12
SERIAL CORRELATION

12.1 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 12.1 Money Demand Equation

Table F5.1 contains quarterly data from 1950.1 to 2000.4 on the U.S. money stock (M1) and output (real GDP) and the price level (CPI_U). Consider a simple (extremely) model of money demand,[1]

$$\ln \mathrm{M1}_t = \beta_1 + \beta_2 \ln \mathrm{GDP}_t + \beta_3 \ln CPI_t + \varepsilon_t$$

A plot of the least squares residuals is shown in Figure 12.1. 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 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 is evident in Figure 12.1. One might argue that this pattern is the result of an obviously naive 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 12.2 shows an obvious case.

Example 12.2 Autocorrelation Induced by Misspecification of the Model

In Examples 2.3 and 7.6, we examined yearly time-series data on the U.S. gasoline market from 1960 to 1995. 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. Other 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 12.2 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 1960 to 1974 and 1975 to 1995 are produced by separate unrestricted regressions. This latter set of residuals is almost nonautocorrelated. (Note also that the range of variation of the residuals falls as

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

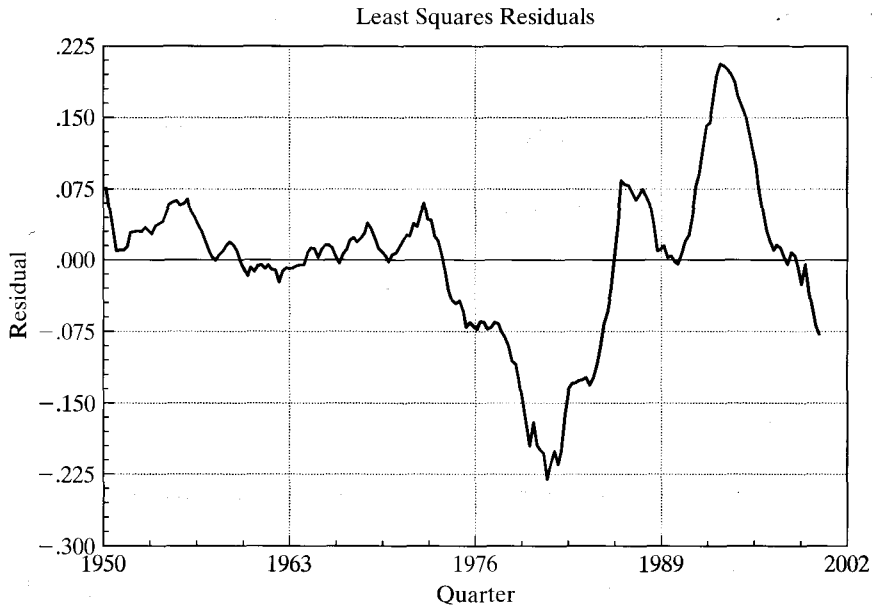


FIGURE 12.1 Autocorrelated Residuals.

the model is improved, i.e., as its fit improves.) The full equation is

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

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

Example 12.3 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. Current research [e.g., Staiger et al. (1996)] 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. We will revisit this in some detail in Example 19.2. For the present, we'll

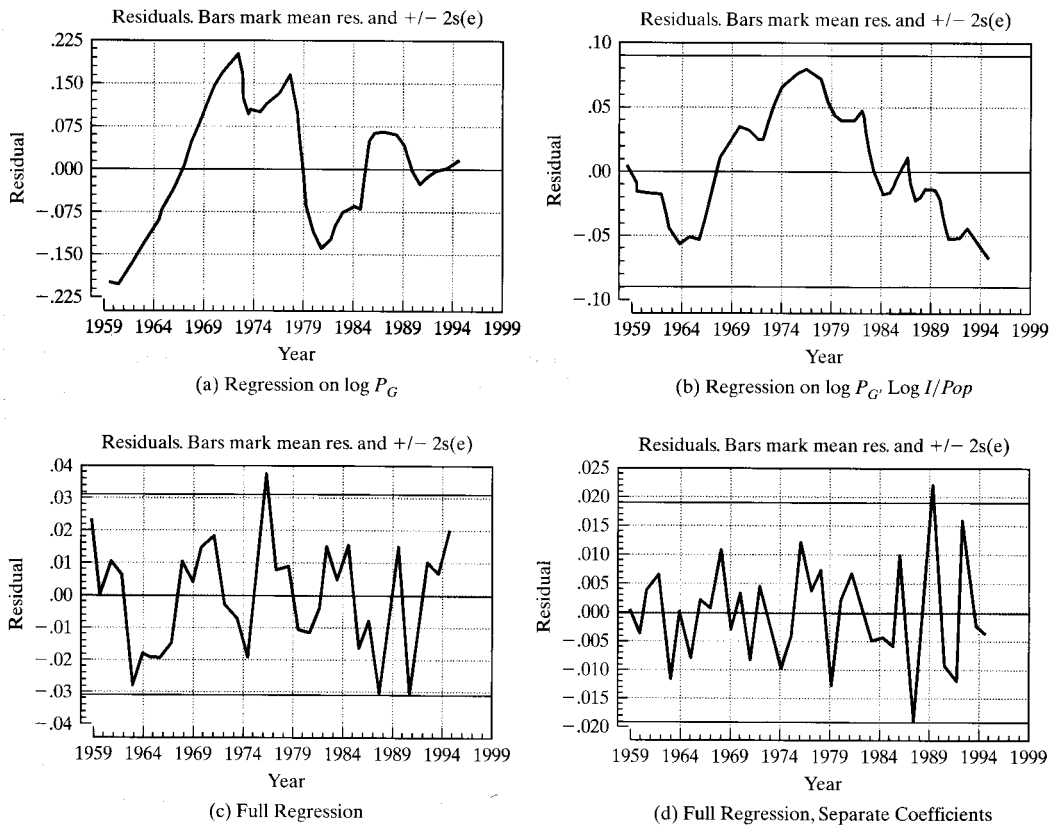


FIGURE 12.2 Residual Plots for Misspecified Models.

