = \beta_1 + \beta_3(\beta_1 + \beta_3 y_{t-2} + \varepsilon_{t-1}) + \varepsilon_t, \tag{20-1}$$

and so on. We see that, in fact, the current y_t is an accumulation of the entire history of the innovations, ε_t. So if we wish to characterize the distribution of y_t, then we might do so in terms of sums of random variables. By continuing to substitute for y_{t-2}, then $y_{t-3}, \ldots$ in (20-1), we obtain an explicit representation of this idea,

$$y_t = \sum_{i=0}^{\infty} \beta_3^i(\beta_1 + \varepsilon_{t-i}).$$

Do sums that reach back into infinite past make any sense? We might view the process as having begun generating data at some remote, effectively infinite past. As long as distant observations become progressively less important, the extension to an infinite past is merely a mathematical convenience. The diminishing importance of past observations is implied by $|\beta_3| < 1$. Notice that, not coincidentally, this requirement is the same as that needed to solve for γ_0 in the preceding paragraphs. A second possibility is to assume that the *observation* of *this* time series begins at some time 0 [with (x_0, ε_0) called the *initial conditions*], by which time the underlying process has reached a state such that the mean and variance of y_t are not (or are no longer) changing over time. The mathematics is slightly different, but we are led to the same characterization of the random process generating y_t. In fact, the same weak stationarity assumption ensures both of them.

Except in very special cases, we would expect all the elements in the T component random vector $(y_1, \ldots, y_T)$ to be correlated. In this instance, said correlation is called *autocorrelation*. As such, the results pertaining to estimation with independent or uncorrelated observations that we used in the previous chapters are no longer usable. In point of fact, we have a sample of but one observation on the multivariate random variable $[y_t, t = 1, \ldots, T]$. There is a counterpart to the cross-sectional notion of parameter estimation, but only under assumptions (e.g., weak stationarity) that establish that parameters in the familiar sense even exist. Even with stationarity, it will emerge that for

[4] The current literature in macroeconometrics and time series analysis is dominated by analysis of cases in which $\beta_3 = 1$ (or counterparts in different models). We will return to this subject in Chapter 21.

[5] See Section 20.4.1 on the stationarity assumption.

estimation and inference, none of our earlier finite-sample results are usable. Consistency and asymptotic normality of estimators are somewhat more difficult to establish in time-series settings because results that require independent observations, such as the central limit theorems, are no longer usable. Nonetheless, counterparts to our earlier results have been established for most of the estimation problems we consider here.

20.3 DISTURBANCE PROCESSES

The preceding section has introduced a bit of the vocabulary and aspects of time-series specification. To obtain the theoretical results, we need to draw some conclusions about autocorrelation and add some details to that discussion.

20.3.1 CHARACTERISTICS OF DISTURBANCE PROCESSES

In the usual time-series setting, the disturbances are assumed to be homoscedastic but correlated across observations, so that

$$E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}] = \sigma^2\boldsymbol{\Omega},$$

where $\sigma^2\boldsymbol{\Omega}$ is a full, positive definite matrix with a constant $\sigma^2 = \text{Var}[\varepsilon_t|\mathbf{X}]$ on the diagonal. As will be clear in the following discussion, we shall also assume that $\boldsymbol{\Omega}_{ts}$ is a function of $|t - s|$, but not of t or s alone, which is a **stationarity** assumption. (See the preceding section.) It implies that the covariance between observations t and s is a function only of $|t - s|$, the distance apart in time of the observations. Because σ^2 is not restricted, we normalize $\boldsymbol{\Omega}_{tt} = 1$. We define the **autocovariances,**

$$\text{Cov}[\varepsilon_t, \varepsilon_{t-s}|\mathbf{X}] = \text{Cov}[\varepsilon_{t+s}, \varepsilon_t|\mathbf{X}] = \sigma^2\boldsymbol{\Omega}_{t,t-s} = \gamma_s = \gamma_{-s}.$$

Note that $\sigma^2\boldsymbol{\Omega}_{tt} = \gamma_0$. The correlation between ε_t and ε_{t-s} is their autocorrelation,

$$\text{Corr}[\varepsilon_t, \varepsilon_{t-s}|\mathbf{X}] = \frac{\text{Cov}[\varepsilon_t, \varepsilon_{t-s}|\mathbf{X}]}{\sqrt{\text{Var}[\varepsilon_t|\mathbf{X}]\text{Var}[\varepsilon_{t-s}|\mathbf{X}]}} = \frac{\gamma_s}{\gamma_0} = \rho_s = \rho_{-s}.$$

We can then write

$$E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}] = \boldsymbol{\Gamma} = \gamma_0\mathbf{R},$$

where $\boldsymbol{\Gamma}$ is an **autocovariance matrix** and $\mathbf{R}$ is an **autocorrelation matrix**—the *ts* element is an **autocorrelation coefficient**,

$$\rho_s = \frac{\gamma_{|t-s|}}{\gamma_0}.$$

(*Note:* The matrix $\boldsymbol{\Gamma} = \gamma_0\mathbf{R}$ is the same as $\sigma^2\,\boldsymbol{\Omega}$.) We will usually use the abbreviation ρ_s to denote the autocorrelation between observations s periods apart.

Different types of processes imply different patterns in $\mathbf{R}$. For example, the most frequently analyzed process is a **first-order autoregression** or **AR(1)** process,

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t,$$

where u_t is a stationary, nonautocorrelated (**white noise**) process and ρ is a parameter. We will verify later that for this process, $\rho_s = \rho^s$. Higher-order **autoregressive processes** of the form

$$\varepsilon_t = \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \cdots + \theta_p\varepsilon_{t-p} + u_t$$

imply more involved patterns, including, for some values of the parameters, cyclical behavior of the autocorrelations.[6] Stationary autoregressions are structured so that the influence of a given disturbance fades as it recedes into the more distant past but vanishes only asymptotically. For example, for the AR(1), $\text{Cov}[\varepsilon_t, \varepsilon_{t-s}]$ is never zero, but it does become negligible if $|\rho|$ is less than 1. **Moving-average processes**, conversely, have a short memory. For the MA(1) process,

$$\varepsilon_t = u_t - \lambda u_{t-1},$$

the memory in the process is only one period: $\gamma_0 = \sigma_u^2(1 + \lambda^2)$, $\gamma_1 = -\lambda\sigma_u^2$, but $\gamma_s = 0$ if $s > 1$.

Example 20.4 Autocorrelation Function for the Rate of Inflation

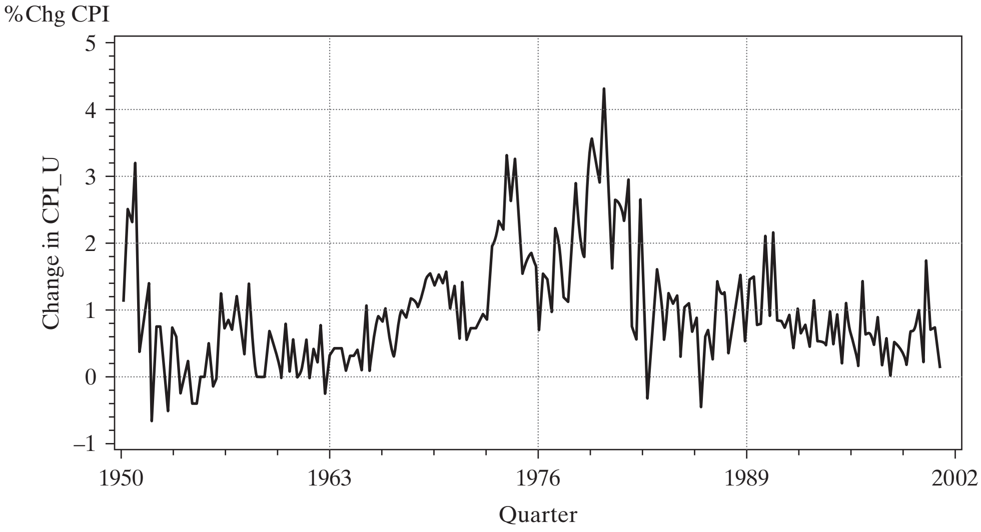
The autocorrelation function for a time series is a useful statistic for describing the nature of the underlying process. The function is computed as

$$ACF(s) = r_s = \frac{c_s}{c_0} = \frac{(1/(T - S))\Sigma_{t=s+1}^{T}(z_t - \bar{z})(z_{t-s} - \bar{z})}{(1/T)\Sigma_{t=s+1}^{T}(z_t - \bar{z})^2}, s = 1, \ldots.$$

The pattern of values of the *ACF* will help reveal the form of the time-series process. For an AR(1) process, the autocorrelations r_s will tend to appear like a geometric series, r^s. For a moving average series such as the MA(1), r_s will show one or a few significant values, then fall sharply to (approximately) zero. The characteristic pattern of an MA(1) process is $r_s = r$ for $s = 1$ and $r_s = 0$ for $s > 1$.

Figure 20.4 shows the quarterly percentage change in the U.S. Consumer Price Index from 1950 to 2000. (We will examine these data in some detail in Chapter 21.) The first 10 autocorrelations for this series are as follows:

FIGURE 20.4 Rate of Inflation in the Consumer Price Index.



[6] This model is considered in more detail in Section 20.9.2.

Lag	1	2	3	4	5	6	7	8	9	10
ACF	0.657	0.602	0.624	0.599	0.469	0.418	0.390	0.360	0.302	0.260

The persistence of the autocorrelations indicates a strongly autoregressive process.

20.3.2 AR(1) DISTURBANCES

Time-series processes such as the ones listed here can be characterized by their order, the values of their parameters, and the behavior of their autocorrelations.[7] We shall consider various forms at different points. The received empirical literature is overwhelmingly dominated by the AR(1) model, which is partly a matter of convenience. Processes more involved than this model are usually extremely difficult to analyze. There is, however, a more practical reason. It is very optimistic to expect to know precisely the correct form of the appropriate model for the disturbance in any given situation. The first-order autoregression has withstood the test of time and experimentation as a reasonable model for underlying processes that probably, in truth, are impenetrably complex. AR(1) works as a first pass—higher-order models are often constructed as a refinement.

The first-order autoregressive disturbance, or AR(1) process, is represented in the **autoregressive form** as

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t, \tag{20-2}$$

where

$$E[u_t|\mathbf{X}] = 0,$$
$$E[u_t^2|\mathbf{X}] = \sigma_u^2,$$

and

$$\text{Cov}[u_t, u_s|\mathbf{X}] = 0 \quad \text{if } t \neq s.$$

Because u_t is white noise, the conditional moments equal the unconditional moments. Thus $E[\varepsilon_t|\mathbf{X}] = E[\varepsilon_t]$ and so on.

By repeated substitution, we have

$$\varepsilon_t = u_t + \rho u_{t-1} + \rho^2 u_{t-2} + \cdots. \tag{20-3}$$

From the preceding **moving-average form**, it is evident that each disturbance ε_t embodies the entire past history of the u's, with the most recent observations receiving greater weight than those in the distant past. Depending on the sign of ρ, the series will exhibit clusters of positive and then negative observations or, if ρ is negative, regular oscillations of sign (as in Example 20.3).

Because the successive values of u_t are uncorrelated, the variance of ε_t is the variance of the right-hand side of (20-3):

$$\text{Var}[\varepsilon_t] = \sigma_u^2 + \rho^2\sigma_u^2 + \rho^4\sigma_u^2 + \cdots. \tag{20-4}$$

To proceed, a restriction must be placed on ρ,

$$|\rho| < 1, \tag{20-5}$$

[7] See Box and Jenkins (1984) for an authoritative study.

because otherwise, the right-hand side of (20-4) will become infinite. This result is the stationarity assumption discussed earlier. With (20-5), which implies that $\lim_{s\to\infty} \rho^s = 0$, $E[\varepsilon_t] = 0$ and

$$\text{Var}[\varepsilon_t] = \frac{\sigma_u^2}{1-\rho^2} = \sigma_\varepsilon^2. \tag{20-6}$$

With the stationarity assumption, there is an easier way to obtain the variance

$$\text{Var}[\varepsilon_t] = \rho^2\,\text{Var}[\varepsilon_{t-1}] + \sigma_u^2$$

because $\text{Cov}[u_t, \varepsilon_s] = 0$ if $t > s$. With stationarity, $\text{Var}[\varepsilon_{t-1}] = \text{Var}[\varepsilon_t]$, which implies (20-6). Proceeding in the same fashion,

$$\text{Cov}[\varepsilon_t, \varepsilon_{t-1}] = E[\varepsilon_t\varepsilon_{t-1}] = E[\varepsilon_{t-1}(\rho\varepsilon_{t-1} + u_t)] = \rho\,\text{Var}[\varepsilon_{t-1}] = \frac{\rho\sigma_u^2}{1-\rho^2}. \tag{20-7}$$

By repeated substitution in (20-2), we see that for any s,

$$\varepsilon_t = \rho^s\varepsilon_{t-s} + \sum_{i=0}^{s-1}\rho^i u_{t-i}$$

(e.g., $\varepsilon_t = \rho^3\varepsilon_{t-3} + \rho^2 u_{t-2} + \rho u_{t-1} + u_t$). Therefore, because ε_s is not correlated with any u_t for which $t > s$ (i.e., any subsequent u_t), it follows that

$$\text{Cov}[\varepsilon_t, \varepsilon_{t-s}] = E[\varepsilon_t\varepsilon_{t-s}] = \frac{\rho^s\sigma_u^2}{1-\rho^2}. \tag{20-8}$$

Dividing by $\gamma_0 = \sigma_u^2/(1-\rho^2)$ provides the autocorrelations,

$$\text{Corr}[\varepsilon_t, \varepsilon_{t-s}] = \rho_s = \rho^s. \tag{20-9}$$

With the stationarity assumption, the autocorrelations fade over time. Depending on the sign of ρ, they will either be declining in geometric progression or alternating in sign if ρ is negative. Collecting terms, we have

$$\sigma^2\boldsymbol{\Omega} = \frac{\sigma_u^2}{1-\rho^2}\begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \cdots & \rho^{T-1} \\ \rho & 1 & \rho & \rho^2 & \cdots & \rho^{T-2} \\ \rho^2 & \rho & 1 & \rho & \cdots & \rho^{T-3} \\ \vdots & \vdots & \vdots & \vdots & \cdots & \rho \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \cdots & \rho & 1 \end{bmatrix}. \tag{20-10}$$

20.4 SOME ASYMPTOTIC RESULTS FOR ANALYZING TIME-SERIES DATA

Because $\boldsymbol{\Omega}$ is not equal to $\mathbf{I}$, the now-familiar complications will arise in establishing the properties of estimators of $\boldsymbol{\beta}$, in particular of the least squares estimator. The finite sample properties of the OLS and GLS estimators remain intact. Least squares will continue to be unbiased. The earlier general proof allows for autocorrelated disturbances. The Aitken theorem (Theorem 9.4) and the distributional results for normally distributed disturbances can still be established conditionally on $\mathbf{X}$. (However, even these will be complicated when $\mathbf{X}$ contains lagged values of the dependent variable.) But finite sample properties are of very limited usefulness in time-series contexts. Nearly all that can be said about estimators involving time-series data is based on their asymptotic properties.

