12

AUTOCORRELATION: WHAT HAPPENS IF THE ERROR TERMS ARE CORRELATED?

The reader may recall that there are generally three types of data that are available for empirical analysis: (1) cross section, (2) time series, and (3) combination of cross section and time series, also known as pooled data. In developing the classical linear regression model (CLRM) in **Part I** we made several assumptions, which were discussed in Section 7.1. However, we noted that *not* all these assumptions would hold in every type of data. As a matter of fact, we saw in the previous chapter that the assumption of homoscedasticity, or equal error variance, may not be always tenable in cross-sectional data. In other words, cross-sectional data are often plagued by the problem of heteroscedasticity.

However, in cross-section studies, data are often collected on the basis of a random sample of cross-sectional units, such as households (in a consumption function analysis) or firms (in an investment study analysis) so that there is no prior reason to believe that the error term pertaining to one household or a firm is correlated with the error term of another household or firm. If by chance such a correlation is observed in cross-sectional units, it is called **spatial autocorrelation,** that is, correlation in space rather than over time. However, it is important to remember that, in cross-sectional analysis, the ordering of the data must have some logic, or economic interest, to make sense of any determination of whether (spatial) autocorrelation is present or not.

The situation, however, is likely to be very different if we are dealing with time series data, for the observations in such data follow a natural ordering over time so that successive observations are likely to exhibit intercorrelations, especially if the time interval between successive observations is

short, such as a day, a week, or a month rather than a year. If you observe stock price indexes, such as the Dow Jones or S&P 500 over successive days, it is not unusual to find that these indexes move up or down for several days in succession. Obviously, in situations like this, the assumption of **no auto,** or **serial, correlation** in the error terms that underlies the CLRM will be violated.

In this chapter we take a critical look at this assumption with a view to answering the following questions:

1. What is the nature of autocorrelation?

2. What are the theoretical and practical consequences of autocorrelation?

3. Since the assumption of no autocorrelation relates to the unobservable disturbances u_t, how does one know that there is autocorrelation in any given situation? Notice that we now use the subscript t to emphasize that we are dealing with time series data.

4. How does one remedy the problem of autocorrelation?

The reader will find this chapter in many ways similar to the preceding chapter on heteroscedasticity in that **under both heteroscedasticity and autocorrelation the usual OLS estimators, although linear, unbiased, and asymptotically (i.e., in large samples) normally distributed,[1] are no longer minimum variance among all linear unbiased estimators. In short, they are not efficient relative to other linear and unbiased estimators. Put differently, they may not be BLUE. As a result, the usual, *t*, *F*, and χ^2 may not be valid.**

12.1 THE NATURE OF THE PROBLEM

The term **autocorrelation** may be defined as "correlation between members of series of observations ordered in time [as in time series data] or space [as in cross-sectional data]."[2] In the regression context, the classical linear regression model assumes that such autocorrelation does not exist in the disturbances u_i. Symbolically,

$$E(u_i u_j) = 0 \qquad i \neq j \tag{3.2.5}$$

Put simply, the classical model assumes that the disturbance term relating to any observation is not influenced by the disturbance term relating to any other observation. For example, if we are dealing with quarterly time series data involving the regression of output on labor and capital inputs and if,

[1] On this, see William H. Greene, *Econometric Analysis,* 4th ed., Prentice Hall, N.J., 2000, Chap. 11, and Paul A. Rudd, *An Introduction to Classical Econometric Theory,* Oxford University Press, 2000, Chap. 19.

[2] Maurice G. Kendall and William R. Buckland, *A Dictionary of Statistical Terms,* Hafner Publishing Company, New York, 1971, p. 8.

say, there is a labor strike affecting output in one quarter, there is no reason to believe that this disruption will be carried over to the next quarter. That is, if output is lower this quarter, there is no reason to expect it to be lower next quarter. Similarly, if we are dealing with cross-sectional data involving the regression of family consumption expenditure on family income, the effect of an increase of one family's income on its consumption expenditure is not expected to affect the consumption expenditure of another family.

However, if there is such a dependence, we have autocorrelation. Symbolically,

$$E(u_i u_j) \neq 0 \qquad i \neq j \tag{12.1.1}$$

In this situation, the disruption caused by a strike this quarter may very well affect output next quarter, or the increases in the consumption expenditure of one family may very well prompt another family to increase its consumption expenditure if it wants to keep up with the Joneses.

Before we find out why autocorrelation exists, it is essential to clear up some terminological questions. Although it is now a common practice to treat the terms **autocorrelation** and **serial correlation** synonymously, some authors prefer to distinguish the two terms. For example, Tintner defines autocorrelation as "lag correlation of a given series with itself, lagged by a number of time units," whereas he reserves the term serial correlation to "lag correlation between two different series."[3] Thus, correlation between two time series such as $u_1, u_2, \ldots, u_{10}$ and $u_2, u_3, \ldots, u_{11}$, where the former is the latter series lagged by one time period, is *autocorrelation,* whereas correlation between time series such as $u_1, u_2, \ldots, u_{10}$ and $v_2, v_3, \ldots, v_{11}$, where u and v are two different time series, is called *serial correlation*. Although the distinction between the two terms may be useful, in this book we shall treat them synonymously.

Let us visualize some of the plausible patterns of auto- and nonautocorrelation, which are given in Figure 12.1. Figure 12.1*a* to *d* shows that there is a discernible pattern among the u's. Figure 12.1*a* shows a cyclical pattern; Figure 12.1*b* and *c* suggests an upward or downward linear trend in the disturbances; whereas Figure 12.1*d* indicates that both linear and quadratic trend terms are present in the disturbances. Only Figure 12.1*e* indicates no systematic pattern, supporting the nonautocorrelation assumption of the classical linear regression model.

The natural question is: Why does serial correlation occur? There are several reasons, some of which are as follows:

Inertia. A salient feature of most economic time series is inertia, or sluggishness. As is well known, time series such as GNP, price indexes, production, employment, and unemployment exhibit (business) cycles.

[3]Gerhard Tintner, *Econometrics*, John Wiley & Sons, New York, 1965.

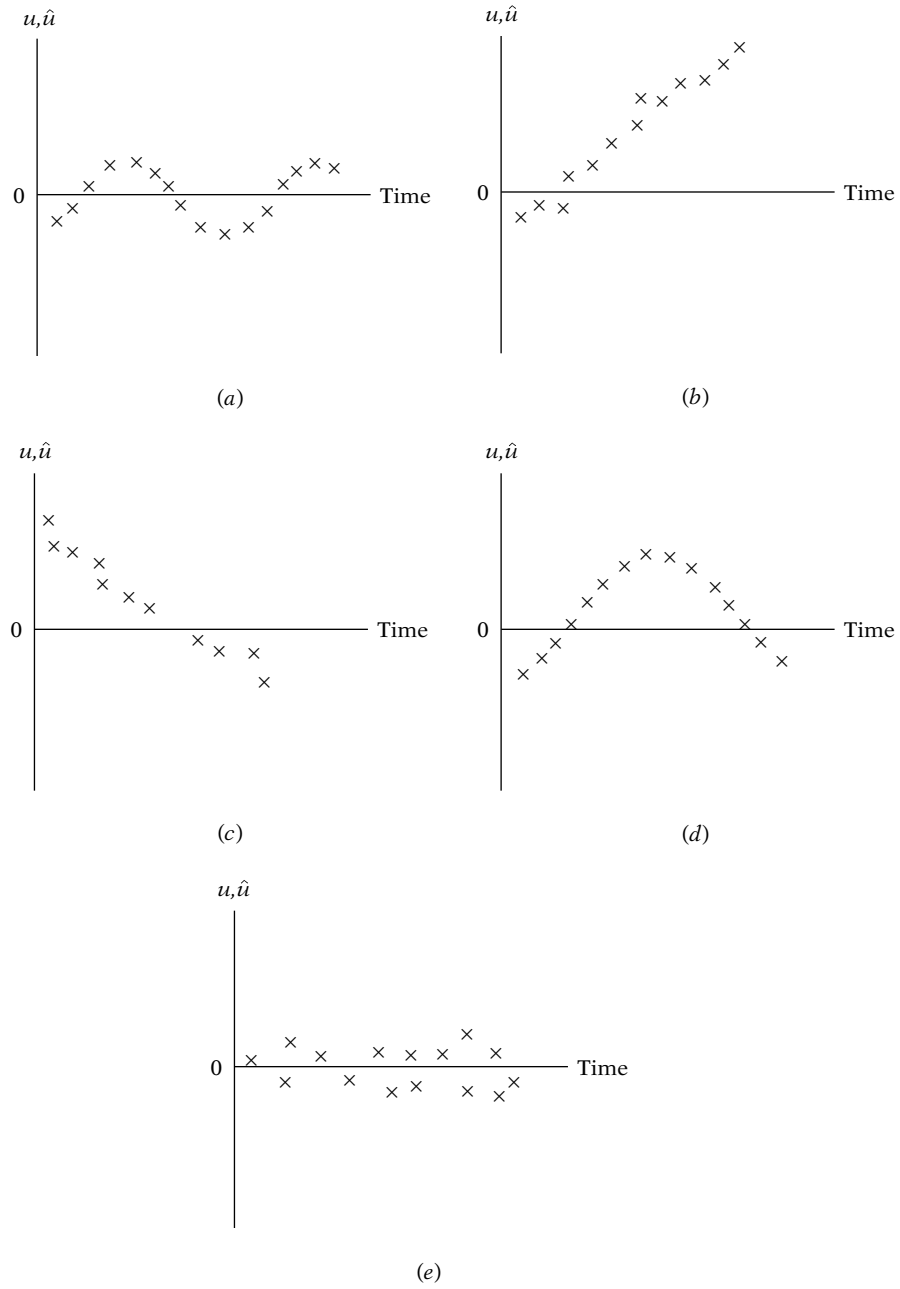

FIGURE 12.1 Patterns of autocorrelation and nonautocorrelation.

Starting at the bottom of the recession, when economic recovery starts, most of these series start moving upward. In this upswing, the value of a series at one point in time is greater than its previous value. Thus there is a "momentum" built into them, and it continues until something happens (e.g., increase in interest rate or taxes or both) to slow them down. Therefore, in regressions involving time series data, successive observations are likely to be interdependent.

Specification Bias: Excluded Variables Case. In empirical analysis the researcher often starts with a plausible regression model that may not be the most "perfect" one. After the regression analysis, the researcher does the postmortem to find out whether the results accord with a priori expectations. If not, surgery is begun. For example, the researcher may plot the residuals $\hat{u}_i$ obtained from the fitted regression and may observe patterns such as those shown in Figure 12.1*a* to *d*. These residuals (which are proxies for u_i) may suggest that some variables that were originally candidates but were not included in the model for a variety of reasons should be included. This is the case of **excluded variable** specification bias. Often the inclusion of such variables removes the correlation pattern observed among the residuals. For example, suppose we have the following demand model:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + u_t \qquad \textbf{(12.1.2)}$$

where Y = quantity of beef demanded, X_2 = price of beef, X_3 = consumer income, X_4 = price of pork, and t = time.[4] However, for some reason we run the following regression:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + v_t \qquad \textbf{(12.1.3)}$$

Now if (12.1.2) is the "correct" model or the "truth" or true relation, running (12.1.3) is tantamount to letting $v_t = \beta_4 X_{4t} + u_t$. And to the extent the price of pork affects the consumption of beef, the error or disturbance term v will reflect a systematic pattern, thus creating (false) autocorrelation. A simple test of this would be to run both (12.1.2) and (12.1.3) and see whether autocorrelation, if any, observed in model (12.1.3) disappears when (12.1.2) is run.[5] The actual mechanics of detecting autocorrelation will be discussed in Section 12.6 where we will show that a plot of the residuals from regressions (12.1.2) and (12.1.3) will often shed considerable light on serial correlation.

[4]As a matter of convention, we shall use the subscript t to denote time series data and the usual subscript i for cross-sectional data.

[5]If it is found that the real problem is one of specification bias, not autocorrelation, then as will be shown in Chap. 13, the OLS estimators of the parameters (12.1.3) may be biased as well as inconsistent.

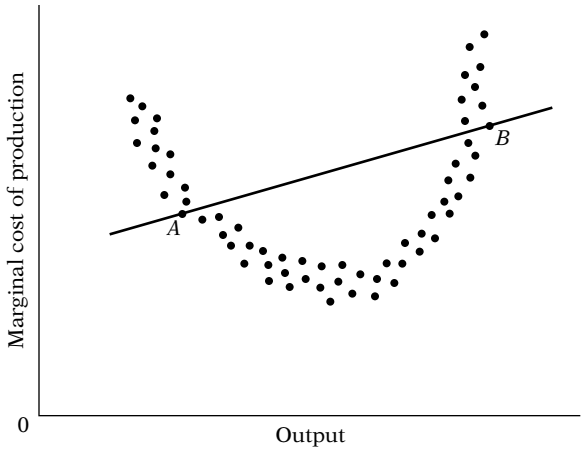

FIGURE 12.2 Specification bias: incorrect functional form.

Specification Bias: Incorrect Functional Form. Suppose the "true" or correct model in a cost-output study is as follows:

$$\text{Marginal cost}_i = \beta_1 + \beta_2\ \text{output}_i + \beta_3\ \text{output}_i^2 + u_i \qquad \textbf{(12.1.4)}$$

but we fit the following model:

$$\text{Marginal cost}_i = \alpha_1 + \alpha_2\ \text{output}_i + v_i \qquad \textbf{(12.1.5)}$$

The marginal cost curve corresponding to the "true" model is shown in Figure 12.2 along with the "incorrect" linear cost curve.

As Figure 12.2 shows, between points A and B the linear marginal cost curve will consistently overestimate the true marginal cost, whereas beyond these points it will consistently underestimate the true marginal cost. This result is to be expected, because the disturbance term v_i is, in fact, equal to $\text{output}^2 + u_i$, and hence will catch the systematic effect of the output^2 term on marginal cost. In this case, v_i will reflect autocorrelation because of the use of an incorrect functional form. In Chapter 13 we will consider several methods of detecting specification bias.

Cobweb Phenomenon. The supply of many agricultural commodities reflects the so-called cobweb phenomenon, where supply reacts to price with a lag of one time period because supply decisions take time to implement (the gestation period). Thus, at the beginning of this year's planting of crops, farmers are influenced by the price prevailing last year, so that their supply function is

$$\text{Supply}_t = \beta_1 + \beta_2 P_{t-1} + u_t \qquad \textbf{(12.1.6)}$$

Suppose at the end of period t, price P_t turns out to be lower than P_{t-1}. Therefore, in period $t + 1$ farmers may very well decide to produce less than

they did in period t. Obviously, in this situation the disturbances u_t are not expected to be random because if the farmers overproduce in year t, they are likely to reduce their production in $t + 1$, and so on, leading to a Cobweb pattern.

Lags. In a time series regression of consumption expenditure on income, it is not uncommon to find that the consumption expenditure in the current period depends, among other things, on the consumption expenditure of the previous period. That is,

$$\text{Consumption}_t = \beta_1 + \beta_2\ \text{income}_t + \beta_3\ \text{consumption}_{t-1} + u_t \qquad \textbf{(12.1.7)}$$

A regression such as (12.1.7) is known as **autoregression** because one of the explanatory variables is the lagged value of the dependent variable. (We shall study such models in Chapter 17.) The rationale for a model such as (12.1.7) is simple. Consumers do not change their consumption habits readily for psychological, technological, or institutional reasons. Now if we neglect the lagged term in (12.1.7), the resulting error term will reflect a systematic pattern due to the influence of lagged consumption on current consumption.