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 1950.1–2000.4 data in Table F5.1 that were used in Example 12.1 (minus two quarters for the change in the rate of inflation). Least squares estimates (with standard errors in parentheses) are as follows:

$$\Delta p_t - \Delta p_{t-1} = 0.49189 - 0.090136\, u_t + e_t$$
$$(0.7405) \quad (0.1257) \quad R^2 = 0.002561, \; T = 201.$$

The implied estimate of the natural rate of unemployment is 5.46 percent, which is in line with other recent estimates. The estimated asymptotic covariance of b_1 and b_2 is −0.08973. Using the delta method, we obtain a standard error of 2.2062 for this estimate, so a confidence interval for the natural rate is 5.46 percent ± 1.96 (2.21 percent) = (1.13 percent, 9.79 percent) (which 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 −0.4263 with a highly significant t ratio of −6.725. We thus conclude that the

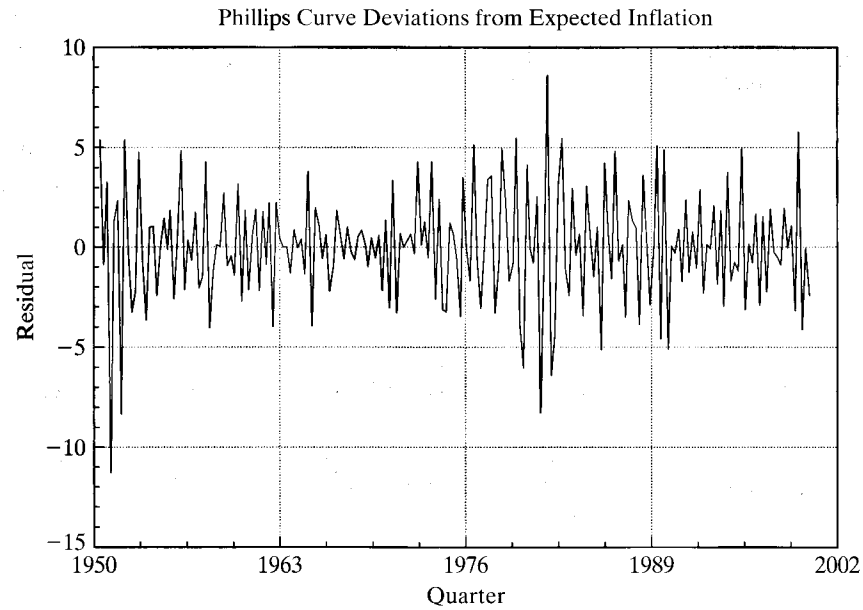


FIGURE 12.3 Negatively Autocorrelated Residuals.

residuals (and, apparently, the disturbances) in this model are highly negatively autocorrelated. This is consistent with the striking pattern in Figure 12.3.

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 Chapter 11, we examined 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—see, for example, "A Simple Message to Autocorrelation Correctors: Don't" [Mizon (1995).]

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.

12.2 THE ANALYSIS OF TIME-SERIES DATA

The treatment in this chapter will be the first structured analysis of time series data in the text. (We had a brief encounter in Section 5.3 where we established some conditions

under which moments of time series data would converge.) Time-series analysis requires some revision of the interpretation of both data generation and sampling that we have maintained thus far.

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 12.1 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

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

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

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

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.[2] Finally, the same results can be obtained for nonzero β_2 if it is further assumed that x_t is a weakly stationary series.[3]

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

$$y_t = \beta_1 + \beta_3(\beta_1 + \beta_3 y_{t-2} + \varepsilon_{t-1}) + \varepsilon_t \qquad \textbf{(12-1)}$$

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 (12-1), 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 are 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

[2]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 20.

[3]See Section 12.4.1 on the stationarity assumption.

"**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 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 and in Chapters 19 and 20.

12.3 DISTURBANCE PROCESSES

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

12.3.1 CHARACTERISTICS OF DISTURBANCE PROCESSES

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

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

where $\sigma^2\boldsymbol{\Omega}$ is a full, positive definite matrix with a constant $\sigma^2 = \text{Var}[\varepsilon_t \mid \mathbf{X}]$ 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. We define the **autocovariances:**

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

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

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

We can then write

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

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

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

(Note that the matrix $\Gamma = \gamma_0 \mathbf{R}$ is the same as $\sigma^2 \mathbf{\Omega}$. The name change conforms to standard usage in the literature.) 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.[4] 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 $s > 1$.

12.3.2 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.[5] 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—as in the example below.

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

$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t, \tag{12-2}$$

where

$$E[u_t] = 0,$$

$$E[u_t^2] = \sigma_u^2,$$

[4] This model is considered in more detail in Chapter 20.

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

and

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

By repeated substitution, we have

$$\varepsilon_t = u_t + \rho u_{t-1} + \rho^2 u_{t-2} + \cdots. \tag{12-3}$$

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 12.3).

Since the successive values of u_t are uncorrelated, the variance of ε_t is the variance of the right-hand side of (12-3):

$$\text{Var}[\varepsilon_t] = \sigma_u^2 + \rho^2\sigma_u^2 + \rho^4\sigma_u^2 + \cdots. \tag{12-4}$$

To proceed, a restriction must be placed on ρ,

$$|\rho| < 1, \tag{12-5}$$

because otherwise, the right-hand side of (12-4) will become infinite. This result is the stationarity assumption discussed earlier. With (12-5), which implies that $\lim_{s\to\infty} \rho^s = 0$, $E[\varepsilon_t] = 0$ and

$$\text{Var}[\varepsilon_t] = \frac{\sigma_u^2}{1-\rho^2} = \sigma_\varepsilon^2. \tag{12-6}$$

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

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

as $\text{Cov}[u_t, \varepsilon_s] = 0$ if $t > s$. With stationarity, $\text{Var}[\varepsilon_{t-1}] = \text{Var}[\varepsilon_t]$, which implies (12-6). Proceeding in the same fashion,

$$\text{Cov}[\varepsilon_t, \varepsilon_{t-1}] = E[\varepsilon_t\varepsilon_{t-1}] = E[\varepsilon_{t-1}(\rho\varepsilon_{t-1} + u_t)] = \rho \text{Var}[\varepsilon_{t-1}] = \frac{\rho\sigma_u^2}{1-\rho^2}. \tag{12-7}$$

By repeated substitution in (12-2), 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, since ε_s is not correlated with any u_t for which $t > s$ (i.e., any subsequent u_t), it follows that

$$\text{Cov}[\varepsilon_t, \varepsilon_{t-s}] = E[\varepsilon_t\varepsilon_{t-s}] = \frac{\rho^s\sigma_u^2}{1-\rho^2}. \tag{12-8}$$

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

$$\text{Corr}[\varepsilon_t, \varepsilon_{t-s}] = \rho_s = \rho^s. \tag{12-9}$$

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

$$\sigma^2\mathbf{\Omega} = \frac{\sigma_u^2}{1-\rho^2}\begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \cdots & \rho^{T-1} \\ \rho & 1 & \rho & \rho^2 & \cdots & \rho^{T-2} \\ \rho^2 & \rho & 1 & \rho & \cdots & \rho^{T-3} \\ \vdots & \vdots & \vdots & \vdots & \cdots & \rho \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \cdots & \rho & 1 \end{bmatrix}. \tag{12-10}$$

12.4 SOME ASYMPTOTIC RESULTS FOR ANALYZING TIME SERIES DATA

Since $\mathbf{\Omega}$ 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 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

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

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 12.1, is surely not the case here. Thus, we require a different tool for this result. We can expand the matrix $\mathbf{Q}_T^*$ as

$$\mathbf{Q}_T^* = \frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}\rho_{ts}\mathbf{x}_t\mathbf{x}_s', \tag{12-11}$$

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 die off 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 $T \to \infty$.

