Chapter Ten

Basic Regression Analysis with Time Series Data

In this chapter, we begin to study the properties of OLS for estimating linear regression models using time series data. In Section 10.1, we discuss some conceptual differences between time series and cross-sectional data. Section 10.2 provides some examples of time series regressions that are often estimated in the empirical social sciences. We then turn our attention to the finite sample properties of the OLS estimators and state the Gauss-Markov assumptions and the classical linear model assumptions for time series regression. While these assumptions have features in common with those for the cross-sectional case, they also have some significant differences that we will need to highlight.

In addition, we return to some issues that we treated in regression with cross-sectional data, such as how to use and interpret the logarithmic functional form and dummy variables. The important topics of how to incorporate trends and account for seasonality in multiple regression are taken up in Section 10.5.

10.1 THE NATURE OF TIME SERIES DATA

An obvious characteristic of time series data which distinguishes it from cross-sectional data is that a time series data set comes with a temporal ordering. For example, in Chapter 1, we briefly discussed a time series data set on employment, the minimum wage, and other economic variables for Puerto Rico. In this data set, we must know that the data for 1970 immediately precede the data for 1971. For analyzing time series data in the social sciences, we must recognize that the past can effect the future, but not vice versa (unlike in the Star Trek universe). To emphasize the proper ordering of time series data, Table 10.1 gives a partial listing of the data on U.S. inflation and unemployment rates in PHILLIPS.RAW.

Another difference between cross-sectional and time series data is more subtle. In Chapters 3 and 4, we studied statistical properties of the OLS estimators based on the notion that samples were randomly drawn from the appropriate population. Understanding why cross-sectional data should be viewed as random outcomes is fairly straightforward: a different sample drawn from the population will generally yield different values of the independent and dependent variables (such as education, experience, wage, and so on). Therefore, the OLS estimates computed from different random samples will generally differ, and this is why we consider the OLS estimators to be random variables.

Table 10.1

Partial Listing of Data on U.S. Inflation and Unemployment Rates, 1948–1996

Year	Inflation	Unemployment
1948	8.1	3.8
1949	−1.2	5.9
1950	1.3	5.3
1951	7.9	3.3
⋮	⋮	⋮
1994	2.6	6.1
1995	2.8	5.6
1996	3.0	5.4

How should we think about randomness in time series data? Certainly, economic time series satisfy the intuitive requirements for being outcomes of random variables. For example, today we do not know what the Dow Jones Industrial Average will be at its close at the end of the next trading day. We do not know what the annual growth in output will be in Canada during the coming year. Since the outcomes of these variables are not foreknown, they should clearly be viewed as random variables.

Formally, a sequence of random variables indexed by time is called a **stochastic process** or a **time series process**. ("Stochastic" is a synonym for random.) When we collect a time series data set, we obtain one possible outcome, or *realization*, of the stochastic process. We can only see a single realization, because we cannot go back in time and start the process over again. (This is analogous to cross-sectional analysis where we can collect only one random sample.) However, if certain conditions in history had been different, we would generally obtain a different realization for the stochastic process, and this is why we think of time series data as the outcome of random variables. The set of all possible realizations of a time series process plays the role of the population in cross-sectional analysis.

10.2 EXAMPLES OF TIME SERIES REGRESSION MODELS

In this section, we discuss two examples of time series models that have been useful in empirical time series analysis and that are easily estimated by ordinary least squares. We will study additional models in Chapter 11.

Static Models

Suppose that we have time series data available on two variables, say y and z, where y_t and z_t are dated contemporaneously. A **static model** relating y to z is

$$y_t = \beta_0 + \beta_1 z_t + u_t, \; t = 1,2, \ldots, n. \tag{10.1}$$

The name "static model" comes from the fact that we are modeling a contemporaneous relationship between y and z. Usually, a static model is postulated when a change in z at time t is believed to have an immediate effect on y: $\Delta y_t = \beta_1 \Delta z_t$, when $\Delta u_t = 0$. Static regression models are also used when we are interested in knowing the tradeoff between y and z.

An example of a static model is the *static Phillips curve*, given by

$$inf_t = \beta_0 + \beta_1 unem_t + u_t, \tag{10.2}$$

where inf_t is the annual inflation rate and $unem_t$ is the unemployment rate. This form of the Phillips curve assumes a constant *natural rate of unemployment* and constant inflationary expectations, and it can be used to study the contemporaneous tradeoff between them. [See, for example, Mankiw (1994, Section 11.2).]

Naturally, we can have several explanatory variables in a static regression model. Let $mrdrte_t$ denote the murders per 10,000 people in a particular city during year t, let $convrte_t$ denote the murder conviction rate, let $unem_t$ be the local unemployment rate, and let $yngmle_t$ be the fraction of the population consisting of males between the ages of 18 and 25. Then, a static multiple regression model explaining murder rates is

$$mrdrte_t = \beta_0 + \beta_1 convrte_t + \beta_2 unem_t + \beta_3 yngmle_t + u_t. \tag{10.3}$$

Using a model such as this, we can hope to estimate, for example, the ceteris paribus effect of an increase in the conviction rate on criminal activity.

Finite Distributed Lag Models

In a **finite distributed lag (FDL) model**, we allow one or more variables to affect y with a lag. For example, for annual observations, consider the model

$$gfr_t = \alpha_0 + \delta_0 pe_t + \delta_1 pe_{t-1} + \delta_2 pe_{t-2} + u_t, \tag{10.4}$$

where gfr_t is the general fertility rate (children born per 1,000 women of childbearing age) and pe_t is the real dollar value of the personal tax exemption. The idea is to see whether, in the aggregate, the decision to have children is linked to the tax value of having a child. Equation (10.4) recognizes that, for both biological and behavioral reasons, decisions to have children would not immediately result from changes in the personal exemption.

Equation (10.4) is an example of the model

$$y_t = \alpha_0 + \delta_0 z_t + \delta_1 z_{t-1} + \delta_2 z_{t-2} + u_t, \tag{10.5}$$

which is an FDL *of order two*. To interpret the coefficients in (10.5), suppose that z is a constant, equal to c, in all time periods before time t. At time t, z increases by one unit to $c + 1$ and then reverts to its previous level at time $t + 1$. (That is, the increase in z is temporary.) More precisely,

$$\ldots, z_{t-2} = c, z_{t-1} = c, z_t = c + 1, z_{t+1} = c, z_{t+2} = c, \ldots.$$

To focus on the ceteris paribus effect of z on y, we set the error term in each time period to zero. Then,

$$\begin{aligned} y_{t-1} &= \alpha_0 + \delta_0 c + \delta_1 c + \delta_2 c, \\ y_t &= \alpha_0 + \delta_0(c + 1) + \delta_1 c + \delta_2 c, \\ y_{t+1} &= \alpha_0 + \delta_0 c + \delta_1(c + 1) + \delta_2 c, \\ y_{t+2} &= \alpha_0 + \delta_0 c + \delta_1 c + \delta_2(c + 1), \\ y_{t+3} &= \alpha_0 + \delta_0 c + \delta_1 c + \delta_2 c, \end{aligned}$$

and so on. From the first two equations, $y_t - y_{t-1} = \delta_0$, which shows that δ_0 is the immediate change in y due to the one-unit increase in z at time t. δ_0 is usually called the **impact propensity** or **impact multiplier**.

Similarly, $\delta_1 = y_{t+1} - y_{t-1}$ is the change in y one period after the temporary change, and $\delta_2 = y_{t+2} - y_{t-1}$ is the change in y two periods after the change. At time $t + 3$, y has reverted back to its initial level: $y_{t+3} = y_{t-1}$. This is because we have assumed that only two lags of z appear in (10.5). When we graph the δ_j as a function of j, we obtain the **lag distribution**, which summarizes the dynamic effect that a temporary increase in z has on y. A possible lag distribution for the FDL of order two is given in Figure 10.1. (Of course, we would never know the parameters δ_j; instead, we will estimate the δ_j and then plot the estimated lag distribution.)

The lag distribution in Figure 10.1 implies that the largest effect is at the first lag. The lag distribution has a useful interpretation. If we standardize the initial value of y at $y_{t-1} = 0$, the lag distribution traces out all subsequent values of y due to a one-unit, temporary increase in z.

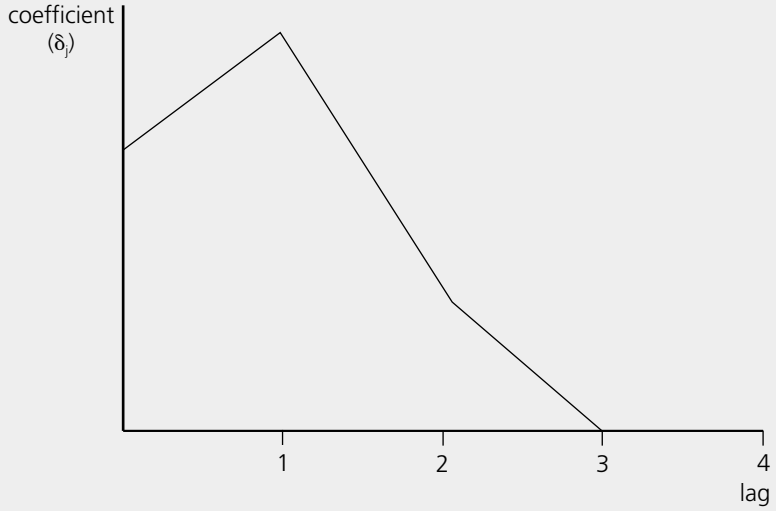
We are also interested in the change in y due to a *permanent* increase in z. Before time t, z equals the constant c. At time t, z increases permanently to $c + 1$: $z_s = c$, $s < t$ and $z_s = c + 1$, $s \geq t$. Again, setting the errors to zero, we have

$$\begin{aligned} y_{t-1} &= \alpha_0 + \delta_0 c + \delta_1 c + \delta_2 c, \\ y_t &= \alpha_0 + \delta_0(c + 1) + \delta_1 c + \delta_2 c, \\ y_{t+1} &= \alpha_0 + \delta_0(c + 1) + \delta_1(c + 1) + \delta_2 c, \\ y_{t+2} &= \alpha_0 + \delta_0(c + 1) + \delta_1(c + 1) + \delta_2(c + 1), \end{aligned}$$

and so on. With the permanent increase in z, after one period, y has increased by $\delta_0 + \delta_1$, and after two periods, y has increased by $\delta_0 + \delta_1 + \delta_2$. There are no further changes in y after two periods. This shows that the sum of the coefficients on current and lagged z, $\delta_0 + \delta_1 + \delta_2$, is the *long-run* change in y given a permanent increase in z and is called the **long-run propensity** (**LRP**) or **long-run multiplier**. The LRP is often of interest in distributed lag models.