"Manipulation" of Data. In empirical analysis, the raw data are often "manipulated." For example, in time series regressions involving quarterly data, such data are usually derived from the monthly data by simply adding three monthly observations and dividing the sum by 3. This averaging introduces smoothness into the data by dampening the fluctuations in the monthly data. Therefore, the graph plotting the quarterly data looks much smoother than the monthly data, and this smoothness may itself lend to a systematic pattern in the disturbances, thereby introducing autocorrelation. Another source of manipulation is **interpolation** or **extrapolation** of data. For example, the Census of Population is conducted every 10 years in this country, the last being in 2000 and the one before that in 1990. Now if there is a need to obtain data for some year within the intercensus period 1990–2000, the common practice is to interpolate on the basis of some ad hoc assumptions. All such data "massaging" techniques might impose upon the data a systematic pattern that might not exist in the original data.[6]

Data Transformation. As an example of this, consider the following model:

$$Y_t = \beta_1 + \beta_2 X_t + u_t \qquad \textbf{(12.1.8)}$$

where, say, Y = consumption expenditure and X = income. Since (12.1.8) holds true at every time period, it holds true also in the previous time

[6]On this, see William H. Greene, op. cit., p. 526.

period, $(t-1)$. So, we can write (12.1.8) as

$$Y_{t-1} = \beta_1 + \beta_2 X_{t-1} + u_{t-1} \tag{12.1.9}$$

Y_{t-1}, X_{t-1}, and u_{t-1} are known as the **lagged values** of Y, X, and u, respectively, here lagged by one period. We will see the importance of the lagged values later in the chapter as well in several places in the text.

Now if we subtract (12.1.9) from (12.1.8), we obtain

$$\Delta Y_t = \beta_2 \Delta X_t + \Delta u_t \tag{12.1.10}$$

where Δ, known as the **first difference operator,** tells us to take successive differences of the variables in question. Thus, $\Delta Y_t = (Y_t - Y_{t-1})$, $\Delta X_t = (X_t - X_{t-1})$, and $\Delta u_t = (u_t - u_{t-1})$. For empirical purposes, we write (12.1.10) as

$$\Delta Y_t = \beta_2 \Delta X_t + v_t \tag{12.1.11}$$

where $v_t = \Delta u_t = (u_t - u_{t-1})$.

Equation (12.1.9) is known as the **level form** and Eq. (12.1.10) is known as the **(first) difference form.** Both forms are often used in empirical analysis. For example, if in (12.1.9) Y and X represent the logarithms of consumption expenditure and income, then in (12.1.10) ΔY and ΔX will represent changes in the logs of consumption expenditure and income. But as we know, a change in the log of a variable is a relative change, or a percentage change, if the former is multiplied by 100. So, instead of studying relationships between variables in the level form, we may be interested in their relationships in the growth form.

Now if the error term in (12.1.8) satisfies the standard OLS assumptions, particularly the assumption of no autocorrelation, it can be shown that the error term v_t in (12.1.11) is autocorrelated. (The proof is given in Appendix 12A, Section 12A.1.) It may be noted here that models like (12.1.11) are known as **dynamic regression models,** that is, models involving lagged regressands. We will study such models in depth in Chapter 17.

The point of the preceding example is that sometimes autocorrelation may be induced as a result of transforming the original model.

Nonstationarity. We mentioned in Chapter 1 that, while dealing with time series data, we may have to find out if a given time series is stationary. Although we will discuss the topic of nonstationary time series more thoroughly in the chapters on time series econometrics in **Part V** of the text, loosely speaking, a time series is stationary if its characteristics (e.g., mean, variance, and covariance) are *time invariant;* that is, they do not change over time. If that is not the case, we have a nonstationary time series.

As we will discuss in **Part V,** in a regression model such as (12.1.8), it is quite possible that both Y and X are nonstationary and therefore the error u is also nonstationary.[7] In that case, the error term will exhibit autocorrelation.

[7] As we will also see in **Part V,** even though Y and X are nonstationary, it is possible to find u to be stationary. We will explore the implication of that later on.

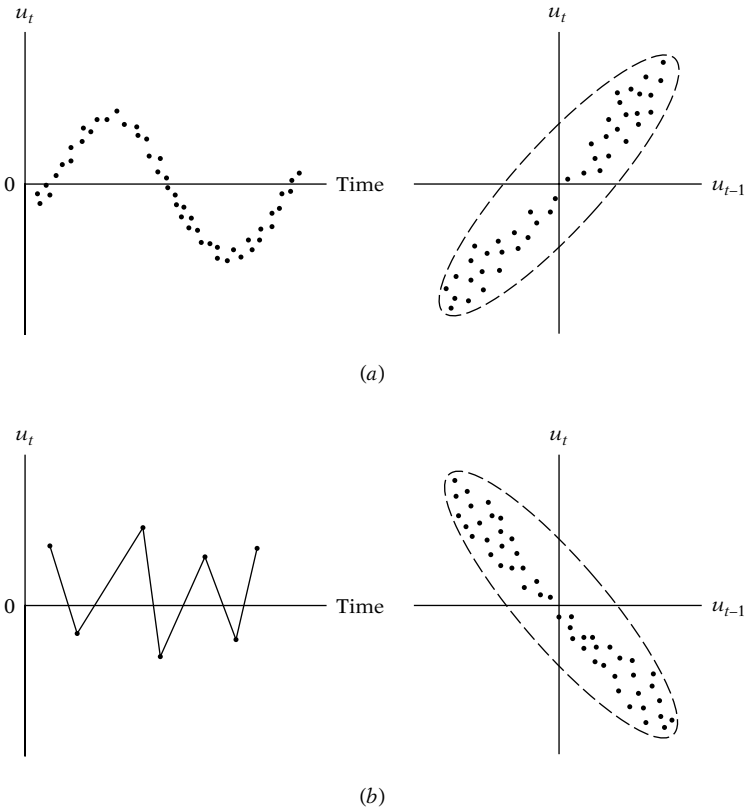

FIGURE 12.3 (*a*) Positive and (*b*) negative autocorrelation.

In summary, then, there are a variety of reasons why the error term in a regression model may be autocorrelated. In the rest of the chapter we investigate in some detail the problems posed by autocorrelation and what can be done about it.

It should be noted also that autocorrelation can be positive (Figure 12.3*a*) as well as negative, although most economic time series generally exhibit positive autocorrelation because most of them ether move upward or downward over extended time periods and do not exhibit a constant up-and-down movement such as that shown in Figure 12.3*b*.

12.2 OLS ESTIMATION IN THE PRESENCE OF AUTOCORRELATION

What happens to the OLS estimators and their variances if we introduce autocorrelation in the disturbances by assuming that $E(u_t u_{t+s}) \neq 0$ $(s \neq 0)$ but retain all the other assumptions of the classical model?[8] Note again that

[8]If $s = 0$, we obtain $E(u_t^2)$. Since $E(u_t) = 0$ by assumption, $E(u_t^2)$ will represent the variance of the error term, which obviously is nonzero (why?).

we are now using the subscript t on the disturbances to emphasize that we are dealing with time series data.

We revert once again to the two-variable regression model to explain the basic ideas involved, namely, $Y_t = \beta_1 + \beta_2 X_t + u_t$. To make any headway, we must assume the mechanism that generates u_t, for $E(u_t u_{t+s}) \neq 0$ $(s \neq 0)$ is too general an assumption to be of any practical use. As a starting point, or first approximation, one can assume that the disturbance, or error, terms are generated by the following mechanism.

$$u_t = \rho u_{t-1} + \varepsilon_t \qquad -1 < \rho < 1 \tag{12.2.1}$$

where ρ (= rho) is known as the **coefficient of autocovariance** and where ε_t is the stochastic disturbance term such that it satisfied the standard OLS assumptions, namely,

$$\begin{aligned} E(\varepsilon_t) &= 0 \\ \text{var}\,(\varepsilon_t) &= \sigma_\varepsilon^2 \\ \text{cov}\,(\varepsilon_t, \varepsilon_{t+s}) &= 0 \qquad s \neq 0 \end{aligned} \tag{12.2.2}$$

In the engineering literature, an error term with the preceding properties is often called a **white noise error term.** What (12.2.1) postulates is that the value of the disturbance term in period t is equal to rho times its value in the previous period plus a purely random error term.

The scheme (12.2.1) is known as **Markov first-order autoregressive scheme,** or simply **a first-order autoregressive scheme,** usually denoted as **AR(1).** The name *autoregressive* is appropriate because (12.2.1) can be interpreted as the regression of u_t on itself lagged one period. It is first order because u_t and its immediate past value are involved; that is, the maximum lag is 1. If the model were $u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \varepsilon_t$, it would be an AR(2), or second-order, autoregressive scheme, and so on. We will examine such higher-order schemes in the chapters on time series econometrics in **Part V.**

In passing, note that rho, the coefficient of autocovariance in (12.2.1), can also be interpreted as the **first-order coefficient of autocorrelation,** or more accurately, **the coefficient of autocorrelation at lag 1.**[9]

[9]This name can be easily justified. By definition, the (population) coefficient of correlation between u_t and u_{t-1} is

$$\begin{aligned} \rho &= \frac{E\{[u_t - E(u_t)][u_{t-1} - E(u_{t-1})]\}}{\sqrt{\text{var}\,(u_t)}\sqrt{\text{var}\,(u_{t-1})}} \\ &= \frac{E(u_t u_{t-1})}{\text{var}\,(u_{t-1})} \end{aligned}$$

since $E(u_t) = 0$ for each t and $\text{var}\,(u_t) = \text{var}\,(u_{t-1})$ because we are retaining the assumption of homoscedasticity. The reader can see that ρ is also the slope coefficient in the regression of u_t on u_{t-1}.

Given the AR(1) scheme, it can be shown that (see Appendix 12A, Section 12A.2)

$$\text{var}\,(u_t) = E\big(u_t^2\big) = \frac{\sigma_\varepsilon^2}{1-\rho^2} \tag{12.2.3}$$

$$\text{cov}\,(u_t, u_{t+s}) = E(u_t u_{t-s}) = \rho^s \frac{\sigma_\varepsilon^2}{1-\rho^2} \tag{12.2.4}$$

$$\text{cor}\,(u_t, u_{t+s}) = \rho^s \tag{12.2.5}$$

where $\text{cov}\,(u_t, u_{t+s})$ means covariance between error terms s periods apart and where $\text{cor}\,(u_t, u_{t+s})$ means correlation between error terms s periods apart. Note that because of the symmetry property of covariances and correlations, $\text{cov}\,(u_t, u_{t+s}) = \text{cov}\,(u_t, u_{t-s})$ and $\text{cor}\,(u_t, u_{t+s}) = \text{cor}\,(u_t, u_{t-s})$.

Since ρ is a constant between -1 and $+1$, (12.2.3) shows that under the AR(1) scheme, the variance of u_t is *still homoscedastic,* but u_t is correlated not only with its immediate past value but its values several periods in the past. It is *critical* to note that $|\rho| < 1$, that is, the absolute value of rho is less than one. If, for example, rho is one, the variances and covariances listed above are not defined. If $|\rho| < 1$, we say that the AR(1) process given in (12.2.1) is *stationary;* that is, the mean, variance, and covariance of u_t do not change over time. If $|\rho|$ is less than one, then it is clear from (12.2.4) that the value of the covariance will decline as we go into the distant past. We will see the utility of the preceding results shortly.

One reason we use the AR(1) process is not only because of its simplicity compared to higher-order AR schemes, but also because in many applications it has proved to be quite useful. Additionally, a considerable amount of theoretical and empirical work has been done on the AR(1) scheme.

Now return to our two-variable regression model: $Y_t = \beta_1 + \beta_2 X_t + u_t$. We know from Chapter 3 that the OLS estimator of the slope coefficient is

$$\hat{\beta}_2 = \frac{\sum x_t y_t}{\sum x_t^2} \tag{12.2.6}$$

and its variance is given by

$$\text{var}\,(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_t^2} \tag{12.2.7}$$

where the small letters as usual denote deviation from the mean values.

Now under the AR(1) scheme, it can be shown that the variance of this estimator is:

$$\text{var}\,(\hat{\beta}_2)_{\text{AR1}} = \frac{\sigma^2}{\sum x_t^2}\left[1 + 2\rho\frac{\sum x_t x_{t-1}}{\sum x_t^2} + 2\rho^2\frac{\sum x_t x_{t-2}}{\sum x_t^2} + \cdots + 2\rho^{n-1}\frac{x_1 x_n}{\sum x_t^2}\right] \tag{12.2.8}$$

where $\text{var}\,(\hat{\beta}_2)_{\text{AR1}}$ means the variance of $\hat{\beta}_2$ under first-order autoregressive scheme.

A comparison of (12.2.8) with (12.2.7) shows the former is equal to the latter times a term that depends on ρ as well as the sample autocorrelations between the values taken by the regressor X at various lags.[10] And in general we cannot foretell whether $\text{var}(\hat{\beta}_2)$ is less than or greater than $\text{var}(\hat{\beta}_2)_{\text{AR1}}$ [but see Eq. (12.4.1) below]. Of course, if rho is zero, the two formulas will coincide, as they should (why?). Also, if the correlations among the successive values of the regressor are very small, the usual OLS variance of the slope estimator will not be seriously biased. But, as a general principle, the two variances will not be the same.

To give some idea about the difference between the variances given in (12.2.7) and (12.2.8), assume that the regressor X also follows the first-order autoregressive scheme with a coefficient of autocorrelation of r. Then it can be shown that (12.2.8) reduces to:

$$\text{var}(\hat{\beta}_2)_{\text{AR(1)}} = \frac{\sigma^2}{\sum x_t^2}\left(\frac{1+r\rho}{1-r\rho}\right) = \text{var}(\hat{\beta}_2)_{\text{OLS}}\left(\frac{1+r\rho}{1-r\rho}\right) \qquad \textbf{(12.2.9)}$$

If, for example, $r = 0.6$ and $\rho = 0.8$, using (12.2.9) we can check that $\text{var}(\hat{\beta}_2)_{\text{AR1}} = 2.8461\ \text{var}(\hat{\beta}_2)_{\text{OLS}}$. To put it another way, $\text{var}(\hat{\beta}_2)_{\text{OLS}} = \frac{1}{2.8461}\text{var}(\hat{\beta}_2)_{\text{AR1}} = 0.3513\ \text{var}(\hat{\beta}_2)_{\text{AR1}}$. That is, the usual OLS formula [i.e., (12.2.7)] will underestimate the variance of $(\hat{\beta}_2)_{\text{AR1}}$ by about 65 percent. As you will realize, this answer is specific for the given values of r and ρ. But the point of this exercise is to warn you that a blind application of the usual OLS formulas to compute the variances and standard errors of the OLS estimators could give seriously misleading results.

Suppose we continue to use the OLS estimator $\hat{\beta}_2$ and adjust the usual variance formula by taking into account the AR(1) scheme. That is, we use $\hat{\beta}_2$ given by (12.2.6) but use the variance formula given by (12.2.8). What now are the properties of $\hat{\beta}_2$? It is easy to prove that $\hat{\beta}_2$ is still linear and unbiased. As a matter of fact, as shown in Appendix 3A, Section 3A.2, the assumption of no serial correlation, like the assumption of no heteroscedasticity, is not required to prove that $\hat{\beta}_2$ is unbiased. Is $\hat{\beta}_2$ still BLUE? Unfortunately, it is not; in the class of linear unbiased estimators, it does not have minimum variance. In short, $\hat{\beta}_2$, although linear-unbiased, is not efficient (relatively speaking, of course). The reader will notice that this finding is quite similar to the finding that $\hat{\beta}_2$ is less efficient in the presence of heteroscedasticity. There we saw that it was the weighted least-square estimator $\hat{\beta}_2^*$ given in (11.3.8), a special case of the generalized least-squares (GLS) estimator, that was efficient. In the case of autocorrelation can we find an estimator that is BLUE? The answer is yes, as can be seen from the discussion in the following section.

[10] Note that the term $r = \sum x_t x_{t+1} / \sum x_t^2$ is the correlation between X_t and X_{t+1} (or X_{t-1}, since the correlation coefficient is symmetric); $r^2 = \sum x_t x_{t+2} / \sum x_t^2$ is the correlation between the X's lagged two periods, and so on.

12.3 THE BLUE ESTIMATOR IN THE PRESENCE OF AUTOCORRELATION

Continuing with the two-variable model and assuming the AR(1) process, we can show that the BLUE estimator of β_2 is given by the following expression[11]:

$$\hat{\beta}_2^{\text{GLS}} = \frac{\sum_{t=2}^{n}(x_t - \rho x_{t-1})(y_t - \rho y_{t-1})}{\sum_{t=2}^{n}(x_t - \rho x_{t-1})^2} + C \qquad \textbf{(12.3.1)}$$

where C is a correction factor that may be disregarded in practice. Note that the subscript t now runs from $t = 2$ to $t = n$. And its variance is given by

$$\text{var}\,\hat{\beta}_2^{\text{GLS}} = \frac{\sigma^2}{\sum_{t=2}^{n}(x_t - \rho x_{t-1})^2} + D \qquad \textbf{(12.3.2)}$$

where D too is a correction factor that may also be disregarded in practice. (See exercise 12.18.)