Asymptotic normality of the least squares and GLS estimators will depend on the behavior of sums such as

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

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 which are used to reach these conclusions.

12.4.1 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, $\varepsilon_t = \rho\varepsilon_{t-1} + u_t$, in order 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 12.1 Strong Stationarity
A time series process, $\{z_t\}_{t=-\infty}^{t=\infty}$ is strongly stationary, or "stationary" if the joint probability distribution of any set of k observations in the sequence, $[z_t, z_{t+1}, \ldots, z_{t+k}]$ is the same regardless of the origin, t, in the time scale.

For example, in (12-2), 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 12.2 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, though it requires less since 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 12.3 Ergodicity

A time series process, $\{\mathbf{z_t}\}_{t=-\infty}^{t=\infty}$ *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* $g: \boldsymbol{R}^b \rightarrow \boldsymbol{R}^1$,

$$\lim_{k\to\infty} |E[f(z_t, z_{t+1}, \ldots, z_{t+a})g(z_{t+k}, z_{t+k+1}, \ldots, z_{t+k+b})|$$
$$= |E[f(z_t, z_{t+1}, \ldots, z_{t+a})]\,|E[g(z_{t+k}, z_{t+k+1}, \ldots, z_{t+k+b})]|\,.$$

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.[6] The analysis relies heavily on the following theorem:

THEOREM 12.1 The Ergodic Theorem

If $\{z_t\}_{t=-\infty}^{t=\infty}$ *is a time-series process which is stationary and ergodic and* $E[|z_t|]$ *is a finite constant and* $E[z_t] = \mu$, *and if* $\bar{z}_T = (1/T)\sum_{t=1}^{T} z_t$, *then* $\bar{z}_T \xrightarrow{a.s.} \mu$. *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

[6]Much of the analysis in later chapters will encounter 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 which 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 (1979). White (2001) provides a concise analysis of many of these concepts as used in econometrics, and some useful commentary.

large numbers have required sums of independent observations. But, in this context, by design, observations are distinctly not independent.

In order for this result to be useful, we will require an extension.

THEOREM 12.2 Ergodicity of Functions
If $\{z_t\}_{t=-\infty}^{t=\infty}$ is a time series process which is 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 $\{\mathbf{z}_t\}_{t=-\infty}^{t=\infty}$. (See White (2001, pp. 44–45) for discussion.)

Theorem 12.2 produces the results we need to characterize the least squares (and other) estimators. In particular, our minimal assumptions about the data are

ASSUMPTION 12.1 **Ergodic Data Series:** In the regression model, $y_t = \mathbf{x}_t'\boldsymbol{\beta} + \varepsilon_t$, $[\mathbf{x}_t, \varepsilon_t]_{t=-\infty}^{t=\infty}$ is a jointly stationary and ergodic process.

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

12.4.2 CONVERGENCE TO NORMALITY—A CENTRAL LIMIT THEOREM

In order 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

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

As noted earlier, we cannot invoke the familiar central limit theorems (Lindberg–Levy, Lindberg–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 12.4 Martingale Sequence
A vector sequence $\mathbf{z}_t$ is a martingale sequence if $E[\mathbf{z}_t \mid \mathbf{z}_{t-1}, \mathbf{z}_{t-2}, \ldots] = \mathbf{z}_{t-1}$.

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

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

DEFINITION 12.5 Martingale Difference Sequence
A vector sequence $\mathbf{z}_t$ *is a martingale difference sequence if* $E[\mathbf{z}_t \mid \mathbf{z}_{t-1}, \mathbf{z}_{t-2}, \ldots] = \mathbf{0}$.

With Definition 12.5, we have the following broadly encompassing result:

THEOREM 12.3 Martingale Difference Central Limit Theorem
If $\mathbf{z}_t$ *is a vector valued stationary and ergodic martingale difference sequence, with* $E[\mathbf{z}_t\mathbf{z}_t'] = \boldsymbol{\Sigma}$, *where* $\boldsymbol{\Sigma}$ *is a finite positive definite matrix, and if* $\bar{\mathbf{z}}_T = (1/T)\sum_{t=1}^{T} \mathbf{z}_t$, *then* $\sqrt{T}\,\bar{\mathbf{z}}_T \xrightarrow{d} N[\mathbf{0}, \boldsymbol{\Sigma}]$. *(For discussion, see Davidson and MacKinnon (1993, Sections. 4.7 and 4.8.)*[7]

Theorem 12.3 is a generalization of the Lindberg–Levy Central Limit Theorem. It is not yet broad enough to cover cases of autocorrelation, but it does go beyond Lindberg–Levy, for example, in extending to the GARCH model of Section 11.8. [Forms of the theorem which surpass Lindberg–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.] 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 which 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. [See, e.g., Hamilton (1994, pp. 208–212).]

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

[7]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. 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).

autocorrelation in ε_t.[8] 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:[9]

1. **Summability of autocovariances:** With dependent observations,

$$\lim_{T\to\infty} \text{Var}[\sqrt{T}\bar{\mathbf{z}}] = \sum_{t=0}^{\infty}\sum_{s=0}^{\infty} \text{Cov}[\mathbf{z}_t, \mathbf{z}_s'] = \sum_{k=-\infty}^{\infty} \boldsymbol{\Gamma}_k = \boldsymbol{\Gamma}^*$$

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 (12-11). If the sum is to be finite, then the $k = 0$ term must be finite, which gives us a necessary condition

$$E[\mathbf{z}_t\mathbf{z}_t'] = \boldsymbol{\Gamma}_0, \text{ a finite matrix.}$$

2. **Asymptotic uncorrelatedness:** $E[\mathbf{z}_t \mid \mathbf{z}_{t-k}, \mathbf{z}_{t-k-1}, \ldots]$ 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 $E[\mathbf{z}_t] = \mathbf{0}$.