Figure 10.1

A lag distribution with two nonzero lags. The maximum effect is at the first lag.



As an example, in equation (10.4), δ_0 measures the immediate change in fertility due to a one-dollar increase in *pe*. As we mentioned earlier, there are reasons to believe that δ_0 is small, if not zero. But δ_1 or δ_2, or both, might be positive. If *pe permanently* increases by one dollar, then, after two years, *gfr* will have changed by $\delta_0 + \delta_1 + \delta_2$. This model assumes that there are no further changes after two years. Whether or not this is actually the case is an empirical matter.

A finite distributed lag model of order q is written as

$$y_t = \alpha_0 + \delta_0 z_t + \delta_1 z_{t-1} + \ldots + \delta_q z_{t-q} + u_t. \qquad \textbf{(10.6)}$$

This contains the static model as a special case by setting $\delta_1, \delta_2, \ldots, \delta_q$ equal to zero. Sometimes, a primary purpose for estimating a distributed lag model is to test whether z has a lagged effect on y. The impact propensity is always the coefficient on the contemporaneous z, δ_0. Occasionally, we omit z_t from (10.6), in which case the impact propensity is zero. The lag distribution is again the δ_j graphed as a function of j. The long-run propensity is the sum of all coefficients on the variables z_{t-j}:

$$\text{LRP} = \delta_0 + \delta_1 + \ldots + \delta_q. \qquad \textbf{(10.7)}$$

Because of the often substantial correlation in z at different lags—that is, due to multicollinearity in (10.6)—it can be difficult to obtain precise estimates of the individual δ_j. Interestingly, even when the δ_j cannot be precisely estimated, we can often get good estimates of the LRP. We will see an example later.

QUESTION 10.1

In an equation for annual data, suppose that

$$int_t = 1.6 + .48\, inf_t - .15\, inf_{t-1} + .32\, inf_{t-2} + u_t,$$

where *int* is an interest rate and *inf* is the inflation rate, what are the impact and long-run propensities?

We can have more than one explanatory variable appearing with lags, or we can add contemporaneous variables to an FDL model. For example, the average education level for women of childbearing age could be added to (10.4), which allows us to account for changing education levels for women.

A Convention About the Time Index

When models have lagged explanatory variables (and, as we will see in the next chapter, models with lagged y), confusion can arise concerning the treatment of initial observations. For example, if in (10.5), we assume that the equation holds, starting at $t = 1$, then the explanatory variables for the first time period are z_1, z_0, and z_{-1}. Our convention will be that these are the initial values in our sample, so that we can always start the time index at $t = 1$. In practice, this is not very important because regression packages automatically keep track of the observations available for estimating models with lags. But for this and the next few chapters, we need some convention concerning the first time period being represented by the regression equation.

10.3 FINITE SAMPLE PROPERTIES OF OLS UNDER CLASSICAL ASSUMPTIONS

In this section, we give a complete listing of the finite sample, or small sample, properties of OLS under standard assumptions. We pay particular attention to how the assumptions must be altered from our cross-sectional analysis to cover time series regressions.

Unbiasedness of OLS

The first assumption simply states that the time series process follows a model which is linear in its parameters.

ASSUMPTION TS.1 (LINEAR IN PARAMETERS)

The stochastic process $\{(x_{t1}, x_{t2}, \ldots, x_{tk}, y_t): t = 1,2,\ldots,n\}$ follows the linear model

$$y_t = \beta_0 + \beta_1 x_{t1} + \ldots + \beta_k x_{tk} + u_t, \qquad \textbf{(10.8)}$$

where $\{u_t: t = 1,2,\ldots,n\}$ is the sequence of errors or disturbances. Here, n is the number of observations (time periods).

Table 10.2

Example of **X** for the Explanatory Variables in Equation (10.3)

t	*convrte*	*unem*	*yngmle*
1	.46	.074	.12
2	.42	.071	.12
3	.42	.063	.11
4	.47	.062	.09
5	.48	.060	.10
6	.50	.059	.11
7	.55	.058	.12
8	.56	.059	.13

In the notation x_{tj}, t denotes the time period, and j is, as usual, a label to indicate one of the k explanatory variables. The terminology used in cross-sectional regression applies here: y_t is the dependent variable, explained variable, or regressand; the x_{tj} are the independent variables, explanatory variables, or regressors.

We should think of Assumption TS.1 as being essentially the same as Assumption MLR.1 (the first cross-sectional assumption), but we are now specifying a linear model for time series data. The examples covered in Section 10.2 can be cast in the form of (10.8) by appropriately defining x_{tj}. For example, equation (10.5) is obtained by setting $x_{t1} = z_t$, $x_{t2} = z_{t-1}$, and $x_{t3} = z_{t-2}$.

In order to state and discuss several of the remaining assumptions, we let $\boldsymbol{x}_t = (x_{t1}, x_{t2}, \ldots, x_{tk})$ denote the set all independent variables in the equation at time t. Further, $\boldsymbol{X}$ denotes the collection of all independent variables for all time periods. It is useful to think of $\boldsymbol{X}$ as being an array, with n rows and k columns. This reflects how time series data are stored in econometric software packages: the t^{th} row of $\boldsymbol{X}$ is $\boldsymbol{x}_t$, consisting of all independent variables for time period t. Therefore, the first row of $\boldsymbol{X}$ corresponds to $t = 1$, the second row to $t = 2$, and the last row to $t = n$. An example is given in Table 10.2, using $n = 8$ and the explanatory variables in equation (10.3).

The next assumption is the time series analog of Assumption MLR.3, and it also drops the assumption of random sampling in Assumption MLR.2.

ASSUMPTION TS.2 (ZERO CONDITIONAL MEAN)

For each t, the expected value of the error u_t, given the explanatory variables for *all* time periods, is zero. Mathematically,

$$E(u_t|\boldsymbol{X}) = 0, t = 1,2,\ldots,n. \quad \textbf{(10.9)}$$

This is a crucial assumption, and we need to have an intuitive grasp of its meaning. As in the cross-sectional case, it is easiest to view this assumption in terms of uncorrelatedness: Assumption TS.2 implies that the error at time t, u_t, is uncorrelated with each explanatory variable in *every* time period. The fact that this is stated in terms of the conditional expectation means that we must also correctly specify the functional relationship between y_t and the explanatory variables. If u_t is independent of $\boldsymbol{X}$ and $E(u_t) = 0$, then Assumption TS.2 automatically holds.

Given the cross-sectional analysis from Chapter 3, it is not surprising that we require u_t to be uncorrelated with the explanatory variables also dated at time t: in conditional mean terms,

$$E(u_t|x_{t1},\ldots,x_{tk}) = E(u_t|\boldsymbol{x}_t) = 0. \quad \textbf{(10.10)}$$

When (10.10) holds, we say that the x_{tj} are **contemporaneously exogenous**. Equation (10.10) implies that u_t and the explanatory variables are contemporaneously uncorrelated: $\text{Corr}(x_{tj},u_t) = 0$, for all j.

Assumption TS.2 requires more than contemporaneous exogeneity: u_t must be uncorrelated with x_{sj}, even when $s \neq t$. This is a strong sense in which the explanatory variables must be exogenous, and when TS.2 holds, we say that the explanatory variables are **strictly exogenous**. In Chapter 11, we will demonstrate that (10.10) is sufficient for proving consistency of the OLS estimator. But to show that OLS is unbiased, we need the strict exogeneity assumption.

In the cross-sectional case, we did not explicitly state how the error term for, say, person i, u_i, is related to the explanatory variables for *other* people in the sample. The reason this was unnecessary is that, with random sampling (Assumption MLR.2), u_i is *automatically* independent of the explanatory variables for observations other than i. In a time series context, random sampling is almost never appropriate, so we must explicitly assume that the expected value of u_t is not related to the explanatory variables in any time periods.

It is important to see that Assumption TS.2 puts no restriction on correlation in the independent variables or in the u_t across time. Assumption TS.2 only says that the average value of u_t is unrelated to the independent variables in all time periods.

Anything that causes the unobservables at time t to be correlated with any of the explanatory variables in any time period causes Assumption TS.2 to fail. Two leading candidates for failure are omitted variables and measurement error in some of the regressors. But, the strict exogeneity assumption can also fail for other, less obvious reasons. In the simple static regression model

$$y_t = \beta_0 + \beta_1 z_t + u_t,$$

Assumption TS.2 requires not only that u_t and z_t are uncorrelated, but that u_t is also uncorrelated with past and future values of z. This has two implications. First, z can have no lagged effect on y. If z does have a lagged effect on y, then we should estimate a distributed lag model. A more subtle point is that strict exogeneity excludes the pos-

sibility that changes in the error term today can cause future changes in z. This effectively rules out feedback from y on future values of z. For example, consider a simple static model to explain a city's murder rate in terms of police officers per capita:

$$mrdrte_t = \beta_0 + \beta_1 polpc_t + u_t.$$

It may be reasonable to assume that u_t is uncorrelated with $polpc_t$ and even with past values of $polpc_t$; for the sake of argument, assume this is the case. But suppose that the city adjusts the size of its police force based on past values of the murder rate. This means that, say, $polpc_{t+1}$ might be correlated with u_t (since a higher u_t leads to a higher $mrdrte_t$). If this is the case, Assumption TS.2 is generally violated.

There are similar considerations in distributed lag models. Usually we do not worry that u_t might be correlated with past z because we are controlling for past z in the model. But feedback from u to future z is always an issue.

Explanatory variables that are strictly exogenous cannot react to what has happened to y in the past. A factor such as the amount of rainfall in an agricultural production function satisfies this requirement: rainfall in any future year is not influenced by the output during the current or past years. But something like the amount of labor input might not be strictly exogenous, as it is chosen by the farmer, and the farmer may adjust the amount of labor based on last year's yield. Policy variables, such as growth in the money supply, expenditures on welfare, highway speed limits are often influenced by what has happened to the outcome variable in the past. In the social sciences, many explanatory variables may very well violate the strict exogeneity assumption.