The estimator $\hat{\beta}_2^{\text{GLS}}$, as the superscript suggests, is obtained by the method of GLS. As noted in Chapter 11, in GLS we incorporate any additional information we have (e.g., the nature of the heteroscedasticity or of the autocorrelation) directly into the estimating procedure by transforming the variables, whereas in OLS such side information is not directly taken into consideration. As the reader can see, the GLS estimator of β_2 given in (12.3.1) incorporates the autocorrelation parameter ρ in the estimating formula, whereas the OLS formula given in (12.2.6) simply neglects it. Intuitively, this is the reason why the GLS estimator is BLUE and not the OLS estimator—the GLS estimator makes the most use of the available information.[12] It hardly needs to be added that if $\rho = 0$, there is no additional information to be considered and hence both the GLS and OLS estimators are identical.

In short, under autocorrelation, it is the GLS estimator given in (12.3.1) that is BLUE, and the minimum variance is now given by (12.3.2) and not by (12.2.8) and obviously not by (12.2.7).

A Technical Note. As we noted in the previous chapter, the Gauss–Markov theorem provides only the sufficient condition for OLS to be BLUE. The necessary and sufficient conditions for OLS to be BLUE are given by

[11]For proofs, see Jan Kmenta, *Elements of Econometrics*, Macmillan, New York, 1971, pp. 274–275. The correction factor C pertains to the first observation, (Y_1, X_1). On this point see exercise 12.18.

[12]The formal proof that $\hat{\beta}_2^{\text{GLS}}$ is BLUE can be found in Kmenta, ibid. But the tedious algebraic proof can be simplified considerably using matrix notation. See J. Johnston, *Econometric Methods,* 3d ed., McGraw-Hill, New York, 1984, pp. 291–293.

Krushkal's theorem, mentioned in the previous chapter. Therefore, in some cases it can happen that OLS is BLUE despite autocorrelation. But such cases are infrequent in practice.

What happens if we blithely continue to work with the usual OLS procedure despite autocorrelation? The answer is provided in the following section.

12.4 CONSEQUENCES OF USING OLS IN THE PRESENCE OF AUTOCORRELATION

As in the case of heteroscedasticity, in the presence of autocorrelation the OLS estimators are still linear unbiased as well as consistent and asymptotically normally distributed, but they are no longer efficient (i.e., minimum variance). What then happens to our usual hypothesis testing procedures if we continue to use the OLS estimators? Again, as in the case of heteroscedasticity, we distinguish two cases. For pedagogical purposes we still continue to work with the two-variable model, although the following discussion can be extended to multiple regressions without much trouble.[13]

OLS Estimation Allowing for Autocorrelation

As noted, $\hat{\beta}_2$ is not BLUE, and even if we use $\text{var}(\hat{\beta}_2)_{\text{AR1}}$, the confidence intervals derived from there are likely to be wider than those based on the GLS procedure. As Kmenta shows, this result is likely to be the case even if the sample size increases indefinitely.[14] That is, $\hat{\beta}_2$ *is not asymptotically efficient*. The implication of this finding for hypothesis testing is clear: We are likely to declare a coefficient statistically insignificant (i.e., not different from zero) even though in fact (i.e., based on the correct GLS procedure) it may be. This difference can be seen clearly from Figure 12.4. In this figure we show the 95% OLS [AR(1)] and GLS confidence intervals assuming that true $\beta_2 = 0$. Consider a particular estimate of β_2, say, b_2. Since b_2 lies in the

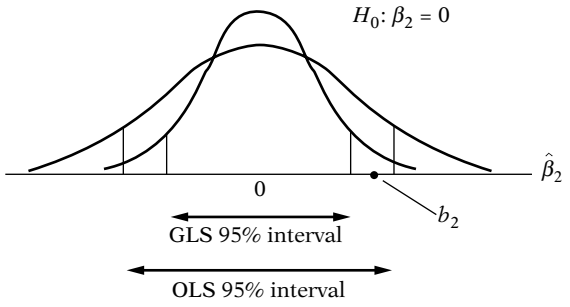

FIGURE 12.4 GLS and OLS 95% confidence intervals.

[13]But matrix algebra becomes almost a necessity to avoid tedious algebraic manipulations.
[14]See Kmenta, op. cit., pp. 277–278.

OLS confidence interval, we could accept the hypothesis that true β_2 is zero with 95% confidence. But if we were to use the (correct) GLS confidence interval, we could reject the null hypothesis that true β_2 is zero, for b_2 lies in the region of rejection.

The message is: To establish confidence intervals and to test hypotheses, one should use GLS and not OLS even though the estimators derived from the latter are unbiased and consistent. (However, see Section 12.11 later.)

OLS Estimation Disregarding Autocorrelation

The situation is potentially very serious if we not only use $\hat{\beta}_2$ but also continue to use $\text{var}(\hat{\beta}_2) = \sigma^2/\sum x_t^2$, which completely disregards the problem of autocorrelation, that is, we mistakenly believe that the usual assumptions of the classical model hold true. Errors will arise for the following reasons:

1. The residual variance $\hat{\sigma}^2 = \sum \hat{u}_i^2/(n-2)$ is likely to underestimate the true σ^2.

2. As a result, we are likely to overestimate R^2.

3. Even if σ^2 is not underestimated, $\text{var}(\hat{\beta}_2)$ may underestimate $\text{var}(\hat{\beta}_2)_{\text{AR1}}$ [Eq. (12.2.8)], its variance under (first-order) autocorrelation, even though the latter is inefficient compared to $\text{var}(\hat{\beta}_2)^{\text{GLS}}$.

4. Therefore, the usual t and F tests of significance are no longer valid, and if applied, are likely to give seriously misleading conclusions about the statistical significance of the estimated regression coefficients.

To establish some of these propositions, let us revert to the two-variable model. We know from Chapter 3 that under the classical assumption

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{(n-2)}$$

provides an unbiased estimator of σ^2, that is, $E(\hat{\sigma}^2) = \sigma^2$. But if there is autocorrelation, given by AR(1), it can be shown that

$$E(\hat{\sigma}^2) = \frac{\sigma^2\{n - [2/(1-\rho)] - 2\rho r\}}{n-2} \qquad \textbf{(12.4.1)}$$

where $r = \sum_{t=1}^{n-1} x_t x_{t-1} / \sum_{t=1}^{n} x_t^2$, which can be interpreted as the (sample) correlation coefficient between successive values of the X's.[15] If ρ and r are both positive (not an unlikely assumption for most economic time series), it is apparent from (12.4.1) that $E(\hat{\sigma}^2) < \sigma^2$; that is, the usual residual variance

[15] See S. M. Goldfeld and R. E. Quandt, *Nonlinear Methods in Econometrics*, North Holland Publishing Company, Amsterdam, 1972, p. 183. In passing, note that if the errors are positively autocorrelated, the R^2 value tends to have an upward bias, that is, it tends to be larger than the R^2 in the absence of such correlation.

formula, on average, will underestimate the true σ^2. In other words, $\hat{\sigma}^2$ will be biased downward. Needless to say, this bias in $\hat{\sigma}^2$ will be transmitted to $\text{var}(\hat{\beta}_2)$ because in practice we estimate the latter by the formula $\hat{\sigma}^2/\sum x_t^2$.

But even if σ^2 is not underestimated, $\text{var}(\hat{\beta}_2)$ is a *biased* estimator of $\text{var}(\hat{\beta}_2)_{\text{AR1}}$, which can be readily seen by comparing (12.2.7) with (12.2.8),[16] for the two formulas are not the same. As a matter of fact, if ρ is positive (which is true of most economic time series) and the X's are positively correlated (also true of most economic time series), then it is clear that

$$\text{var}(\hat{\beta}_2) < \text{var}(\hat{\beta}_2)_{\text{AR1}} \tag{12.4.2}$$

that is, the usual OLS variance of $\hat{\beta}_2$ underestimates its variance under AR(1) [see Eq. (12.2.9)]. Therefore, if we use $\text{var}(\hat{\beta}_2)$, we shall inflate the precision or accuracy (i.e., underestimate the standard error) of the estimator $\hat{\beta}_2$. As a result, in computing the t ratio as $t = \hat{\beta}_2/\text{se}(\hat{\beta}_2)$ (under the hypothesis that $\beta_2 = 0$), we shall be overestimating the t value and hence the statistical significance of the estimated β_2. The situation is likely to get worse if additionally σ^2 is underestimated, as noted previously.

To see how OLS is likely to underestimate σ^2 and the variance of $\hat{\beta}_2$, let us conduct the following **Monte Carlo experiment.** Suppose in the two-variable model we "know" that the true $\beta_1 = 1$ and $\beta_2 = 0.8$. Therefore, the stochastic PRF is

$$Y_t = 1.0 + 0.8X_t + u_t \tag{12.4.3}$$

Hence,

$$E(Y_t \mid X_t) = 1.0 + 0.8X_t \tag{12.4.4}$$

which gives the true population regression line. Let us assume that u_t are generated by the first-order autoregressive scheme as follows:

$$u_t = 0.7u_{t-1} + \varepsilon_t \tag{12.4.5}$$

where ε_t satisfy all the OLS assumptions. We assume further for convenience that the ε_t are normally distributed with zero mean and unit ($= 1$) variance. Equation (12.4.5) postulates that the successive disturbances are positively correlated, with a coefficient of autocorrelation of +0.7, a rather high degree of dependence.

Now, using a table of random normal numbers with zero mean and unit variance, we generated 10 random numbers shown in Table 12.1 and then by the scheme (12.4.5) we generated u_t. To start off the scheme, we need to specify the initial value of u, say, $u_0 = 5$.

Plotting the u_t generated in Table 12.1, we obtain Figure 12.5, which shows that initially each successive u_t is higher than its previous value and

[16] For a formal proof, see Kmenta, op. cit., p. 281.

TABLE 12.1 A HYPOTHETICAL EXAMPLE OF POSITIVELY AUTOCORRELATED ERROR TERMS

	ε_t^*	$u_t = 0.7u_{t-1} + \varepsilon_t$
0	0	$u_0 = 5$ (assumed)
1	0.464	$u_1 = 0.7(5) + 0.464 = 3.964$
2	2.026	$u_2 = 0.7(3.964) + 2.0262 = 4.8008$
3	2.455	$u_3 = 0.7(4.8010) + 2.455 = 5.8157$
4	−0.323	$u_4 = 0.7(5.8157) - 0.323 = 3.7480$
5	−0.068	$u_5 = 0.7(3.7480) - 0.068 = 2.5556$
6	0.296	$u_6 = 0.7(2.5556) + 0.296 = 2.0849$
7	−0.288	$u_7 = 0.7(2.0849) - 0.288 = 1.1714$
8	1.298	$u_8 = 0.7(1.1714) + 1.298 = 2.1180$
9	0.241	$u_9 = 0.7(2.1180) + 0.241 = 1.7236$
10	−0.957	$u_{10} = 0.7(1.7236) - 0.957 = 0.2495$

*Obtained from *A Million Random Digits and One Hundred Thousand Deviates,* Rand Corporation, Santa Monica, Calif., 1950.

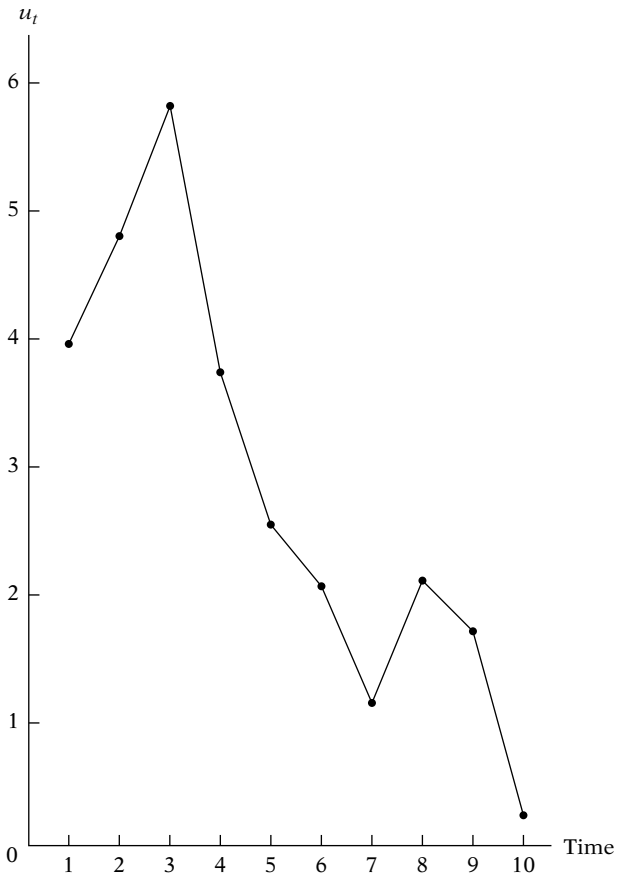

FIGURE 12.5 Correlation generated by the scheme $u_t = 0.7u_{t-1} + \varepsilon_t$ (Table 12.1).

TABLE 12.2 GENERATION OF *Y* SAMPLE VALUES

X_t	u_t^*	$Y_t = 1.0 + 0.8X_t + u_t$
1	3.9640	$Y_1 = 1.0 + 0.8(1) + 3.9640 = 5.7640$
2	4.8010	$Y_2 = 1.0 + 0.8(2) + 4.8008 = 7.4008$
3	5.8157	$Y_3 = 1.0 + 0.8(3) + 5.8157 = 9.2157$
4	3.7480	$Y_4 = 1.0 + 0.8(4) + 3.7480 = 7.9480$
5	2.5556	$Y_5 = 1.0 + 0.8(5) + 2.5556 = 7.5556$
6	2.0849	$Y_6 = 1.0 + 0.8(6) + 2.0849 = 7.8849$
7	1.1714	$Y_7 = 1.0 + 0.8(7) + 1.1714 = 7.7714$
8	2.1180	$Y_8 = 1.0 + 0.8(8) + 2.1180 = 9.5180$
9	1.7236	$Y_9 = 1.0 + 0.8(9) + 1.7236 = 9.9236$
10	0.2495	$Y_{10} = 1.0 + 0.8(10) + 0.2495 = 9.2495$

*Obtained from Table 12.1.

subsequently it is generally smaller than its previous value showing, in general, a positive autocorrelation.

Now suppose the values of X are fixed at 1, 2, 3, . . . , 10. Then, given these X's, we can generate a sample of 10 Y values from (12.4.3) and the values of u_t given in Table 12.1. The details are given in Table 12.2. Using the data of Table 12.2, if we regress Y on X, we obtain the following (sample) regression:

$$\begin{aligned} \hat{Y}_t = \quad & 6.5452 \; + \; 0.3051X_t \\ & (0.6153) \quad (0.0992) \\ t = \; & (10.6366) \quad (3.0763) \\ & r^2 = 0.5419 \qquad \hat{\sigma}^2 = 0.8114 \end{aligned} \tag{12.4.6}$$

whereas the true regression line is as given by (12.4.4). Both the regression lines are given in Figure 12.6, which shows clearly how much the fitted regression line distorts the true regression line; it seriously underestimates the true slope coefficient but overestimates the true intercept. (But note that the OLS estimators are still unbiased.)

Figure 12.6 also shows why the true variance of u_i is likely to be underestimated by the estimator $\hat{\sigma}^2$, which is computed from the $\hat{u}_i$. The $\hat{u}_i$ are generally close to the fitted line (which is due to the OLS procedure) but deviate substantially from the true PRF. Hence, they do not give a correct picture of u_i. To gain some insight into the extent of underestimation of true σ^2, suppose we conduct another sampling experiment. Keeping the X_t and ε_t given in Tables 12.1 and 12.2, let us assume $\rho = 0$, that is, no autocorrelation. The new sample of Y values thus generated is given in Table 12.3.