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

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

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 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 to illustrate this point. If $z_t = \rho z_{t-1} + u_t$, then

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

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 (12-3). You can verify that if $|\rho| < 1$, the negligibility condition will be met.

[8] 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).

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

With all this machinery in place, we now have the theorem we will need:

THEOREM 12.4 Gordin's Central Limit Theorem
If conditions (1) – (3) listed above are met, then $\sqrt{T}\bar{\mathbf{z}}_T \xrightarrow{d} N[\mathbf{0}, \mathbf{\Gamma}^*]$.

We will be able to employ these tools when we consider the least squares, IV and GLS estimators in the discussion to follow.

12.5 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 10 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 which have long periods and little cyclical variation, such as aggregate output series. As might be expected, the greater is 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.

12.5.1 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 12.1 and 12.2, $(1/T)(\mathbf{X}'\mathbf{X})$ 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 12.1 and 12.2 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 12.3 for $\sqrt{T}\,\bar{\mathbf{w}}$. 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 12.4 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 12.4 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 least squares will no longer be unbiased or consistent. To take the simplest case, suppose that

$$\begin{aligned} y_t &= \beta y_{t-1} + \varepsilon_t, \\ \varepsilon_t &= \rho \varepsilon_{t-1} + u_t. \end{aligned} \tag{12-12}$$

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 (12-12) by subtracting ρy_{t-1} from y_t. Then,

$$y_t = (\beta + \rho) y_{t-1} - \beta\rho y_{t-2} + u_t \tag{12-13}$$

which is a classical regression with stochastic regressors. Since 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 (12-13) imply the following relationships.

$$\begin{aligned} \gamma_0 &= \rho_1\gamma_1 + \rho_2\gamma_2 + \sigma_u^2 \\ \gamma_1 &= \rho_1\gamma_0 + \rho_2\gamma_1 \\ \gamma_2 &= \rho_1\gamma_1 + \rho_2\gamma_0 \end{aligned} \tag{12-14}$$

(These are the **Yule Walker equations** for this model. See Section 20.2.3.) 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 (12-14) for γ_0, γ_1 and γ_2 in terms of ρ_1, ρ_2 and σ_u^2 produces

$$\text{plim } b = \frac{\beta + \rho}{1 + \beta\rho}. \tag{12-15}$$

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.

12.5.2 ESTIMATING THE VARIANCE OF THE LEAST SQUARES ESTIMATOR

As usual, $s^2(\mathbf{X}'\mathbf{X})^{-1}$ is an inappropriate estimator of $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{\Omega}\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}$, both because s^2 is a biased estimator of σ^2 and because the matrix is incorrect. Generalities

TABLE 12.1 Robust Covariance Estimation

Variable	*OLS Estimate*	*OLS SE*	*Corrected SE*
Constant	0.7746	0.0335	0.0733
ln Output	0.2955	0.0190	0.0394
ln CPI	0.5613	0.0339	0.0708

$R^2 = 0.99655$, $d = 0.15388$, $r = 0.92331$.

are scarce, but in general, for economic time series which 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 $\mathbf{\Omega}$ 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. Suppose that the form of the autocorrelation is unknown. Then, a direct estimator of $\mathbf{\Omega}$ or $\mathbf{\Omega}(\boldsymbol{\theta})$ is not available. The problem is estimation of

$$\mathbf{\Sigma} = \frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}\rho_{|t-s|}\mathbf{x}_t\mathbf{x}_s'. \tag{12-16}$$

Following White's suggestion for heteroscedasticity, Newey and West's (1987a) robust, consistent estimator for autocorrelated disturbances with an unspecified structure is

$$\mathbf{S}_* = \mathbf{S}_0 + \frac{1}{T}\sum_{j=1}^{L}\sum_{t=j+1}^{T}\left(1 - \frac{j}{L+1}\right)e_t e_{t-j}[\mathbf{x}_t\mathbf{x}_{t-j}' + \mathbf{x}_{t-j}\mathbf{x}_t'], \tag{12-17}$$

[See (10-16) in Section 10.3.] The maximum lag L must be determined in advance to be large enough that autocorrelations at lags longer than L are small enough to ignore. For a moving-average process, this value can be expected to be a relatively small number. For autoregressive processes or mixtures, however, the autocorrelations are never zero, and the researcher must make a judgment as to how far back it is necessary to go.[10]

Example 12.4 Autocorrelation Consistent Covariance Estimation

For the model shown in Example 12.1, the regression results with the uncorrected standard errors and the Newey-West autocorrelation robust covariance matrix for lags of 5 quarters are shown in Table 12.1. The effect of the very high degree of autocorrelation is evident.

[10]Davidson and MacKinnon (1993) give further discussion. Current practice is to use the smallest integer greater than or equal to $T^{1/4}$.

12.6 GMM ESTIMATION

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

$$\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t(y_t - \mathbf{x}_t'\hat{\boldsymbol{\beta}}_{GMM}) = \frac{1}{T}\mathbf{X}'\hat{\boldsymbol{\varepsilon}}(\hat{\boldsymbol{\beta}}_{GMM}) = \bar{\mathbf{m}}(\hat{\boldsymbol{\beta}}_{GMM}) = \mathbf{0}. \tag{12-18}$$

The estimator is obtained by minimizing

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

where $\mathbf{W}$ is a positive definite weighting matrix. The optimal weighting matrix would be

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

which is the inverse of

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

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 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 12.5.2 to construct the asymptotic covariance matrix. We will require the assumptions in Section 12.4 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 12.9.4.

12.7 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

$$\begin{aligned} e_t &= re_{t-1} + v_t, \\ e_t &= y_t - \mathbf{x}_t'\mathbf{b}. \\ r &= \left(\sum_{t=2}^{T} e_t e_{t-1}\right) \Big/ \left(\sum_{t=1}^{T} e_t^2\right) \end{aligned} \tag{12-19}$$

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 (12-19) 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 12.3. The tests we consider here are refinements of this approach.

12.7.1 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

$$\text{LM} = T\left(\frac{\mathbf{e}'\mathbf{X}_0(\mathbf{X}_0'\mathbf{X}_0)^{-1}\mathbf{X}_0'\mathbf{e}}{\mathbf{e}'\mathbf{e}}\right) = TR_0^2 \tag{12-20}$$

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.[11] Since $\mathbf{X}'\mathbf{e} = \mathbf{0}$, 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.