As we saw in our analysis of heteroscedasticity, whether least squares is consistent or not depends on the matrices

$$\mathbf{Q}_T = (1/T)\mathbf{X}'\mathbf{X}$$

and

$$\mathbf{Q}_T^* = (1/T)\mathbf{X}'\boldsymbol{\Omega}\mathbf{X}.$$

In our earlier analyses, we were able to argue for convergence of $\mathbf{Q}_T$ to a positive definite matrix of constants, $\mathbf{Q}$, by invoking laws of large numbers. But these theorems assume that the observations in the sums are independent, which as suggested in Section 20.2, is surely not the case here. Thus, we require a different tool for this result. We can expand the matrix $\mathbf{Q}_T^*$ as

$$\mathbf{Q}_T^* = \frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}\rho_{ts}\mathbf{x}_t\mathbf{x}_s', \tag{20-11}$$

where $\mathbf{x}_t'$ and $\mathbf{x}_s'$ are rows of $\mathbf{X}$ and ρ_{ts} is the autocorrelation between ε_t and ε_s. Sufficient conditions for this matrix to converge are that $\mathbf{Q}_T$ converge and that the correlations between disturbances diminish reasonably rapidly as the observations become further apart in time. For example, if the disturbances follow the AR(1) process described earlier, then $\rho_{ts} = \rho^{|t-s|}$ and if $\mathbf{x}_t$ is sufficiently well behaved, $\mathbf{Q}_T^*$ will converge to a positive definite matrix $\mathbf{Q}^*$ as $T \rightarrow \infty$. **Asymptotic normality** of the least squares and GLS estimators will depend on the behavior of sums such as

$$\sqrt{T}\overline{\mathbf{w}}_T = \sqrt{T}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\varepsilon_t\right) = \sqrt{T}\left(\frac{1}{T}\mathbf{X}'\boldsymbol{\varepsilon}\right).$$

Asymptotic normality of least squares is difficult to establish for this general model. The central limit theorems we have relied on thus far do not extend to sums of *dependent* observations. The results of Amemiya (1985), Mann and Wald (1943), and Anderson (1971) do carry over to most of the familiar types of autocorrelated disturbances, including those that interest us here, so we shall ultimately conclude that ordinary least squares, GLS, and instrumental variables continue to be consistent and asymptotically normally distributed, and, in the case of OLS, inefficient. This section will provide a brief introduction to some of the underlying principles that are used to reach these conclusions.

20.4.1 CONVERGENCE OF MOMENTS—THE ERGODIC THEOREM

The discussion thus far has suggested (appropriately) that stationarity (or its absence) is an important characteristic of a process. The points at which we have encountered this notion concerned requirements that certain sums converge to finite values. In particular, for the AR(1) model, $\varepsilon_t = \rho\varepsilon_{t-1} + u_t$, for the variance of the process to be finite, we require $|\rho| < 1$, which is a sufficient condition. However, this result is only a byproduct. Stationarity (at least, the weak stationarity we have examined) is only a characteristic of the sequence of moments of a distribution.

DEFINITION 20.1 Strong Stationarity
A time-series process, $\{z_t\}_{t=-\infty}^{t=\infty}$, is strongly stationary, or "stationary," if the joint probability distribution of any adjacent set of k observations in the sequence $[z_t, z_{t+1}, \ldots, z_{t+k-1}]$ is the same regardless of the origin, t , in the time scale.

For example, in (20-2), if we add $u_t \sim N[0, \sigma_u^2]$, then the resulting process, $\{\varepsilon_t\}_{t=-\infty}^{t=\infty}$, can easily be shown to be strongly stationary.

DEFINITION 20.2 Weak Stationarity
A time-series process, $\{z_t\}_{t=-\infty}^{t=\infty}$, is weakly stationary (or covariance stationary) if $E[z_t]$ is finite and is the same for all t and if the covariances between any two observations (labeled their autocovariance), $\text{Cov}[z_t, z_{t-k}]$, is a finite function only of model parameters and their distance apart in time, k, but not of the absolute location of either observation on the time scale.

Weak stationary is obviously implied by strong stationary, although it requires less because the distribution can, at least in principle, be changing on the time axis. The distinction is rarely necessary in applied work. In general, save for narrow theoretical examples, it will be difficult to come up with a process that is weakly but not strongly stationary. The reason for the distinction is that in much of our work, only weak stationary is required, and, as always, when possible, econometricians will dispense with unnecessary assumptions.

As we will discover shortly, stationarity is a crucial characteristic at this point in the analysis. If we are going to proceed to parameter estimation in this context, we will also require another characteristic of a time series, **ergodicity**. There are various ways to delineate this characteristic, none of them particularly intuitive. We borrow one definition from Davidson and MacKinnon (1993, p. 132) which comes close:

DEFINITION 20.3 Ergodicity
A strongly stationary time-series process, $\{z_t\}_{t=-\infty}^{t=\infty}$, is ergodic if for any two bounded functions that map vectors in the a and b dimensional real vector spaces to real scalars, $f: \boldsymbol{R}^a \rightarrow \boldsymbol{R}^1$ and $g: \boldsymbol{R}^b \rightarrow \boldsymbol{R}^1$,

$$\begin{aligned}&\lim_{k\rightarrow\infty} |E[f(z_t, z_{t+1}, \ldots, z_{t+a-1})g(z_{t+k}, z_{t+k+1}, \ldots, z_{t+k+b-1})]| \\ &= |E[f(z_t, z_{t+1}, \ldots, z_{t+a-1})]|\,|E[g(z_{t+k}, z_{t+k+1}, \ldots, z_{t+k+b-1})]|.\end{aligned}$$

The definition states essentially that if events are separated far enough in time, then they are *asymptotically independent*. An implication is that in a time series, every observation will contain at least some unique information. Ergodicity is a crucial element of our theory of estimation. When a time series has this property (with stationarity), then we can consider estimation of parameters in a meaningful sense.[8] The analysis relies heavily on the following theorem:

THEOREM 20.1 The Ergodic Theorem
If $\{z_t\}_{t=-\infty}^{t=\infty}$ is a time-series process that is strongly stationary and ergodic and $E[|z_t|]$ is a finite constant, and if $\bar{z}_T = (1/T)\sum_{t=1}^{T} z_t$, then $\bar{z}_T \xrightarrow{a.s.} \mu$, where $\mu = E[z_t]$. Note that the convergence is almost surely not in probability (which is implied) or in mean square (which is also implied). [See White (2001, p. 44) and Davidson and MacKinnon (1993, p. 133).]

What we have in the ergodic theorem is, for sums of dependent observations, a counterpart to the laws of large numbers that we have used at many points in the preceding chapters. Note, once again, the need for this extension is that to this point, our laws of large numbers have required sums of independent observations. But, in this context, by design, observations are distinctly not independent.

For this result to be useful, we will require an extension.

THEOREM 20.2 Ergodicity of Functions
If $\{z_t\}_{t=-\infty}^{t=\infty}$ is a time-series process that is strongly stationary and ergodic and if $y_t = f\{z_t\}$ is a measurable function in the probability space that defines z_t, then y_t is also stationary and ergodic. Let $\{\mathbf{z}_t\}_{t=-\infty}^{t=\infty}$ define a $K \times 1$ vector valued stochastic process—each element of the vector is an ergodic and stationary series, and the characteristics of ergodicity and stationarity apply to the joint distribution of the elements of $\{\mathbf{z}_t\}_{t=-\infty}^{t=\infty}$. Then, the ergodic theorem applies to functions of $\{\mathbf{z}_t\}_{t=-\infty}^{t=\infty}$.[9]

Theorem 20.2 produces the results we need to characterize the least squares (and other) estimators. In particular, by applying the assumptions of Theorem 20.2 to the data series, $[\mathbf{x}_t, \varepsilon_t]_{t=-\infty}^{t=\infty}$ we obtain that $y_t = \mathbf{x}_t'\boldsymbol{\beta} + \varepsilon_t$ is a stationary and ergodic process.

[8] Much of the analysis to follow will involve nonstationary series, which are the focus of most of the current literature—tests for nonstationarity largely dominate the recent study in time-series analysis. Ergodicity is a much more subtle and difficult concept. For any process that we will consider, ergodicity will have to be a given, at least at this level. A classic reference on the subject is Doob (1953). Another authoritative treatise is Billingsley (1995). White (2001) provides a concise analysis of many of these concepts as used in econometrics, and some useful commentary.

[9] See White (2001, pp. 44–45) for discussion.

By analyzing terms element by element we can use these results directly to assert that averages of $\mathbf{w}_t = \mathbf{x}_t\varepsilon_t$, $\mathbf{Q}_{tt} = \mathbf{x}_t\mathbf{x}_t'$, and $\mathbf{Q}_{tt}^* = \varepsilon_t^2\mathbf{x}_t\mathbf{x}_t'$ will converge to their population counterparts, $\mathbf{0}$, $\mathbf{Q}$ and $\mathbf{Q}^*$.

20.4.2 CONVERGENCE TO NORMALITY—A CENTRAL LIMIT THEOREM

To form a distribution theory for least squares, GLS, ML, and GMM, we will need a counterpart to the central limit theorem. In particular, we need to establish a large sample distribution theory for quantities of the form

$$\sqrt{T}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\varepsilon_t\right) = \sqrt{T}\overline{\mathbf{w}}.$$

As noted earlier, we cannot invoke the familiar central limit theorems (Lindeberg–Levy, Lindeberg–Feller, Liapounov) because the observations in the sum are not independent. But, with the assumptions already made, we do have an alternative result. Some needed preliminaries are as follows:

DEFINITION 20.4 Martingale Sequence
A vector sequence $\mathbf{z}_t$ is a martingale sequence if $E[\mathbf{z}_t | \mathbf{z}_{t-1}, \mathbf{z}_{t-2}, \ldots] = \mathbf{z}_{t-1}$.

An important example of a martingale sequence is the **random walk**,

$$z_t = z_{t-1} + u_t,$$

where $\text{Cov}[u_t, u_s] = 0$ for all $t \neq s$. Then

$$E[z_t | z_{t-1}, z_{t-2}, \ldots] = E[z_{t-1} | z_{t-1}, z_{t-2}, \ldots] + E[u_t | z_{t-1}, z_{t-2}, \ldots] = z_{t-1} + 0 = z_{t-1}.$$

DEFINITION 20.5 Martingale Difference Sequence
A vector sequence $\mathbf{z}_t$ is a martingale difference sequence if $E[\mathbf{z}_t | \mathbf{z}_{t-1}, \mathbf{z}_{t-2}, \ldots] = \mathbf{0}$.

With Definition 20.5, we have the following broadly encompassing result:

THEOREM 20.3 Martingale Difference Central Limit Theorem
If $\mathbf{z}_t$ is a vector valued stationary and ergodic martingale difference sequence, with $E[\mathbf{z}_t\mathbf{z}_t'] = \mathbf{\Sigma}$, where $\mathbf{\Sigma}$ is a finite positive definite matrix, and if $\overline{\mathbf{z}}_T = (1/T)\sum_{t=1}^{T}\mathbf{z}_t$, then $\sqrt{T}\,\overline{\mathbf{z}}_T \xrightarrow{d} N[\mathbf{0}, \mathbf{\Sigma}]$. [For discussion, see Davidson and MacKinnon (1993, Sections. 4.7 and 4.8).][10]

[10] For convenience, we are bypassing a step in this discussion: establishing multivariate normality requires that the result first be established for the marginal normal distribution of each component, then that every linear combination of the variables also be normally distributed. (See Theorems D.17 and D.18A.) Our interest at this point is merely to collect the useful end results. Interested users may find the detailed discussions of the many subtleties and narrower points in White (2001) and Davidson and MacKinnon (1993, Chapter 4).

Theorem 20.3 is a generalization of the Lindeberg–Levy central limit theorem. It is not yet broad enough to cover cases of autocorrelation, but it does go beyond Lindeberg–Levy, for example, in extending to the GARCH model of Section 20.13.3.[11] But, looking ahead, this result encompasses what will be a very important application. Suppose in the classical linear regression model, $\{\mathbf{x}_t\}_{t=-\infty}^{t=\infty}$ is a stationary and ergodic multivariate stochastic process and $\{\varepsilon_t\}_{t=-\infty}^{t=\infty}$ is an i.i.d. process—that is, not autocorrelated and not heteroscedastic. Then, this is the most general case of the classical model that still maintains the assumptions about ε_t that we made in Chapter 2. In this case, the process $\{\mathbf{w}_t\}_{t=-\infty}^{t=\infty} = \{\mathbf{x}_t\varepsilon_t\}_{t=-\infty}^{t=\infty}$ is a martingale difference sequence, so that with sufficient assumptions on the moments of $\mathbf{x}_t$ we could use this result to establish consistency and asymptotic normality of the least squares estimator.[12]

We now consider a central limit theorem that is broad enough to include the case that interested us at the outset, stochastically dependent observations on $\mathbf{x}_t$ and autocorrelation in ε_t.[13] Suppose as before that $\{\mathbf{z}_t\}_{t=-\infty}^{t=\infty}$ is a stationary and ergodic stochastic process. We consider $\sqrt{T}\,\bar{\mathbf{z}}_T$. The following conditions are assumed:[14]

1. **Asymptotic uncorrelatedness:** $E[\mathbf{z}_t|\mathbf{z}_{t-k}, \mathbf{z}_{t-k-1}, \ldots]$ converges in mean square to zero as $k \to \infty$. Note that is similar to the condition for ergodicity. White (2001) demonstrates that a (nonobvious) implication of this assumption is $E[\mathbf{z}_t] = \mathbf{0}$.
2. **Summability of autocovariances:** With dependent observations,

$$\lim_{T\to\infty} \operatorname{Var}[\sqrt{T}\,\bar{\mathbf{z}}_T] = \sum_{t=1}^{\infty}\sum_{s=1}^{\infty}\operatorname{Cov}[\mathbf{z}_t, \mathbf{z}_s'] = \sum_{k=-\infty}^{\infty}\boldsymbol{\Gamma}_k = \boldsymbol{\Gamma}^*.$$

   To begin, we will need to assume that this matrix is finite, a condition called **summability**. Note this is the condition needed for convergence of $\mathbf{Q}_T^*$ in (20-11). If the sum is to be finite, then the $k = 0$ term must be finite, which gives us a necessary condition,

$$E[\mathbf{z}_t\mathbf{z}_t'] = \Gamma_0, \quad \text{a finite matrix.}$$

3. **Asymptotic negligibility of innovations:** Let

$$\mathbf{r}_{tk} = E[\mathbf{z}_t|\mathbf{z}_{t-k}, \mathbf{z}_{t-k-1}, \ldots] - E[\mathbf{z}_t|\mathbf{z}_{t-k-1}, \mathbf{z}_{t-k-2}, \ldots].$$

An observation $\mathbf{z}_t$ may be viewed as the accumulated information that has entered the process since it began up to time t. Thus, it can be shown that

$$\mathbf{z}_t = \sum_{s=0}^{\infty}\mathbf{r}_{ts}.$$

The vector $\mathbf{r}_{tk}$ can be viewed as the information in this accumulated sum that entered the process at time $t - k$. The condition imposed on the process is that $\sum_{s=0}^{\infty}\sqrt{E[\mathbf{r}_{ts}'\mathbf{r}_{ts}]}$ be

[11] Forms of the theorem that surpass Lindeberg–Feller (D.19) and Liapounov (Theorem D.20) by allowing for different variances at each time, t, appear in Ruud (2000, p. 479) and White (2001, p. 133). These variants extend beyond our requirements in this treatment.