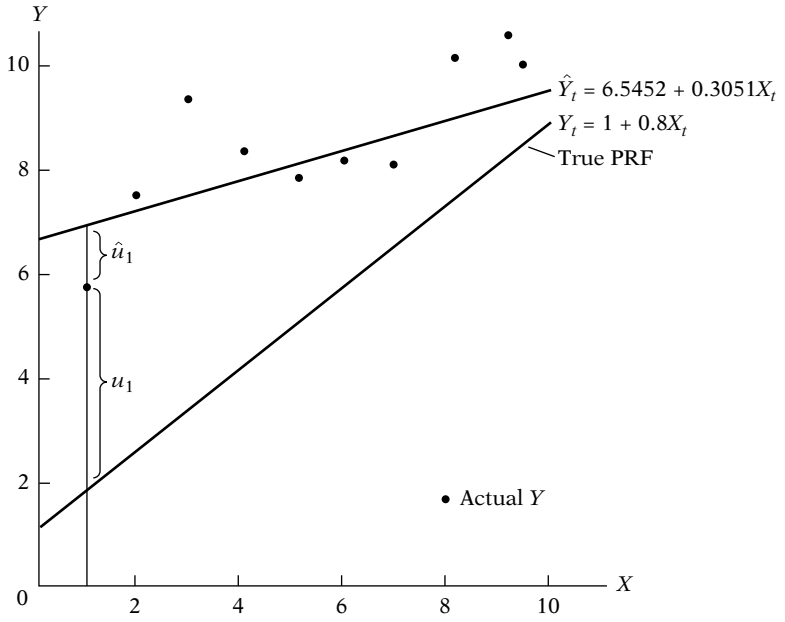

FIGURE 12.6 True PRF and the estimated regression line for the data of Table 12.2.

TABLE 12.3 SAMPLE OF *Y* VALUES WITH ZERO SERIAL CORRELATION

X_t	$\varepsilon_t = u_t^*$	$Y_t = 1.0 + 0.8X_t + \varepsilon_t$
1	0.464	2.264
2	2.026	4.626
3	2.455	5.855
4	−0.323	3.877
5	−0.068	4.932
6	0.296	6.096
7	−0.288	6.312
8	1.298	8.698
9	0.241	8.441
10	−0.957	8.043

*Since there is no autocorrelation, the u_t and ε_t are identical. The ε_t are from Table 12.1.

The regression based on Table 12.3 is as follows:

$$\begin{aligned} \hat{Y}_t &= 2.5345 + 0.6145X_t \\ &\quad (0.6796) \quad (0.1087) \\ t &= (3.7910) \quad (5.6541) \\ r^2 &= 0.7997 \qquad \hat{\sigma}^2 = 0.9752 \end{aligned} \tag{12.4.7}$$

This regression is much closer to the "truth" because the Y's are now essentially random. Notice that $\hat{\sigma}^2$ has increased from 0.8114 ($\rho = 0.7$) to 0.9752 ($\rho = 0$). Also notice that the standard errors of $\hat{\beta}_1$ and $\hat{\beta}_2$ have increased. This result is in accord with the theoretical results considered previously.

12.5 RELATIONSHIP BETWEEN WAGES AND PRODUCTIVITY IN THE BUSINESS SECTOR OF THE UNITED STATES, 1959–1998

Now that we have discussed the consequences of autocorrelation, the obvious question is, How do we detect it and how do we correct for it? Before we turn to these topics, it is useful to consider a concrete example. Table 12.4 gives data on indexes of real compensation per hour (Y) and output per hour (X) in the business sector of the U.S. economy for the period 1959–1998, the base of the indexes being 1992 = 100.

First plotting the data on Y and X, we obtain Figure 12.7. Since the relationship between real compensation and labor productivity is expected to be positive, it is not surprising that the two variables are positively related. What is surprising is that the relationship between the two is almost linear,

TABLE 12.4 INDEXES OF REAL COMPENSATION AND PRODUCTIVITY, UNITED STATES, 1959–1998

Observation	Y	X	Observation	Y	X
1959	58.5	47.2	1979	90.0	79.7
1960	59.9	48.0	1980	89.7	79.8
1961	61.7	49.8	1981	89.8	81.4
1962	63.9	52.1	1982	91.1	81.2
1963	65.3	54.1	1983	91.2	84.0
1964	67.8	54.6	1984	91.5	86.4
1965	69.3	58.6	1985	92.8	88.1
1966	71.8	61.0	1986	95.9	90.7
1967	73.7	62.3	1987	96.3	91.3
1968	76.5	64.5	1988	97.3	92.4
1969	77.6	64.8	1989	95.8	93.3
1970	79.0	66.2	1990	96.4	94.5
1971	80.5	68.8	1991	97.4	95.9
1972	82.9	71.0	1992	100.0	100.0
1973	84.7	73.1	1993	99.9	100.1
1974	83.7	72.2	1994	99.7	101.4
1975	84.5	74.8	1995	99.1	102.2
1976	87.0	77.2	1996	99.6	105.2
1977	88.1	78.4	1997	101.1	107.5
1978	89.7	79.5	1998	105.1	110.5

Notes: X = index of output per hour, business sector (1992 = 100)
Y = index of real compensation per hour, business sector (1992 = 100)
Source: Economic Report of the President, 2000, Table B–47, p. 362.

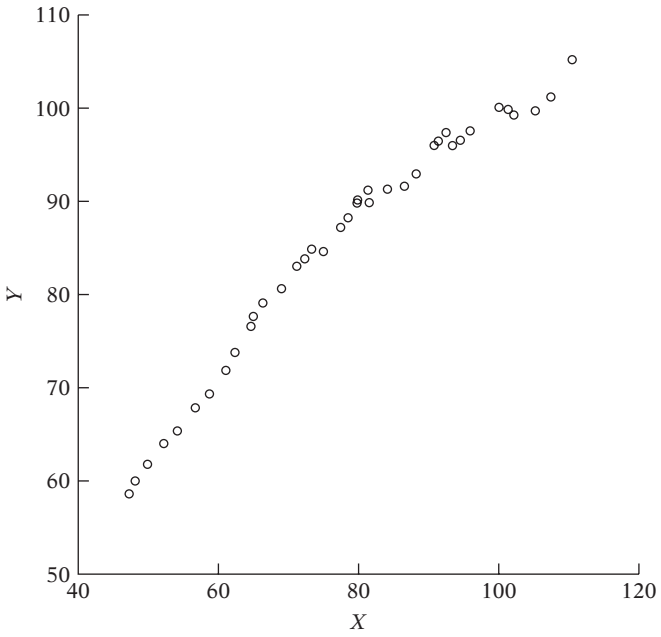

FIGURE 12.7 Index of compensation (Y) and index of productivity (X), United States, 1959–1998.

although there is some hint that at higher values of productivity the relationship between the two may be slightly nonlinear. Therefore, we decided to estimate a linear as well as a log–linear model, with the following results:

$$\begin{aligned} \hat{Y}_t &= 29.5192 + 0.7136X_t \\ \text{se} &= (1.9423) \quad (0.0241) \\ t &= (15.1977) \quad (29.6066) \end{aligned} \tag{12.5.1}$$

$$r^2 = 0.9584 \qquad d = 0.1229 \qquad \hat{\sigma} = 2.6755$$

where d is the Durbin–Watson statistic, which will be discussed shortly.

$$\begin{aligned} \widehat{\ln Y_t} &= 1.5239 + 0.6716 \ln X_t \\ \text{se} &= (0.0762) \quad (0.0175) \\ t &= (19.9945) \quad (38.2892) \end{aligned} \tag{12.5.2}$$

$$r^2 = 0.9747 \qquad d = 0.1542 \qquad \hat{\sigma} = 0.0260$$

For discussion purposes, we will call (12.5.1) and (12.5.2) wages–productivity regressions.

Qualitatively, both the models give similar results. In both cases the estimated coefficients are "highly" significant, as indicated by the high t values. In the linear model, if the index of productivity goes up by a unit, on average, the index of compensation goes up by about 0.71 units. In the log–linear model, the slope coefficient being elasticity (why?), we find that if the index of productivity goes up by 1 percent, on average, the index of real compensation goes up by about 0.67 percent.

How reliable are the results given in (12.5.1) and (12.5.2) if there is autocorrelation? As stated previously, if there is autocorrelation, the estimated standard errors are biased, as a result of which the estimated t ratios are unreliable. We obviously need to find out if our data suffer from autocorrelation. In the following section we discuss several methods of detecting autocorrelation. We will illustrate these methods with the linear model (12.5.1) only, leaving the log–linear model (12.5.2) as an exercise.

12.6 DETECTING AUTOCORRELATION

I. Graphical Method

Recall that the assumption of nonautocorrelation of the classical model relates to the population disturbances u_t, which are not directly observable. What we have instead are their proxies, the residuals $\hat{u}_t$, which can be obtained by the usual OLS procedure. Although the $\hat{u}_t$ are not the same thing as u_t,[17] very often a visual examination of the $\hat{u}$'s gives us some clues about the likely presence of autocorrelation in the u's. Actually, a visual examination of $\hat{u}_t$ or $(\hat{u}_t^2)$ can provide useful information not only about autocorrelation but also about heteroscedasticity (as we saw in the preceding chapter), model inadequacy, or specification bias, as we shall see in the next chapter. As one author notes:

> The importance of producing and analyzing plots of [residuals] as a standard part of statistical analysis cannot be overemphasized. Besides occasionally providing an easy to understand summary of a complex problem, they allow the simultaneous examination of the data as an aggregate while clearly displaying the behavior of individual cases.[18]

There are various ways of examining the residuals. We can simply plot them against time, the **time sequence plot,** as we have done in Figure 12.8, which shows the residuals obtained from the wages–productivity regression (12.5.1). The values of these residuals are given in Table 12.5 along with some other data.

---

[17]Even if the disturbances u_t are homoscedastic and uncorrelated, their estimators, the residuals, $\hat{u}_t$, are heteroscedastic and autocorrelated. On this, see G. S. Maddala, *Introduction to Econometrics*, 2d ed., Macmillan, New York, 1992, pp. 480–481. However, it can be shown that as the sample size increases indefinitely, the residuals tend to converge to their true values, the u_t's. On this see, E. Malinvaud, *Statistical Methods of Econometrics,* 2d ed., North-Holland Publishers, Amsterdam, 1970, p. 88.

[18]Stanford Weisberg, *Applied Linear Regression,* John Wiley & Sons, New York, 1980, p. 120.

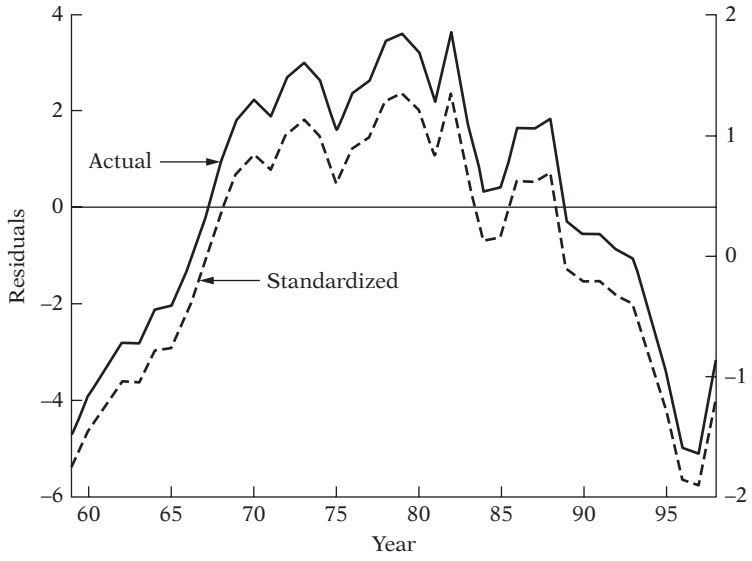

FIGURE 12.8 Residuals and standardized residuals from the wages–productivity regression (12.5.1).

TABLE 12.5 RESIDUALS: ACTUAL, STANDARDIZED, AND LAGGED

Observation	RES1	SRES1	RES1(−1)	Observation	RES1	SRES1	RES1(−1)
1959	−4.703979	−1.758168		1979	3.602089	1.346324	3.444821
1960	−3.874907	−1.448293	−4.703979	1980	3.230723	1.207521	3.602089
1961	−3.359494	−1.255651	−3.874907	1981	2.188868	0.818116	3.230723
1962	−2.800911	−1.046874	−3.359494	1982	3.631600	1.357354	2.188868
1963	−2.828229	−1.057084	−2.800911	1983	1.733354	0.647862	3.631600
1964	−2.112378	−0.789526	−2.828229	1984	0.320571	0.119817	1.733354
1965	−2.039697	−0.762361	−2.112378	1985	0.407350	0.152252	0.320571
1966	−1.252480	−0.468129	−2.039697	1986	1.651836	0.617393	0.407350
1967	−0.280237	−0.104742	−1.252480	1987	1.623640	0.606855	1.651836
1968	0.949713	0.354966	−0.280237	1988	1.838615	0.687204	1.623640
1969	1.835615	0.686083	0.949713	1989	−0.303679	−0.113504	1.838615
1970	2.236492	0.835915	1.835615	1990	−0.560070	−0.209333	−0.303679
1971	1.880977	0.703038	2.236492	1991	−0.559193	−0.209005	−0.560070
1972	2.710926	1.013241	1.880977	1992	−0.885197	−0.330853	−0.559193
1973	3.012241	1.125861	2.710926	1993	−1.056563	−0.394903	−0.885197
1974	2.654535	0.992164	3.012241	1994	−2.184320	−0.816416	−1.056563
1975	1.599020	0.597653	2.654535	1995	−3.355248	−1.254064	−2.184320
1976	2.386238	0.891885	1.599020	1996	−4.996226	−1.867399	−3.355248
1977	2.629847	0.982936	2.386238	1997	−5.137643	−1.920255	−4.996226
1978	3.444821	1.287543	2.629847	1998	−3.278621	−1.225424	−5.137643

Notes: RES 1 = residuals from regression (12.5.1).
SRES 1 = standardized residuals = RES1/2.6755.
RES(−1) = residuals lagged one period.

Alternatively, we can plot the **standardized residuals** against time, which are also shown in Figure 12.8 and Table 12.5. The standardized residuals are simply the residuals ($\hat{u}_t$) divided by the standard error of the regression ($\sqrt{\hat{\sigma}^2}$), that is, they are ($\hat{u}_t/\hat{\sigma}$). Notice that $\hat{u}_t$ and $\hat{\sigma}$ are measured in the units in which the regressand Y is measured. The values of the standardized residuals will therefore be pure numbers (devoid of units of measurement) and can be compared with the standardized residuals of other regressions. Moreover, the standardized residuals, like $\hat{u}_t$, have zero mean (why?) and *approximately* unit variance.[19] In large samples ($\hat{u}_t/\hat{\sigma}$) is approximately normally distributed with zero mean and unit variance. For our example, $\hat{\sigma} = 2.6755$.

Examining the time sequence plot given in Figure 12.8, we observe that both $\hat{u}_t$ and the standardized $\hat{u}_t$ exhibit a pattern observed in Figure 12.1*d*, suggesting that perhaps u_t are not random.

To see this differently, we can plot $\hat{u}_t$ against $\hat{u}_{t-1}$, that is, plot the residuals at time t against their value at time $(t - 1)$, a kind of empirical test of the AR(1) scheme. If the residuals are nonrandom, we should obtain pictures similar to those shown in Figure 12.3. This plot for our wages–productivity regression is as shown in Figure 12.9; the underlying data are given in

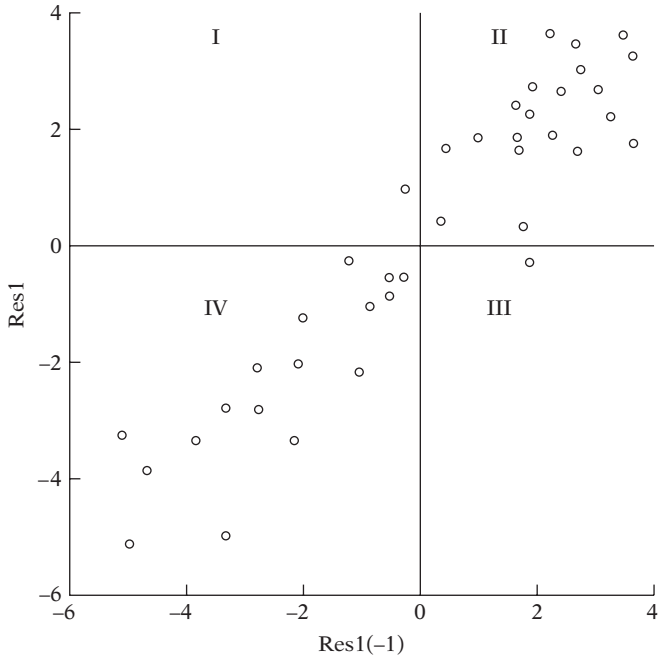

FIGURE 12.9 Current residuals versus lagged residuals.