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

An alternative test which 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

$$Q = T\sum_{j=1}^{P} r_j^2, \tag{12-21}$$

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

$$Q' = T(T+2)\sum_{j=1}^{P} \frac{r_j^2}{T-j}. \tag{12-22}$$

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, since it does not condition on $\mathbf{x}_t$, the

[11] 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.

Box–Pierce test is less powerful than the LM test when the null hypothesis is false, as intuition might suggest.

12.7.3 THE DURBIN–WATSON TEST

The Durbin–Watson statistic[12] was the first formal procedure developed for testing for autocorrelation using the least squares residuals. The test statistic is

$$d = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2} = 2(1 - r) - \frac{e_1^2 + e_T^2}{\sum_{t=1}^{T} e_t^2} \tag{12-23}$$

where r is the same first order autocorrelation which 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. Useable critical values which depend only on T and K are presented in tables such as that at the end of this book. 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.

12.7.4 TESTING IN THE PRESENCE OF A LAGGED DEPENDENT VARIABLES

The Durbin–Watson test is not likely to be valid when there is a lagged dependent variable in the equation.[13] 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. 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

$$h = r\sqrt{T/(1 - Ts_c^2)}, \tag{12-24}$$

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.

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

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

12.7.5 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.

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 β 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.

12.8 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,

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y}), \tag{12-25}$$

and the estimate of its sampling variance,

$$\text{Est. Var}[\hat{\boldsymbol{\beta}}] = \hat{\sigma}_\varepsilon^2[\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}]^{-1}, \tag{12-26}$$

where

$$\hat{\sigma}_\varepsilon^2 = \frac{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\boldsymbol{\Omega}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{T} \tag{12-27}$$

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

$$\mathbf{y}_* = \begin{bmatrix} \sqrt{1-\rho^2}\,y_1 \\ y_2 - \rho y_1 \\ y_3 - \rho y_2 \\ \vdots \\ y_T - \rho y_{T-1} \end{bmatrix}, \quad \mathbf{X}_* = \begin{bmatrix} \sqrt{1-\rho^2}\,\mathbf{x}_1 \\ \mathbf{x}_2 - \rho \mathbf{x}_1 \\ \mathbf{x}_3 - \rho \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_T - \rho \mathbf{x}_{T-1} \end{bmatrix}. \tag{12-28}$$

These transformations are variously labeled **partial differences, quasi differences,** or **pseudodifferences.** 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 (12-6)]. 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,

$$\varepsilon_t = \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + u_t, \tag{12-29}$$

the transformed data for generalized least squares are obtained by

$$\begin{aligned} \mathbf{z}_{*1} &= \left[\frac{(1+\theta_2)\left[(1-\theta_2)^2 - \theta_1^2\right]}{1-\theta_2}\right]^{1/2} \mathbf{z}_1, \\ \mathbf{z}_{*2} &= \left(1-\theta_2^2\right)^{1/2}\mathbf{z}_2 - \frac{\theta_1\left(1-\theta_1^2\right)^{1/2}}{1-\theta_2}\mathbf{z}_1, \\ \mathbf{z}_{*t} &= \mathbf{z}_t - \theta_1\mathbf{z}_{t-1} - \theta_2\mathbf{z}_{t-2}, \quad t > 2, \end{aligned} \tag{12-30}$$

where $\mathbf{z}_t$ is used for y_t or $\mathbf{x}_t$. The transformation becomes progressively more complex for higher-order processes.[14]

Note that in both the AR(1) and AR(2) models, the transformation to 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. Since 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 12.4. We begin by assuming that $\{\mathbf{x}_t, \varepsilon_t\}$ are

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

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 series. 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.

12.9 ESTIMATION WHEN Ω IS UNKNOWN

For an unknown Ω, there are a variety of approaches. Any consistent estimator of $\Omega(\rho)$ will suffice—recall from Theorem (10.8) in Section 10.5.2, all that is needed for efficient estimation of $\boldsymbol{\beta}$ is a consistent estimator of $\Omega(\rho)$. The complication arises, as might be expected, in estimating the autocorrelation parameter(s).

12.9.1 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. Since $\mathbf{b}$ 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 (12-25)–(12-28). 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. Since 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 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 (12-31)].

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

$$\ln L = -\frac{\sum_{t=1}^{T} u_t^2}{2\sigma_u^2} + \frac{1}{2}\ln(1-\rho^2) - \frac{T}{2}\left(\ln 2\pi + \ln \sigma_u^2\right), \qquad \textbf{(12-31)}$$

where, as before, the first observation is computed differently from the others using (12-28). For a given value of ρ, the maximum likelihood estimators of $\boldsymbol{\beta}$ and σ_u^2 are the usual ones, GLS and the mean squared residual using the transformed data. The problem is estimation of ρ. One possibility is to search the range $-1 < \rho < 1$ for the value that with the implied estimates of the other parameters maximizes $\ln L$. [This is Hildreth and Lu's (1960) approach.] Beach and MacKinnon (1978a) argue that this way to do the search is very inefficient and have devised a much faster algorithm. Omitting the first observation and adding an approximation at the lower right corner produces

the standard approximations to the asymptotic variances of the estimators,

$$\begin{aligned}\text{Est.Asy. Var}[\hat{\boldsymbol{\beta}}_{ML}] &= \hat{\sigma}^2_{\varepsilon,ML}\left[\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}_{ML}\mathbf{X}\right]^{-1},\\ \text{Est.Asy. Var}[\hat{\sigma}^2_{u,ML}] &= 2\hat{\sigma}^4_{u,ML}/T,\\ \text{Est.Asy. Var}[\hat{\rho}_{ML}] &= \left(1-\hat{\rho}^2_{ML}\right)/T.\end{aligned} \tag{12-32}$$

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 Beach and MacKinnon's maximum likelihood estimator is probably the most common choice.