Even though Assumption TS.2 can be unrealistic, we begin with it in order to conclude that the OLS estimators are unbiased. Most treatments of static and finite distributed lag models assume TS.2 by making the stronger assumption that the explanatory variables are nonrandom, or fixed in repeated samples. The nonrandomness assumption is obviously false for time series observations; Assumption TS.2 has the advantage of being more realistic about the random nature of the x_{tj}, while it isolates the necessary assumption about how u_t and the explanatory variables are related in order for OLS to be unbiased.

The last assumption needed for unbiasedness of OLS is the standard no perfect collinearity assumption.

ASSUMPTION TS.3 (NO PERFECT COLLINEARITY)
In the sample (and therefore in the underlying time series process), no independent variable is constant or a perfect linear combination of the others.

We discussed this assumption at length in the context of cross-sectional data in Chapter 3. The issues are essentially the same with time series data. Remember, Assumption TS.3 does allow the explanatory variables to be correlated, but it rules out *perfect* correlation in the sample.

THEOREM 10.1 (UNBIASEDNESS OF OLS)
Under Assumptions TS.1, TS.2, and TS.3, the OLS estimators are unbiased conditional on $\boldsymbol{X}$, and therefore unconditionally as well: $E(\hat{\beta}_j) = \beta_j$, $j = 0,1,\ldots,k$.

QUESTION 10.2

In the FDL model $y_t = \alpha_0 + \delta_0 z_t + \delta_1 z_{t-1} + u_t$, what do we need to assume about the sequence $\{z_0, z_1, \ldots, z_n\}$ in order for Assumption TS.3 to hold?

The proof of this theorem is essentially the same as that for Theorem 3.1 in Chapter 3, and so we omit it. When comparing Theorem 10.1 to Theorem 3.1, we have been able to drop the random sampling assumption by assuming that, for each t, u_t has zero mean given the explanatory variables at all time periods. If this assumption does not hold, OLS cannot be shown to be unbiased.

The analysis of omitted variables bias, which we covered in Section 3.3, is essentially the same in the time series case. In particular, Table 3.2 and the discussion surrounding it can be used as before to determine the directions of bias due to omitted variables.

The Variances of the OLS Estimators and the Gauss-Markov Theorem

We need to add two assumptions to round out the Gauss-Markov assumptions for time series regressions. The first one is familiar from cross-sectional analysis.

ASSUMPTION TS.4 (HOMOSKEDASTICITY)
Conditional on $\boldsymbol{X}$, the variance of u_t is the same for all t: $\text{Var}(u_t|\boldsymbol{X}) = \text{Var}(u_t) = \sigma^2$, $t = 1,2,\ldots,n$.

This assumption means that $\text{Var}(u_t|\boldsymbol{X})$ cannot depend on $\boldsymbol{X}$—it is sufficient that u_t and $\boldsymbol{X}$ are independent—*and* that $\text{Var}(u_t)$ must be constant over time. When TS.4 does not hold, we say that the errors are *heteroskedastic*, just as in the cross-sectional case. For example, consider an equation for determining three-month, T-bill rates ($i3_t$) based on the inflation rate (inf_t) and the federal deficit as a percentage of gross domestic product (def_t):

$$i3_t = \beta_0 + \beta_1 inf_t + \beta_2 def_t + u_t. \tag{10.11}$$

Among other things, Assumption TS.4 requires that the unobservables affecting interest rates have a constant variance over time. Since policy regime changes are known to affect the variability of interest rates, this assumption might very well be false. Further, it could be that the variability in interest rates depends on the level of inflation or relative size of the deficit. This would also violate the homoskedasticity assumption.

When $\text{Var}(u_t|\boldsymbol{X})$ does depend on $\boldsymbol{X}$, it often depends on the explanatory variables at time t, $\boldsymbol{x}_t$. In Chapter 12, we will see that the tests for heteroskedasticity from Chapter 8 can also be used for time series regressions, at least under certain assumptions.

The final Gauss-Markov assumption for time series analysis is new.

ASSUMPTION TS.5 (NO SERIAL CORRELATION)
Conditional on $\boldsymbol{X}$, the errors in two different time periods are uncorrelated: $\text{Corr}(u_t,u_s|\boldsymbol{X}) = 0$, for all $t \neq s$.

The easiest way to think of this assumption is to ignore the conditioning on $\boldsymbol{X}$. Then, Assumption TS.5 is simply

$$\text{Corr}(u_t,u_s) = 0, \text{ for all } t \neq s. \tag{10.12}$$

(This is how the no serial correlation assumption is stated when $\boldsymbol{X}$ is treated as nonrandom.) When considering whether Assumption TS.5 is likely to hold, we focus on equation (10.12) because of its simple interpretation.

When (10.12) is false, we say that the errors in (10.8) suffer from **serial correlation**, or **autocorrelation**, because they are correlated across time. Consider the case of errors from adjacent time periods. Suppose that, when $u_{t-1} > 0$ then, on average, the error in the next time period, u_t, is also positive. Then $\text{Corr}(u_t,u_{t-1}) > 0$, and the errors suffer from serial correlation. In equation (10.11) this means that, if interest rates are unexpectedly high for this period, then they are likely to be above average (for the given levels of inflation and deficits) for the next period. This turns out to be a reasonable characterization for the error terms in many time series applications, which we will see in Chapter 12. For now, we assume TS.5.

Importantly, Assumption TS.5 assumes nothing about temporal correlation in the *independent* variables. For example, in equation (10.11), inf_t is almost certainly correlated across time. But this has nothing to do with whether TS.5 holds.

A natural question that arises is: In Chapters 3 and 4, why did we not assume that the errors for different cross-sectional observations are uncorrelated? The answer comes from the random sampling assumption: under random sampling, u_i and u_h are independent for any two observations i and h. It can also be shown that this is true, conditional on all explanatory variables in the sample. Thus, for our purposes, serial correlation is only an issue in time series regressions.

Assumptions TS.1 through TS.5 are the appropriate Gauss-Markov assumptions for time series applications, but they have other uses as well. Sometimes, TS.1 through TS.5 are satisfied in cross-sectional applications, even when random sampling is not a reasonable assumption, such as when the cross-sectional units are large relative to the population. It is possible that correlation exists, say, across cities within a state, but as long as the errors are uncorrelated across those cities, Assumption TS.5 holds. But we are primarily interested in applying these assumptions to regression models with time series data.

THEOREM 10.2 (OLS SAMPLING VARIANCES)

Under the time series Gauss-Markov assumptions TS.1 through TS.5, the variance of $\hat{\beta}_j$, conditional on $\boldsymbol{X}$, is

$$\text{Var}(\hat{\beta}_j|\boldsymbol{X}) = \sigma^2/[\text{SST}_j(1 - R_j^2)], j = 1, \ldots, k, \tag{10.13}$$

where SST_j is the total sum of squares of x_{tj} and R_j^2 is the R-squared from the regression of x_j on the other independent variables.

Equation (10.13) is the exact variance we derived in Chapter 3 under the cross-sectional Gauss-Markov assumptions. Since the proof is very similar to the one for Theorem 3.2, we omit it. The discussion from Chapter 3 about the factors causing large variances, including multicollinearity among the explanatory variables, applies immediately to the time series case.

The usual estimator of the error variance is also unbiased under Assumptions TS.1 through TS.5, and the Gauss-Markov theorem holds.

THEOREM 10.3 (UNBIASED ESTIMATION OF σ^2)
Under Assumptions TS.1 through TS.5, the estimator $\hat{\sigma}^2 = \text{SSR}/df$ is an unbiased estimator of σ^2, where $df = n - k - 1$.

THEOREM 10.4 (GAUSS-MARKOV THEOREM)
Under Assumptions TS.1 through TS.5, the OLS estimators are the best linear unbiased estimators conditional on $\boldsymbol{X}$.

QUESTION 10.3

In the FDL model $y_t = \alpha_0 + \delta_0 z_t + \delta_1 z_{t-1} + u_t$, explain the nature of any multicollinearity in the explanatory variables.

The bottom line here is that OLS has the same desirable finite sample properties under TS.1 through TS.5 that it has under MLR.1 through MLR.5.

Inference Under the Classical Linear Model Assumptions

In order to use the usual OLS standard errors, t statistics, and F statistics, we need to add a final assumption that is analogous to the normality assumption we used for cross-sectional analysis.

ASSUMPTION TS.6 (NORMALITY)
The errors u_t are independent of $\boldsymbol{X}$ and are independently and identically distributed as Normal$(0,\sigma^2)$.

Assumption TS.6 implies TS.3, TS.4, and TS.5, but it is stronger because of the independence and normality assumptions.

THEOREM 10.5 (NORMAL SAMPLING DISTRIBUTIONS)
Under Assumptions TS.1 through TS.6, the CLM assumptions for time series, the OLS estimators are normally distributed, conditional on $\boldsymbol{X}$. Further, under the null hypothesis, each t statistic has a t distribution, and each F statistic has an F distribution. The usual construction of confidence intervals is also valid.

The implications of Theorem 10.5 are of utmost importance. It implies that, when Assumptions TS.1 through TS.6 hold, everything we have learned about estimation and inference for cross-sectional regressions applies directly to time series regressions. Thus, t statistics can be used for testing statistical significance of individual explanatory variables, and F statistics can be used to test for joint significance.

Just as in the cross-sectional case, the usual inference procedures are only as good as the underlying assumptions. The classical linear model assumptions for time series data are much more restrictive than those for the cross-sectional data—in particular, the strict exogeneity and no serial correlation assumptions can be unrealistic. Nevertheless, the CLM framework is a good starting point for many applications.

EXAMPLE 10.1
(Static Phillips Curve)

To determine whether there is a tradeoff, on average, between unemployment and inflation, we can test H_0: $\beta_1 = 0$ against H_0: $\beta_1 < 0$ in equation (10.2). If the classical linear model assumptions hold, we can use the usual OLS t statistic. Using annual data for the United States in PHILLIPS.RAW, for the years 1948 through 1996, we obtain

$$\begin{gathered} \widehat{inf}_t = 1.42 + .468\, unem_t \\ (1.72)\quad (.289) \\ n = 49,\ R^2 = .053,\ \bar{R}^2 = .033. \end{gathered} \tag{10.14}$$

This equation does not suggest a tradeoff between *unem* and *inf*: $\hat{\beta}_1 > 0$. The t statistic for $\hat{\beta}_1$ is about 1.62, which gives a p-value against a two-sided alternative of about .11. Thus, if anything, there is a positive relationship between inflation and unemployment.