[12] See, for example, Hamilton (1994, pp. 208–212).

[13] Detailed analysis of this case is quite intricate and well beyond the scope of this book. Some fairly terse analysis may be found in White (2001, pp. 122–133) and Hayashi (2000).

[14] See Hayashi (2000, p. 405) who attributes the results to Gordin (1969).

finite. In words, condition 3 states that information eventually becomes negligible as it fades far back in time from the current observation. The AR(1) model (as usual) helps illustrate this point. If $z_t = \rho z_{t-1} + u_t$, then

$$\begin{aligned} r_{t0} &= E[z_t | z_t, z_{t-1}, \ldots] - E[z_t | z_{t-1}, z_{t-2}, \ldots] = z_t - \rho z_{t-1} = u_t, \\ r_{t1} &= E[z_t | z_{t-1}, z_{t-2} \ldots] - E[z_t | z_{t-2}, z_{t-3} \ldots] \\ &= E[\rho z_{t-1} + u_t | z_{t-1}, z_{t-2} \ldots] - E[\rho(\rho z_{t-2} + u_{t-1}) + u_t | z_{t-2}, z_{t-3}, \ldots] \\ &= \rho(z_{t-1} - \rho z_{t-2}) \\ &= \rho u_{t-1}. \end{aligned}$$

By a similar construction, $r_{tk} = \rho^k u_{t-k}$ from which it follows that $z_t = \sum_{s=0}^{\infty} \rho^s u_{t-s}$, which we saw earlier in (20-3). You can verify that if $|\rho| < 1$, the negligibility condition will be met.

THEOREM 20.4 Gordin's Central Limit Theorem
If $\mathbf{z}_t$ *is strongly stationary and ergodic and if conditions* **1–3** *are met, then* $\sqrt{T}\,\bar{\mathbf{z}}_T \xrightarrow{d} N[\mathbf{0}, \boldsymbol{\Gamma}^*]$.

With all this machinery in place, we now have the theorem we will need. We will be able to employ these tools when we consider the least squares, IV, and GLS estimators in the discussion to follow.

20.5 LEAST SQUARES ESTIMATION

The least squares estimator is

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \boldsymbol{\beta} + \left(\frac{\mathbf{X}'\mathbf{X}}{T}\right)^{-1}\left(\frac{\mathbf{X}'\boldsymbol{\varepsilon}}{T}\right).$$

Unbiasedness follows from the results in Chapter 4—no modification is needed. We know from Chapter 9 that the Gauss–Markov theorem has been lost—assuming it exists (that remains to be established), the GLS estimator is efficient, and OLS is not. How much information is lost by using least squares instead of GLS depends on the data. Broadly, least squares fares better in data that have long periods and little cyclical variation, such as aggregate output series. As might be expected, the greater the autocorrelation in ε, the greater will be the benefit to using generalized least squares (when this is possible). Even if the disturbances are normally distributed, the usual F and t statistics do not have those distributions. So, not much remains of the finite sample properties we obtained in Chapter 4. The asymptotic properties remain to be established.

20.5.1 ASYMPTOTIC PROPERTIES OF LEAST SQUARES

The asymptotic properties of **b** are straightforward to establish given our earlier results. If we assume that the process generating $\mathbf{x}_t$ is stationary and ergodic, then by Theorems 20.1 and 20.2, $(1/T)(\mathbf{X}'\mathbf{X})$ converges to $\mathbf{Q}$ and we can apply the Slutsky theorem to the inverse. If ε_t is not serially correlated, then $\mathbf{w}_t = \mathbf{x}_t\varepsilon_t$ is a martingale difference sequence, so $(1/T)(\mathbf{X}'\boldsymbol{\varepsilon})$ converges to zero. This establishes consistency for the simple case. On the other hand, if $[\mathbf{x}_t, \varepsilon_t]$ are jointly stationary and ergodic, then we can invoke the ergodic theorems 20.1 and 20.2 for both moment matrices and establish consistency. Asymptotic normality is a bit more subtle. For the case without serial correlation in ε_t, we can employ Theorem 20.3 for $\sqrt{T}\,\overline{\mathbf{w}}$. The involved case is the one that interested us at the outset of this discussion, that is, where there is autocorrelation in ε_t and dependence in $\mathbf{x}_t$. Theorem 20.4 is in place for this case. Once again, the conditions described in the preceding section must apply and, moreover, the assumptions needed will have to be established both for $\mathbf{x}_t$ and ε_t. Commentary on these cases may be found in Davidson and MacKinnon (1993), Hamilton (1994), White (2001), and Hayashi (2000). Formal presentation extends beyond the scope of this text, so at this point, we will proceed, and assume that the conditions underlying Theorem 20.4 are met. The results suggested here are quite general, albeit only sketched for the general case. For the remainder of our examination, at least in this chapter, we will confine attention to fairly simple processes in which the necessary conditions for the asymptotic distribution theory will be fairly evident.

There is an important exception to the results in the preceding paragraph. If the regression contains any lagged values of the dependent variable, then in most cases, least squares will no longer be unbiased or consistent. (We will examine the exceptions in Section 20.9.3.) To take the simplest case, suppose that

$$\begin{aligned} y_t &= \beta y_{t-1} + \varepsilon_t, \\ \varepsilon_t &= \rho\varepsilon_{t-1} + u_t, \end{aligned} \qquad \textbf{(20-12)}$$

and assume $|\beta| < 1, |\rho| < 1$. In this model, the regressor and the disturbance are correlated. There are various ways to approach the analysis. One useful way is to rearrange (20-12) by subtracting ρy_{t-1} from y_t. Then,

$$y_t = (\beta + \rho)y_{t-1} - \beta\rho y_{t-2} + u_t, \qquad \textbf{(20-13)}$$

which is a classical regression with stochastic regressors. Because u_t is an innovation in period t, it is uncorrelated with both regressors, and least squares regression of y_t on (y_{t-1}, y_{t-2}) estimates $\rho_1 = (\beta + \rho)$ and $\rho_2 = -\beta\rho$. What is estimated by regression of y_t on y_{t-1} alone? Let $\gamma_k = \text{Cov}[y_t, y_{t-k}] = \text{Cov}[y_t, y_{t+k}]$. By stationarity, $\text{Var}[y_t] = \text{Var}[y_{t-1}]$, and $\text{Cov}[y_t, y_{t-1}] = \text{Cov}[y_{t-1}, y_{t-2}]$, and so on. These and (20-13) imply the following relationships:

$$\begin{aligned} \gamma_0 &= \rho_1\gamma_1 + \rho_2\gamma_2 + \sigma_u^2, \\ \gamma_1 &= \rho_1\gamma_0 + \rho_2\gamma_1, \\ \gamma_2 &= \rho_1\gamma_1 + \rho_2\gamma_0. \end{aligned} \qquad \textbf{(20-14)}$$

(These are the **Yule–Walker equations** for this model.) The slope in the simple regression estimates γ_1/γ_0, which can be found in the solutions to these three equations. (An alternative approach is to use the left-out variable formula, which is a useful way to interpret this estimator.) In this case, we see that the slope in the short regression is an estimator of $(\beta + \rho) - \beta\rho(\gamma_1/\gamma_0)$. In either case, solving the three equations in (20-14) for γ_0, γ_1, and γ_2 in terms of ρ_1, ρ_2, and σ_u^2 produces

$$\text{plim } b = \frac{\beta + \rho}{1 + \beta\rho}. \qquad \textbf{(20-15)}$$

This result is between β (when $\rho = 0$) and 1 (when both β and $\rho = 1$). Therefore, least squares is inconsistent unless ρ equals zero. The more general case that includes regressors, $\mathbf{x}_t$, involves more complicated algebra but gives essentially the same result. This is a general result; when the equation contains a lagged dependent variable in the presence of autocorrelation, OLS and GLS are inconsistent. The problem can be viewed as one of an omitted variable.

20.5.2 ESTIMATING THE VARIANCE OF THE LEAST SQUARES ESTIMATOR

As usual, $s^2(\mathbf{X}'\mathbf{X})^{-1}$ is an inappropriate estimator of $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\Omega\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}$, both because s^2 is a biased estimator of σ^2 and because the matrix is incorrect. Generalities are scarce, but in general, for economic time series that are positively related to their past values, the standard errors conventionally *estimated* by least squares are likely to be too small. For slowly changing, trending aggregates such as output and consumption, this is probably the norm. For highly variable data such as inflation, exchange rates, and market returns, the situation is less clear. Nonetheless, as a general proposition, one would normally not want to rely on $s^2(\mathbf{X}'\mathbf{X})^{-1}$ as an estimator of the asymptotic covariance matrix of the least squares estimator.

In view of this situation, if one is going to use least squares, then it is desirable to have an appropriate estimator of the covariance matrix of the least squares estimator. There are two approaches. If the form of the autocorrelation is known, then one can estimate the parameters of $\boldsymbol{\Omega}$ directly and compute a consistent estimator. Of course, if so, then it would be more sensible to use feasible generalized least squares instead and not waste the sample information on an inefficient estimator. The second approach parallels the use of the White estimator for heteroscedasticity.

The extension of White's result to the more general case of autocorrelation is much more difficult than in the heteroscedasticity case. The natural counterpart for estimating

$$\mathbf{Q}_* = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\sigma_{ij}\mathbf{x}_i\mathbf{x}_j' \tag{20-16}$$

in (9-3) would be

$$\mathbf{Q}_* = \frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}e_t e_s \mathbf{x}_t\mathbf{x}_s'.$$

But there are two problems with this estimator, one theoretical and one practical.

Unlike the heteroscedasticity case, the matrix in (20-16) is $1/T$ times a sum of T^2 terms, so it is difficult to conclude yet that it will converge to anything at all. This application is most likely to arise in a time-series setting. To obtain convergence, it is necessary to assume that the terms involving unequal subscripts in (20-16) diminish in importance as T grows. A sufficient condition is that terms with subscript pairs $|t - s|$ grow smaller as the distance between them grows larger. In practical terms, observation pairs are progressively less correlated as their separation in time grows. Intuitively, if one can think of weights with the diagonal elements getting a weight of 1.0, then in the sum, the weights in the sum grow smaller as we move away from the diagonal. If we think of the sum of the weights rather than just the number of terms, then this sum falls off sufficiently rapidly that as n grows large, the sum is of order T rather than T^2. Thus, we achieve convergence of $\mathbf{Q}^*$ by assuming that the rows of $\mathbf{X}$ are well behaved and that the correlations diminish with increasing separation in time. (See Section 9.2. for a more formal statement of this condition.)

TABLE 20.1 Robust Covariance Estimation

Variable	*OLS Estimate*	*OLS SE*	*Corrected SE*
Constant	−1.6331	0.2286	0.3335
ln *Output*	0.2871	0.04738	0.07806
ln *CPI*	0.9718	0.03377	0.06585
$R^2 = 0.98952, r = 0.98762$			

The practical problem is that $\hat{\mathbf{Q}}_*$ need not be positive definite. Newey and West (1987a) have devised an estimator that overcomes this difficulty,

$$\hat{\mathbf{Q}}_* = \mathbf{S}_0 + \frac{1}{T}\sum_{l=1}^{L}\sum_{t=l+1}^{T} w_l e_t e_{t-l}(\mathbf{x}_t\mathbf{x}_{t-l}' + \mathbf{x}_{t-l}\mathbf{x}_t'),$$
$$w_l = 1 - \frac{l}{(L+1)}. \qquad \textbf{(20-17)}$$

[See (9-5).] [The weight in (20-17) is the Bartlett weight.] The **Newey–West autocorrelation consistent covariance estimator** is surprisingly simple and relatively easy to implement.[15] There is a final problem to be solved. It must be determined in advance how large L is to be. In general, there is little theoretical guidance. Current practice specifies $L \approx T^{1/4}$. Unfortunately, the result is not quite as crisp as that for the heteroscedasticity consistent estimator.

We have the result that $\mathbf{b}$ and $\mathbf{b}_{\text{IV}}$ are asymptotically normally distributed, and we have an appropriate estimator for the asymptotic covariance matrix. We have not specified the distribution of the disturbances, however. Thus, for inference purposes, the F statistic is approximate at best. Moreover, for more involved hypotheses, the likelihood ratio and Lagrange multiplier tests are unavailable. That leaves the Wald statistic, including asymptotic t ratios, as the main tool for statistical inference. We will examine a number of applications in the chapters to follow.

The White and Newey–West estimators are standard in the econometrics literature. We will encounter them at many points in the discussion to follow.

Example 20.5 Autocorrelation Consistent Covariance Estimation

For the model shown in Example 20.1, the regression results with the uncorrected standard errors and the Newey–West autocorrelation robust covariance matrix for lags of five quarters are shown in Table 20.1. The effect of the very high degree of autocorrelation is evident.

20.6 GMM ESTIMATION

The GMM estimator in the regression model with autocorrelated disturbances is produced by the empirical moment equations,

$$\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t(y_t - \mathbf{x}_t'\hat{\boldsymbol{\beta}}_{GMM}) = \frac{1}{T}\mathbf{X}'\hat{\varepsilon}(\hat{\boldsymbol{\beta}}_{GMM}) = \overline{\mathbf{m}}(\hat{\boldsymbol{\beta}}_{GMM}) = \mathbf{0}. \qquad \textbf{(20-18)}$$

[15] Both estimators are now standard features in modern econometrics computer programs. Further results on different weighting schemes may be found in Hayashi (2000, pp. 406–410).