[19]Actually, it is the so-called **Studentized** residuals that have a unit variance. But in practice the standardized residuals will give the same picture, and hence we may rely on them. On this, see Norman Draper and Harry Smith, *Applied Regression Analysis,* 3d ed., John Wiley & Sons, New York, 1998, pp. 207–208.

Table 12.5. As this figure reveals, most of the residuals are bunched in the second (northeast) and the fourth (southwest) quadrants, suggesting a strong positive correlation in the residuals.

The graphical method we have just discussed, although powerful and suggestive, is subjective or qualitative in nature. But there are several quantitative tests that one can use to supplement the purely qualitative approach. We now consider some of these tests.

II. The Runs Test

If we carefully examine Figure 12.8, we notice a peculiar feature: Initially, we have several residuals that are negative, then there is a series of positive residuals, and then there are several residuals that are negative. If these residuals were purely random, could we observe such a pattern? Intuitively, it seems unlikely. This intuition can be checked by the so-called **runs test,** sometimes also know as the **Geary test,** a nonparametric test.[20]

To explain the runs test, let us simply note down the signs (+ or −) of the residuals obtained from the wages–productivity regression, which are given in the first column of Table 12.5.

$$(---------)(+++++++++++++++++++++)(----------) \tag{12.6.1}$$

Thus there are 9 negative residuals, followed by 21 positive residuals, followed by 10 negative residuals, for a total of 40 observations.

We now define a **run** as an uninterrupted sequence of one symbol or attribute, such as + or −. We further define the **length of a run** as the number of elements in it. In the sequence shown in (12.6.1), there are 3 runs: a run of 9 minuses (i.e., of length 9), a run of 21 pluses (i.e., of length 21) and a run of 10 minuses (i.e., of length 10). For a better visual effect, we have presented the various runs in parentheses.

By examining how runs behave in a strictly random sequence of observations, one can derive a test of randomness of runs. We ask this question: Are the 3 runs observed in our illustrative example consisting of 40 observations too many or too few compared with the number of runs expected in a strictly random sequence of 40 observations? If there are too many runs, it would mean that in our example the residuals change sign frequently, thus indicating negative serial correlation (cf. Figure 12.3*b*). Similarly, if there are too few runs, they may suggest positive autocorrelation, as in Figure 12.3*a*. A priori, then, Figure 12.8 would indicate positive correlation in the residuals.

[20]In **nonparametric** tests we make no assumptions about the (probability) distribution from which the observations are drawn. On the Geary test, see R. C. Geary, "Relative Efficiency of Count Sign Changes for Assessing Residual Autoregression in Least Squares Regression," *Biometrika*, vol. 57, 1970, pp. 123–127.

Now let

$$\begin{aligned} N &= \text{total number of observations} = N_1 + N_2 \\ N_1 &= \text{number of } + \text{ symbols (i.e., } + \text{ residuals)} \\ N_2 &= \text{number of } - \text{ symbols (i.e., } - \text{ residuals)} \\ R &= \text{number of runs} \end{aligned}$$

Then under the null hypothesis that the successive outcomes (here, residuals) are independent, and assuming that $N_1 > 10$ and $N_2 > 10$, the number of runs is (*asymptotically*) *normally distributed* with

$$\begin{aligned} \text{Mean:} \quad & E(R) = \frac{2N_1N_2}{N} + 1 \\ \text{Variance:} \quad & \sigma_R^2 = \frac{2N_1N_2(2N_1N_2 - N)}{(N)^2(N-1)} \end{aligned} \tag{12.6.2}$$

Note: $N = N_1 + N_2$.

If the null hypothesis of randomness is sustainable, following the properties of the normal distribution, we should expect that

$$\text{Prob}\,[E(R) - 1.96\sigma_R \leq R \leq E(R) + 1.96\sigma_R] = 0.95 \tag{12.6.3}$$

That is, the probability is 95 percent that the preceding interval will include R. Therefore we have this rule:

> **Decision Rule.** Do not reject the null hypothesis of randomness with 95% confidence if R, the number of runs, lies in the preceding confidence interval; reject the null hypothesis if the estimated R lies outside these limits. (*Note:* You can choose any level of confidence you want.)

Returning to our example, we know that N_1, the number of minuses, is 19 and N_2, the number of pluses, is 21 and $R = 3$. Using the formulas given in (12.6.2), we obtain:

$$\begin{aligned} E(R) &= 10.975 \\ \sigma_R^2 &= 9.6936 \\ \sigma_R &= 3.1134 \end{aligned} \tag{12.6.4}$$

The 95% confidence interval for R in our example is thus:

$$[10.975 \pm 1.96(3.1134)] = (4.8728, 17.0722)$$

Obviously, this interval does not include 3. Hence, we can *reject* the hypothesis that the residuals in our wages–productivity regression are random

with 95% confidence. In other words, the residuals exhibit autocorrelation. As a general rule, if there is positive autocorrelation, the number of runs will be few, whereas if there is negative autocorrelation, the number of runs will be many. Of course, from (12.6.2) we can find out whether we have too many runs or too few runs.

Swed and Eisenhart have developed special tables that give critical values of the runs expected in a random sequence of N observations if N_1 or N_2 is smaller than 20. These tables are given in **Appendix D,** Table D.6. Using these tables, the reader can verify that the residuals in our wages–productivity regression are indeed nonrandom; actually they are positively correlated.

III. Durbin–Watson *d* Test[21]

The most celebrated test for detecting serial correlation is that developed by statisticians Durbin and Watson. It is popularly known as the **Durbin–Watson *d* statistic,** which is defined as

$$d = \frac{\sum_{t=2}^{t=n}(\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^{t=n}\hat{u}_t^2} \tag{12.6.5}$$

which is simply the ratio of the sum of squared differences in successive residuals to the RSS. Note that in the numerator of the d statistic the number of observations is $n - 1$ because one observation is lost in taking successive differences.

A great advantage of the d statistic is that it is based on the estimated residuals, which are routinely computed in regression analysis. Because of this advantage, it is now a common practice to report the Durbin–Watson d along with summary measures, such as R^2, adjusted R^2, t, and F. Although it is now routinely used, it is **important to note the assumptions underlying the *d* statistic.**

1. The regression model includes the intercept term. If it is not present, as in the case of the regression through the origin, it is essential to rerun the regression including the intercept term to obtain the RSS.[22]

2. The explanatory variables, the X's, are nonstochastic, or fixed in repeated sampling.

3. The disturbances u_t are generated by the first-order autoregressive scheme: $u_t = \rho u_{t-1} + \varepsilon_t$. Therefore, it cannot be used to detect higher-order autoregressive schemes.

4. The error term u_t is assumed to be normally distributed.

[21] J. Durbin and G. S. Watson, "Testing for Serial Correlation in Least-Squares Regression," *Biometrika*, vol. 38, 1951, pp. 159–171.

[22] However, R. W. Farebrother has calculated d values when the intercept term is absent from the model. See his "The Durbin–Watson Test for Serial Correlation When There Is No Intercept in the Regression," *Econometrica,* vol. 48, 1980, pp. 1553–1563.

5. The regression model does not include the lagged value(s) of the dependent variable as one of the explanatory variables. Thus, the test is inapplicable in models of the following type:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \cdots + \beta_k X_{kt} + \gamma Y_{t-1} + u_t \qquad \textbf{(12.6.6)}$$

where Y_{t-1} is the one period lagged value of Y. Such models are known as **autoregressive models,** which we will study in Chapter 17.

6. There are no missing observations in the data. Thus, in our wages–productivity regression for the period 1959–1998, if observations for, say, 1978 and 1982 were missing for some reason, the d statistic makes no allowance for such missing observations.[23]

The exact sampling or probability distribution of the d statistic given in (12.6.5) is difficult to derive because, as Durbin and Watson have shown, it depends in a complicated way on the X values present in a given sample.[24] This difficulty should be understandable because d is computed from $\hat{u}_t$, which are, of course, dependent on the given X's. Therefore, unlike the *t, F,* or χ^2 tests, there is no unique critical value that will lead to the rejection or the acceptance of the null hypothesis that there is no first-order serial correlation in the disturbances u_i. However, Durbin and Watson were successful in deriving a lower bound d_L and an upper bound d_U such that if the computed d from (12.6.5) lies outside these critical values, a decision can be made regarding the presence of positive or negative serial correlation. Moreover, these limits depend only on the number of observations n and the number of explanatory variables and do not depend on the values taken by these explanatory variables. These limits, for n going from 6 to 200 and up to 20 explanatory variables, have been tabulated by Durbin and Watson and are reproduced in **Appendix D,** Table D.5 (up to 20 explanatory variables).

The actual test procedure can be explained better with the aid of Figure 12.10, which shows that the limits of d are 0 and 4. These can be established as follows. Expand (12.6.5) to obtain

$$d = \frac{\sum \hat{u}_t^2 + \sum \hat{u}_{t-1}^2 - 2\sum \hat{u}_t \hat{u}_{t-1}}{\sum \hat{u}_t^2} \qquad \textbf{(12.6.7)}$$

Since $\sum \hat{u}_t^2$ and $\sum \hat{u}_{t-1}^2$ differ in only one observation, they are approximately equal. Therefore, setting $\sum \hat{u}_{t-1}^2 \approx \sum \hat{u}_t^2$, (12.6.7) may be written as

$$d \approx 2\left(1 - \frac{\sum \hat{u}_t \hat{u}_{t-1}}{\sum \hat{u}_t^2}\right) \qquad \textbf{(12.6.8)}$$

where $\approx$ means approximately.

[23]For further details, see Gabor Korosi, Laszlo Matyas, and Istvan P. Szekey, *Practical Econometrics,* Avebury Press, England, 1992, pp. 88–89.

[24]But see the discussion on the "exact" Durbin–Watson test given later in the section.

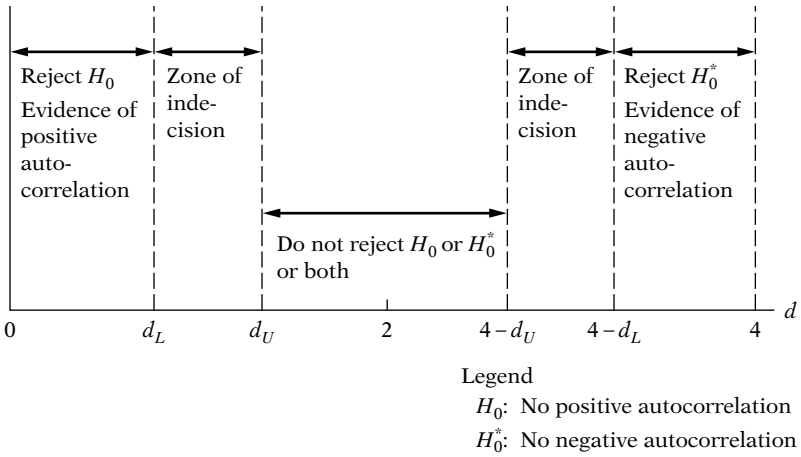

FIGURE 12.10 Durbin–Watson *d* statistic.

Now let us define

$$\hat{\rho} = \frac{\sum \hat{u}_t \hat{u}_{t-1}}{\sum \hat{u}_t^2} \tag{12.6.9}$$

as the sample first-order coefficient of autocorrelation, an estimator of ρ. (See footnote 9.) Using (12.6.9), we can express (12.6.8) as

$$d \approx 2(1 - \hat{\rho}) \tag{12.6.10}$$

But since $-1 \leq \rho \leq 1$, (12.6.10) implies that

$$0 \leq d \leq 4 \tag{12.6.11}$$

These are the bounds of d; any estimated d value must lie within these limits.

It is apparent from Eq. (12.6.10) that if $\hat{\rho} = 0$, $d = 2$; that is, if there is no serial correlation (of the first-order), d is expected to be about 2. *Therefore, as a rule of thumb, if d is found to be 2 in an application, one may assume that there is no first-order autocorrelation, either positive or negative.* If $\hat{\rho} = +1$, indicating perfect positive correlation in the residuals, $d \approx 0$. Therefore, the closer d is to 0, the greater the evidence of positive serial correlation. This relationship should be evident from (12.6.5) because if there is positive autocorrelation, the $\hat{u}_t$'s will be bunched together and their differences will therefore tend to be small. As a result, the numerator sum of squares will be smaller in comparison with the denominator sum of squares, which remains a unique value for any given regression.

If $\hat{\rho} = -1$, that is, there is perfect negative correlation among successive residuals, $d \approx 4$. Hence, the closer d is to 4, the greater the evidence of negative serial correlation. Again, looking at (12.6.5), this is understandable. For if there is negative autocorrelation, a positive $\hat{u}_t$ will tend to be followed by a negative $\hat{u}_t$ and vice versa so that $|\hat{u}_t - \hat{u}_{t-1}|$ will usually be greater than $|\hat{u}_t|$. Therefore, the numerator of d will be comparatively larger than the denominator.

The mechanics of the Durbin–Watson test are as follows, assuming that the assumptions underlying the test are fulfilled:

1. Run the OLS regression and obtain the residuals.

2. Compute d from (12.6.5). (Most computer programs now do this routinely.)

3. For the given sample size and given number of explanatory variables, find out the critical d_L and d_U values.

4. Now follow the decision rules given in Table 12.6. For ease of reference, these decision rules are also depicted in Figure 12.10.

To illustrate the mechanics, let us return to our wages–productivity regression. From the data given in Table 12.5 the estimated d value can be shown to be 0.1229, suggesting that there is positive serial correlation in the residuals. From the Durbin–Watson tables, we find that for 40 observations and one explanatory variable, $d_L = 1.44$ and $d_U = 1.54$ at the 5 percent level. Since the computed d of 0.1229 lies below d_L, we cannot reject the hypothesis that there is positive serial correlations in the residuals.

Although extremely popular, the d test has one great drawback in that, if it falls in the **indecisive zone,** one cannot conclude that (first-order) autocorrelation does or does not exist. To solve this problem, several authors have proposed modifications of the d test but they are rather involved and beyond the scope of this book.[25] In many situations, however, it has been found that the upper limit d_U is approximately the true significance limit and therefore in case d lies in the indecisive zone, one can use the following **modified *d* test:** Given the level of significance α,

1. H_0: $\rho = 0$ versus H_1: $\rho > 0$. Reject H_0 at α level if $d < d_U$. That is, there is statistically significant positive autocorrelation.

TABLE 12.6 DURBIN–WATSON *d* TEST: DECISION RULES

Null hypothesis	Decision	If
No positive autocorrelation	Reject	$0 < d < d_L$
No positive autocorrelation	No decision	$d_L \leq d \leq d_U$
No negative correlation	Reject	$4 - d_L < d < 4$
No negative correlation	No decision	$4 - d_U \leq d \leq 4 - d_L$
No autocorrelation, positive or negative	Do not reject	$d_U < d < 4 - d_U$

[25]For details, see Thomas B. Fomby, R. Carter Hill, and Stanley R. Johnson, *Advanced Econometric Methods,* Springer Verlag, New York, 1984, pp. 225–228.

2. $H_0: \rho = 0$ versus $H_1: \rho < 0$. Reject H_0 at α level if the estimated $(4 - d) < d_U$, that is, there is statistically significant evidence of negative autocorrelation.

3. $H_0: \rho = 0$ versus $H_1: \rho \neq 0$. Reject H_0 at 2α level if $d < d_U$ or $(4 - d) < d_U$, that is, there is statistically significant evidence of autocorrelation, positive or negative.

It may be pointed out that the indecisive zone narrows as the sample size increases, which can be seen clearly from the Durbin–Watson tables. For example, with 4 regressors and 20 observations, the 5 percent lower and upper d values are 0.894 and 1.828, respectively, but these values are 1.515 and 1.739 if the sample size is 75.

The computer program Shazam performs an *exact d test,* that is, it gives the p value, the exact probability of the computed d value. With modern computing facilities, it is no longer difficult to find the p value of the computed d statistic. Using SHAZAM (version 9) for our wages–productivity regression, we find the p value of the computed d of 0.1229 is practically zero, thereby reconfirming our earlier conclusion based on the Durbin–Watson tables.