12.9.2 AR(2) DISTURBANCES

Maximum likelihood procedures for most other disturbance processes are exceedingly complex. Beach and MacKinnon (1978b) have derived an algorithm for AR(2) disturbances. For higher-order autoregressive models, maximum likelihood estimation is presently impractical, but the two-step estimators can easily be extended. For models of the form

$$\varepsilon_t = \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \cdots + \theta_p\varepsilon_{t-p} + u_t, \tag{12-33}$$

a simple approach for estimation of the autoregressive parameters is to use the following method: Regress e_t on $e_{t-1}, \ldots, e_{t-p}$, to obtain consistent estimates of the autoregressive parameters. With the estimates of $\rho_1, \ldots, \rho_p$ in hand, the Cochrane–Orcutt estimator can be obtained. If the model is an AR(2), the full FGLS procedure can be used instead. The least squares computations for the transformed data provide (at least asymptotically) the appropriate estimates of σ_u^2 and the covariance matrix of $\hat{\boldsymbol{\beta}}$. As before, iteration is possible but brings no gains in efficiency.

12.9.3 APPLICATION: ESTIMATION OF A MODEL WITH AUTOCORRELATION

A restricted version of the model for the U.S. gasoline market that appears in Example 12.2 is

$$\ln\frac{G_t}{pop_t} = \beta_1 + \beta_2 \ln P_{G,t} + \beta_3 \ln\frac{I_t}{pop_t} + \beta_4 \ln P_{NC,t} + \beta_5 \ln P_{UC,t} + \varepsilon_t.$$

The results in Figure 12.2 suggest that the specification above may be incomplete, and, if so, there may be autocorrelation in the disturbance in this specification. Least squares estimation of the equation produces the results in the first row of Table 12.2. The first 5 autocorrelations of the least squares residuals are 0.674, 0.207, −0.049, −0.159, and −0.158. This produces Box–Pierce and Box–Ljung statistics of 19.816 and 21.788, 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

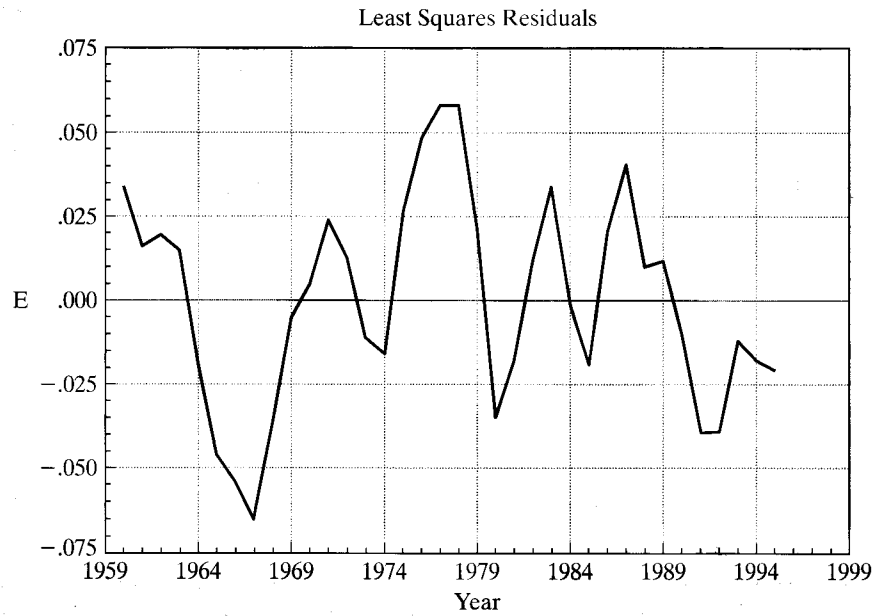
TABLE 12.2 Parameter Estimates (Standard Errors in Parentheses)

	β_1	β_2	β_3	β_4	β_5	ρ
OLS	−7.736	−0.0591	1.373	−0.127	−0.119	0.000
$R^2 = 0.95799$	(0.674)	(0.0325)	(0.0756)	(0.127)	(0.0813)	(0.000)
Prais–Winsten	−6.782	−0.152	1.267	−0.0308	−0.0638	0.862
	(−0.955)	(0.0370)	(0.107)	(0.127)	(0.0758)	(0.0855)
Cochrane–Orcutt	−7.147	−0.149	1.307	−0.0599	−0.0563	0.849
	(1.297)	(0.0382)	(0.144)	(0.146)	(0.0789)	(−.0893)
Maximum Likelihood	−5.159	−0.208	1.0828	0.0878	−0.0351	0.930
	(1.132)	(0.0349)	(0.127)	(0.125)	(0.0659)	(0.0620)
AR(2)	−11.828	−0.0310	1.415	−0.192	−0.114	0.760
	(0.888)	(0.0292)	(0.0682)	(0.133)	(0.0846)	(r_1)
			$\theta_1 = 0.9936319$, $\theta_2 = -4620284$			

five lags of the residuals. The coefficients on the lagged residuals and the associated t statistics are 1.075 (5.493), −0.712 (−2.488), 0.310 (0.968), −0.227 (−0.758), 0.000096 (0.000). The R^2 in this regression is 0.598223, which produces a chi-squared value of 21.536. The conclusion is the same. Finally, the Durbin–Watson statistic is 0.60470. For four regressors and 36 observations, the critical value of d_l is 1.24, so on this basis as well, the hypothesis $\rho = 0$ would be rejected. The plot of the residuals shown in Figure 12.4 seems consistent with this conclusion.

The Prais and Winsten FGLS estimates appear in the second row of Table 12.4, followed by the Cochrane and Orcutt results then the maximum likelihood estimates.

FIGURE 12.4 Least Squares Residuals.



In each of these cases, the autocorrelation coefficient is reestimated using the FGLS residuals. This recomputed value is what appears in the table.

One might want to examine the residuals after estimation to ascertain whether the AR(1) model is appropriate. In the results above, 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 0.865. The associated Durbin–Watson statistic is 0.278. 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 the adjusted residuals is 0.438 with a Durbin–Watson statistic of 1.125. It appears on this basis that, in fact, the AR(1) model has not completed the specification.

The results noted earlier suggest that an AR(2) process might better characterize the disturbances in this model. Simple regression of the least squares residuals on a constant and two lagged values (the two period counterpart to a method of obtaining r in the AR(1) model) produces slope coefficients of 0.9936319 and −0.4620284.[15] The GLS transformations for the AR(2) model are given in (12-30). We recomputed the regression using the AR(2) transformation and these two coefficients. These are the final results shown in Table 12.2. They do bring a substantial change in the results. As an additional check on the adequacy of the model, we now computed the corrected FGLS residuals from the AR(2) model,

$$\hat{u}_t = \hat{\varepsilon}_t - \hat{\theta}_1\hat{\varepsilon}_{t-1} - \hat{\theta}_2\hat{\varepsilon}_{t-2}$$

The first five autocorrelations of these residuals are 0.132, 0.134, 0.016, 0.022, and −0.118. The Box–Pierce and Box–Ljung statistics are 1.605 and 1.857, which are far from statistically significant. We thus conclude that the AR(2) model accounts for the autocorrelation in the data.