The estimator is obtained by minimizing

$$q = \overline{\mathbf{m}}'(\hat{\boldsymbol{\beta}}_{GMM})\mathbf{W}\,\overline{\mathbf{m}}(\hat{\boldsymbol{\beta}}_{GMM})$$

where **W** is a positive definite weighting matrix. The optimal weighting matrix would be

$$\mathbf{W} = \{\text{Asy. Var}[\sqrt{T}\,\overline{\mathbf{m}}(\boldsymbol{\beta})]\}^{-1},$$

which is the inverse of

$$\text{Asy. Var}\,[\sqrt{T}\,\overline{\mathbf{m}}(\boldsymbol{\beta})] = \text{Asy. Var}\left[\frac{1}{\sqrt{T}}\sum_{i=1}^{n}\mathbf{x}_i\varepsilon_i\right] = \underset{T\to\infty}{\text{plim}}\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}\sigma^2\rho_{ts}\mathbf{x}_t\mathbf{x}_s' = \sigma^2\mathbf{Q}^*.$$

The optimal weighting matrix would be $[\sigma^2\mathbf{Q}^*]^{-1}$. As in the heteroscedasticity case, this minimization problem is an exactly identified case, so, the weighting matrix is actually irrelevant to the solution. *The GMM estimator for the regression model with autocorrelated disturbances is ordinary least squares*. We can use the results in Section 20.5.2 to construct the asymptotic covariance matrix. We will require the assumptions in Section 20.4 to obtain convergence of the moments and asymptotic normality. We will wish to extend this simple result in one instance. In the common case in which $\mathbf{x}_t$ contains lagged values of y_t, we will want to use an instrumental variable estimator. We will return to that estimation problem in Section 20.9.3.

20.7 TESTING FOR AUTOCORRELATION

The available tests for autocorrelation are based on the principle that if the true disturbances are autocorrelated, then this fact can be detected through the autocorrelations of the least squares residuals. The simplest indicator is the slope in the artificial regression

$$\begin{aligned} e_t &= re_{t-1} + v_t, \\ e_t &= y_t - \mathbf{x}_t'\mathbf{b}, \\ r &= \left(\sum_{t=2}^{T}e_te_{t-1}\right)\Big/\left(\sum_{t=1}^{T-1}e_t^2\right). \end{aligned} \qquad \textbf{(20-19)}$$

If there is autocorrelation, then the slope in this regression will be an estimator of $\rho = \text{Corr}[\varepsilon_t, \varepsilon_{t-1}]$. The complication in the analysis lies in determining a formal means of evaluating when the estimator is *large*, that is, on what statistical basis to reject the null hypothesis that ρ equals zero. As a first approximation, treating (20-19) as a classical linear model and using a t or F (squared t) test to test the hypothesis is a valid way to proceed based on the Lagrange multiplier principle. We used this device in Example 20.3. The tests we consider here are refinements of this approach.

20.7.1 LAGRANGE MULTIPLIER TEST

The Breusch (1978)–Godfrey (1978) test is a Lagrange multiplier test of H_0: no autocorrelation versus H_1: $\varepsilon_t = \text{AR}(P)$ or $\varepsilon_t = \text{MA}(P)$. The same test is used for either structure. The test statistic is

$$LM = T\left(\frac{\mathbf{e}'\mathbf{X}_0(\mathbf{X}_0'\mathbf{X}_0)^{-1}\mathbf{X}_0'\mathbf{e}}{\mathbf{e}'\mathbf{e}}\right) = TR_0^2, \qquad \textbf{(20-20)}$$

where $\mathbf{X}_0$ is the original $\mathbf{X}$ matrix augmented by P additional columns containing the lagged OLS residuals, $e_{t-1}, \ldots, e_{t-P}$. The test can be carried out simply by regressing the ordinary least squares residuals e_t on $\mathbf{x}_{t0}$ (filling in missing values for lagged residuals with zeros) and referring TR_0^2 to the tabled critical value for the chi-squared distribution with P degrees of freedom.[16] Because $\mathbf{X}'\mathbf{e} = \mathbf{0}$, the test is equivalent to regressing e_t on the part of the lagged residuals that is unexplained by $\mathbf{X}$. There is therefore a compelling logic to it; if any fit is found, then it is due to correlation between the current and lagged residuals. The test is a joint test of the first P autocorrelations of ε_t, not just the first.

Example 20.6 Test for Autocorrelation

For the model shown in Examples 20.1 and 20.4, the regression of the least squares residuals on a constant, ln*GDP*, ln*CPI* and two lagged values of the residuals (with initial values filled with zeros) produces $R^2 = 0.97632$. With $T = 204$, the Lagrange multiplier statistic is 199.17. The critical value from the chi-squared table for 2 degrees of freedom is 5.99. The hypothesis that there is no second (or greater) degree autocorrelation is rejected.

20.7.2 BOX AND PIERCE'S TEST AND LJUNG'S REFINEMENT

An alternative test that is asymptotically equivalent to the LM test when the null hypothesis, $\rho = 0$, is true and when $\mathbf{X}$ does not contain lagged values of y is due to Box and Pierce (1970). The Q **test** is carried out by referring

$$Q = T\sum_{j=1}^{P} r_j^2, \tag{20-21}$$

where $r_j = \left(\sum_{t=j+1}^{T} e_t e_{t-j}\right) \Big/ \left(\sum_{t=1}^{T} e_t^2\right)$, to the critical values of the chi-squared table with P degrees of freedom. A refinement suggested by Ljung and Box (1979) is

$$Q' = T(T+2)\sum_{j=1}^{P} \frac{r_j^2}{T-j}. \tag{20-22}$$

The essential difference between the Godfrey–Breusch and the Box–Pierce tests is the use of partial correlations (controlling for $\mathbf{X}$ and the other variables) in the former and simple correlations in the latter. Under the null hypothesis, there is no autocorrelation in ε_t, and no correlation between $\mathbf{x}_t$ and ε_s in any event, so the two tests are asymptotically equivalent. On the other hand, because it does not condition on $\mathbf{x}_t$, the Box–Pierce test is less powerful than the LM test when the null hypothesis is false, as intuition might suggest.

20.7.3 THE DURBIN–WATSON TEST

The Durbin–Watson statistic[17] was the first formal procedure developed for testing for autocorrelation using the least squares residuals. The test statistic is

$$d = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2} = 2(1-r) - \frac{e_1^2 + e_T^2}{\sum_{t=1}^{T} e_t^2}, \tag{20-23}$$

[16] A warning to practitioners: Current software varies on whether the lagged residuals are filled with zeros or the first P observations are simply dropped when computing this statistic. In the interest of replicability, users should determine which is the case before reporting results.

[17] Durbin and Watson (1950, 1951, 1971).

where r is the same first-order autocorrelation that underlies the preceding two statistics. If the sample is reasonably large, then the last term will be negligible, leaving $d \approx 2(1 - r)$. The statistic takes this form because the authors were able to determine the exact distribution of this transformation of the autocorrelation and could provide tables of critical values for specific values of T and K. The one-sided test for H_0: $\rho = 0$ against H_1: $\rho > 0$ is carried out by comparing d to values $d_L(T, K)$ and $d_U(T, K)$. If $d < d_L$, the null hypothesis is rejected; if $d > d_U$, the hypothesis is not rejected. If d lies between d_L and d_U, then no conclusion is drawn.

20.7.4 TESTING IN THE PRESENCE OF A LAGGED DEPENDENT VARIABLE

The Durbin–Watson test is not likely to be valid when there is a lagged dependent variable in the equation.[18] The statistic will usually be biased toward a finding of no autocorrelation. Three alternatives have been devised. The LM and Q tests can be used whether or not the regression contains a lagged dependent variable. (In the absence of a lagged dependent variable, they are asymptotically equivalent.) As an alternative to the standard test, Durbin (1970) derived a Lagrange multiplier test that is appropriate in the presence of a lagged dependent variable. The test may be carried out by referring

$$h = r\sqrt{T/(1 - Ts_c^2)}, \tag{20-24}$$

where s_c^2 is the estimated variance of the least squares regression coefficient on y_{t-1}, to the standard normal tables. Large values of h lead to rejection of H_0. The test has the virtues that it can be used even if the regression contains additional lags of y_t, and it can be computed using the standard results from the initial regression without any further regressions. If $s_c^2 > 1/T$, however, then it cannot be computed. An alternative is to regress e_t on $\mathbf{x}_t, y_{t-1}, \ldots, e_{t-1}$, and any additional lags that are appropriate for e_t and then to test the joint significance of the coefficient(s) on the lagged residual(s) with the standard F test. This method is a minor modification of the Breusch–Godfrey test. Under H_0, the coefficients on the remaining variables will be zero, so the tests are the same asymptotically.

20.7.5 SUMMARY OF TESTING PROCEDURES

The preceding has examined several testing procedures for locating autocorrelation in the disturbances. In all cases, the procedure examines the least squares residuals. We can summarize the procedures as follows:

LM test. $LM = TR^2$ in a regression of the least squares residuals on $[\mathbf{x}_t, e_{t-1}, \ldots\ e_{t-P}]$. Reject H_0 if $LM > \chi^2_*[P]$. This test examines the covariance of the residuals with lagged values, controlling for the intervening effect of the independent variables.

Q test. $Q = T(T + 2)\sum_{j=1}^{P} r_j^2/(T - j)$. Reject H_0 if $Q > \chi^2_*[P]$. This test examines the raw correlations between the residuals and P lagged values of the residuals.

Durbin–Watson test. $d = 2(1 - r)$. Reject H_0: $\rho = 0$ if $d < d_L^*$. This test looks directly at the first-order autocorrelation of the residuals.

Durbin's test. $F_D =$ the F statistic for the joint significance of P lags of the residuals in the regression of the least squares residuals on $[\mathbf{x}_t, y_{t-1}, \ldots y_{t-R}, e_{t-1}, \ldots e_{t-P}]$. Reject H_0 if $F_D > F_*[P, T - K - P]$. This test examines the partial correlations between the residuals and the lagged residuals, controlling for the intervening effect of the independent variables and the lagged dependent variable.

[18] This issue has been studied by Nerlove and Wallis (1966), Durbin (1970), and Dezhbaksh (1990).

The Durbin–Watson test has some major shortcomings. The inconclusive region is large if T is small or moderate. The bounding distributions, while free of the parameters $\boldsymbol{\beta}$ and σ, do depend on the data (and assume that $\mathbf{X}$ is nonstochastic). An exact version based on an algorithm developed by Imhof (1980) avoids the inconclusive region, but is rarely used. The *LM* and Box–Pierce statistics do not share these shortcomings—their limiting distributions are chi squared independently of the data and the parameters. For this reason, the *LM* test has become the standard method in applied research.

20.8 EFFICIENT ESTIMATION WHEN Ω IS KNOWN

As a prelude to deriving feasible estimators for $\boldsymbol{\beta}$ in this model, we consider full generalized least squares estimation assuming that $\boldsymbol{\Omega}$ is known. In the next section, we will turn to the more realistic case in which $\boldsymbol{\Omega}$ must be estimated as well.

If the parameters of $\boldsymbol{\Omega}$ are known, then the GLS estimator,

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y}), \tag{20-25}$$

and the estimate of its sampling variance,

$$\text{Est.Asy.Var}[\hat{\boldsymbol{\beta}}] = \hat{\sigma}_\varepsilon^2[\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}]^{-1}, \tag{20-26}$$

where

$$\hat{\sigma}_\varepsilon^2 = \frac{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\boldsymbol{\Omega}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{T} \tag{20-27}$$

can be computed in one step. For the AR(1) case, data for the transformed model are

$$\mathbf{y}_* = \begin{bmatrix} \sqrt{1 - \rho^2}y_1 \\ y_2 - \rho y_1 \\ y_3 - \rho y_2 \\ \vdots \\ y_T - \rho y_{T-1} \end{bmatrix}, \qquad \mathbf{X}_* = \begin{bmatrix} \sqrt{1 - \rho^2}\mathbf{x}_1 \\ \mathbf{x}_2 - \rho \mathbf{x}_1 \\ \mathbf{x}_3 - \rho \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_T - \rho \mathbf{x}_{T-1} \end{bmatrix}. \tag{20-28}$$

These transformations are variously labeled **partial differences, quasi differences**, or **pseudo-differences**. Note that in the transformed model, every observation except the first contains a constant term. What was the column of 1s in $\mathbf{X}$ is transformed to $[(1 - \rho^2)^{1/2}, (1 - \rho), (1 - \rho), \ldots]$. Therefore, if the sample is relatively small, then the problems with measures of fit noted in Section 3.5 will reappear.

The variance of the transformed disturbance is

$$\text{Var}[\varepsilon_t - \rho\varepsilon_{t-1}] = \text{Var}[u_t] = \sigma_u^2.$$

The variance of the first disturbance is also σ_u^2; [see (20-6)]. This can be estimated using $(1 - \rho^2)\hat{\sigma}_\varepsilon^2$.

Corresponding results have been derived for higher-order autoregressive processes. For the AR(2) model,

$$\varepsilon_t = \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + u_t, \tag{20-29}$$

the transformed data for generalized least squares are obtained by

$$\mathbf{z}_{*1} = \left[\frac{(1 + \theta_2)[(1 - \theta_2)^2 - \theta_1^2]}{1 - \theta_2}\right]^{1/2}\mathbf{z}_1,$$

$$\mathbf{z}_{*2} = (1 - \theta_2^2)^{1/2}\mathbf{z}_2 - \frac{\theta_1(1 - \theta_1^2)^{1/2}}{1 - \theta_2}\mathbf{z}_1,$$

$$\mathbf{z}_{*t} = \mathbf{z}_t - \theta_1\mathbf{z}_{t-1} - \theta_2\mathbf{z}_{t-2}, \quad t > 2, \qquad \textbf{(20-30)}$$

where $\mathbf{z}_t$ is used for y_t or $\mathbf{x}_t$. The transformation becomes progressively more complex for higher-order processes.[19]

Note that in both the AR(1) and AR(2) models, the transformation to $\mathbf{y}_*$ and $\mathbf{X}_*$ involves starting values for the processes that depend only on the first one or two observations. We can view the process as having begun in the infinite past. Because the sample contains only T observations, however, it is convenient to treat the first one or two (or P) observations as shown and consider them as initial values. Whether we view the process as having begun at time $t = 1$ or in the infinite past is ultimately immaterial in regard to the asymptotic properties of the estimators.

The asymptotic properties for the GLS estimator are quite straightforward given the apparatus we assembled in Section 20.4. We begin by assuming that $\{\mathbf{x}_t, \varepsilon_t\}$ are jointly an ergodic, stationary process. Then, after the GLS transformation, $\{\mathbf{x}_{*t}, \varepsilon_{*t}\}$ is also stationary and ergodic. Moreover, ε_{*t} is nonautocorrelated by construction. In the transformed model, then, $\{\mathbf{w}_{*t}\} = \{\mathbf{x}_{*t}\varepsilon_{*t}\}$ is a stationary and ergodic martingale difference sequence. We can use the ergodic theorem to establish consistency and the central limit theorem for martingale difference sequences to establish asymptotic normality for GLS in this model. Formal arrangement of the relevant results is left as an exercise.

20.9 ESTIMATION WHEN Ω IS UNKNOWN

For an unknown $\boldsymbol{\Omega}$, there are a variety of approaches. Any consistent estimator of $\boldsymbol{\Omega}(\rho)$ will suffice—recall from Theorem 9.5 in Section 9.4.2, all that is needed for efficient estimation of $\boldsymbol{\beta}$ is a consistent estimator of $\boldsymbol{\Omega}(\rho)$. The complication arises, as might be expected, in estimating the autocorrelation parameter(s).