The Durbin–Watson d test has become so venerable that practitioners often forget the assumptions underlying the test. In particular, the assumptions that (1) the explanatory variables, or regressors, are nonstochastic; (2) the error term follows the normal distribution; and (3) that the regression models do not include the lagged value(s) of the regressand are very important for the application of the d test.

If a regression model contains lagged value(s) of the regressand, the d value in such cases is often around 2, which would suggest that there is no (first-order) autocorrelation in such models. Thus, there is a built-in bias against discovering (first-order) autocorrelation in such models. This does not mean that autoregressive models do not suffer from the autocorrelation problem. As a matter of fact, Durbin has developed the so-called ***h* test** to test serial correlation in such models. But this test is not as powerful, in a statistical sense, as the **Breusch–Godfrey test** to be discussed shortly, so there is no need to use the ***h* test.** However, because of its historical importance, it is discussed in exercise 12.36.

Also, if the error term u_t are not NIID, the routinely used d test may not be reliable.[26] In this respect the **runs test** discussed earlier has an advantage in that it does not make any (probability) distributional assumption about the error term. However, if the sample is large (technically infinite), we can use the Durbin–Watson d, for it can be shown that[27]

$$\sqrt{n}\left(1 - \frac{1}{2}d\right) \approx N(0, 1) \qquad \textbf{(12.6.12)}$$

[26]For an advanced discussion, see Ron C. Mittelhammer, George G. Judge, and Douglas J. Miller, *Econometric Foundations,* Cambridge University Press, New York, 2000, p. 550.

[27]See James Davidson, *Econometric Theory,* Blackwell Publishers, New York, 2000, p. 161.

That is, in large samples the d statistic as transformed in (12.6.12) follows the standard normal distribution. Incidentally, in view of the relationship between d and $\hat{\rho}$, the estimated first-order autocorrelation coefficient, shown in (12.6.10), it follows that

$$\sqrt{n}\hat{\rho} \approx N(0, 1) \qquad \textbf{(12.6.13)}$$

that is, in large samples, the square root of the sample size times the estimated first-order autocorrelation coefficient also follows the standard normal distribution.

As an illustration of the test, for our wages–productivity example, we found that $d = 0.1229$ with $n = 40$. Therefore, from (12.6.12) we find that

$$\sqrt{40}\left(1 - \frac{0.1229}{2}\right) \approx 5.94$$

Asymptotically, if the null hypothesis of zero (first-order) autocorrelation were true, the probability of obtaining a Z value (i.e., a standardized normal variable) of as much as 5.94 or greater is extremely small. Recall that for a standard normal distribution, the (two-tail) critical 5 percent Z value is only 1.96 and the 1 percent critical Z value is about 2.58. Although our sample size is only 40, for practical purposes it may be large enough to use the normal approximation. The conclusion remains the same, namely, that the residuals from the wages–productivity regression suffer from autocorrelation.

But the most serious problem with the d test is the assumption that the regressors are nonstochastic, that is, their values are fixed in repeated sampling. If this is not the case, then the d test is not valid either in finite, or small, samples or in large samples.[28] And since this assumption is usually difficult to maintain in economic models involving time series data, one author contends that the Durbin–Watson statistic may not be useful in econometrics involving time series data.[29] In his view, more useful tests of autocorrelation are available, but they are all based on large samples. We discuss one such test below, the **Breusch–Godfrey test.**

IV. A General Test of Autocorrelation: The Breusch–Godfrey (BG) Test[30]

To avoid some of the pitfalls of the Durbin–Watson d test of autocorrelation, statisticians Breusch and Godfrey have developed a test of autocorrelation that is general in the sense that it allows for (1) nonstochastic regressors, such as the lagged values of the regressand; (2) higher-order autoregressive

[28]Ibid., p. 161.

[29]Fumio Hayashi, *Econometrics,* Princeton University Press, Princeton, N.J., 2000, p. 45.

[30]See, L. G. Godfrey, "Testing Against General Autoregressive and Moving Average Error Models When the Regressor include Lagged Dependent Variables," *Econometrica*, vol. 46, 1978, pp. 1293–1302, and T. S. Breusch, "Testing for Autocorrelation in Dynamic Linear Models," *Australian Economic Papers,* vol. 17, 1978, pp. 334–355.

schemes, such as AR(1), AR(2), etc.; and (3) simple or higher-order **moving averages** of white noise error terms, such as ε_t in (12.2.1).[31]

Without going into the mathematical details, which can be obtained from the references, the **BG test,** which is also known as the **LM test,**[32] proceeds as follows: We use the two-variable regression model to illustrate the test, although many regressors can be added to the model. Also, lagged values of the regressand can be added to the model. Let

$$Y_t = \beta_1 + \beta_2 X_t + u_t \tag{12.6.14}$$

Assume that the error term u_t follows the pth-order autoregressive, AR(p), scheme as follows:

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \cdots + \rho_p u_{t-p} + \varepsilon_t \tag{12.6.15}$$

where ε_t is a white noise error term as discussed previously. As you will recognize, this is simply the extension of the AR(1) scheme.

The null hypothesis H_0 to be tested is that

$$H_0: \rho_1 = \rho_2 = \cdots = \rho_p = 0 \tag{12.6.16}$$

That is, there is no serial correlation of any order. The BG test involves the following steps:

1. Estimate (12.6.14) by OLS and obtain the residuals, $\hat{u}_t$.

2. Regress $\hat{u}_t$ on the original X_t (if there is more than one X variable in the original model, include them also) and $\hat{u}_{t-1}, \hat{u}_{t-2}, \ldots, \hat{u}_{t-p}$, where the latter are the lagged values of the estimated residuals in step 1. Thus, if $p = 4$, we will introduce four lagged values of the residuals as additional regressor in the model. Note that to run this regression we will have only $(n - p)$ observations (why?). In short, run the following regression:

$$\hat{u}_t = \alpha_1 + \alpha_2 X_t + \hat{\rho}_1 \hat{u}_{t-1} + \hat{\rho}_2 \hat{u}_{t-2} + \cdots + \hat{\rho}_p \hat{u}_{t-p} + \varepsilon_t \tag{12.6.17}$$

and obtain R^2 from this (auxiliary) regression.[33]

3. If the sample size is large (technically, infinite), Breusch and Godfrey have shown that

$$(n - p)R^2 \sim \chi^2_p \tag{12.6.18}$$

[31]For example, in the regression $Y_t = \beta_1 + \beta_2 X_t + u_t$ the error term can be represented as $u_t = \varepsilon_t + \lambda_1 \varepsilon_{t-1} + \lambda_2 \varepsilon_{t-2}$, which represents a three-period moving average of the white noise error term ε_t.

[32]The test is based on the **Lagrange Multiplier principle** briefly mentioned in Chap. 8.

[33]The reason that the original regressor X is included in the model is to allow for the fact that X may not be strictly nonstochastic. But if it is strictly nonstochastic, it may be omitted from the model. On this, see Jeffrey M. Wooldridge, *Introductory Econometrics: A Modern Approach,* South-Western Publishing Co., 200, p. 386.

That is, asymptotically, $n - p$ times the R^2 value obtained from the auxiliary regression (12.6.17) follows the chi-square distribution with p df. If in an application, $(n - p)R^2$ exceeds the critical chi-square value at the chosen level of significance, we reject the null hypothesis, in which case at least one rho in (12.6.15) is statistically significantly different from zero.

The following *practical points* about the BG test may be noted:

1. The regressors included in the regression model may contain lagged values of the regressand Y, that is, Y_{t-1}, Y_{t-2}, etc., may appear as explanatory variables. Contrast this model with the Durbin–Watson test restriction that there be no lagged values of the regressand among the regressors.

2. As noted earlier, the BG test is applicable even if the disturbances follow a pth-order **moving average (MA)** process, that is, the u_t are generated as follows:

$$u_t = \varepsilon_t + \lambda_1\varepsilon_{t-1} + \lambda_2\varepsilon_{t-2} + \cdots + \lambda_p\varepsilon_{t-p} \qquad \textbf{(12.6.19)}$$

where ε_t is a white noise error term, that is, the error term that satisfies all the classical assumptions.

In the chapters on time series econometrics, we will study in some detail the pth-order autoregressive and moving average processes.

3. If in (12.6.15) $p = 1$, meaning first-order autoregression, then the BG test is known as **Durbin's *M* test.**

4. A drawback of the BG test is that the value of p, the length of the lag, cannot be specified a priori. Some experimentation with the p value is inevitable. Sometimes one can use the so-called **Akaike** and **Schwarz** information criteria to select the lag length. We will discuss these criteria in Chapter 13 and later in the chapters on time series econometrics.

ILLUSTRATION OF THE BG TEST:
THE WAGES–PRODUCTIVITY RELATION

To illustrate the test, we will apply it to our illustrative example. Using an AR(6) scheme, we obtained the results shown in exercise 12.25. From the regression results given there, it can be seen that $(n - p) = 34$ and $R^2 = 0.8920$. Therefore, multiplying these two, we obtain a chi-square value of 30.328. For 6 df (why?), the probability of obtaining a chi-square value of as much as 30.328 or greater is extremely small; the chi-square table in Appendix D.4 shows that the probability of obtaining a chi-square value of as much as 18.5476 or greater is only 0.005. Therefore, for the same df, the probability of obtaining a chi-square value of about 30 must be extremely small. As a matter of fact, the actual p value is almost zero.

Therefore, the conclusion is that, for our example, at least one of the six autocorrelations must be nonzero.

Trying varying lag lengths from 1 to 6, we find that only the AR(1) coefficient is significant, suggesting that there is no need to consider more than one lag. In essence the BG test in this case turns out to be **Durbin's *m* test.**

Why So Many Tests of Autocorrelation?

The answer to this question is that "... no particular test has yet been judged to be unequivocally best [i.e., more powerful in the statistical sense], and thus the analyst is still in the unenviable position of considering a

varied collection of test procedures for detecting the presence or structure, or both, of autocorrelation."[34] Of course, a similar argument can be made about the various tests of heteroscedasticity discussed in the previous chapter.

12.7 WHAT TO DO WHEN YOU FIND AUTOCORRELATION: REMEDIAL MEASURES

If after applying one or more of the diagnostic tests of autocorrelation discussed in the previous section, we find that there is autocorrelation, what then? We have four options:

1. Try to find out if the autocorrelation is **pure autocorrelation** and not the result of mis-specification of the model. As we discussed in Section 12.1, sometimes we observe patterns in residuals because the model is mis-specified—that is, it has excluded some important variables—or because its functional form is incorrect.

2. If it is pure autocorrelation, one can use appropriate transformation of the original model so that in the transformed model we do not have the problem of (pure) autocorrelation. As in the case of heteroscedasticity, we will have to use some type of **generalized least-square (GLS) method.**

3. In large samples, we can use the **Newey–West** method to obtain standard errors of OLS estimators that are corrected for autocorrelation. This method is actually an extension of White's heteroscedasticity-consistent standard errors method that we discussed in the previous chapter.

4. In some situations we can continue to use the OLS method.

Because of the importance of each of these topics, we devote a section to each one.

12.8 MODEL MIS-SPECIFICATION VERSUS PURE AUTOCORRELATION

Let us return to our wages–productivity regression given in (12.5.1). There we saw that the d value was 0.1229 and based on the Durbin–Watson d test we concluded that there was positive correlation in the error term. Could this correlation have arisen because our model was not correctly specified? Since the data underlying regression (12.5.1) is time series data, it is quite possible that both wages and productivity exhibit trends. If that is the case,

[34]Ron C. Mittelhammer et al., op. cit., p. 547. Recall that the **power of a statistical test** is one minus the probability of committing a Type II error, that is, one minus the probability of accepting a false hypothesis. The maximum power of a test is 1 and the minimum is 0. The closer of the power of a test is to zero, the worse is that test, and the closer it is to 1, the more powerful is that test. What these authors are essentially saying is that there is no single most powerful test of autocorrelation.

then we need to include the time or trend, t, variable in the model to see the relationship between wages and productivity net of the trends in the two variables.

To test this, we included the trend variable in (12.5.1) and obtained the following results

$$\begin{aligned} \hat{Y}_t = \quad & 1.4752 \; + \; 1.3057X_t - \quad 0.9032t \\ \text{se} = (13.18) \quad & (0.2765) \quad (0.4203) \\ t = \; (0.1119) \quad & (4.7230) \quad (-2.1490) \\ & R^2 = 0.9631; \qquad d = 0.2046 \end{aligned} \tag{12.8.1}$$

The interpretation of this model is straightforward: Over time, the index of real wages has been decreasing by about 0.90 units per year. After allowing for this, if the productivity index went up by a unit, on average, the real wage index went up by about 1.30 units, although this number is not statistically different from one (why?). What is interesting to note is that even allowing for the trend variable, the d value is still very low, suggesting that (12.8.1) suffers from pure autocorrelation and not necessarily specification error.

How do we know that (12.8.1) is the correct specification? To test this, we regress Y on X and X^2 to test for the possibility that the real wage index may be nonlinearly related to the productivity index. The results of this regression are as follows:

$$\begin{aligned} \hat{Y}_t = -16.2181 \; + \quad & 1.9488X_t - \quad 0.0079X_t^2 \\ t = \; (-5.4891) \quad & (24.9868) \quad (-15.9363) \\ & R^2 = 0.9947 \qquad d = 1.02 \end{aligned} \tag{12.8.2}$$

These results are interesting. All the coefficients are statistically highly significant, the p values being extremely small. From the negative quadratic term, it seems that although the real wage index increases as the productivity index increases, it increases at a decreasing rate. But look at the d value. It still suggests positive autocorrelation in the residuals, for $d_L = 1.391$ and $d_U = 1.60$ and the estimated d value lies below d_L.

It may be safe to conclude from the preceding analysis that our wages–productivity regression probably suffers from pure autocorrelation and not necessarily from specification bias. Knowing the consequences of autocorrelation, we may therefore want to take some corrective action. We will do so shortly.

Incidentally, for all the wages–productivity regressions that we have presented above, we applied the **Jarque–Bera test of normality** and found that the residuals were normally distributed, which is comforting because the d test assumes normality of the error term.

12.9 CORRECTING FOR (PURE) AUTOCORRELATION: THE METHOD OF GENERALIZED LEAST SQUARES (GLS)

Knowing the consequences of autocorrelation, especially the lack of efficiency of OLS estimators, we may need to remedy the problem. The remedy depends on the knowledge one has about the nature of interdependence among the disturbances, that is, knowledge about the structure of autocorrelation.

As a starter, consider the two-variable regression model:

$$Y_t = \beta_1 + \beta_2 X_t + u_t \tag{12.9.1}$$

and assume that the error term follows the AR(1) scheme, namely,

$$u_t = \rho u_{t-1} + \varepsilon_t \quad -1 < \rho < 1 \tag{12.9.2}$$

Now we consider two cases: (1) ρ is known and (2) ρ is not known but has to be estimated.

When ρ Is Known

If the coefficient of first-order autocorrelation is known, the problem of autocorrelation can be easily solved. If (12.9.1) holds true at time t, it also holds true at time $(t - 1)$. Hence,

$$Y_{t-1} = \beta_1 + \beta_2 X_{t-1} + u_{t-1} \tag{12.9.3}$$

Multiplying (12.9.3) by ρ on both sides, we obtain

$$\rho Y_{t-1} = \rho\beta_1 + \rho\beta_2 X_{t-1} + \rho u_{t-1} \tag{12.9.4}$$

Subtracting (12.9.4) from (12.9.1) gives

$$(Y_t - \rho Y_{t-1}) = \beta_1(1 - \rho) + \beta_2(X_t - \rho X_{t-1}) + \varepsilon_t \tag{12.9.5}$$

where $\varepsilon_t = (u_t - \rho u_{t-1})$

We can express (12.9.5) as

$$Y_t^* = \beta_1^* + \beta_2^* X_t^* + \varepsilon_t \tag{12.9.6}$$

where $\beta_1^* = \beta_1(1 - \rho)$, $Y_t^* = (Y_t - \rho Y_{t-1})$, $X_t^* = (X_t - \rho X_{t-1})$, and $\beta_2^* = \beta_2$.

Since the error term in (12.9.6) satisfies the usual OLS assumptions, we can apply OLS to the transformed variables Y^* and X^* and obtain estimators with all the optimum properties, namely, BLUE. In effect, running (12.9.6) is tantamount to using generalized least squares (GLS) discussed in the previous chapter—recall that GLS is nothing but OLS applied to the transformed model that satisfies the classical assumptions.