The preceding suggests how one might discover the appropriate model for autocorrelation in a regression model. However, it is worth keeping in mind that the source of the autocorrelation might itself be discernible in the data. The finding of an AR(2) process may still suggest that the regression specification is incomplete or inadequate in some way.

[15] In fitting an AR(1) model, the stationarity condition is obvious; $|r|$ must be less than one. For an AR(2) process, the condition is less than obvious. We will examine this issue in Chapter 20. For the present, we merely state the result; the two values $(1/2)[\theta_1 \pm (\theta_1^2 + 4\theta_2)^{1/2}]$ must be less than one in absolute value. Since the term in parentheses might be negative, the "roots" might be a complex pair $a \pm bi$, in which case $a^2 + b^2$ must be less than one. You can verify that the two complex roots for our process above are indeed "inside the unit circle."

12.9.4 ESTIMATION WITH A LAGGED DEPENDENT VARIABLE

In Section 12.5.1, we considered the problem of estimation by least squares when the model contains both autocorrelation and lagged dependent variable(s). Since 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 5.4. 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 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, is 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

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

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 $\mathbf{Z}$ corresponding to y_{t-1} is the only one that need be different from that of $\mathbf{X}$. 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

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

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

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

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. Hatanaka shows that these estimators are asymptotically equivalent to maximum likelihood estimators.

12.10 COMMON FACTORS

We saw in Example 12.2 that misspecification of an equation could create the appearance of serially correlated disturbances when, in fact, there are none. An orthodox (perhaps somewhat optimistic) purist might argue that autocorrelation is *always* an artifact of misspecification. Although this view might be extreme [see, e.g., Hendry (1980) for a more moderate, but still strident statement], it does suggest a useful point. It might be useful if we could examine the specification of a model statistically with this consideration in mind. The test for **common factors** is such a test. [See, as well, the aforementioned paper by Mizon (1995).]

The assumption that the correctly specified model is

$$y_t = \mathbf{x}_t'\boldsymbol{\beta} + \varepsilon_t, \quad \varepsilon_t = \rho\varepsilon_{t-1} + u_t, \quad t = 1, \ldots, T$$

implies the "reduced form,"

$$M_0: y_t = \rho y_{t-1} + (\mathbf{x}_t - \rho\mathbf{x}_{t-1})'\boldsymbol{\beta} + u_t, \quad t = 2, \ldots, T,$$

where u_t is free from serial correlation. The second of these is actually a restriction on the model

$$M_1: y_t = \rho y_{t-1} + \mathbf{x}_t'\boldsymbol{\beta} + \mathbf{x}_{t-1}'\boldsymbol{\alpha} + u_t, \quad t = 2, \ldots, T,$$

in which, once again, u_t is a classical disturbance. The second model contains $2K + 1$ parameters, but if the model is correct, then $\boldsymbol{\alpha} = -\rho\boldsymbol{\beta}$ and there are only $K + 1$ parameters and K restrictions. Both M_0 and M_1 can be estimated by least squares, although M_0 is a nonlinear model. One might then test the restrictions of M_0 using an F test. This test will be valid asymptotically, although its exact distribution in finite samples will not be precisely F. In large samples, KF will converge to a chi-squared statistic, so we use the F distribution as usual to be conservative. There is a minor practical complication in implementing this test. Some elements of $\boldsymbol{\alpha}$ may not be estimable. For example, if $\mathbf{x}_t$ contains a constant term, then the one in $\boldsymbol{\alpha}$ is unidentified. If $\mathbf{x}_t$ contains both current and lagged values of a variable, then the one period lagged value will appear twice in M_1, once in $\mathbf{x}_t$ as the lagged value and once in $\mathbf{x}_{t-1}$ as the current value. There are other combinations that will be problematic, so the actual number of restrictions that appear in the test is reduced to the number of identified parameters in $\boldsymbol{\alpha}$.

Example 12.5 Tests for Common Factors

We will examine the gasoline demand model of Example 12.2 and consider a simplified version of the equation

$$\ln\frac{G_t}{pop_t} = \beta_1 + \beta_2 \ln P_{G,t} + \beta_3 \ln\frac{I_t}{pop_t} + \beta_4 \ln P_{NC,t} + \beta_5 \ln P_{UC,t} + \varepsilon_t.$$

If the AR(1) model is appropriate for ε_t, then the restricted model,

$$\begin{aligned}\ln\frac{G_t}{pop_t} = {} & \beta_1 + \beta_2(\ln P_{G,t} - \rho \ln P_{G,t-1}) + \beta_3\left(\ln\frac{I_t}{pop_t} - \rho \ln\frac{I_{t-1}}{pop_{t-1}}\right) \\ & + \beta_4(\ln P_{NC,t} - \rho \ln P_{NC,t-1}) + \beta_5(\ln P_{UC,t} - \rho \ln P_{UC,t-1}) \\ & + \rho \ln G_{t-1}/pop_{t-1} + u_t,\end{aligned}$$

with six free coefficients will not significantly degrade the fit of the unrestricted model, which has 10 free coefficients. The F statistic, with 4 and 25 degrees of freedom, for this test equals

4.311, which is larger than the critical value of 2.76. Thus, we would conclude that the AR(1) model would not be appropriate for this specification and these data. Note that we reached the same conclusion after a more conventional analysis of the residuals in the application in Section 12.9.3.