20.9.1 AR(1) DISTURBANCES

The AR(1) model is the one most widely used and studied. The most common procedure is to begin FGLS with a natural estimator of ρ, the autocorrelation of the residuals. Because $\mathbf{b}$ is consistent, we can use r. Others that have been suggested include Theil's (1971) estimator, $r[(T - K)/(T - 1)]$ and Durbin's (1970), the slope on y_{t-1} in a regression of y_t on y_{t-1}, $\mathbf{x}_t$ and $\mathbf{x}_{t-1}$. The second step is FGLS based on (20-25)–(20-28). This is the **Prais and Winsten** (1954) **estimator**. The **Cochrane and Orcutt** (1949) **estimator** (based on computational ease) omits the first observation.

It is possible to iterate any of these estimators to convergence. Because the estimator is asymptotically efficient at every iteration, nothing is gained by doing so. Unlike the heteroscedastic model, iterating when there is autocorrelation does not produce the

[19] See Box and Jenkins (1984) and Fuller (1976).

maximum likelihood estimator. The iterated FGLS estimator, regardless of the estimator of ρ, does not account for the term (1/2) $\ln(1 - \rho^2)$ in the log-likelihood function [see the following (20-31)].

Maximum likelihood estimators can be obtained by maximizing the log likelihood with respect to $\boldsymbol{\beta}$, σ_u^2, and ρ. The log-likelihood function may be written

$$\ln L = -\frac{\sum_{t=1}^{T} u_t^2}{2\sigma_u^2} + \frac{1}{2}\ln(1 - \rho^2) - \frac{T}{2}(\ln 2\pi + \ln \sigma_u^2), \tag{20-31}$$

where, as before, the first observation is computed differently from the others using (20-28). Based on the MLE, the standard approximations to the asymptotic variances of the estimators are

$$\begin{aligned} \text{Est.Asy.Var}[\hat{\boldsymbol{\beta}}_{ML}] &= \hat{\sigma}^2_{\varepsilon,ML}[\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}_{ML}\mathbf{X}]^{-1}, \\ \text{Est.Asy.Var}[\hat{\sigma}^2_{u,ML}] &= 2\hat{\sigma}^4_{u,ML}/T, \\ \text{Est.Asy.Var}[\hat{\rho}_{ML}] &= (1 - \hat{\rho}^2_{ML})/T. \end{aligned} \tag{20-32}$$

All the foregoing estimators have the same asymptotic properties. The available evidence on their small-sample properties comes from Monte Carlo studies and is, unfortunately, only suggestive. Griliches and Rao (1969) find evidence that if the sample is relatively small and ρ is not particularly large, say, less than 0.3, then least squares is as good as or better than FGLS. The problem is the additional variation introduced into the sampling variance by the variance of r. Beyond these, the results are rather mixed. Maximum likelihood seems to perform well in general, but the Prais–Winsten estimator is evidently nearly as efficient. Both estimators have been incorporated in all contemporary software. In practice, the Prais and Winsten (1954) and Beach and MacKinnon (1978a) maximum likelihood estimators are probably the most common choices.

20.9.2 APPLICATION: ESTIMATION OF A MODEL WITH AUTOCORRELATION

The model of the U.S. gasoline market that appears in Example 6.20 is

$$\ln\left(\frac{G}{Pop}\right)_t = \beta_1 + \beta_2 \ln\left(\frac{Income}{Pop}\right)_t + \beta_3 \ln P_{G,t} + \beta_4 \ln P_{NC,t} + \beta_5 \ln P_{UC,t} + \beta_6 t + \varepsilon_t.$$

The results in Figure 20.2 suggest that the specification may be incomplete, and, if so, there may be autocorrelation in the disturbances in this specification. Least squares estimates of the parameters using the data in Appendix Table F2.2 appear in the first row of Table 20.2. [The dependent variable is ln (*Gas expenditure*/(*price* × *population*)). These are the OLS results reported in Example 6.20.] The first five autocorrelations of the least squares residuals are 0.667, 0.438, 0.142, −0.018, and −0.198. This produces Box–Pierce and Box–Ljung statistics of 36.217 and 38.789, respectively, both of which are larger than the critical value from the chi-squared table of 11.07. We regressed the least squares residuals on the independent variables and five lags of the residuals. (The missing values in the first five years were filled with zeros.) The coefficients on the lagged residuals and the associated t statistics are 0.741(4.635), 0.153(0.789), −0.246(−1.262), 0.0942(0.472), and −0.125(−0.658). The R^2 in this regression is 0.549086, which produces a chi-squared value of 28.55. This is larger than the critical value of 11.07, so once again,

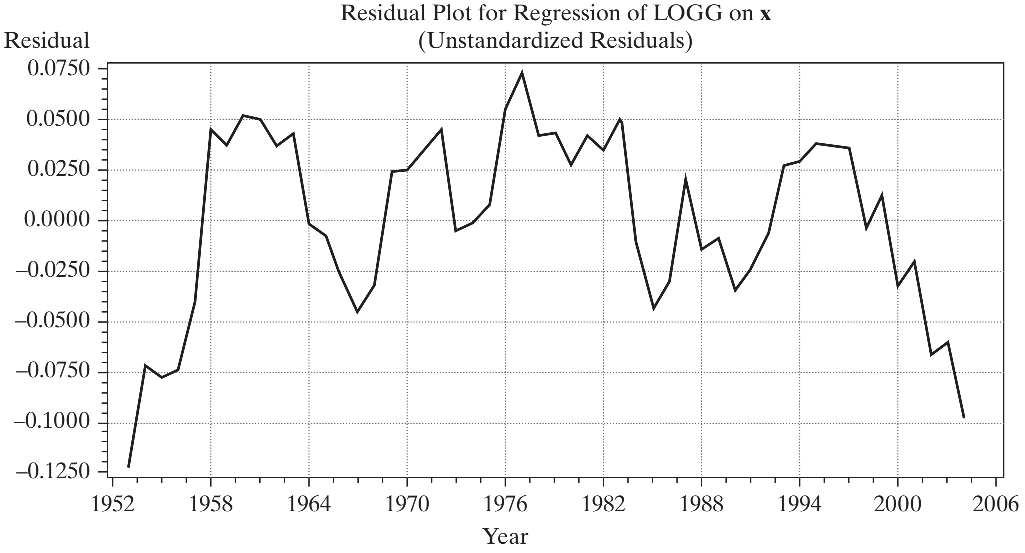
TABLE 20.2 Parameter Estimates (Standard errors in parentheses)

	β_1	β_2	β_3	β_4	β_5	β_6	ρ
OLS	−26.68	1.6250	−0.05392	−0.0834	−0.08467	−0.01393	0.0000
$R^2 = 0.96493$	(2.000)	(0.1952)	(0.04216)	(0.1765)	(0.1024)	(0.00477)	(0.0000)
Prais–	−18.58	0.7447	−0.1138	−0.1364	−0.08956	0.006689	0.9567
Winsten	(1.768)	(0.1761)	(0.03689)	(0.1528)	(0.07213)	(0.004974)	(0.04078)
Cochrane–	−18.76	0.7300	−0.1080	−0.06675	0.04190	−0.0001653	0.9695
Orcutt	(1.382)	(0.1377)	(0.02885)	(0.1201)	(0.05713)	(0.004082)	(0.03434)
Maximum	−16.25	0.4690	−0.1387	−0.09682	−0.001485	0.01280	0.9792
Likelihood	(1.391)	(0.1350)	(0.02794)	(0.1270)	(0.05198)	(0.004427)	(0.02816)
AR(2)	−19.45	0.8116	−0.09538	−0.09099	0.04091	−0.001374	0.8610
	(1.495)	(0.1502)	(0.03117)	(0.1297)	(0.06558)	(0.004227)	(0.07053)

the null hypothesis of zero autocorrelation is rejected. The plot of the residuals shown in Figure 20.5 seems consistent with this conclusion.

The Prais and Winsten FGLS estimates appear in the second row of Table 20.2, followed by the Cochrane and Orcutt results, then the maximum likelihood estimates. [The autocorrelation coefficient computed using $(1 - d/2)$ (see Section 20.7.3) is 0.78750. The MLE is computed using the Beach and MacKinnon algorithm.] Finally, we fit the AR(2) model by first regressing the least squares residuals, e_t, on e_{t-1} and e_{t-2} (without a constant term and filling the first two observations with zeros). The two estimates are 0.751941 and −0.022464, respectively. With the estimates of θ_1 and θ_2, we transformed the data using $y_t^* = y_t - \theta_1 y_{t-1} - \theta_2 y_{t-2}$ and likewise for each regressor. Two observations are then discarded, so the AR(2) regression uses 50 observations while

FIGURE 20.5 Least Squares Residuals.



the Prais–Winsten estimator uses 52 and the Cochrane–Orcutt regression uses 51. In each case, the autocorrelation of the FGLS residuals is computed and reported in the last column of the table.

One might want to examine the residuals after estimation to ascertain whether the AR(1) model is appropriate. In the results just presented, there are two large autocorrelation coefficients listed with the residual-based tests, and in computing the LM statistic, we found that the first two coefficients were statistically significant. If the AR(1) model is appropriate, then one should find that only the coefficient on the first lagged residual is statistically significant in this auxiliary, second-step regression. Another indicator is provided by the FGLS residuals themselves. After computing the FGLS regression, the estimated residuals,

$$\hat{\varepsilon} = y_t - \mathbf{x}_t'\hat{\boldsymbol{\beta}},$$

will still be autocorrelated. In our results using the Prais–Winsten estimates, the autocorrelation of the FGLS residuals is 0.957. This is to be expected. However, if the model is correct, then the transformed residuals,

$$\hat{u}_t = \hat{\varepsilon}_t - \hat{\rho}\hat{\varepsilon}_{t-1},$$

should be at least close to nonautocorrelated. But, for our data, the autocorrelation of these adjusted residuals is only 0.292. It appears on this basis that, in fact, the AR(1) model has largely completed the specification.

20.9.3 ESTIMATION WITH A LAGGED DEPENDENT VARIABLE

In Section 20.5.1, we encountered the problem of estimation by least squares when the model contains both autocorrelation and lagged dependent variable(s). Because the OLS estimator is inconsistent, the residuals on which an estimator of ρ would be based are likewise inconsistent. Therefore, $\hat{\rho}$ will be inconsistent as well. The consequence is that the FGLS estimators described earlier are not usable in this case. There is, however, an alternative way to proceed, based on the method of instrumental variables. The method of instrumental variables was introduced in Section 8.3.2. To review, the general problem is that in the regression model, if

$$\text{plim}(1/T)\mathbf{X}'\boldsymbol{\varepsilon} \neq \mathbf{0},$$

then the least squares estimator is not consistent. A consistent estimator is

$$\mathbf{b}_{\text{IV}} = (\mathbf{Z}'\mathbf{X})^{-1}(\mathbf{Z}'\mathbf{y}),$$

where $\mathbf{Z}$ is a set of K variables chosen such that $\text{plim}(1/T)\mathbf{Z}'\boldsymbol{\varepsilon} = \mathbf{0}$ but $\text{plim}(1/T)\mathbf{Z}'\mathbf{X} \neq \mathbf{0}$. For the purpose of consistency only, any such set of instrumental variables will suffice. The relevance of that here is that the obstacle to consistent FGLS is, at least for the present, the lack of a consistent estimator of ρ. By using the technique of instrumental variables, we may estimate $\boldsymbol{\beta}$ consistently, then estimate ρ and proceed.

Hatanaka (1974, 1976) has devised an efficient two-step estimator based on this principle. To put the estimator in the current context, we consider estimation of the model

$$\begin{aligned} y_t &= \mathbf{x}_t'\boldsymbol{\beta} + \gamma y_{t-1} + \varepsilon_t, \\ \varepsilon_t &= \rho\varepsilon_{t-1} + u_t. \end{aligned}$$

To get to the second step of FGLS, we require a consistent estimator of the slope parameters. These estimates can be obtained using an IV estimator, where the column of **Z** corresponding to y_{t-1} is the only one that need be different from that of **X**. An appropriate instrument can be obtained by using the fitted values in the regression of y_t on $\mathbf{x}_t$ and $\mathbf{x}_{t-1}$. The residuals from the IV regression are then used to construct

$$\hat{\rho} = \frac{\sum_{t=3}^{T} \hat{\varepsilon}_t \hat{\varepsilon}_{t-1}}{\sum_{t=3}^{T} \hat{\varepsilon}_t^2}, \tag{20-33}$$

where

$$\hat{\varepsilon}_t = y_t - \mathbf{b}_{\text{IV}}'\mathbf{x}_t - c_{\text{IV}} y_{t-1}.$$

FGLS estimates may now be computed by regressing $y_{*_t} = y_t - \hat{\rho} y_{t-1}$ on

$$\begin{aligned} \mathbf{x}_{*_t} &= \mathbf{x}_t - \hat{\rho}\mathbf{x}_{t-1}, \\ y_{*_{t-1}} &= y_{t-1} - \hat{\rho} y_{t-2}, \\ \hat{\varepsilon}_{t-1} &= y_{t-1} - \mathbf{b}_{\text{IV}}'\mathbf{x}_{t-1} - c_{\text{IV}} y_{t-2}. \end{aligned}$$

Let d be the coefficient on $\hat{\varepsilon}_{t-1}$ in this regression. The efficient estimator of ρ is

$$\hat{\hat{\rho}} = \hat{\rho} + d.$$

Appropriate asymptotic standard errors for the estimators, including $\hat{\hat{\rho}}$, are obtained from the $s^2[\mathbf{X}_*'\mathbf{X}_*]^{-1}$ computed at the second step. These estimators are asymptotically equivalent to maximum likelihood estimators.[20]

One could argue that the concern about the bias of least squares is misdirected. Consider, again, the model in (20-12),

$$\begin{aligned} y_t &= \beta y_{t-1} + \varepsilon_t, \\ \varepsilon_t &= \rho \varepsilon_{t-1} + u_t. \end{aligned}$$

We established that linear regression of y_t on y_{t-1} estimates not β, but $\gamma = (\beta + \rho)/(1 - \beta\rho)$. It would follow that

$$E[y_t | y_{t-1}] = \gamma y_{t-1},$$

and this is what was of interest from the outset. If so, then the existence of autocorrelation in ε_t is a moot point. In a more completely specified model,

$$y_t = \mathbf{x}_t'\boldsymbol{\beta} + \gamma y_{t-1} + \varepsilon_t,$$

what is likely to be of interest is $E[y_t | \mathbf{x}_t, y_{t-1}] = \mathbf{x}_t'\lambda + \delta y_{t-1}$, and the question of autocorrelation of ε_t is a side issue. The nature of the autocorrelation in ε_t will determine whether $\boldsymbol{\beta} = \boldsymbol{\lambda}$ and $\gamma = \delta$. In the simplest case, as we saw earlier, if $\text{Cov}[\varepsilon_t, \varepsilon_{t-s}] = 0$ for all s, then these equalities will hold. If ε_t is autocorrelated, then they will not. There is a fundamental ambiguity in this treatment, however. In the simple model, we also found earlier that $\text{E}[y_t | y_{t-1}, y_{t-2}] = \gamma_1 y_{t-1} + \gamma_2 y_{t-2}$. There is no argument that the second-order equation is more or less correct than the first. They are two different

[20] See Hatanaka (2000).

representations of the same time series.[21] This idea calls into question the notion of "correcting" for autocorrelation in a regression. We saw in Example 20.2 another implication. The objective of the model builder would be to build residual autocorrelation out of the model. The presence of autocorrelation in the disturbance suggests that the regression part of the equation is incomplete.