There are some problems with this analysis that we cannot address in detail now. In Chapter 12, we will see that the CLM assumptions do not hold. In addition, the static Phillips curve is probably not the best model for determining whether there is a short-run tradeoff between inflation and unemployment. Macroeconomists generally prefer the expectations augmented Phillips curve, a simple example of which is given in Chapter 11.

As a second example, we estimate equation (10.11) using anual data on the U.S. economy.

EXAMPLE 10.2
(Effects of Inflation and Deficits on Interest Rates)

The data in INTDEF.RAW come from the 1997 *Economic Report of the President* and span the years 1948 through 1996. The variable *i3* is the three-month T-bill rate, *inf* is the annual inflation rate based on the consumer price index (CPI), and *def* is the federal budget deficit as a percentage of GDP. The estimated equation is

$$\widehat{i3}_t = 1.25 + .613\ inf_t + .700\ def_t \quad \textbf{(10.15)}$$
$$(0.44)\quad (.076)\qquad (.118)$$
$$n = 49,\ R^2 = .697,\ \bar{R}^2 = .683.$$

These estimates show that increases in inflation and the relative size of the deficit work together to increase short-term interest rates, both of which are expected from basic economics. For example, a ceteris paribus one percentage point increase in the inflation rate increases *i3* by .613 points. Both *inf* and *def* are very statistically significant, assuming, of course, that the CLM assumptions hold.

10.4 FUNCTIONAL FORM, DUMMY VARIABLES, AND INDEX NUMBERS

All of the functional forms we learned about in earlier chapters can be used in time series regressions. The most important of these is the natural logarithm: time series regressions with constant percentage effects appear often in applied work.

EXAMPLE 10.3
(Puerto Rican Employment and the Minimum Wage)

Annual data on the Puerto Rican employment rate, minimum wage, and other variables are used by Castillo-Freedman and Freedman (1992) to study the effects of the U.S. minimum wage on employment in Puerto Rico. A simplified version of their model is

$$\log(prepop_t) = \beta_0 + \beta_1 \log(mincov_t) + \beta_2 \log(usgnp_t) + u_t, \quad \textbf{(10.16)}$$

where $prepop_t$ is the employment rate in Puerto Rico during year t (ratio of those working to total population), $usgnp_t$ is real U.S. gross national product (in billions of dollars), and *mincov* measures the importance of the minimum wage relative to average wages. In particular, *mincov* = (*avgmin*/*avgwage*)·*avgcov*, where *avgmin* is the average minimum wage, *avgwage* is the average overall wage, and *avgcov* is the average coverage rate (the proportion of workers actually covered by the minimum wage law).

Using data for the years 1950 through 1987 gives

$$\widehat{\log(prepop_t)} = -1.05 - .154 \log(mincov_t) - .012 \log(usgnp_t) \quad \textbf{(10.17)}$$
$$(0.77)\quad (.065)\qquad\qquad (.089)$$
$$n = 38,\ R^2 = .661,\ \bar{R}^2 = .641.$$

The estimated elasticity of *prepop* with respect to *mincov* is $-.154$, and it is statistically significant with $t = -2.37$. Therefore, a higher minimum wage lowers the employment rate, something that classical economics predicts. The GNP variable is not statistically significant, but this changes when we account for a time trend in the next section.

We can use logarithmic functional forms in distributed lag models, too. For example, for quarterly data, suppose that money demand (M_t) and gross domestic product (GDP_t) are related by

$$\log(M_t) = \alpha_0 + \delta_0\log(GDP_t) + \delta_1\log(GDP_{t-1}) + \delta_2\log(GDP_{t-2}) + \delta_3\log(GDP_{t-3}) + \delta_4\log(GDP_{t-4}) + u_t.$$

The impact propensity in this equation, δ_0, is also called the **short-run elasticity**: it measures the immediate percentage change in money demand given a 1% increase in *GDP*. The long-run propensity, $\delta_0 + \delta_1 + \ldots + \delta_4$, is sometimes called the **long-run elasticity**: it measures the percentage increase in money demand after four quarters given a permanent 1% increase in *GDP*.

Binary or dummy independent variables are also quite useful in time series applications. Since the unit of observation is time, a dummy variable represents whether, in each time period, a certain event has occurred. For example, for annual data, we can indicate in each year whether a Democrat or a Republican is president of the United States by defining a variable $democ_t$, which is unity if the president is a Democrat, and zero otherwise. Or, in looking at the effects of capital punishment on murder rates in Texas, we can define a dummy variable for each year equal to one if Texas had capital punishment during that year, and zero otherwise.

Often dummy variables are used to isolate certain periods that may be systematically different from other periods covered by a data set.

EXAMPLE 10.4

(Effects of Personal Exemption on Fertility Rates)

The general fertility rate (*gfr*) is the number of children born to every 1,000 women of childbearing age. For the years 1913 through 1984, the equation,

$$gfr_t = \beta_0 + \beta_1 pe_t + \beta_2 ww2_t + \beta_3 pill_t + u_t,$$

explains *gfr* in terms of the average real dollar value of the personal tax exemption (*pe*) and two binary variables. The variable *ww2* takes on the value unity during the years 1941 through 1945, when the United States was involved in World War II. The variable *pill* is unity from 1963 on, when the birth control pill was made available for contraception.

Using the data in FERTIL3.RAW, which were taken from the article by Whittington, Alm, and Peters (1990), gives

$$\widehat{gfr}_t = \underset{(3.21)}{98.68} + \underset{(.030)}{.083}\, pe_t - \underset{(7.46)}{24.24}\, ww2_t - \underset{(4.08)}{31.59}\, pill_t \qquad (10.18)$$
$$n = 72,\ R^2 = .473,\ \bar{R}^2 = .450.$$

Each variable is statistically significant at the 1% level against a two-sided alternative. We see that the fertility rate was lower during World War II: given *pe*, there were about 24 fewer births for every 1,000 women of childbearing age, which is a large reduction. (From 1913 through 1984, *gfr* ranged from about 65 to 127.) Similarly, the fertility rate has been substantially lower since the introduction of the birth control pill.

The variable of economic interest is *pe*. The average *pe* over this time period is $100.40, ranging from zero to $243.83. The coefficient on *pe* implies that a 12-dollar increase in *pe* increases *gfr* by about one birth per 1,000 women of childbearing age. This effect is hardly trivial.

In Section 10.2, we noted that the fertility rate may react to changes in *pe* with a lag. Estimating a distributed lag model with two lags gives

$$\begin{aligned} \widehat{gfr}_t = 95.87 &+ .073\ pe_t - .0058\ pe_{t-1} + .034\ pe_{t-2} \\ &\quad (3.28)\ \ (.126)\ \ \ (.1557)\ \ \ \ (.126) \\ &- 22.13\ ww2_t - 31.30\ pill_t \\ &\quad (10.73)\ \ \ \ (3.98) \\ n = 70,\ R^2 = .499,\ \bar{R}^2 &= .459. \end{aligned} \tag{10.19}$$

In this regression, we only have 70 observations because we lose two when we lag *pe* twice. The coefficients on the *pe* variables are estimated very imprecisely, and each one is individually insignificant. It turns out that there is substantial correlation between pe_t, pe_{t-1}, and pe_{t-2}, and this multicollinearity makes it difficult to estimate the effect at each lag. However, pe_t, pe_{t-1}, and pe_{t-2} are jointly significant: the *F* statistic has a *p*-value = .012. Thus, *pe* does have an effect on *gfr* [as we already saw in (10.18)], but we do not have good enough estimates to determine whether it is contemporaneous or with a one- or two-year lag (or some of each). Actually, pe_{t-1} and pe_{t-2} are jointly insignificant in this equation (*p*-value = .95), so at this point, we would be justified in using the static model. But for illustrative purposes, let us obtain a confidence interval for the long-run propensity in this model.

The estimated LRP in (10.19) is $.073 - .0058 + .034 \approx .101$. However, we do not have enough information in (10.19) to obtain the standard error of this estimate. To obtain the standard error of the estimated LRP, we use the trick suggested in Section 4.4. Let $\theta_0 = \delta_0 + \delta_1 + \delta_2$ denote the LRP and write δ_0 in terms of θ_0, δ_1, and δ_2 as $\delta_0 = \theta_0 - \delta_1 - \delta_2$. Next, substitute for δ_0 in the model

$$gfr_t = \alpha_0 + \delta_0 pe_t + \delta_1 pe_{t-1} + \delta_2 pe_{t-2} + \ldots$$

to get

$$\begin{aligned} gfr_t &= \alpha_0 + (\theta_0 - \delta_1 - \delta_2)pe_t + \delta_1 pe_{t-1} + \delta_2 pe_{t-2} + \ldots \\ &= \alpha_0 + \theta_0 pe_t + \delta_1(pe_{t-1} - pe_t) + \delta_2(pe_{t-2} - pe_t) + \ldots. \end{aligned}$$

From this last equation, we can obtain $\hat{\theta}_0$ and its standard error by regressing gfr_t on pe_t, $(pe_{t-1} - pe_t)$, $(pe_{t-2} - pe_t)$, $ww2_t$, and $pill_t$. The coefficient and associated standard error on pe_t are what we need. Running this regression gives $\hat{\theta}_0 = .101$ as the coefficient on pe_t (as we already knew from above) and $\text{se}(\hat{\theta}_0) = .030$ [which we could not compute from (10.19)]. Therefore, the *t* statistic for $\hat{\theta}_0$ is about 3.37, so $\hat{\theta}_0$ is statistically different from zero at small significance levels. Even though none of the $\hat{\delta}_j$ is individually significant, the LRP is very significant. The 95% confidence interval for the LRP is about .041 to .160.

Whittington, Alm, and Peters (1990) allow for further lags but restrict the coefficients to help alleviate the multicollinearity problem that hinders estimation of the individual δ_j. (See Problem 10.6 for an example of how to do this.) For estimating the LRP, which would

seem to be of primary interest here, such restrictions are unnecessary. Whittington, Alm, and Peters also control for additional variables, such as average female wage and the unemployment rate.

Binary explanatory variables are the key component in what is called an **event study**. In an event study, the goal is to see whether a particular event influences some outcome. Economists who study industrial organization have looked at the effects of certain events on firm stock prices. For example, Rose (1985) studied the effects of new trucking regulations on the stock prices of trucking companies.