Regression (12.9.5) is known as the **generalized,** or **quasi, difference equation.** It involves regressing Y on X, not in the original form, but in the **difference form,** which is obtained by subtracting a proportion ($= \rho$) of the value of a variable in the previous time period from its value in the current time period. In this differencing procedure we lose one observation because the first observation has no antecedent. To avoid this loss of one observation, the first observation on Y and X is transformed as follows[35]: $Y_1\sqrt{1-\rho^2}$ and $X_1\sqrt{1-\rho^2}$. This transformation is known as the **Prais–Winsten transformation.**

When ρ Is Not Known

Although conceptually straightforward to apply, the method of generalized difference given in (12.9.5) is difficult to implement because ρ is rarely known in practice. Therefore, we need to find ways of estimating ρ. We have several possibilities.

The First-Difference Method. Since ρ lies between 0 and ± 1, one could start from two extreme positions. At one extreme, one could assume that $\rho = 0$, that is, no (first-order) serial correlation, and at the other extreme we could let $\rho = \pm 1$, that is, perfect positive or negative correlation. As a matter of fact, when a regression is run, one generally assumes that there is no autocorrelation and then lets the Durbin–Watson or other test show whether this assumption is justified. If, however, $\rho = +1$, the generalized difference equation (12.9.5) reduces to the **first-difference equation:**

$$Y_t - Y_{t-1} = \beta_2(X_t - X_{t-1}) + (u_t - u_{t-1})$$

or

$$\Delta Y_t = \beta_2 \Delta X_t + \varepsilon_t \qquad \textbf{(12.9.7)}$$

where Δ is the first-difference operator introduced in (12.1.10)

Since the error term in (12.9.7) is free from (first-order) serial correlation (why?), to run the regression (12.9.7) all one has to do is form the first differences of both the regressand and regressor(s) and run the regression on these first differences.

The first difference transformation may be appropriate if the coefficient of autocorrelation is very high, say in excess of 0.8, or the Durbin–Watson d is quite low. Maddala has proposed this rough rule of thumb: *Use the first difference form whenever* $d < R^2$.[36] This is the case in our wages–productivity

[35]The loss of one observation may not be very serious in large samples but can make a substantial difference in the results in small samples. Without transforming the first observation as indicated, the error variance will not be homoscedastic. On this see, Jeffrey Wooldridge, op. cit., p. 388. For some Monte Carlo results on the importance of the first observation, see Russell Davidson and James G. MacKinnon, *Estimation and Inference in Econometrics,* Oxford University Press, New York, 1993, Table 10.1, p. 349.

[36]Maddala, op. cit., p. 232.

regression (12.5.1), where we found that $d = 0.1229$ and $r^2 = 0.9584$. The first-difference regression for our illustrative example will be presented shortly.

An interesting feature of the first-difference model (12.9.7) is that **there is no intercept in it.** Hence, to estimate (12.9.7), you have to use the **regression through the origin** routine (that is, suppress the intercept term), which is now available in most software packages. If, however, you forget to drop the intercept term in the model and estimate the following model that includes the intercept term

$$\Delta Y_t = \beta_1 + \beta_2 \Delta X_t + \varepsilon_t \tag{12.9.8}$$

then the original model must have a *trend* in it and β_1 represents the coefficient of the trend variable.[37] Therefore, one "accidental" benefit of introducing the intercept term in the first-difference model is to test for the presence of a trend variable in the original model.

Returning to our wages–productivity regression (12.5.1), and given the AR(1) scheme and a low d value in relation to r^2, we rerun (12.5.1) in the first-difference form without the intercept term; remember that (12.5.1) is in the *level form*. The results are as follows[38]:

$$\begin{aligned} \widehat{\Delta Y}_t &= \ 0.7199 \Delta X_t \\ &\quad t = (9.2073) \qquad r^2 = 0.3610 \qquad d = 1.5096 \end{aligned} \tag{12.9.9}$$

Compared with the level form regression (12.5.1), we see that the slope coefficient has not changed much, but the r^2 value has dropped considerably. This is generally the case because by taking the first differences we are essentially studying the behavior of variables around their (linear) trend values. Of course, we cannot compare the r^2 of (12.9.9) directly with that of the r^2 of (12.5.1) because the dependent variables in the two models are different.[39] Also, notice that compared with the original regression, the d value has increased dramatically, perhaps indicating that there is little autocorrelation in the first-difference regression.[40]

Another interesting aspect of the first-difference transformation relates to the *stationarity* properties of the underlying time series. Return to Eq. (12.2.1), which describes the AR(1) scheme. Now if in fact $\rho = 1$, then it is clear from Eqs. (12.2.3) and (12.2.4) that the series u_t is *nonstationary,* for the variances and covariances become infinite. That is why, when we

[37]This is easy to show. Let $Y_t = \alpha_1 + \beta_1 t + \beta_2 X_t + u_t$. Therefore, $Y_{t-1} = \alpha + \beta_1(t-1) + \beta_2 X_{t-1} + u_{t-1}$. Subtracting the latter from the former, you will obtain: $\Delta Y_t = \beta_1 + \beta_2 \Delta X_t + \varepsilon_t$, which shows that the intercept term in this equation is indeed the coefficient of the trend variable in the original model. Remember that we are assuming that $\rho = 1$.

[38]In exercise 12.38 you are asked to run this model, including the constant term.

[39]The comparison of r^2 in the level and first-difference form is slightly involved. For an extended discussion on this, see Maddala, op. cit., Chap. 6.

[40]It is not clear whether the computed d in the first-difference regression can be interpreted in the same way as it was in the original, level form regression. However, applying the runs test, it can be seen that there is no evidence of autocorrelation in the residuals of the first-difference regression.

discussed this topic, we put the restriction that $|\rho| < 1$. But it is clear from (12.2.1) that if the autocorrelation coefficient is in fact 1, then (12.2.1) becomes

$$u_t = u_{t-1} + \varepsilon_t$$

or

$$(u_t - u_{t-1}) = \Delta u_t = \varepsilon_t \tag{12.9.10}$$

That is, it is the first-differenced u_t that becomes stationary, for it is equal to ε_t, which is a white noise error term.

The point of the preceding discussion is that if the original time series are nonstationary, very often their first differences become stationary. And, therefore, first-difference transformation serves a dual purpose in that it might get rid of (first-order) autocorrelation and also render the time series stationary. We will revisit this topic in **Part V,** where we discuss the econometrics of time series analysis in some depth.

We mentioned that the first-difference transformation may be appropriate if ρ is high or d is low. Strictly speaking, the first-difference transformation is valid only if $\rho = 1$. As a matter of fact, there is a test, called the **Berenblutt–Webb test,**[41] to test the hypothesis that $\rho = 1$. The test statistic they use is called the ***g* statistic,** which is defined as follows:

$$g = \frac{\sum_2^n \hat{e}_t^2}{\sum_1^n \hat{u}_t^2} \tag{12.9.11}$$

where $\hat{u}_t$ are the OLS residuals from the original (i.e., level form) regression and e_t are the OLS residuals from the first-difference regression. Keep in mind that in the first-difference form there is no intercept.

To test the significance of the g statistic, assuming that the level form regression contains the intercept term, we can use the Durbin–Watson tables except that now the null hypothesis is that $\rho = 1$ rather than the Durbin–Watson hypothesis that $\rho = 0$.

Revisiting our wages–productivity regression, for the original regression (12.5.1) we obtain $\sum \hat{u}_t^2 = 272.0220$ and for the first regression (12.7.11) we obtain $\sum \hat{e}_t^2 = 0.334270$. Putting these values into the g statistic given in (12.9.11), we obtain

$$g = \frac{33.4270}{272.0220} = 0.0012 \tag{12.9.12}$$

Consulting the Durbin–Watson table for 39 observations and 1 explanatory variable, we find that $d_L = 1.435$ and $d_U = 1.540$ (5 percent level). Since the observed g lies below the lower limit of d, we do not reject the hypothesis that true $\rho = 1$. *Keep in mind that although we use the same Durbin–Watson*

[41] I. I. Berenblutt and G. I. Webb, "A New Test for Autocorrelated Errors in the Linear Regression Model," *Journal of the Royal Statistical Society,* Series B, vol. 35, No.1, 1973, pp. 33–50.

tables, now the null hypothesis is that $\rho = 1$ *and not that* $\rho = 0$. In view of this finding, the results given in (12.9.9) may be acceptable.

ρ Based on Durbin–Watson *d* Statistic. If we cannot use the first difference transformation because ρ is not sufficiently close to unity, we have an easy method of estimating it from the relationship between d and ρ established previously in (12.6.10), from which we can estimate ρ as follows:

$$\hat{\rho} \approx 1 - \frac{d}{2} \tag{12.9.13}$$

Thus, in reasonably large samples one can obtain rho from (12.9.13) and use it to transform the data as shown in the generalized difference equation (12.9.5). Keep in mind that the relationship between ρ and d given in (12.9.13) may not hold true in small samples, for which Theil and Nagar have proposed a modification, which is given in exercise 12.6.

In our wages–productivity regression (12.5.1), we obtain a d value of 0.1229. Using this value in (12.9.13), we obtain $\hat{\rho} \approx 0.9386$. Using this estimated rho value, we can estimate regression (12.9.5). All we have to do is subtract 0.9386 times the previous value of Y from its current value and similarly subtract 0.9386 times the previous value of X from its current value and run the OLS regression on the variables thus transformed as in (12.9.6), where $Y_t^* = (Y_t - 0.9386Y_{t-1})$ and $X_t^* = (X_t - 0.9386X_{t-1})$.

ρ Estimated from the Residuals. If the AR(1) scheme $u_t = \rho u_{t-1} + \varepsilon_t$ is valid, a simple way to estimate rho is to regress the residuals $\hat{u}_t$ on $\hat{u}_{t-1}$, for the $\hat{u}_t$ are consistent estimators of the true u_t, as noted previously. That is, we run the following regression:

$$\hat{u}_t = \rho \,.\, \hat{u}_{t-1} + v_t \tag{12.9.14}$$

where $\hat{u}_t$ are the residuals obtained from the original (level form) regression and where v_t are the error term of this regression. Note that there is no need to introduce the intercept term in (12.9.14), for we know the OLS residuals sum to zero.

The residuals from our wages–productivity regression given in (12.5.1) are already shown in Table 12.5. Using these residuals, the following regression results were obtained:

$$\begin{aligned} \widehat{\hat{u}_t} &= \quad 0.9142\hat{u}_{t-1} \\ t &= (16.2281) \qquad\quad r^2 = 0.8736 \end{aligned} \tag{12.9.15}$$

As this regression shows, $\hat{\rho} = 0.9142$. Using this estimate, one can transform the original model as per (12.9.6). Since the rho estimated by this procedure is about the same as that obtained from the Durbin–Watson d, the

regression results using the rho of (12.9.15) should not be very different from those obtained from the rho estimated from the Durbin–Watson d. We leave it to the reader to verify this.

Iterative Methods of Estimating ρ. All the methods of estimating ρ discussed previously provide us with only a single estimate of ρ. But there are the so-called **iterative methods** that estimate ρ iteratively, that is, by successive approximation, starting with some initial value of ρ. Among these methods the following may be mentioned: the **Cochrane–Orcutt iterative procedure,** the **Cochrane–Orcutt two-step procedure,** the **Durbin two–step procedure,** and the **Hildreth–Lu scanning or search procedure.** Of these, the most popular is the Cochran–Orcutt iterative method. To save space, the iterative methods are discussed by way of exercises. Remember that the ultimate objective of these methods is to provide an estimate of ρ that may be used to obtain GLS estimates of the parameters. One advantage of the Cochrane–Orcutt iterative method is that it can be used to estimate not only an AR(1) scheme, but also higher-order autoregressive schemes, such as $\hat{u}_t = \hat{\rho}_1\hat{u}_{t-1} + \hat{\rho}_2\hat{u}_{t-2} + v_t$, which is AR(2). Having obtained the two rhos, one can easily extend the generalized difference equation (12.9.6). Of course, the computer can now do all this.

Returning to our wages–productivity regression, and assuming an AR(1) scheme, we use the Cochrane–Orcutt iterative method, which gives the following estimates of rho: 0.9142, 0.9052, 0.8992, 0.8956, 0.8935, 0.8924, and 0.8919. The last value of 0.8919 can now be used to transform the original model as in (12.9.6) and estimate it by OLS. Of course, OLS on the transformed model is simply the GLS. The results are as follows:

Dropping the First Observation Since the first observation has no antecedent, in estimating (12.9.6), we drop the first observation. The regression results are as follows:

$$\begin{aligned} \hat{Y}_t^* &= 45.105 \ + \ 0.5503X_t^* \\ \text{se} &= (6.190) \quad (0.0652) \\ t &= (7.287) \quad (8.433) \qquad r^2 = 0.9959 \end{aligned} \tag{12.9.16}$$

Comparing the results of this regression with the original regression given in (12.5.1), we see that the slope coefficient has dropped dramatically. Notice two things about (12.9.16). First, the intercept coefficient in (12.9.16) is $\beta_1(1-\rho)$, from which the original β_1 can be easily retrieved, since we know that $\rho = 0.8913$. Secondly, the r^2's of the transformed model (12.9.16) and the original model (12.5.1) cannot be directly compared, since the dependent variables in the two models are different.

Retaining the First Observation à la Prais–Winsten. We cautioned earlier that in small samples keeping the first observation or omitting it can

make a substantial difference in small samples, although in large samples the difference may be inconsequential.

Retaining the first observation à la Prais–Winsten, we obtain the following regression results[42]:

$$\begin{aligned} \hat{Y}_t^* &= 26.454 \quad + \quad 0.7245X_t^* \\ \text{se} &= (5.4520) \quad (0.0612) \\ t &= (4.8521) \quad (11.8382) \qquad r^2 = 0.9949 \end{aligned} \tag{12.9.17}$$

The difference between (12.9.16) and (12.9.17) tells us that the inclusion or exclusion of the first observation can make a substantial difference in the regression results. Also, note that the slope coefficient in (12.9.17) is approximately the same as that in (12.5.1).

General Comments. There are several points about correcting for autocorrelation using the various methods discussed above.

First, since the OLS estimators are consistent despite autocorrelation, in large samples, it makes little difference whether we estimate ρ from the Durbin–Watson d, or from the regression of the residuals in the current period on the residuals in the previous period, or from the Cochrane–Orcutt iterative procedure because they all provide consistent estimates of the true ρ. *Second,* the various methods discussed above are basically two-step methods. In step 1 we obtain an estimate of the unknown ρ and in step 2 we use that estimate to transform the variables to estimate the generalized difference equation, which is basically GLS. But since we use $\hat{\rho}$ instead of the true ρ, all these methods of estimation are known in the literature as **feasible GLS (FGLS)** or **estimated GLS (EGLS)** methods.

Third, it is important to note that whenever we use an **FGLS** or **EGLS** method to estimate the parameters of the transformed model, the estimated coefficients will not necessarily have the usual optimum properties of the classical model, such as BLUE, especially in small samples. Without going into complex technicalities, it may be stated *as a general principle that whenever we use an estimator in place of its true value, the estimated OLS coefficients may have the usual optimum properties asymptotically, that is, in large samples. Also, the conventional hypothesis testing procedures are, strictly speaking, valid asymptotically. In small samples, therefore, one has to be careful in interpreting the estimated results.*

Fourth, in using EGLS, if we do not include the first observation (as was originally the case with the Cochrane–Orcutt procedure), not only the

[42]Including the first observation, the iterated values of rho are: 0.9142, 9.9462, 0.9556, 0.9591, 0.9605, and 0.9610. The last value was used in transforming the data to form the generalized difference equation.

numerical values but also the efficiency of the estimators can be adversely affected, especially if the sample size is small and if the regressors are not strictly speaking nonstochastic.[43] Therefore, in small samples it is important to keep the first observation *à la* Prais–Winsten. Of course, if the sample size is reasonably large, EGLS, with or without the first observation, gives similar results. Incidentally, in the literature EGLS with Prais–Winsten transformation is known as the **full EGLS,** or **FEGLS,** for short.