12.11 FORECASTING IN THE PRESENCE OF AUTOCORRELATION

For purposes of forecasting, we refer first to the transformed model,

$$y_{*_t} = \mathbf{x}'_{*_t}\boldsymbol{\beta} + \varepsilon_{*_t}.$$

Suppose that the process generating ε_t is an AR(1) and that ρ is known. Since this model is a classical regression model, the results of Section 6.6 may be used. The optimal forecast of $y^0_{*_{T+1}}$, given $\mathbf{x}^0_{T+1}$ and $\mathbf{x}_T$ (i.e., $\mathbf{x}^0_{*_{T+1}} = \mathbf{x}^0_{T+1} - \rho\mathbf{x}_T$), is

$$\hat{y}^0_{*_{T+1}} = \mathbf{x}^{0\prime}_{*_{T+1}}\hat{\boldsymbol{\beta}}.$$

Disassembling $\hat{y}^0_{*_{T+1}}$, we find that

$$\hat{y}^0_{T+1} - \rho y_T = \mathbf{x}^{0\prime}_{T+1}\hat{\boldsymbol{\beta}} - \rho\mathbf{x}'_T\hat{\boldsymbol{\beta}}$$

or

$$\begin{aligned}\hat{y}^0_{T+1} &= \mathbf{x}^{0\prime}_{T+1}\hat{\boldsymbol{\beta}} + \rho(y_T - \mathbf{x}'_T\hat{\boldsymbol{\beta}}) \\ &= \mathbf{x}^{0\prime}_{T+1}\hat{\boldsymbol{\beta}} + \rho e_T.\end{aligned} \tag{12-34}$$

Thus, we carry forward a proportion ρ of the estimated disturbance in the preceding period. This step can be justified by reference to

$$E[\varepsilon_{T+1} \mid \varepsilon_T] = \rho\varepsilon_T.$$

It can also be shown that to forecast n periods ahead, we would use

$$\hat{y}^0_{T+n} = \mathbf{x}^{0\prime}_{T+n}\hat{\boldsymbol{\beta}} + \rho^n e_T.$$

The extension to higher-order autoregressions is direct. For a second-order model, for example,

$$\hat{y}^0_{T+n} = \hat{\boldsymbol{\beta}}'\mathbf{x}^0_{T+n} + \theta_1 e_{T+n-1} + \theta_2 e_{T+n-2}. \tag{12-35}$$

For residuals that are outside the sample period, we use the recursion

$$e_s = \theta_1 e_{s-1} + \theta_2 e_{s-2}, \tag{12-36}$$

beginning with the last two residuals within the sample.

Moving average models are somewhat simpler, as the autocorrelation lasts for only Q periods. For an MA(1) model, for the first postsample period,

$$\hat{y}^0_{T+1} = \mathbf{x}^{0\prime}_{T+1}\hat{\boldsymbol{\beta}} + \hat{\varepsilon}_{T+1},$$

where

$$\hat{\varepsilon}_{T+1} = \hat{u}_{T+1} - \lambda\hat{u}_T.$$

Therefore, a forecast of ε_{T+1} will use all previous residuals. One way to proceed is to accumulate $\hat{\varepsilon}_{T+1}$ from the recursion

$$\hat{u}_t = \hat{\varepsilon}_t + \lambda \hat{u}_{t-1}$$

with $\hat{u}_{T+1} = \hat{u}_0 = 0$ and $\hat{\varepsilon}_t = (y_t - \mathbf{x}_t'\hat{\boldsymbol{\beta}})$. After the first postsample period,

$$\hat{\varepsilon}_{T+n} = \hat{u}_{T+n} - \lambda \hat{u}_{T+n-1} = 0.$$

If the parameters of the disturbance process are known, then the variances for the forecast errors can be computed using the results of Section 6.6. For an AR(1) disturbance, the estimated variance would be

$$s_f^2 = \hat{\sigma}_\varepsilon^2 + (\mathbf{x}_t - \rho \mathbf{x}_{t-1})' \{\text{Est. Var}[\hat{\boldsymbol{\beta}}]\} (\mathbf{x}_t - \rho \mathbf{x}_{t-1}). \tag{12-37}$$

For a higher-order process, it is only necessary to modify the calculation of $\mathbf{x}_{*_t}$ accordingly. The forecast variances for an MA(1) process are somewhat more involved. Details may be found in Judge et al. (1985) and Hamilton (1994). If the parameters of the disturbance process, ρ, λ, θ_j, and so on, are estimated as well, then the forecast variance will be greater. For an AR(1) model, the necessary correction to the forecast variance of the n-period-ahead forecast error is $\hat{\sigma}_\varepsilon^2 n^2 \rho^{2(n-1)}/T$. [For a one-period-ahead forecast, this merely adds a term, $\hat{\sigma}_\varepsilon^2/T$, in the brackets in (12-36)]. Higher-order AR and MA processes are analyzed in Baillie (1979). Finally, if the regressors are stochastic, the expressions become more complex by another order of magnitude.

If ρ is known, then (12-34) provides the best linear unbiased forecast of y_{t+1}.[16] If, however, ρ must be estimated, then this assessment must be modified. There is information about ε_{t+1} embodied in e_t. Having to estimate ρ, however, implies that some or all the value of this information is offset by the variation introduced into the forecast by including the stochastic component $\hat{\rho} e_t$.[17] Whether (12-34) is preferable to the obvious expedient $\hat{y}_{T+n}^0 = \hat{\boldsymbol{\beta}}' \mathbf{x}_{T+n}^0$ in a small sample when ρ is estimated remains to be settled.

12.12 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 which 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

[16] See Goldberger (1962).

[17] See Baillie (1979).

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.

Key Terms and Concepts

- AR(1)
- Asymptotic negligibility
- Asymptotic normality
- Autocorrelation
- Autocorrelation matrix
- Autocovariance
- Autocovariance matrix
- Autoregressive form
- Cochrane–Orcutt estimator
- Common factor model
- Covariance stationarity
- Durbin–Watson test
- Ergodicity
- Ergodic Theorem
- First-order autoregression
- Expectations augmented Phillips curve
- GMM estimator
- Initial conditions
- Innovation
- Lagrange multiplier test
- Martingale sequence
- Martingale difference sequence
- Moving average form
- Moving average process
- Partial difference
- Prais–Winsten estimator
- Pseudo differences
- Q test
- Quasi differences
- Stationarity
- 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. The following regression is obtained by ordinary least squares, using 21 observations. (Estimated asymptotic standard errors are shown in parentheses.)

$$y_t = \underset{}{1.3} + \underset{(0.3)}{0.97}y_{t-1} + \underset{(0.18)}{2.31}x_t, \quad \underset{(1.04)}{\text{D} - \text{W} = 1.21.}$$

   Test for the presence of autocorrelation in the disturbances.
4. It is commonly asserted that the Durbin–Watson statistic is only appropriate for testing for first-order autoregressive disturbances. 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.
5. The data used to fit the expectations augmented Phillips curve in Example 12.3 are given in Table F5.1. 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. Since 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?

6. Data for fitting an improved Phillips curve model can be obtained from many sources, including the Bureau of Economic Analysis's (BEA) own website, Economagic.com, and so on. Obtain the necessary data and expand the model of example 12.3. Does adding additional explanatory variables to the model reduce the extreme pattern of the OLS residuals that appears in Figure 12.3?