A simple version of an equation used for such event studies is

$$R_t^f = \beta_0 + \beta_1 R_t^m + \beta_2 d_t + u_t,$$

where R_t^f is the stock return for firm f during period t (usually a week or a month), R_t^m is the market return (usually computed for a broad stock market index), and d_t is a dummy variable indicating when the event occurred. For example, if the firm is an airline, d_t might denote whether the airline experienced a publicized accident or near accident during week t. Including R_t^m in the equation controls for the possibility that broad market movements might coincide with airline accidents. Sometimes, multiple dummy variables are used. For example, if the event is the imposition of a new regulation that might affect a certain firm, we might include a dummy variable that is one for a few weeks before the regulation was publicly announced and a second dummy variable for a few weeks after the regulation was announced. The first dummy variable might detect the presence of inside information.

Before we give an example of an event study, we need to discuss the notion of an **index number** and the difference between nominal and real economic variables. An index number typically aggregates a vast amount of information into a single quantity. Index numbers are used regularly in time series analysis, especially in macroeconomic applications. An example of an index number is the index of industrial production (IIP), computed monthly by the Board of Governors of the Federal Reserve. The IIP is a measure of production across a broad range of industries, and, as such, its magnitude in a particular year has no quantitative meaning. In order to interpret the magnitude of the IIP, we must know the **base period** and the **base value**. In the 1997 *Economic Report of the President* (*ERP*), the base year is 1987, and the base value is 100. (Setting IIP to 100 in the base period is just a convention; it makes just as much sense to set IIP = 1 in 1987, and some indexes are defined with one as the base value.) Because the IIP was 107.7 in 1992, we can say that industrial production was 7.7% higher in 1992 than in 1987. We can use the IIP in any two years to compute the percentage difference in industrial output during those two years. For example, since IIP = 61.4 in 1970 and IIP = 85.7 in 1979, industrial production grew by about 39.6% during the 1970s.

It is easy to change the base period for any index number, and sometimes we must do this to give index numbers reported with different base years a common base year. For example, if we want to change the base year of the IIP from 1987 to 1982, we simply divide the IIP for each year by the 1982 value and then multiply by 100 to make the base period value 100. Generally, the formula is

$$newindex_t = 100(oldindex_t/oldindex_{\text{newbase}}), \tag{10.20}$$

where $oldindex_{\text{newbase}}$ is the original value of the index in the new base year. For example, with base year 1987, the IIP in 1992 is 107.7; if we change the base year to 1982, the IIP in 1992 becomes 100(107.7/81.9) = 131.5 (because the IIP in 1982 was 81.9).

Another important example of an index number is a *price index*, such as the consumer price index (CPI). We already used the CPI to compute annual inflation rates in Example 10.1. As with the industrial production index, the CPI is only meaningful when we compare it across different years (or months, if we are using monthly data). In the 1997 *ERP*, CPI = 38.8 in 1970, and CPI = 130.7 in 1990. Thus, the general price level grew by almost 237% over this twenty-year period. (In 1997, the CPI is defined so that its average in 1982, 1983, and 1984 equals 100; thus, the base period is listed as 1982–1984.)

In addition to being used to compute inflation rates, price indexes are necessary for turning a time series measured in *nominal dollars* (or *current dollars*) into *real dollars* (or *constant dollars*). Most economic behavior is assumed to be influenced by real, not nominal, variables. For example, classical labor economics assumes that labor supply is based on the real hourly wage, not the nominal wage. Obtaining the real wage from the nominal wage is easy if we have a price index such as the CPI. We must be a little careful to first divide the CPI by 100, so that the value in the base year is one. Then, if w denotes the average hourly wage in nominal dollars and p = CPI/100, the *real wage* is simply w/p. This wage is measured in dollars for the base period of the CPI. For example, in Table B-45 in the 1997 *ERP*, average hourly earnings are reported in nominal terms and in 1982 dollars (which means that the CPI used in computing the real wage had the base year 1982). This table reports that the nominal hourly wage in 1960 was \$2.09, but measured in 1982 dollars, the wage was \$6.79. The real hourly wage had peaked in 1973, at \$8.55 in 1982 dollars, and had fallen to \$7.40 by 1995. Thus, there has been a nontrivial decline in real wages over the past 20 years. (If we compare nominal wages from 1973 and 1995, we get a very misleading picture: \$3.94 in 1973 and \$11.44 in 1995. Since the real wage has actually fallen, the increase in the nominal wage is due entirely to inflation.)

Standard measures of economic output are in real terms. The most important of these is *gross domestic product*, or *GDP*. When growth in GDP is reported in the popular press, it is always *real* GDP growth. In the 1997 *ERP*, Table B-9, GDP is reported in billions of 1992 dollars. We used a similar measure of output, real gross national product, in Example 10.3.

Interesting things happen when real dollar variables are used in combination with natural logarithms. Suppose, for example, that average weekly hours worked are related to the real wage as

$$\log(hours) = \beta_0 + \beta_1\log(w/p) + u.$$

Using the fact that $\log(w/p) = \log(w) - \log(p)$, we can write this as

$$\log(hours) = \beta_0 + \beta_1\log(w) + \beta_2\log(p) + u, \tag{10.21}$$

but with the restriction that $\beta_2 = -\beta_1$. Therefore, the assumption that only the real wage influences labor supply imposes a restriction on the parameters of model (10.21).

If $\beta_2 \neq -\beta_1$, then the price level has an effect on labor supply, something that can happen if workers do not fully understand the distinction between real and nominal wages.

There are many practical aspects to the actual computation of index numbers, but it would take us too far afield to cover those here. Detailed discussions of price indexes can be found in most intermediate macroeconomic texts, such as Mankiw (1994, Chapter 2). For us, it is important to be able to use index numbers in regression analysis. As mentioned earlier, since the magnitudes of index numbers are not especially informative, they often appear in logarithmic form, so that regression coefficients have percentage change interpretations.

We now give an example of an event study that also uses index numbers.

EXAMPLE 10.5

(Antidumping Filings and Chemical Imports)

Krupp and Pollard (1996) analyzed the effects of antidumping filings by U.S. chemical industries on imports of various chemicals. We focus here on one industrial chemical, barium chloride, a cleaning agent used in various chemical processes and in gasoline production. In the early 1980s, U.S. barium chloride producers believed that China was offering its U.S. imports at an unfairly low price (an action known as *dumping*), and the barium chloride industry filed a complaint with the U.S. International Trade Commission (ITC) in October 1983. The ITC ruled in favor of the U.S. barium chloride industry in October 1984. There are several questions of interest in this case, but we will touch on only a few of them. First, are imports unusually high in the period immediately preceding the initial filing? Second, do imports change noticeably after an antidumping filing? Finally, what is the reduction in imports after a decision in favor of the U.S. industry?

To answer these questions, we follow Krupp and Pollard by defining three dummy variables: *befile6* is equal to one during the six months before filing, *affile6* indicates the six months after filing, and *afdec6* denotes the six months after the positive decision. The dependent variable is the volume of imports of barium chloride from China, *chnimp*, which we use in logarithmic form. We include as explanatory variables, all in logarithmic form, an index of chemical production, *chempi* (to control for overall demand for barium chloride), the volume of gasoline production, *gas* (another demand variable), and an exchange rate index, *rtwex*, which measures the strength of the dollar against several other currencies. The chemical production index was defined to be 100 in June 1977. The analysis here differs somewhat from Krupp and Pollard in that we use natural logarithms of all variables (except the dummy variables, of course), and we include all three dummy variables in the same regression.

Using monthly data from February 1978 through December 1988 gives the following:

$$\begin{aligned}\widehat{\log(chnimp)} &= \underset{(21.05)}{-17.80} + \underset{(0.48)}{3.12}\log(chempi) + \underset{(.907)}{.196}\log(gas) \\ &+ \underset{(.400)}{.983}\log(rtwex) + \underset{(.261)}{.060}\,befile6 - \underset{(.264)}{.032}\,affile6 - \underset{(.286)}{.566}\,afdec6 \\ n &= 131,\ R^2 = .305,\ \bar{R}^2 = .271.\end{aligned} \tag{10.22}$$

The equation shows that *befile6* is statistically insignificant, so there is no evidence that Chinese imports were unusually high during the six months before the suit was filed. Further, although the estimate on *affile6* is negative, the coefficient is small (indicating about a 3.2% fall in Chinese imports), and it is statistically very insignificant. The coefficient on *afdec6* shows a substantial fall in Chinese imports of barium chloride after the decision in favor of the U.S. industry, which is not surprising. Since the effect is so large, we compute the exact percentage change: $100[\exp(-.566) - 1] \approx -43.2\%$. The coefficient is statistically significant at the 5% level against a two-sided alternative.

The coefficient signs on the control variables are what we expect: an increase in overall chemical production increases the demand for the cleaning agent. Gasoline production does not affect Chinese imports significantly. The coefficient on log(*rtwex*) shows that an increase in the value of the dollar relative to other currencies increases the demand for Chinese imports, as is predicted by economic theory. (In fact, the elasticity is not statistically different from one. Why?)

Interactions among qualitative and quantitative variables are also used in time series analysis. An example with practical importance follows.

EXAMPLE 10.6

(Election Outcomes and Economic Performance)

Fair (1996) summarizes his work on explaining presidential election outcomes in terms of economic performance. He explains the proportion of the two-party vote going to the Democratic candidate using data for the years 1916 through 1992 (every four years) for a total of 20 observations. We estimate a simplified version of Fair's model (using variable names that are more descriptive than his):

$$demvote = \beta_0 + \beta_1 partyWH + \beta_2 incum + \beta_3 partyWH \cdot gnews + \beta_4 partyWH \cdot inf + u,$$

where *demvote* is the proportion of the two-party vote going to the Democratic candidate. The explanatory variable *partyWH* is similar to a dummy variable, but it takes on the value one if a Democrat is in the White House and -1 if a Republican is in the White House. Fair uses this variable to impose the restriction that the effect of a Republican being in the White House has the same magnitude but opposite sign as a Democrat being in the White House. This is a natural restriction since the party shares must sum to one, by definition. It also saves two degrees of freedom, which is important with so few observations. Similarly, the variable *incum* is defined to be one if a Democratic incumbent is running, -1 if a Republican incumbent is running, and zero otherwise. The variable *gnews* is the number of quarters during the current administration's first 15 (out of 16 total), where the quarterly growth in real per capita output was above 2.9% (at an annual rate), and *inf* is the average annual inflation rate over the first 15 quarters of the administration. See Fair (1996) for precise definitions.