Example 20.7 Dynamically Complete Regression

Figure 20.6 shows the residuals from two specifications of the gasoline demand model from Section 20.9.2: a static form,

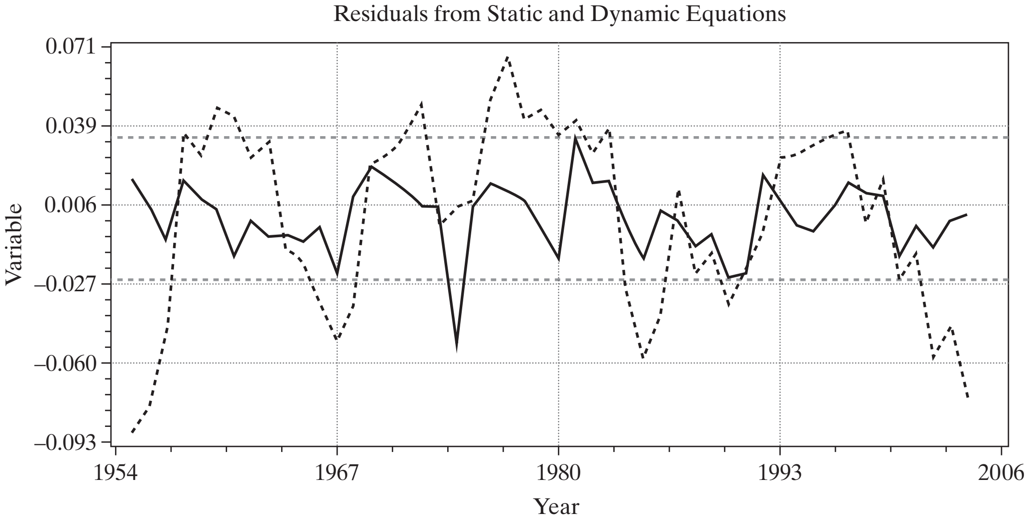
$$\ln\left(\frac{G}{pop}\right)_t = \beta_1 + \beta_2 \ln\left(\frac{Income}{Pop}\right)_t + \beta_3 \ln P_{G,t} + \beta_4 \ln P_{NC,t} + \beta_5 \ln P_{UC,t} + \beta_6 t + \varepsilon_t,$$

and a dynamic form,

$$\ln\left(\frac{G}{Pop}\right)_t = \beta_1 + \beta_2 \ln\left(\frac{Income}{Pop}\right)_t + \beta_3 \ln P_{G,t} + \beta_4 \ln P_{NC,t} + \beta_5 \ln P_{UC,t} + \beta_6 t + \gamma \ln\left(\frac{G}{Pop}\right)_{t-1} + \varepsilon_t.$$

The residuals from the dynamic model are shown with the solid lines. The horizontal bars contain the full range of variation of these residuals. The dashed figure shows the residuals from the static model. The much narrower range of the first set reflects the better fit of the model with the additional (highly significant) regressor. Note, as well, the more substantial amount of fluctuation which suggests less autocorrelation of the residuals from the more dynamically complete regression. To test for autocorrelation of the residuals, we computed the residuals from each regression and regressed them on the lagged residual and the other variables in the equations. For the dynamic model, the LM statistic (TR^2) equaled 1.641. This would be a

FIGURE 20.6 Regression Residuals.



[21] This is an implication of Wold's Decomposition Theorem. See Anderson (1971) or Greene (2003b, p. 619).

TABLE 20.3 Estimated Gasoline Demand Equations

	Dynamic Model				*Static Model*	
		Std.	*Elasticity*			*Std.*
Variable	*Estimate*	*Error*	*S.R.*	*L.R.*	*Estimate*	*Error*
Constant	−5.31920	1.45463	–	–	−26.4319	1.83501
ln *Income*	0.33945	0.10203	0.339	1.642	1.60170	0.17904
ln *Price*	−0.07617	0.01463	−0.076	−0.368	−0.06167	0.03872
ln *P New Cars*	−0.11713	0.06144	−0.117	−0.567	−0.14083	0.16284
ln *P Used Cars*	0.10016	0.03709	0.100	0.484	−0.01293	0.09664
Time trend	−0.00362	0.00180	–	–	−0.01518	0.00439
ln *Demand*[−1]	0.79327	0.04807	–	–	–	–
R^2	0.99552				0.96780	
LM Statistic (1)	1.641				29.787	

chi-squared variable with one degree of freedom. The critical value is 3.84, so the hypothesis of no autocorrelation is not rejected. The equation would appear to be dynamically complete. The same computation for the static model produces a chi-squared value of 29.787.

The estimates of the parameters for the two equations are given in Table 20.3. The fit of the model is high in both cases, but approaches one in the dynamic case. Long-run income and price elasticities are computed as $\eta = \beta_k/(1 - \gamma)$.

20.10 AUTOREGRESSIVE CONDITIONAL HETEROSCEDASTICITY

Heteroscedasticity is often associated with cross-sectional data, whereas time series are usually studied in the context of homoscedastic processes. In analyses of macroeconomic data, Engle (1982, 1983) and Cragg (1982) found evidence that for some kinds of data, the disturbance variances in time-series models were less stable than usually assumed. Engle's results suggested that in models of inflation, large and small forecast errors appeared to occur in clusters, suggesting a form of heteroscedasticity in which the variance of the forecast error depends on the size of the previous disturbance. He suggested the autoregressive, conditionally heteroscedastic, or ARCH, model as an alternative to the usual time-series process. More recent studies of financial markets suggest that the phenomenon is quite common. The ARCH model has proven to be useful in studying the volatility of inflation,[22] the term structure of interest rates,[23] the volatility of stock market returns,[24] and the behavior of foreign exchange markets,[25] to name but a few. This section will describe specification, estimation, and testing, in the basic formulations of the ARCH model and some extensions.[26]

[22] Coulson and Robins (1985).

[23] Engle, Hendry, and Trumble (1985).

[24] Engle, Lilien, and Robins (1987).

[25] Domowitz and Hakkio (1985) and Bollerslev and Ghysels (1996).

[26] Engle and Rothschild (1992) give a survey of this literature which describes many extensions. Mills (1993) also presents several applications. See, as well, Bollerslev (1986) and Li, Ling, and McAleer (2001). See McCullough and Renfro (1999) for discussion of estimation of this model.

Example 20.8 Stochastic Volatility

Figure 20.7 shows Bollerslev and Ghysel's 1974 data on the daily percentage nominal return for the Deutschmark/Pound exchange rate. (These data are given in Appendix Table F20.1.) The variation in the series appears to be fluctuating, with several clusters of large and small movements.

20.10.1 THE ARCH(1) MODEL

The simplest form of this model is the ARCH(1) model,

$$\begin{aligned} y_t &= \mathbf{x}_t'\boldsymbol{\beta} + \varepsilon_t, \\ \varepsilon_t &= u_t\sqrt{\alpha_0 + \alpha_1\varepsilon_{t-1}^2}. \end{aligned} \qquad \textbf{(20-34)}$$

where u_t is distributed as standard normal.[27] It follows that $E[\varepsilon_t \mid \mathbf{x}_t, \varepsilon_{t-1}] = 0$, so that $E[\varepsilon_t \mid \mathbf{x}_t] = 0$ and $E[y_t \mid \mathbf{x}_t] = \mathbf{x}_t'\boldsymbol{\beta}$. Therefore, this model is a classical regression model. But

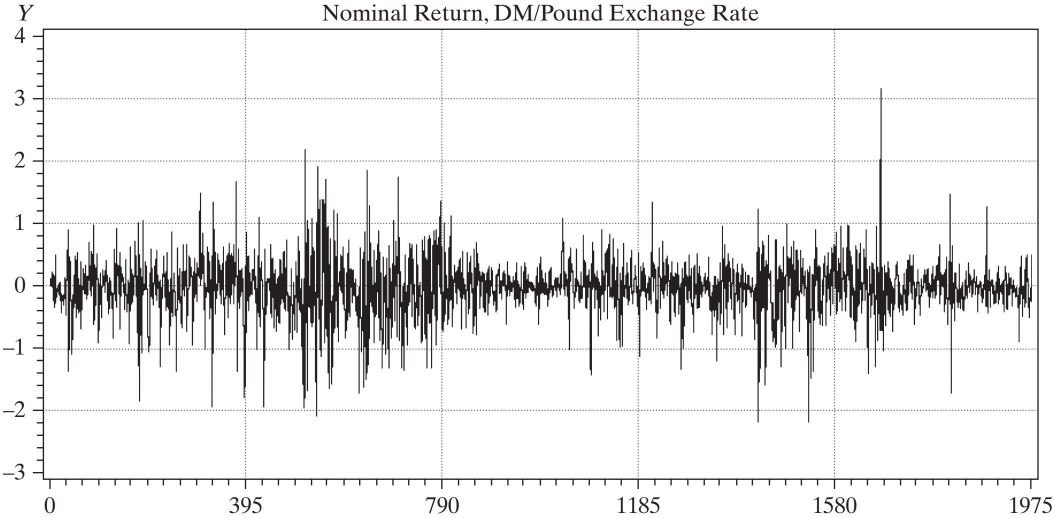
$$\text{Var}[\varepsilon_t \mid \varepsilon_{t-1}] = E[\varepsilon_t^2 \mid \varepsilon_{t-1}] = E[u_t^2][\alpha_0 + \alpha_1\varepsilon_{t-1}^2] = \alpha_0 + \alpha_1\varepsilon_{t-1}^2,$$

so ε_t is *conditionally heteroscedastic*, not with respect to $\mathbf{x}_t$ as we considered in Chapter 9, but with respect to ε_{t-1}. The unconditional variance of ε_t is

$$\text{Var}[\varepsilon_t] = \text{Var}\{E[\varepsilon_t \mid \varepsilon_{t-1}]\} + E\{\text{Var}[\varepsilon_t \mid \varepsilon_{t-1}]\} = \alpha_0 + \alpha_1 E[\varepsilon_{t-1}^2] = \alpha_0 + \alpha_1\text{Var}[\varepsilon_{t-1}].$$

If the process generating the disturbances is weakly (covariance) stationary (see Definition 19.2),[28] then the unconditional variance is not changing over time so

FIGURE 20.7 Nominal Exchange Rate Returns.



[27] The assumption that u_t has unit variance is not a restriction. The scaling implied by any other variance would be absorbed by the other parameters.

[28] This discussion will draw on the results and terminology of time-series analysis in Section 20.3. The reader may wish to peruse this material at this point.

$$\mathrm{Var}[\varepsilon_t] = \mathrm{Var}[\varepsilon_{t-1}] = \alpha_0 + \alpha_1 \mathrm{Var}[\varepsilon_{t-1}] = \frac{\alpha_0}{1 - \alpha_1}.$$

For this ratio to be finite and positive, $|\alpha_1|$ must be less than 1. Then, unconditionally, ε_t is distributed with mean zero and variance $\sigma^2 = \alpha_0/(1 - \alpha_1)$. Therefore, the model obeys the classical assumptions, and ordinary least squares is the most efficient *linear* unbiased estimator of $\boldsymbol{\beta}$.

But there is a more efficient *nonlinear* estimator. The log-likelihood function for this model is given by Engle (1982). Conditioned on starting values y_0 and $\mathbf{x}_0$ (and ε_0), the conditional log likelihood for observations $t = 1, \ldots, T$ is the one we examined in Section 14.9.2.a for the general heteroscedastic regression model [see (14-58)],

$$\ln L = -\frac{T}{2}\ln(2\pi) - \frac{1}{2}\sum_{t=1}^{T}\ln(\alpha_0 + \alpha_1\varepsilon_{t-1}^2) - \frac{1}{2}\sum_{t=1}^{T}\frac{\varepsilon_t^2}{\alpha_0 + \alpha_1\varepsilon_{t-1}^2}, \quad \varepsilon_t = y_t - \boldsymbol{\beta}'\mathbf{x}_t. \quad \textbf{(20-35)}$$

Maximization of log L can be done with the conventional methods, as discussed in Appendix E.[29]

20.10.2 ARCH(q), ARCH-IN-MEAN, AND GENERALIZED ARCH MODELS

The natural extension of the ARCH(1) model presented before is a more general model with longer lags. The ARCH(q) process,

$$\sigma_t^2 = \alpha_0 + \alpha_1\varepsilon_{t-1}^2 + \alpha_2\varepsilon_{t-2}^2 + \cdots + \alpha_q\varepsilon_{t-q}^2,$$

is a qth order moving average [MA(q)] process.[30] This section will generalize the ARCH(q) model, as suggested by Bollerslev (1986), in the direction of an autoregressive-moving average (ARMA) model of Section 21.2. The discussion will parallel his development, although many details are omitted for brevity. The reader is referred to that paper for background and for some of the less critical details.

Among the many variants of the capital asset pricing model (CAPM) is an intertemporal formulation by Merton (1980) that suggests an approximate linear relationship between the return and variance of the market portfolio. One of the possible flaws in this model is its assumption of a constant variance of the market portfolio. In this connection, then, the **ARCH-in-Mean**, or ARCH-M, model suggested by Engle, Lilien, and Robins (1987) is a natural extension. The model states that

$$y_t = \boldsymbol{\beta}'\mathbf{x}_t + \delta\sigma_t^2 + \varepsilon_t,$$

$$\mathrm{Var}[\varepsilon_t|\Psi_t] = \mathrm{ARCH}(q).$$

Among the interesting implications of this modification of the standard model is that under certain assumptions, δ is the coefficient of relative risk aversion. The ARCH-M model has been applied in a wide variety of studies of volatility in asset returns, including

[29] Engle (1982) and Judge et al. (1985, pp. 441–444) suggest a four-step procedure based on the method of scoring that resembles the two-step method we used for the multiplicative heteroscedasticity model in Section 14.10.3. However, the full MLE is now incorporated in most modern software, so the simple regression-based methods, which are difficult to generalize, are less attractive in the current literature. But see McCullough and Renfro (1999) and Fiorentini, Calzolari, and Panattoni (1996) for commentary and some cautions related to maximum likelihood estimation.