12.10 THE NEWEY–WEST METHOD OF CORRECTING THE OLS STANDARD ERRORS

Instead of using the FGLS methods discussed in the previous section, we can still use OLS but correct the standard errors for autocorrelation by a procedure developed by Newey and West.[44] This is an extension of White's heteroscedasticity-consistent standard errors that we discussed in the previous chapter. The corrected standard errors are known as **HAC (heteroscedasticity- and autocorrelation-consistent) standard errors** or simply as **Newey–West standard errors.** We will not present the mathematics behind the Newey–West procedure, for it is involved.[45] But most modern computer packages now calculate the Newey–West standard errors. But it is important to point out that the Newey–West procedure is *strictly speaking valid in large samples* and may not be appropriate in small samples. But in large samples we now have a method that produces autocorrelation-corrected standard errors so that we do not have to worry about the EGLS transformations discussed in the previous chapter. Therefore, if a sample is reasonably large, one should use the Newey–West procedure to correct OLS standard errors not only in situations of autocorrelation only but also in cases of heteroscedasticity, for the HAC method can handle both, unlike the White method, which was designed specifically for heteroscedasticity.

Once again let us return to our wages–productivity regression (12.5.1). We know that this regression suffers from autocorrelation. Our sample of 40 observations is reasonably large, so we can use the HAC procedure. Using *Eviews 4,* we obtain the following regression results:

$$\begin{aligned} \hat{Y}_t &= 29.5192 \;+\; 0.7136\hat{X}_t \\ \text{se} &= (4.1180)^* \quad (0.0512)^* \\ r^2 &= 0.9584 \qquad d = 0.1229 \end{aligned} \tag{12.10.1}$$

where * denotes HAC standard errors.

[43]This is especially so if the regressors exhibit a trend, which is quite common in economic data.

[44]W. K. Newey, and K. West, "A Simple Positive Semi-Definite Heteroscedasticity and Autocorrelation Consistent Covariance Matrix, *Econometrica,* vol. 55, 1987, pp. 703–708.

[45]If you can handle matrix algebra, the method is discussed in Greene, op. cit, 4th ed., pp. 462–463.

Comparing this regression with (12.5.1), we find that in both the equations the estimated coefficients and the r^2 value are the same. But, importantly, note that the HAC standard errors are much greater than the OLS standard errors and therefore the HAC t ratios are much smaller than the OLS t ratios. This shows that OLS had in fact underestimated the true standard errors. Curiously, the d statistics in both (12.5.1) and (12.10.1) is the same. But don't worry, for the HAC procedure has already taken this into account in correcting the OLS standard errors.

12.11 OLS VERSUS FGLS AND HAC

The practical problem facing the researcher is this: In the presence of autocorrelation, OLS estimators, although unbiased, consistent, and asymptotically normally distributed, are not efficient. Therefore, the usual inference procedure based on the t, F, and χ^2 tests is no longer appropriate. On the other hand, FGLS and HAC produce estimators that are efficient, but the finite, or small-sample, properties of these estimators are not well documented. This means in small samples the FGLS and HAC might actually do worse than OLS. As a matter of fact, in a Monte Carlo study Griliches and Rao[46] found that if the sample is relatively small and the coefficient of autocorrelation, ρ, is less than 0.3, OLS is as good or better than FGLS. As a practical matter, then, one may use OLS in small samples in which the estimated rho is, say, less than 0.3. Of course, what is a large and what is a small sample are relative questions, and one has to use some practical judgment. If you have only 15 to 20 observations, the sample may be small, but if you have, say, 50 or more observations, the sample may be reasonably large.

12.12 FORECASTING WITH AUTOCORRELATED ERROR TERMS

In Section 5.10, we introduced the basics of forecasting in the context of the two-variable regression model using the classical framework. How do these basics change if there is autocorrelation? Although this topic is generally covered in a course in economic forecasting, we can provide a glimpse of it here. To be specific, we will continue with the two-variable model and assume an AR(1) scheme. Thus,

$$Y_t = \beta_1 + \beta_2 X_t + u_t \tag{12.12.1}$$

$$u_t = \rho u_{t-1} + \varepsilon_t \qquad -1 < \rho < 1 \tag{12.12.2}$$

where ε_t is a white noise error term.

Substituting (12.12.2) into (12.12.1), we obtain

$$Y_t = \beta_1 + \beta_2 X_t + \rho u_{t-1} + \varepsilon_t \tag{12.12.3}$$

[46] Z. Griliches, and P. Rao, "Small Sample Properties of Several Two-stage Regression Methods in the Context of Autocorrelated Errors," *Journal of the American Statistical Association,* vol. 64, 1969, pp. 253–272.

If we want to forecast Y for the next time period $(t + 1)$, we obtain

$$Y_{t+1} = \beta_1 + \beta_2 X_{t+1} + \rho u_t + \varepsilon_{t+1} \qquad \textbf{(12.12.4)}$$

Thus, the forecast for the next period consists of three parts: (1) its expected value $= (\beta_1 + \beta_2 X_{t+1})$, (2) ρ times the preceding error term, and (3) a purely white noise term, whose expected value is zero. Given the value of X_{t+1}, we estimate (1) by $\hat{\beta}_1 + \hat{\beta} X_{t+1}$, where the OLS estimators are obtained from a given sample, and we estimate (2) as $\hat{\rho}\hat{u}_t$, where $\hat{\rho}$ is estimated by one of the methods discussed in Section 12.9. At time $(t + 1)$, the value of $\hat{u}_t$ is already known. Therefore, the *estimated* value of Y_{t+1} in (12.1.4) is:

$$\hat{Y}_{t+1} = \hat{\beta}_1 + \hat{\beta}_2 X_{t+1} + \hat{\rho}\hat{u}_t \qquad \textbf{(12.12.5)}$$

Following this logic,

$$\hat{Y}_{t+2} = \hat{\beta}_1 + \hat{\beta}_2 X_{t+2} + \hat{\rho}^2\hat{u}_t \qquad \textbf{(12.12.6)}$$

for the second period, and so on.

The forecasting that we did in Section 5.10 is called **statistic forecasting,** whereas that represented by (12.12.5) and (12.12.6) is called **dynamic forecasting,** for in making these forecasts we are taking into account the errors made in the past forecasts.

As in Section 5.10, we will need to compute the forecast (standard) errors of (12.12.5) and (12.12.6). But the formulas become complicated.[47] Since most modern econometrics packages, such as Microfit, Eviews, and Shazam, produce the standard errors of forecast, there is no need to present the computing formulas here.

As an illustration, let us fall back on our wages–productivity regression. Recall that our sample data is for the period 1959–1998. We reestimated this model using the data for 1959–1996 only, saving the last two observations for forecasting purposes. Using Microfit 4.1, we obtained the following forecast values of Y for 1997 and 1998, both static and dynamic, using the estimated regression for 1959–1996.

	Year 1997	Year 1998
Actual Y value	101.1	105.1
Static forecast of Y	107.24 (2.64)	109.45 (2.67)
Static forecast error	−6.14	−4.35
Dynamic forecast	100.75 (1.08)	101.95 (1.64)
Dynamic forecast error	0.35	3.14

Note: Figures in parentheses are the estimated standard errors of forecast values.

[47]For further discussion, see, Robert S. Pindyck and Daniel L. Rubinfeld, *Econometric Models and Economic Forecasts,* McGraw-Hill, 4th ed., 1998, pp. 214–217.

As you can see from the preceding exercise, the dynamic forecasts are closer to their actual values than the static forecasts and the standard errors of dynamic forecasts are smaller than their static counterpart. So, it may be profitable to incorporate the AR(1) scheme (or higher-order schemes) for the purpose of forecasting. However, note that for both types of forecasts the standard errors of forecast for 1998 are greater than that for 1997, which suggests, not surprisingly, that forecasting into the distant future may be hazardous.

12.13 ADDITIONAL ASPECTS OF AUTOCORRELATION

Dummy Variables and Autocorrelation

In Chapter 9 we considered dummy variable regression models. In particular, recall the U.S. savings–income regression model for 1970–1995 that we presented in (9.5.1), which for convenience is reproduced below:

$$Y_t = \alpha_1 + \alpha_2 + \beta_1 X_t + \beta_2 (D_t X_t) + u_t \qquad \textbf{(12.13.1)}$$

where Y = savings
X = income
$D = 1$ for observations in period 1982–1995
$D = 0$ for observations in period 1970–1981

The regression results based on this model are given in (9.5.4). Of course, this model was estimated with the usual OLS assumptions.

But now suppose that u_t follows a first-order autoregressive, AR(1), scheme. That is, $u_t = \rho u_{t-1} + \varepsilon_t$. Ordinarily, if ρ is known or can be estimated by one of the methods discussed above, we can use the generalized difference method to estimate the parameters of the model that is free from (first-order) autocorrelation. However, the presence of the dummy variable D poses a special problem: Note that the dummy variable simply classifies an observation as belonging to the first or second period. How do we transform it? One can follow the following procedure.[48]

1. In (12.13.1), values of D are zero for all observations in the first period; in period 2 the value of D for the **first** observation is $1/(1-\rho)$ instead of 1, and 1 for all other observations.

2. The variable X_t is transformed as $(X_t - \rho X_{t-1})$. Note that we lose one observation in this transformation, unless one resorts to **Prais–Winsten transformation** for the first observation, as noted earlier.

3. The value of $D_t X_t$ is zero for all observations in the first period (*note:* D_t is zero in the first period); in the second period the first observation takes the value of $D_t X_t = X_t$ and the remaining observations in the second period are set to $(D_t X_t - D_t \rho X_{t-1}) = (X_t - \rho X_{t-1})$. (*Note:* the value of D_t in the second period is 1.)

[48]See Maddala, op. cit., pp. 321–322.

As the preceding discussion points out, the *critical observation* is the first observation in the second period. If this is taken care of in the manner just suggested, there should be no problem in estimating regressions like (12.13.1) subject to AR(1) autocorrelation. In exercise 12.37, the reader is asked to carry such a transformation for the data on U.S. savings and income given in Chapter 9.

ARCH and GARCH Models

Just as the error term u at time t can be correlated with the error term at time $(t-1)$ in an AR(1) scheme or with various lagged error terms in a general AR(p) scheme, can there be autocorrelation in the variance σ^2 at time t with its values lagged one or more periods? Such an autocorrelation has been observed by researchers engaged in forecasting financial time series, such as stock prices, inflation rates, and foreign exchange rates. Such autocorrelation is given the rather daunting names **autoregressive conditional heteroscedasticity (ARCH)** if the error variance is related to the squared error term in the previous term and **generalized autoregressive conditional heteroscedasticity (GARH)** if the error variance is related to squared error terms several periods in the past. Since this topic belongs in the general area of time series econometrics, we will discuss it in some depth in the chapters on time series econometrics. Our objective here is to point out that autocorrelation is not confined to relationships between current and past error terms but also with current and past error variances.

Coexistence of Autocorrelation and Heteroscedasticity

What happens if a regression model suffers from both heteroscedasticity and autocorrelation? Can we solve the problem sequentially, that is, take care of heteroscedasticity first and then autocorrelation? As a matter of fact, one author contends that "Autoregression can only be detected after the heteroscedasticity is controlled for."[49] But can we develop an omnipotent test that can solve these and other problems (e.g., model specification) simultaneously? Yes, such tests exist, but their discussion will take us far afield. It is better to leave them for references.[50]

12.14 SUMMARY AND CONCLUSIONS

1. If the assumption of the classical linear regression model—that the errors or disturbances u_t entering into the population regression function (PRF) are random or uncorrelated—is violated, the problem of serial or autocorrelation arises.

[49]Lois W. Sayrs, *Pooled Time Series Analysis,* Sage Publications, California, 1989, p. 19.

[50]See Jeffrey M. Wooldridge, op. cit., pp. 402–403, and A. K. Bera and C. M. Jarque, "Efficient Tests for Normality, Homoscedasticity and Serial Independence of Regression Residuals: Monte Carlo Evidence," *Economic Letters,* vol. 7, 1981, pp. 313–318.

2. Autocorrelation can arise for several reasons, such as inertia or sluggishness of economic time series, specification bias resulting from excluding important variables from the model or using incorrect functional form, the cobweb phenomenon, data massaging, and data transformation. As a result, it is useful to distinguish between pure autocorrelation and "induced" autocorrelation because of one or more factors just discussed.

3. Although in the presence of autocorrelation the OLS estimators remain unbiased, consistent, and asymptotically normally distributed, they are no longer efficient. As a consequence, the usual t, F, and χ^2 tests cannot be legitimately applied. Hence, remedial results may be called for.

4. The remedy depends on the nature of the interdependence among the disturbances u_t. But since the disturbances are unobservable, the common practice is to assume that they are generated by some mechanism.

5. The mechanism that is commonly assumed is the Markov first-order autoregressive scheme, which assumes that the disturbance in the current time period is linearly related to the disturbance term in the previous time period, the coefficient of autocorrelation ρ providing the extent of the interdependence. This mechanism is know as the AR(1) scheme.

6. If the AR(1) scheme is valid and the coefficient of autocorrelation is known, the serial correlation problem can be easily attacked by transforming the data following the generalized difference procedure. The AR(1) scheme can be easily generalized to an AR(p). One can also assume a moving average (MA) mechanism or a mixture of AR and MA schemes, known as ARMA. This topic will be discussed in the chapters on time series econometrics.

7. Even if we use an AR(1) scheme, the coefficient of autocorrelation is not known a priori. We considered several methods of estimating ρ, such as the Durbin–Watson d, Theil–Nagar modified d, Cochrane–Orcutt (C–O) iterative procedure, C–O two-step method, and the Durbin two-step procedure. In large samples, these methods generally yield similar estimates of ρ, although in small samples they perform differently. In practice, the C–O iterative method has become quite popular.

8. Using any of the methods just discussed, we can use the generalized difference method to estimate the parameters of the transformed model by OLS, which essentially amounts to GLS. But since we estimate ρ ($= \hat{\rho}$), we call the method of estimation as feasible, or estimated, GLS, or FGLS or EGLS for short.

9. In using EGLS, one has to be careful in dropping the first observation, for in small samples the inclusion or exclusion of the first observation can make a dramatic difference in the results. Therefore, in small samples it is advisable to transform the first observation according to the Prais–Winsten procedure. In large samples, however, it makes little difference if the first observation is included or not.

10. It is very important to note that the method of EGLS has the usual optimum statistical properties only in large samples. In small samples, OLS may actually do better that EGLS, especially if $\rho < 0.3$.

11. Instead of using EGLS, we can still use OLS but correct the standard errors for autocorrelation by the Newey–West HAC procedure. Strictly speaking, this procedure is valid in large samples. One advantage of the HAC procedure is that it not only corrects for autocorrelation but also for heteroscedasticity, if it is present.

12. Of course, before remediation comes detection of autocorrelation. There are formal and informal methods of detection. Among the informal methods, one can simply plot the actual or standardized residuals, or plot current residuals against past residuals. Among formal methods, one can use the runs test, Durbin–Watson d test, asymptotic normality test, Berenblutt–Webb test, and Breusch–Godfrey (BG) test. Of these, the most popular and routinely used is the Durbin–Watson d test. Despite its hoary past, this test has severe limitations. It is better to use the BG test, for it is much more general in that it allows for both AR and MA error structures as well as the presence of lagged regressand as an explanatory variable. But keep in mind that it is a large sample test.

13. In this chapter we also discussed very briefly the detection of autocorrelation in the presence of dummy regressors, the use of autocorrelated errors for forecasting purposes, and the topic of ARCH and GARCH.

EXERCISES

Questions

12.1. State whether the following statements are true or false. Briefly justify your answer.

a. When autocorrelation is present, OLS estimators are biased as well as inefficient.

b. The Durbin–Watson d test assumes that the variance of the error term u_t is homoscedastic.

c. The first-difference transformation to eliminate autocorrelation assumes that the coefficient of autocorrelation ρ is -1.

d. The R^2 values of two models, one involving regression in the first-difference form and another in the level form, are not directly comparable.

e. A significant Durbin–Watson d does not necessarily mean there is autocorrelation of the first order.

f. In the presence of autocorrelation, the conventionally computed variances and standard errors of forecast values are inefficient.

g. The exclusion of an important variable(s) from a regression model may give a significant d value.

h. In the AR(1) scheme, a test of the hypothesis that $\rho = 1$ can be made by the Berenblutt–Webb g statistic as well as the Durbin–Watson d statistic.

i. In the regression of the first difference of Y on the first differences of X, if there is a constant term and a linear trend term, it means in the original model there is a linear as well as a quadratic trend term.