Economists are most interested in the interaction terms *partyWH·gnews* and *partyWH·inf*. Since *partyWH* equals one when a Democrat is in the White House, β_3 measures the effect of good economic news on the party in power; we expect $\beta_3 > 0$. Similarly,

β_4 measures the effect that inflation has on the party in power. Because inflation during an administration is considered to be bad news, we expect $\beta_4 < 0$.

The estimated equation using the data in FAIR.RAW is

$$\begin{aligned} \widehat{demvote} &= .481 - .0435\ partyWH + .0544\ incum \\ &\quad (.012)\ \ (.0405)\qquad\quad (.0234) \\ &\quad + .0108\ partyWH{\cdot}gnews - .0077\ partyWH{\cdot}inf \\ &\quad (.0041)\qquad\qquad\quad (.0033) \\ n &= 20,\ R^2 = .663,\ \bar{R}^2 = .573. \end{aligned} \tag{10.23}$$

All coefficients, except that on *partyWH*, are statistically significant at the 5% level. Incumbency is worth about 5.4 percentage points in the share of the vote. (Remember, *demvote* is measured as a proportion.) Further, the economic news variable has a positive effect: one more quarter of good news is worth about 1.1 percentage points. Inflation, as expected, has a negative effect: if average annual inflation is, say, two percentage points higher, the party in power loses about 1.5 percentage points of the two-party vote.

We could have used this equation to predict the outcome of the 1996 presidential election between Bill Clinton, the Democrat, and Bob Dole, the Republican. (The independent candidate, Ross Perot, is excluded because Fair's equation is for the two-party vote only.) Since Clinton ran as an incumbent, *partyWH* = 1 and *incum* = 1. To predict the election outcome, we need the variables *gnews* and *inf*. During Clinton's first 15 quarters in office, per capita real GDP exceeded 2.9% three times, so *gnews* = 3. Further, using the GDP price deflator reported in Table B-4 in the 1997 *ERP*, the average annual inflation rate (computed using Fair's formula) from the fourth quarter in 1991 to the third quarter in 1996 was 3.019. Plugging these into (10.23) gives

$$\widehat{demvote} = .481 - .0435 + .0544 + .0108(3) - .0077(3.019) \approx .5011.$$

Therefore, based on information known before the election in November, Clinton was predicted to receive a very slight majority of the two-party vote: about 50.1%. In fact, Clinton won more handily: his share of the two-party vote was 54.65%.

10.5 TRENDS AND SEASONALITY

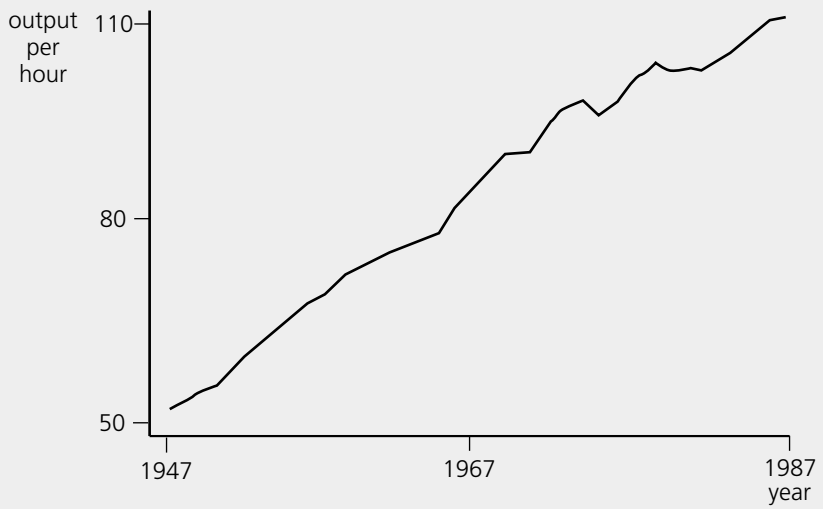
Characterizing Trending Time Series

Many economic time series have a common tendency of growing over time. We must recognize that some series contain a **time trend** in order to draw causal inference using time series data. Ignoring the fact that two sequences are trending in the same or opposite directions can lead us to falsely conclude that changes in one variable are actually caused by changes in another variable. In many cases, two time series processes appear to be correlated only because they are both trending over time for reasons related to other unobserved factors.

Figure 10.2 contains a plot of labor productivity (output per hour of work) in the United States for the years 1947 through 1987. This series displays a clear upward trend, which reflects the fact that workers have become more productive over time.

Figure 10.2

Output per labor hour in the United States during the years 1947–1987; 1977 = 100.



Other series, at least over certain time periods, have clear downward trends. Because positive trends are more common, we will focus on those during our discussion.

What kind of statistical models adequately capture trending behavior? One popular formulation is to write the series $\{y_t\}$ as

$$y_t = \alpha_0 + \alpha_1 t + e_t, \; t = 1,2,\ldots, \tag{10.24}$$

where, in the simplest case, $\{e_t\}$ is an independent, identically distributed (i.i.d.) sequence with $\mathrm{E}(e_t) = 0$, $\mathrm{Var}(e_t) = \sigma_e^2$. Note how the parameter α_1 multiplies time, t, resulting in a **linear time trend**. Interpreting α_1 in (10.24) is simple: holding all other factors (those in e_t) fixed, α_1 measures the change in y_t from one period to the next due to the passage of time: when $\Delta e_t = 0$,

$$\Delta y_t = y_t - y_{t-1} = \alpha_1.$$

Another way to think about a sequence that has a linear time trend is that its average value is a linear function of time:

$$\mathrm{E}(y_t) = \alpha_0 + \alpha_1 t. \tag{10.25}$$

If $\alpha_1 > 0$, then, on average, y_t is growing over time and therefore has an upward trend. If $\alpha_1 < 0$, then y_t has a downward trend. The values of y_t do not fall exactly on the line

QUESTION 10.4

In Example 10.4, we used the general fertility rate as the dependent variable in a finite distributed lag model. From 1950 through the mid-1980s, the *gfr* has a clear downward trend. Can a linear trend with $\alpha_1 < 0$ be realistic for all future time periods? Explain.

in (10.25) due to randomness, but the expected values are on the line. Unlike the mean, the variance of y_t is constant across time: $\text{Var}(y_t) = \text{Var}(e_t) = \sigma_e^2$.

If $\{e_t\}$ is an i.i.d. sequence, then $\{y_t\}$ is an independent, though not identically, distributed sequence. A more realistic characterization of trending time series allows $\{e_t\}$ to be correlated over time, but this does not change the flavor of a linear time trend. In fact, what is important for regression analysis under the classical linear model assumptions is that $\text{E}(y_t)$ is linear in t. When we cover large sample properties of OLS in Chapter 11, we will have to discuss how much temporal correlation in $\{e_t\}$ is allowed.

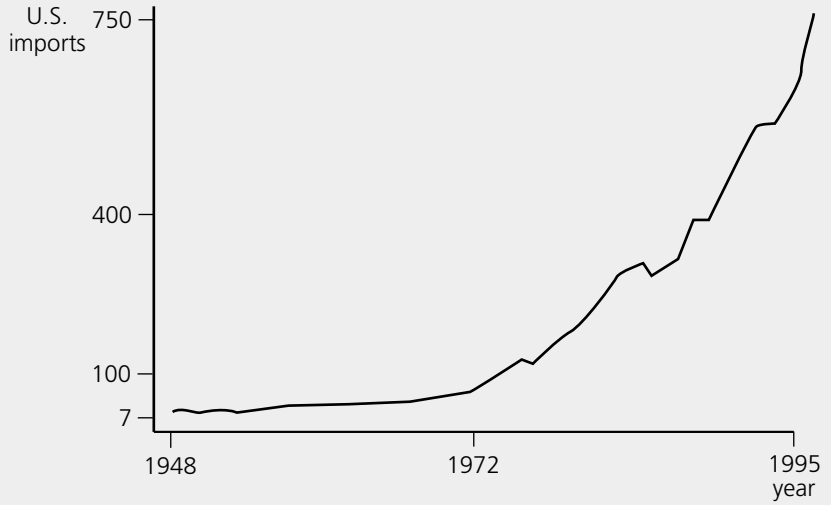
Many economic time series are better approximated by an **exponential trend**, which follows when a series has the same average growth rate from period to period. Figure 10.3 plots data on annual nominal imports for the United States during the years 1948 through 1995 (*ERP* 1997, Table B–101).

In the early years, we see that the change in the imports over each year is relatively small, whereas the change increases as time passes. This is consistent with a *constant average growth rate*: the percentage change is roughly the same in each period.

In practice, an exponential trend in a time series is captured by modeling the natural logarithm of the series as a linear trend (assuming that $y_t > 0$):

Figure 10.3

Nominal U.S. imports during the years 1948–1995 (in billions of U.S. dollars).



$$\log(y_t) = \beta_0 + \beta_1 t + e_t,\ t = 1,2,\ldots. \tag{10.26}$$

Exponentiating shows that y_t itself has an exponential trend: $y_t = \exp(\beta_0 + \beta_1 t + e_t)$. Because we will want to use exponentially trending time series in linear regression models, (10.26) turns out to be the most convenient way for representing such series.

How do we interpret β_1 in (10.26)? Remember that, for small changes, $\Delta\log(y_t) = \log(y_t) - \log(y_{t-1})$ is approximately the proportionate change in y_t:

$$\Delta\log(y_t) \approx (y_t - y_{t-1})/y_{t-1}. \tag{10.27}$$

The right-hand side of (10.27) is also called the **growth rate** in y from period $t - 1$ to period t. To turn the growth rate into a percent, we simply multiply by 100. If y_t follows (10.26), then, taking changes and setting $\Delta e_t = 0$,

$$\Delta\log(y_t) = \beta_1, \text{ for all } t. \tag{10.28}$$

In other words, β_1 is approximately the average per period growth rate in y_t. For example, if t denotes year and $\beta_1 = .027$, then y_t grows about 2.7% per year on average.