[30] Once again, see Engle (1982).

the daily Standard & Poor's Index[31] and weekly New York Stock Exchange returns.[32] A lengthy list of applications is given in Bollerslev, Chou, and Kroner (1992).

The ARCH-M model has several noteworthy statistical characteristics. Unlike the standard regression model, misspecification of the variance function does affect the consistency of estimators of the parameters of the mean.[33] Recall that in the classical regression setting, weighted least squares is consistent even if the weights are misspecified as long as the weights are uncorrelated with the disturbances. That is not true here. If the ARCH part of the model is misspecified, then conventional estimators of $\boldsymbol{\beta}$ and δ will not be consistent. Bollerslev, Chou, and Kroner (1992) list a large number of studies that called into question the specification of the ARCH-M model, and they subsequently obtained quite different results after respecifying the model. A closely related practical problem is that the mean and variance parameters in this model are no longer uncorrelated. In analysis up to this point, we made quite profitable use of the block diagonality of the Hessian of the log-likelihood function for the model of heteroscedasticity. But the Hessian for the ARCH-M model is not block diagonal. In practical terms, the estimation problem cannot be segmented as we have done previously with the heteroscedastic regression model. All the parameters must be estimated simultaneously.

The generalized autoregressive conditional heteroscedasticity (GARCH) model is defined as follows.[34] The underlying regression is the usual one in (20-34). Conditioned on an information set at time t, denoted Ψ_t, the distribution of the disturbance is assumed to be

$$\varepsilon_t \,|\, \Psi_t \sim N[0, \sigma_t^2],$$

where the conditional variance is

$$\sigma_t^2 = \alpha_0 + \delta_1\sigma_{t-1}^2 + \delta_2\sigma_{t-2}^2 + \cdots + \delta_p\sigma_{t-p}^2 + \alpha_1\varepsilon_{t-1}^2 + \alpha_2\varepsilon_{t-2}^2 + \cdots + \alpha_q\varepsilon_{t-q}^2. \tag{20-36}$$

Define

$$\mathbf{z}_t = [1, \sigma_{t-1}^2, \sigma_{t-2}^2, \ldots, \sigma_{t-p}^2, \varepsilon_{t-1}^2, \varepsilon_{t-2}^2, \ldots, \varepsilon_{t-q}^2]'$$

and

$$\boldsymbol{\gamma} = [\alpha_0, \delta_1, \delta_2, \ldots, \delta_p, \alpha_1, \ldots, \alpha_q]' = [\alpha_0, \boldsymbol{\delta}', \boldsymbol{\alpha}']'.$$

Then,

$$\sigma_t^2 = \boldsymbol{\gamma}'\mathbf{z}_t.$$

Notice that the conditional variance is defined by an autoregressive-moving average [ARMA (p, q)] process in the innovations ε_t^2. The difference here is that the *mean* of the

[31] See French, Schwert, and Stambaugh (1987).

[32] See Chou (1988).

[33] See Pagan and Ullah (1988) for a formal analysis of this point.

[34] As have most areas in time-series econometrics, the line of literature on GARCH models has progressed rapidly in recent years and will surely continue to do so. We have presented Bollerslev's model in some detail, despite many recent extensions, not only to introduce the topic as a bridge to the literature, but also because it provides a convenient and interesting setting in which to discuss several related topics such as double-length regression and pseudo-maximum likelihood estimation.

random variable of interest y_t is described completely by a heteroscedastic, but otherwise ordinary, regression model. The *conditional variance*, however, evolves over time in what might be a very complicated manner, depending on the parameter values and on p and q. The model in (20-36) is a GARCH(p,q) model, where p refers, as before, to the order of the autoregressive part.[35] As Bollerslev (1986) demonstrates with an example, the virtue of this approach is that a GARCH model with a small number of terms appears to perform as well as or better than an ARCH model with many.

The **stationarity conditions** are important in this context to ensure that the moments of the normal distribution are finite. The reason is that higher moments of the normal distribution are finite powers of the variance. A normal distribution with variance σ_t^2 has fourth moment $3\sigma_t^4$, sixth moment $15\sigma_t^6$, and so on. [The precise relationship of the even moments of the normal distribution to the variance is $\mu_{2k} = (\sigma^2)^k(2k)!/(k!2^k)$.] Simply ensuring that σ_t^2 is stable does not ensure that higher powers are as well.[36] Bollerslev presents a useful figure that shows the conditions needed to ensure stability for moments up to order 12 for a GARCH(1,1) model and gives some additional discussion. For example, for a GARCH(1,1) process, for the fourth moment to exist, $3\alpha_1^2 + 2\alpha_1\delta_1 + \delta_1^2$ must be less than 1.

It is convenient to write (20-36) in terms of polynomials in the lag operator,

$$\sigma_t^2 = \alpha_0 + D(L)\sigma_t^2 + A(L)\varepsilon_t^2.$$

The stationarity condition for such an equation is that the roots of the characteristic equation, $1 - D(z) = 0$, must lie outside the unit circle. For the present, we will assume that this case is true for the model we are considering and that $A(1) + D(1) < 1$. [This assumption is stronger than that needed to ensure stationarity in a higher-order autoregressive model, which would depend only on $D(L)$.] The implication is that the GARCH process is covariance stationary with $E[\varepsilon_t] = 0$ (unconditionally), $\text{Var}[\varepsilon_t] = \alpha_0/[1 - A(1) - D(1)]$, and $\text{Cov}[\varepsilon_t, \varepsilon_s] = 0$ for all $t \neq s$. Thus, unconditionally the model is the classical regression model that we examined in Chapters 2–6.

The usefulness of the GARCH specification is that it allows the variance to evolve over time in a way that is much more general than the simple specification of the ARCH model. For the example discussed in his paper, Bollerslev reports that although Engle and Kraft's (1983) ARCH(8) model for the rate of inflation in the GNP deflator appears to remove all ARCH effects, a closer look reveals GARCH effects at several lags. By fitting a GARCH(1,1) model to the same data, Bollerslev finds that the ARCH effects out to the same eight-period lag as fit by Engle and Kraft and his observed GARCH effects are all satisfactorily accounted for.

20.10.3 MAXIMUM LIKELIHOOD ESTIMATION OF THE GARCH MODEL

Bollerslev describes a method of estimation based on the BHHH algorithm. As he shows, the method is relatively simple, although with the line search and first derivative

[35] We have changed Bollerslev's notation slightly so as not to conflict with our previous presentation. He used $\boldsymbol{\beta}$ instead of our $\boldsymbol{\delta}$ in (20-36) and $\mathbf{b}$ instead of our $\boldsymbol{\beta}$ in (20-34).

[36] The conditions cannot be imposed a priori. In fact, there is no nonzero set of parameters that guarantees stability of *all* moments, even though the normal distribution has finite moments of all orders. As such, the normality assumption must be viewed as an approximation.

method that he suggests, it probably involves more computation and more iterations than necessary. Following the suggestions of Harvey (1976), it turns out that there is a simpler way to estimate the GARCH model that is also very illuminating. This model is actually very similar to the more conventional model of multiplicative heteroscedasticity that we examined in Section 14.10.3.

For normally distributed disturbances, the log likelihood for a sample of T observations is[37]

$$\ln L = \sum_{t=1}^{T} -\frac{1}{2}\left[\ln(2\pi) + \ln\sigma_t^2 + \frac{\varepsilon_t^2}{\sigma_t^2}\right] = \sum_{t=1}^{T}\ln f_t(\boldsymbol{\theta}) = \sum_{t=1}^{T} l_t(\boldsymbol{\theta}),$$

where $\varepsilon_t = y_t - \mathbf{x}_t'\boldsymbol{\beta}$ and $\boldsymbol{\theta} = (\boldsymbol{\beta}', \alpha_0, \boldsymbol{\alpha}', \boldsymbol{\delta}')' = (\boldsymbol{\beta}', \boldsymbol{\gamma}')'$. Derivatives of $\ln L$ are obtained by summation. Let l_t denote $\ln f_t(\boldsymbol{\theta})$. The first derivatives with respect to the variance parameters are

$$\frac{\partial l_t}{\partial\boldsymbol{\gamma}} = -\frac{1}{2}\left[\frac{1}{\sigma_t^2} - \frac{\varepsilon_t^2}{(\sigma_t^2)^2}\right]\frac{\partial\sigma_t^2}{\partial\boldsymbol{\gamma}} = \frac{1}{2}\left(\frac{1}{\sigma_t^2}\right)\frac{\partial\sigma_t^2}{\partial\boldsymbol{\gamma}}\left(\frac{\varepsilon_t^2}{\sigma_t^2} - 1\right) = \frac{1}{2}\left(\frac{1}{\sigma_t^2}\right)\mathbf{g}_t v_t = \mathbf{b}_t v_t. \qquad \textbf{(20-37)}$$

Note that $E[v_t] = 0$. Suppose, for now, that there are no regression parameters. Newton's method for estimating the variance parameters would be

$$\hat{\boldsymbol{\gamma}}^{i+1} = \hat{\boldsymbol{\gamma}}^{i} - \mathbf{H}^{-1}\mathbf{g}, \qquad \textbf{(20-38)}$$

where $\mathbf{H}$ indicates the Hessian and $\mathbf{g}$ is the first derivatives vector. Following Harvey's suggestion (see Section 14.10.3), we will use the method of scoring instead. To do this, we make use of $E[v_t] = 0$ and $E[\varepsilon_t^2/\sigma_t^2] = 1$. After taking expectations in (20-37), the iteration reduces to a linear regression of $v_{*_t} = (1/\sqrt{2})v_t$ on regressors $\mathbf{w}_{*_t} = (1/\sqrt{2})\mathbf{g}_t/\sigma_t^2$. That is,

$$\hat{\boldsymbol{\gamma}}^{i+1} = \hat{\boldsymbol{\gamma}}^{i} + [\mathbf{W}_*'\mathbf{W}_*]^{-1}\mathbf{W}_*'\mathbf{v}_* = \hat{\boldsymbol{\gamma}}^{i} + [\mathbf{W}_*'\mathbf{W}_*]^{-1}\left(\frac{\partial\ln L}{\partial\boldsymbol{\gamma}}\right), \qquad \textbf{(20-39)}$$

where row t of $\mathbf{W}_*$ is $\mathbf{w}_{*t}'$. The iteration has converged when the slope vector is zero, which happens when the first derivative vector is zero. When the iterations are complete, the estimated asymptotic covariance matrix is simply

$$\text{Est.Asy.Var}[\hat{\boldsymbol{\gamma}}] = [\hat{\mathbf{W}}_*'\mathbf{W}_*]^{-1}$$

based on the estimated parameters.

The usefulness of the result just given is that $E[\partial^2\ln L/\partial\boldsymbol{\gamma}\partial\boldsymbol{\beta}']$ is, in fact, zero. Because the expected Hessian is block diagonal, applying the method of scoring to the full parameter vector can proceed in two parts, exactly as it did in Section 14.10.3 for the multiplicative heteroscedasticity model. That is, the updates for the mean and variance

[37] There are three minor errors in Bollerslev's derivation that we note here to avoid the apparent inconsistencies. In his (22), $\frac{1}{2}h_t$ should be $\frac{1}{2}h_t^{-1}$. In (23), $-2h_t^{-2}$ should be $-h_t^{-2}$. In (28), $h\,\partial h/\partial\omega$ should, in each case, be $(1/h)\,\partial h/\partial\omega$. [In his (8), $\alpha_0\alpha_1$ should be $\alpha_0 + \alpha_1$, but this has no implications for our derivation.]

parameter vectors can be computed separately. Consider then the slope parameters, $\boldsymbol{\beta}$. The same type of modified scoring method as used earlier produces the iteration

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{i+1} &= \hat{\boldsymbol{\beta}}^{i} + \left[\sum_{t=1}^{T}\frac{\mathbf{x}_t\mathbf{x}_t'}{\sigma_t^2} + \frac{1}{2}\left(\frac{\mathbf{d}_t}{\sigma_t^2}\right)\left(\frac{\mathbf{d}_t}{\sigma_t^2}\right)'\right]^{-1}\left[\sum_{t=1}^{T}\frac{\mathbf{x}_t\varepsilon_t}{\sigma_t^2} + \frac{1}{2}\left(\frac{\mathbf{d}_t}{\sigma_t^2}\right)v_t\right] \\ &= \hat{\boldsymbol{\beta}}^{i} + \left[\sum_{t=1}^{T}\frac{\mathbf{x}_t\mathbf{x}_t'}{\sigma_t^2} + \frac{1}{2}\left(\frac{\mathbf{d}_t}{\sigma_t^2}\right)\left(\frac{\mathbf{d}_t}{\sigma_t^2}\right)'\right]^{-1}\left(\frac{\partial \ln L}{\partial \boldsymbol{\beta}}\right) \\ &= \hat{\boldsymbol{\beta}}^{i} + \mathbf{h}^{i}, \qquad \textbf{(20-40)}\end{aligned}$$

which has been referred to as a **double-length regression**.[38] The update vector $\mathbf{h}^i$ is the vector of slopes in an augmented or double-length generalized regression,

$$\mathbf{h}^i = [\mathbf{C}'\boldsymbol{\Omega}^{-1}\mathbf{C}]^{-1}[\mathbf{C}'\boldsymbol{\Omega}^{-1}\mathbf{a}], \qquad \textbf{(20-41)}$$

where $\mathbf{C}$ is a $2T \times K$ matrix whose first T rows are the $\mathbf{X}$ from the original regression model and whose next T rows are $(1/\sqrt{2})\mathbf{d}_t'/\sigma_t^2, t = 1, \ldots, T$; $\mathbf{a}$ is a $2T \times 1$ vector whose first T elements are ε_t and whose next T elements are $(1/\sqrt{2})v_t/\sigma_t^2, t = 1, \ldots, T$; and $\boldsymbol{\Omega}$ is a diagonal matrix with $1/\sigma_t^2$ in positions $1, \ldots, T$ and ones below observation T. At convergence, $[\mathbf{C}'\boldsymbol{\Omega}^{-1}\mathbf{C}]^{-1}$ provides the asymptotic covariance matrix for the MLE. The resemblance to the familiar result for the generalized regression model is striking, but note that this result is based on the double-length regression.

The iteration is done simply by computing the update vectors to the current parameters as defined earlier.[39] An important consideration is that to apply the scoring method, the estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are updated simultaneously. That is, one does not use the updated estimate of $\boldsymbol{\gamma}$ in (20-39) to update the weights for the GLS regression to compute the new $\boldsymbol{\beta}$ in (20-40). The same estimates (the results of the prior iteration) are used on the right-hand sides of both (20-39) and (20-40). The remaining problem is to obtain starting values for the iterations. One obvious choice is $\mathbf{b}$, the OLS estimator, for $\boldsymbol{\beta}$, $\mathbf{e}'\mathbf{e}/T = s^2$ for α_0, and zero for all the remaining parameters. The OLS slope vector will be consistent under all specifications. A useful alternative in this context would be to start $\boldsymbol{\alpha}$ at the vector of slopes in the least squares regression of e_t^2, the squared OLS residual, on a constant and q lagged values.[40] As discussed later, an LM test for the presence of GARCH effects is then a byproduct of the first iteration. In principle, the updated result of the first iteration is an **efficient two-step estimator** of all the parameters. But having gone to the full effort to set up the iterations, nothing is gained by not iterating to convergence. One virtue of allowing the procedure to iterate to convergence is that the resulting log-likelihood function can be used in likelihood ratio tests.