Although linear and exponential trends are the most common, time trends can be more complicated. For example, instead of the linear trend model in (10.24), we might have a quadratic time trend:

$$y_t = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + e_t. \tag{10.29}$$

If α_1 and α_2 are positive, then the slope of the trend is increasing, as is easily seen by computing the approximate slope (holding e_t fixed):

$$\frac{\Delta y_t}{\Delta t} \approx \alpha_1 + 2\alpha_2 t. \tag{10.30}$$

[If you are familiar with calculus, you recognize the right-hand side of (10.30) as the derivative of $\alpha_0 + \alpha_1 t + \alpha_2 t^2$ with respect to t.] If $\alpha_1 > 0$, but $\alpha_2 < 0$, the trend has a hump shape. This may not be a very good description of certain trending series because it requires an increasing trend to be followed, eventually, by a decreasing trend. Nevertheless, over a given time span, it can be a flexible way of modeling time series that have more complicated trends than either (10.24) or (10.26).

Using Trending Variables in Regression Analysis

Accounting for explained or explanatory variables that are trending is fairly straightforward in regression analysis. First, nothing about trending variables necessarily violates the classical linear model assumptions, TS.1 through TS.6. However, we must be careful to allow for the fact that unobserved, trending factors that affect y_t might also be correlated with the explanatory variables. If we ignore this possibility, we may find a spurious relationship between y_t and one or more explanatory variables. The phenomenon of finding a relationship between two or more trending variables simply

because each is growing over time is an example of **spurious regression**. Fortunately, adding a time trend eliminates this problem.

For concreteness, consider a model where two observed factors, x_{t1} and x_{t2}, affect y_t. In addition, there are unobserved factors that are systematically growing or shrinking over time. A model that captures this is

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \beta_3 t + u_t. \tag{10.31}$$

This fits into the multiple linear regression framework with $x_{t3} = t$. Allowing for the trend in this equation explicitly recognizes that y_t may be growing ($\beta_3 > 0$) or shrinking ($\beta_3 < 0$) over time for reasons essentially unrelated to x_{t1} and x_{t2}. If (10.31) satisfies assumptions TS.1, TS.2, and TS.3, then omitting t from the regression and regressing y_t on x_{t1}, x_{t2} will generally yield biased estimators of β_1 and β_2: we have effectively omitted an important variable, t, from the regression. This is especially true if x_{t1} and x_{t2} are themselves trending, because they can then be highly correlated with t. The next example shows how omitting a time trend can result in spurious regression.

EXAMPLE 10.7
(Housing Investment and Prices)

The data in HSEINV.RAW are annual observations on housing investment and a housing price index in the United States for 1947 through 1988. Let *invpc* denote real per capita housing investment (in thousands of dollars) and let *price* denote a housing price index (equal to one in 1982). A simple regression in constant elasticity form, which can be thought of as a supply equation for housing stock, gives

$$\begin{gathered}\widehat{\log(invpc)} = -.550 + 1.241 \log(price) \\ (.043) \quad (0.382) \\ n = 42, R^2 = .208, \bar{R}^2 = .189.\end{gathered} \tag{10.32}$$

The elasticity of per capita investment with respect to price is very large and statistically significant; it is not statistically different from one. We must be careful here. Both *invpc* and *price* have upward trends. In particular, if we regress log(*invpc*) on *t*, we obtain a coefficient on the trend equal to .0081 (standard error = .0018); the regression of log(*price*) on *t* yields a trend coefficient equal to .0044 (standard error = .0004). While the standard errors on the trend coefficients are not necessarily reliable—these regressions tend to contain substantial serial correlation—the coefficient estimates do reveal upward trends.

To account for the trending behavior of the variables, we add a time trend:

$$\begin{gathered}\widehat{\log(invpc)} = -.913 - .381 \log(price) + .0098\, t \\ (.136) \quad (.679) \qquad\qquad (.0035) \\ n = 42, R^2 = .341, \bar{R}^2 = .307.\end{gathered} \tag{10.33}$$

The story is much different now: the estimated price elasticity is negative and not statistically different from zero. The time trend is statistically significant, and its coefficient implies

an approximate 1% increase in *invpc* per year, on average. From this analysis, we cannot conclude that real per capita housing investment is influenced at all by price. There are other factors, captured in the time trend, that affect *invpc*, but we have not modeled these. The results in (10.32) show a spurious relationship between *invpc* and *price* due to the fact that price is also trending upward over time.

In some cases, adding a time trend can make a key explanatory variable *more* significant. This can happen if the dependent and independent variables have different kinds of trends (say, one upward and one downward), but movement in the independent variable *about* its trend line causes movement in the dependent variable away from its trend line.

EXAMPLE 10.8
(Fertility Equation)

If we add a linear time trend to the fertility equation (10.18), we obtain

$$\widehat{gfr}_t = \underset{(3.36)}{111.77} + \underset{(.040)}{.279}\, pe_t - \underset{(6.30)}{35.59}\, ww2_t + \underset{(6.626)}{.997}\, pill_t - \underset{(0.19)}{1.15}\, t \qquad \textbf{(10.34)}$$
$$n = 72,\ R^2 = .662,\ \bar{R}^2 = .642.$$

The coefficient on *pe* is more than triple the estimate from (10.18), and it is much more statistically significant. Interestingly, *pill* is not significant once an allowance is made for a linear trend. As can be seen by the estimate, *gfr* was falling, on average, over this period, other factors being equal.

Since the general fertility rate exhibited both upward and downward trends during the period from 1913 through 1984, we can see how robust the estimated effect of *pe* is when we use a quadratic trend:

$$\widehat{gfr}_t = \underset{(4.36)}{124.09} + \underset{(.040)}{.348}\, pe_t - \underset{(5.71)}{35.88}\, ww2_t - \underset{(6.34)}{10.12}\, pill_t$$
$$- \underset{(0.39)}{2.53}\, t + \underset{(.0050)}{.0196}\, t^2 \qquad \textbf{(10.35)}$$
$$n = 72,\ R^2 = .727,\ \bar{R}^2 = .706.$$

The coefficient on *pe* is even larger and more statistically significant. Now, *pill* has the expected negative effect and is marginally significant, and both trend terms are statistically significant. The quadratic trend is a flexible way to account for the unusual trending behavior of *gfr*.

You might be wondering in Example 10.8: Why stop at a quadratic trend? Nothing prevents us from adding, say, t^3 as an independent variable, and, in fact, this might be

warranted (see Exercise 10.12). But we have to be careful not to get carried away when including trend terms in a model. We want relatively simple trends that capture broad movements in the dependent variable that are not explained by the independent variables in the model. If we include enough polynomial terms in t, then we can track any series pretty well. But this offers little help in finding which explanatory variables affect y_t.

A Detrending Interpretation of Regressions with a Time Trend

Including a time trend in a regression model creates a nice interpretation in terms of **detrending** the original data series before using them in regression analysis. For concreteness, we focus on model (10.31), but our conclusions are much more general.

When we regress y_t on x_{t1}, x_{t2} and t, we obtain the fitted equation

$$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x_{t1} + \hat{\beta}_2 x_{t2} + \hat{\beta}_3 t. \tag{10.36}$$

We can extend the results on the partialling out interpretation of OLS that we covered in Chapter 3 to show that $\hat{\beta}_1$ and $\hat{\beta}_2$ can be obtained as follows.

(i) Regress each of y_t, x_{t1} and x_{t2} on a constant and the time trend t and save the residuals, say $\ddot{y}_t$, $\ddot{x}_{t1}$, $\ddot{x}_{t2}$, $t = 1,2, \ldots, n$. For example,

$$\ddot{y}_t = y_t - \hat{\alpha}_0 - \hat{\alpha}_1 t.$$

Thus, we can think of $\ddot{y}_t$ as being *linearly detrended.* In detrending y_t, we have estimated the model

$$y_t = \alpha_0 + \alpha_1 t + e_t$$

by OLS; the residuals from this regression, $\hat{e}_t = \ddot{y}_t$, have the time trend removed (at least in the sample). A similar interpretation holds for $\ddot{x}_{t1}$ and $\ddot{x}_{t2}$.

(ii) Run the regression of

$$\ddot{y}_t \text{ on } \ddot{x}_{t1}, \ddot{x}_{t2}. \tag{10.37}$$

(No intercept is necessary, but including an intercept affects nothing: the intercept will be estimated to be zero.) This regression exactly yields $\hat{\beta}_1$ and $\hat{\beta}_2$ from (10.36).

This means that the estimates of primary interest, $\hat{\beta}_1$ and $\hat{\beta}_2$, can be interpreted as coming from a regression *without* a time trend, but where we first detrend the dependent variable and all other independent variables. The same conclusion holds with any number of independent variables and if the trend is quadratic or of some other polynomial degree.

If t is omitted from (10.36), then no detrending occurs, and y_t might seem to be related to one or more of the x_{tj} simply because each contains a trend; we saw this in Example 10.7. If the trend term is statistically significant, and the results change in important ways when a time trend is added to a regression, then the initial results without a trend should be treated with suspicion.

The interpretation of $\hat{\beta}_1$ and $\hat{\beta}_2$ shows that it is a good idea to include a trend in the regression if any independent variable is trending, even if y_t is not. If y_t has no noticeable trend, but, say, x_{t1} is growing over time, then excluding a trend from the regression

may make it look as if x_{t1} has no effect on y_t, even though movements of x_{t1} about its trend may affect y_t. This will be captured if t is included in the regression.

EXAMPLE 10.9
(Puerto Rican Employment)

When we add a linear trend to equation (10.17), the estimates are

$$\widehat{\log(prepop_t)} = \underset{(1.30)}{-8.70} - \underset{(.044)}{.169}\log(mincov_t) + \underset{(0.18)}{1.06}\log(usgnp_t) - \underset{(.005)}{.032}\,t \qquad (10.38)$$
$$n = 38,\ R^2 = .847,\ \bar{R}^2 = .834.$$

The coefficient on log(*usgnp*) has changed dramatically: from −.012 and insignificant to 1.06 and very significant. The coefficient on the minimum wage has changed only slightly, although the standard error is notably smaller, making log(*mincov*) more significant than before.

The variable $prepop_t$ displays no clear upward or downward trend, but log(*usgnp*) has an upward, linear trend. (A regression of log(*usgnp*) on *t* gives an estimate of about .03, so that *usgnp* is growing by about 3% per year over the period.) We can think of the estimate 1.06 as follows: when *usgnp* increases by 1% *above* its long-run trend, *prepop* increases by about 1.06%.