[38] See Orme (1990) and Davidson and MacKinnon (1993, Chapter 14).

[39] See Fiorentini et al. (1996) on computation of derivatives in GARCH models.

[40] A test for the presence of ARCH(q) effects against none can be carried out by carrying TR^2 from this regression into a table of critical values for the chi-squared distribution. But in the presence of GARCH effects, this procedure loses its validity.

20.10.4 TESTING FOR GARCH EFFECTS

The preceding development appears fairly complicated. In fact, it is not, because at each step, nothing more than a linear least squares regression is required. The intricate part of the computation is setting up the derivatives. On the other hand, it does take a fair amount of programming to get this far.[41] As Bollerslev suggests, it might be useful to test for GARCH effects first.

The simplest approach is to examine the squares of the least squares residuals. The autocorrelations (correlations with lagged values) of the squares of the residuals provide evidence about ARCH effects. An LM test of ARCH(q) against the hypothesis of no ARCH effects [ARCH(0), the classical model] can be carried out by computing $\chi^2 = TR^2$ in the regression of e_t^2 on a constant and q lagged values. Under the null hypothesis of no ARCH effects, the statistic has a limiting chi-squared distribution with q degrees of freedom. Values larger than the critical table value give evidence of the presence of ARCH (or GARCH) effects.

Bollerslev suggests a Lagrange multiplier statistic that is, in fact, surprisingly simple to compute. The LM test for GARCH(p,0) against GARCH(p,q) can be carried out by referring T times the R^2 in the linear regression defined in (20-42) to the chi-squared critical value with q degrees of freedom. There is, unfortunately, an indeterminacy in this test procedure. The test for ARCH(q) against GARCH(p,q) is exactly the same as that for ARCH(p) against ARCH($p + q$). For carrying out the test, one can use as starting values a set of estimates that includes $\boldsymbol{\delta} = \mathbf{0}$ and any consistent estimators for $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. Then TR^2 for the regression at the initial iteration provides the test statistic.[42] A number of recent papers have questioned the use of test statistics based solely on normality. Wooldridge (1991) is a useful summary with several examples.

Example 20.9 GARCH Model for Exchange Rate Volatility

Bollerslev and Ghysels analyzed the exchange rate data in Appendix Table F20.1 using a GARCH(1,1) model,

$$\begin{aligned} y_t &= \mu + \varepsilon_t, \\ E[\varepsilon_t \mid \varepsilon_{t-1}] &= 0, \\ \text{Var}[\varepsilon_t \mid \varepsilon_{t-1}] &= \sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \delta\sigma_{t-1}^2. \end{aligned}$$

The least squares residuals for this model are simply $e_t = y_t - \bar{y}$. Regression of the squares of these residuals on a constant and 10 lagged squared values using observations 11–1974 produces an $R^2 = 0.09795$. With $T = 1964$, the chi-squared statistic is 192.37, which is larger than the critical value from the table of 18.31. We conclude that there is evidence of GARCH effects in these residuals. The maximum likelihood estimates of the GARCH model are given in Table 20.4. Note the resemblance between the OLS unconditional variance (0.221128) and the estimated equilibrium variance from the GARCH model, 0.2631.

[41] Because this procedure is available as a preprogrammed procedure in many computer programs, including *EViews, Stata, RATS, NLOGIT, Shazam*, and other programs, this warning might itself be overstated.

[42] Bollerslev argues that, in view of the complexity of the computations involved in estimating the GARCH model, it is useful to have a test for GARCH effects. This case is one (as are many other maximum likelihood problems) in which the apparatus for carrying out the test is the same as that for estimating the model. Having computed the LM statistic for GARCH effects, one can proceed to estimate the model just by allowing the program to iterate to convergence. There is no additional cost beyond waiting for the answer.

TABLE 20.4 Maximum Likelihood Estimates of a GARCH(1,1) Model[43]

	μ	α_0	α_1	δ	$\alpha_0/(1-\alpha_1-\delta)$
Estimate	−0.006190	0.01076	0.1531	0.8060	0.2631
Std. Error	0.00873	0.00312	0.0273	0.0302	0.594
t ratio	−0.709	3.445	5.605	26.731	0.443
$\ln L = -1106.61, \ln L_{\text{OLS}} = -1311.09, \bar{y} = -0.01642, s^2 = 0.221128$					

20.10.5 PSEUDO–MAXIMUM LIKELIHOOD ESTIMATION

We now consider an implication of nonnormality of the disturbances. Suppose that the assumption of normality is weakened to only

$$E[\varepsilon_t | \Psi_t] = 0, \qquad E\left[\frac{\varepsilon_t^2}{\sigma_t^2}\middle|\Psi_t\right] = 1, \qquad E\left[\frac{\varepsilon_t^4}{\sigma_t^4}\middle|\Psi_t\right] = \kappa < \infty,$$

where σ_t^2 is as defined earlier. Now the normal log-likelihood function is inappropriate. In this case, the nonlinear (ordinary or weighted) least squares estimator would have the properties discussed in Chapter 7. It would be more difficult to compute than the MLE discussed earlier, however. It has been shown[44] that the pseudo-MLE obtained by maximizing the same log likelihood as if it were correct produces a consistent estimator despite the misspecification.[45] The asymptotic covariance matrices for the parameter estimators must be adjusted, however.

The general result for cases such as this one[46] is that the appropriate asymptotic covariance matrix for the pseudo-MLE of a parameter vector $\boldsymbol{\theta}$ would be

$$\text{Asy. Var}[\hat{\boldsymbol{\theta}}] = \mathbf{H}^{-1}\mathbf{F}\mathbf{H}^{-1}, \tag{20-42}$$

where

$$\mathbf{H} = -E\left[\frac{\partial^2 \ln L}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'}\right],$$

and

$$\mathbf{F} = E\left[\left(\frac{\partial \ln L}{\partial\boldsymbol{\theta}}\right)\left(\frac{\partial \ln L}{\partial\boldsymbol{\theta}'}\right)\right],$$

(i.e., the BHHH estimator), and ln L is the used but inappropriate log-likelihood function. For current purposes, **H** and **F** are still block diagonal, so we can treat the mean and variance parameters separately. In addition, $E[v_t]$ is still zero, so the second derivative terms in both blocks are quite simple. (The parts involving $\partial^2\sigma_t^2/\partial\boldsymbol{\gamma}\,\partial\boldsymbol{\gamma}'$ and

[43] These data have become a standard data set for the evaluation of software for estimating GARCH models. The values given are the benchmark estimates. Standard errors differ substantially from one method to the next. Those given are the Bollerslev and Wooldridge (1992) results. See McCullough and Renfro (1999).

[44] See White (1982a) and Weiss (1982).

[45] White (1982a) gives some additional requirements for the true underlying density of ε_t. Gourieroux, Monfort, and Trognon (1984) also consider the issue. Under the assumptions given, the expectations of the matrices in (20-36) and (20-41) remain the same as under normality. The consistency and asymptotic normality of the pseudo-MLE can be argued under the logic of GMM estimators.

[46] See Gourieroux, Monfort, and Trognon (1984).

$\partial^2\sigma_t^2/\partial\boldsymbol{\beta}\,\partial\boldsymbol{\beta}'$ fall out of the expectation.) Taking expectations and inserting the parts produces the corrected asymptotic covariance matrix for the variance parameters,

$$\text{Est.Asy.Var}[\hat{\boldsymbol{\gamma}}_{\text{PMLE}}] = [\mathbf{W}_*'\mathbf{W}_*]^{-1}\mathbf{B}'\mathbf{B}[\mathbf{W}_*'\mathbf{W}_*]^{-1},$$

where the rows of **W*** are defined in (20-39) and those of **B** are in (20-37). For the slope parameters, the adjusted asymptotic covariance matrix would be

$$\text{Est.Asy.Var}[\hat{\boldsymbol{\beta}}_{\text{PMLE}}] = [\mathbf{C}'\boldsymbol{\Omega}^{-1}\mathbf{C}]^{-1}\left[\sum_{t=1}^{T}\mathbf{b}_t\,\mathbf{b}_t'\right][\mathbf{C}'\boldsymbol{\Omega}^{-1}\mathbf{C}]^{-1},$$

where the outer matrix is defined in (20-41) and, from the first derivatives given in (20-37) and (20-40),[47]

$$\mathbf{b}_t = \frac{\mathbf{x}_t\varepsilon_t}{\sigma_t^2} + \frac{1}{2}\left(\frac{v_t}{\sigma_t^2}\right)\mathbf{d}_t.$$

20.11 SUMMARY AND CONCLUSIONS

This chapter has examined the generalized regression model with serial correlation in the disturbances. We began with some general results on analysis of time-series data. When we consider dependent observations and serial correlation, the laws of large numbers and central limit theorems used to analyze independent observations no longer suffice. We presented some useful tools that extend these results to time-series settings. We then considered estimation and testing in the presence of autocorrelation. As usual, OLS is consistent but inefficient. The Newey–West estimator is a robust estimator for the asymptotic covariance matrix of the OLS estimator. This pair of estimators also constitute the GMM estimator for the regression model with autocorrelation. We then considered two-step feasible generalized least squares and maximum likelihood estimation for the special case usually analyzed by practitioners, the AR(1) model. The model with a correction for autocorrelation is a restriction on a more general model with lagged values of both dependent and independent variables. We considered a means of testing this specification as an alternative to fixing the problem of autocorrelation. The final section, on ARCH and GARCH effects, describes an extension of the models of autoregression to the conditional variance of ε as opposed to the conditional mean. This model embodies elements of both autocorrelation and heteroscedasticity. The set of methods plays a fundamental role in the modern analysis of volatility in financial data.

Key Terms and Concepts

- AR(1)
- ARCH
- ARCH-in-mean
- Asymptotic negligibility
- Asymptotic normality
- Autocorrelation coefficient
- Autocorrelation function
- Autocorrelation matrix
- Autocovariance
- Autocovariance matrix
- Autoregressive form
- Autoregressive processes

[47] McCullough and Renfro (1999) examined several approaches to computing an appropriate asymptotic covariance matrix for the GARCH model, including the conventional Hessian and BHHH estimators and three sandwich-style estimators, including the one suggested earlier and two based on the method of scoring suggested by Bollerslev and Wooldridge (1992). None stands out as obviously better, but the Bollerslev and QMLE estimator based on an actual Hessian appears to perform well in Monte Carlo studies.

- Cochrane–Orcutt estimator
- Covariance stationarity
- Double-length regression
- Durbin–Watson test
- Efficient two-step estimator
- Ergodicity
- Expectations-augmented Phillips curve
- First-order autoregression
- Innovation
- LM test
- Martingale sequence
- Martingale difference sequence
- Moving-average form
- Moving-average process
- Newey–West autocorrelation consistent covariance estimator
- Partial difference
- Prais–Winsten estimator
- Pseudo-differences
- *Q* test
- Quasi differences
- Random walk
- Stationarity
- Stationarity conditions
- Summability
- Time-series process
- Time window
- Weakly stationary
- White noise
- Yule–Walker equations

Exercises

1. Does first differencing reduce autocorrelation? Consider the models $y_t = \boldsymbol{\beta}'\mathbf{x}_t + \varepsilon_t$, where $\varepsilon_t = \rho\varepsilon_{t-1} + u_t$ and $\varepsilon_t = u_t - \lambda u_{t-1}$. Compare the autocorrelation of ε_t in the original model with that of v_t in $y_t - y_{t-1} = \boldsymbol{\beta}'(\mathbf{x}_t - \mathbf{x}_{t-1}) + v_t$, where $v_t = \varepsilon_t - \varepsilon_{t-1}$.
2. Derive the disturbance covariance matrix for the model

$$y_t = \boldsymbol{\beta}'\mathbf{x}_t + \varepsilon_t,$$
$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t - \lambda u_{t-1}.$$

   What parameter is estimated by the regression of the OLS residuals on their lagged values?
3. It is commonly asserted that the Durbin–Watson statistic is only appropriate for testing for first-order autoregressive disturbances. The Durbin–Watson statistic estimates $2(1 - \rho)$ where ρ is the first-order autocorrelation of the residuals. What combination of the coefficients of the model is estimated by the Durbin–Watson statistic in each of the following cases: AR(1), AR(2), MA(1)? In each case, assume that the regression model does not contain a lagged dependent variable. Comment on the impact on your results of relaxing this assumption.

Applications

1. The data used to fit the expectations augmented Phillips curve in Example 20.3 are given in Appendix Table F5.2. Using these data, reestimate the model given in the example. Carry out a formal test for first-order autocorrelation using the LM statistic. Then, reestimate the model using an AR(1) model for the disturbance process. Because the sample is large, the Prais–Winsten and Cochrane–Orcutt estimators should give essentially the same answer. Do they? After fitting the model, obtain the transformed residuals and examine them for first-order autocorrelation. Does the AR(1) model appear to have adequately fixed the problem?
2. Data for fitting an improved Phillips curve model can be obtained from many sources, including the Bureau of Economic Analysis's (BEA) own Web site, www.economagic.com, and so on. Obtain the necessary data and expand the model of

Example 20.3. Does adding additional explanatory variables to the model reduce the extreme pattern of the OLS residuals that appears in Figure 20.3?

3. (This exercise requires appropriate computer software. The computations required can be done with *RATS*, *EViews*, *Stata*, *LIMDEP*, and a variety of other software using only preprogrammed procedures.) Quarterly data on the consumer price index for 1950.1 to 2000.4 are given in Appendix Table F5.2. Use these data to fit the model proposed by Engle and Kraft (1983). The model is

$$\pi_t = \beta_0 + \beta_1\pi_{t-1} + \beta_2\pi_{t-2} + \beta_3\pi_{t-3} + \beta_4\pi_{t-4} + \varepsilon_t,$$

where $\pi_t = 100 \ln[p_t/p_{t-1}]$ and p_t is the price index.

a. Fit the model by ordinary least squares, then use the tests suggested in the text to see if ARCH effects appear to be present.

b. The authors fit an ARCH(8) model with declining weights,

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{8}\left(\frac{9-i}{36}\right)\varepsilon_{t-i}^2.$$

Fit this model. If the software does not allow constraints on the coefficients, you can still do this with a two-step least squares procedure, using the least squares residuals from the first step. What do you find?

c. Bollerslev (1986) recomputed this model as a GARCH(1, 1). Use the GARCH(1, 1) to form and refit your model.