Computing *R*-squared when the Dependent Variable is Trending

R-squareds in time series regressions are often very high, especially compared with typical *R*-squareds for cross-sectional data. Does this mean that we learn more about factors affecting *y* from time series data? Not necessarily. On one hand, time series data often come in aggregate form (such as average hourly wages in the U.S. economy), and aggregates are often easier to explain than outcomes on individuals, families, or firms, which is often the nature of cross-sectional data. But the usual and adjusted *R*-squares for time series regressions can be artificially high when the dependent variable is trending. Remember that R^2 is a measure of how large the error variance is relative to the variance of *y*. The formula for the adjusted *R*-squared shows this directly:

$$\bar{R}^2 = 1 - (\hat{\sigma}_u^2/\hat{\sigma}_y^2),$$

where $\hat{\sigma}_u^2$ is the unbiased estimator of the error variance, $\hat{\sigma}_y^2 = \text{SST}/(n-1)$, and $\text{SST} = \sum_{t=1}^{n}(y_t - \bar{y})^2$. Now, estimating the error variance when y_t is trending is no problem, provided a time trend is included in the regression. However, when $\text{E}(y_t)$ follows, say, a linear time trend [see (10.24)], $\text{SST}/(n-1)$ is no longer an unbiased or consistent estimator of $\text{Var}(y_t)$. In fact, $\text{SST}/(n-1)$ can substantially overestimate the variance in y_t, because it does not account for the trend in y_t.

When the dependent variable satisfies linear, quadratic, or any other polynomial trends, it is easy to compute a goodness-of-fit measure that first nets out the effect of any time trend on y_t. The simplest method is to compute the usual *R*-squared in a regression where the dependent variable has already been detrended. For example, if the model is (10.31), then we first regress y_t on t and obtain the residuals $\ddot{y}_t$. Then, we regress

$$\ddot{y}_t \text{ on } x_{t1}, x_{t2}, \text{ and } t. \tag{10.39}$$

The *R*-squared from this regression is

$$1 - \frac{\text{SSR}}{\sum_{t=1}^{n} \ddot{y}_t^2}, \tag{10.40}$$

where SSR is identical to the sum of squared residuals from (10.36). Since $\sum_{t=1}^{n} \ddot{y}_t^2 \leq \sum_{t=1}^{n} (y_t - \bar{y})^2$ (and usually the inequality is strict), the *R*-squared from (10.40) is no greater than, and usually less than, the *R*-squared from (10.36). (The sum of squared residuals is identical in both regressions.) When y_t contains a strong linear time trend, (10.40) can be much less than the usual *R*-squared.

The *R*-squared in (10.40) better reflects how well x_{t1} and x_{t2} explain y_t, because it nets out the effect of the time trend. After all, we can always explain a trending variable with some sort of trend, but this does not mean we have uncovered any factors that cause movements in y_t. An adjusted *R*-squared can also be computed based on (10.40): divide SSR by $(n - 4)$ because this is the *df* in (10.36) and divide $\sum_{t=1}^{n} \ddot{y}_t^2$ by $(n - 2)$, as there are two trend parameters estimated in detrending y_t. In general, SSR is divided by the *df* in the usual regression (that includes any time trends), and $\sum_{t=1}^{n} \ddot{y}_t^2$ is divided by $(n - p)$, where p is the number of trend parameters estimated in detrending y_t. See Wooldridge (1991a) for further discussion on computing goodness-of-fit measures with trending variables.

EXAMPLE 10.10
(Housing Investment)

In Example 10.7, we saw that including a linear time trend along with log(*price*) in the housing investment equation had a substantial effect on the price elasticity. But the *R*-squared from regression (10.33), taken literally, says that we are "explaining" 34.1% of the variation in log(*invpc*). This is misleading. If we first detrend log(*invpc*) and regress the detrended variable on log(*price*) and *t*, the *R*-squared becomes .008, and the adjusted *R*-squared is actually negative. Thus, movements in log(*price*) about its trend have virtually no explanatory power for movements in log(*invpc*) about its trend. This is consistent with the fact that the *t* statistic on log(*price*) in equation (10.33) is very small.

Before leaving this subsection, we must make a final point. In computing the *R*-squared form of an *F* statistic for testing multiple hypotheses, we just use the usual *R*-squareds without any detrending. Remember, the *R*-squared form of the *F* statistic is just a computational device, and so the usual formula is always appropriate.

Seasonality

If a time series is observed at monthly or quarterly intervals (or even weekly or daily), it may exhibit **seasonality**. For example, monthly housing starts in the Midwest are strongly influenced by weather. While weather patterns are somewhat random, we can be sure that the weather during January will usually be more inclement than in June, and so housing starts are generally higher in June than in January. One way to model this phenomenon is to allow the expected value of the series, y_t, to be different in each month. As another example, retail sales in the fourth quarter are typically higher than in the previous three quarters because of the Christmas holiday. Again, this can be captured by allowing the average retail sales to differ over the course of a year. This is in addition to possibly allowing for a trending mean. For example, retail sales in the most recent first quarter were higher than retail sales in the fourth quarter from 30 years ago, because retail sales have been steadily growing. Nevertheless, if we compare average sales within a typical year, the seasonal holiday factor tends to make sales larger in the fourth quarter.

Even though many monthly and quarterly data series display seasonal patterns, not all of them do. For example, there is no noticeable seasonal pattern in monthly interest or inflation rates. In addition, series that do display seasonal patterns are often **seasonally adjusted** before they are reported for public use. A seasonally adjusted series is one that, in principle, has had the seasonal factors removed from it. Seasonal adjustment can be done in a variety of ways, and a careful discussion is beyond the scope of this text. [See Harvey (1990) and Hylleberg (1986) for detailed treatments.]

Seasonal adjustment has become so common that it is not possible to get seasonally unadjusted data in many cases. Quarterly U.S. GDP is a leading example. In the annual *Economic Report of the President*, many macroeconomic data sets reported at monthly frequencies (at least for the most recent years) and those that display seasonal patterns are all seasonally adjusted. The major sources for macroeconomic time series, including *Citibase*, also seasonally adjust many of the series. Thus, the scope for using our own seasonal adjustment is often limited.

Sometimes, we do work with seasonally unadjusted data, and it is useful to know that simple methods are available for dealing with seasonality in regression models. Generally, we can include a set of **seasonal dummy variables** to account for seasonality in the dependent variable, the independent variables, or both.

The approach is simple. Suppose that we have monthly data, and we think that seasonal patterns within a year are roughly constant across time. For example, since Christmas always comes at the same time of year, we can expect retail sales to be, on average, higher in months late in the year than in earlier months. Or, since weather patterns are broadly similar across years, housing starts in the Midwest will be higher on average during the summer months than the winter months. A general model for monthly data that captures these phenomena is

$$y_t = \beta_0 + \delta_1 feb_t + \delta_2 mar_t + \delta_3 apr_t + \ldots + \delta_{11} dec_t + \beta_1 x_{t1} + \ldots + \beta_k x_{tk} + u_t, \quad \textbf{(10.41)}$$

where feb_t, mar_t, …, dec_t are dummy variables indicating whether time period t corresponds to the appropriate month. In this formulation, January is the base month, and β_0 is the intercept for January. If there is no seasonality in y_t, once the x_{tj} have been controlled for, then δ_1 through δ_{11} are all zero. This is easily tested via an F test.

QUESTION 10.5

In equation (10.41), what is the intercept for March? Explain why seasonal dummy variables satisfy the strict exogeneity assumption.

EXAMPLE 10.11
(Effects of Antidumping Filings)

In Example 10.5, we used monthly data that have not been seasonally adjusted. Therefore, we should add seasonal dummy variables to make sure none of the important conclusions changes. It could be that the months just before the suit was filed are months where imports are higher or lower, on average, than in other months. When we add the 11 monthly dummy variables as in (10.41) and test their joint significance, we obtain p-value = .59, and so the seasonal dummies are jointly insignificant. In addition, nothing important changes in the estimates once statistical significance is taken into account. Krupp and Pollard (1996) actually used three dummy variables for the seasons (fall, spring, and summer, with winter as the base season), rather than a full set of monthly dummies; the outcome is essentially the same.

If the data are quarterly, then we would include dummy variables for three of the four quarters, with the omitted category being the base quarter. Sometimes, it is useful to interact seasonal dummies with some of the x_{tj} to allow the effect of x_{tj} on y_t to differ across the year.

Just as including a time trend in a regression has the interpretation of initially detrending the data, including seasonal dummies in a regression can be interpreted as **deseasonalizing** the data. For concreteness, consider equation (10.41) with $k = 2$. The OLS slope coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$ on x_1 and x_2 can be obtained as follows:

(i) Regress each of y_t, x_{t1} and x_{t2} on a constant and the monthly dummies, feb_t, mar_t, …, dec_t, and save the residuals, say $\ddot{y}_t$, $\ddot{x}_{t1}$ and $\ddot{x}_{t2}$, for all $t = 1,2, \ldots, n$. For example,

$$\ddot{y}_t = y_t - \hat{\alpha}_0 - \hat{\alpha}_1 feb_t - \hat{\alpha}_2 mar_t - \ldots - \hat{\alpha}_{11} dec_t.$$

This is one method of deseasonalizing a monthly time series. A similar interpretation holds for $\ddot{x}_{t1}$ and $\ddot{x}_{t2}$.

(ii) Run the regression, without the monthly dummies, of $\ddot{y}_t$ on $\ddot{x}_{t1}$ and $\ddot{x}_{t2}$ [just as in (10.37)]. This gives $\hat{\beta}_1$ and $\hat{\beta}_2$.

In some cases, if y_t has pronounced seasonality, a better goodness-of-fit measure is an R-squared based on the deseasonalized y_t. This nets out any seasonal effects that are

not explained by the x_{tj}. Specific degrees of freedom ajustments are discussed in Wooldridge (1991a).

Time series exhibiting seasonal patterns can be trending as well, in which case, we should estimate a regression model with a time trend and seasonal dummy variables. The regressions can then be interpreted as regressions using both detrended and deseasonalized series. Goodness-of-fit statistics are discussed in Wooldridge (1991a): essentially, we detrend and deasonalize y_t by regressing on both a time trend and seasonal dummies before computing *R*-squared.

SUMMARY

In this chapter, we have covered basic regression analysis with time series data. Under assumptions that parallel those for cross-sectional analysis, OLS is unbiased (under TS.1 through TS.3), OLS is BLUE (under TS.1 through TS.5), and the usual OLS standard errors, *t* statistics, and *F* statistics can be used for statistical inference (under TS.1 through TS.6). Because of the temporal correlation in most time series data, we must explicitly make assumptions about how the errors are related to the explanatory variables in all time periods and about the temporal correlation in the errors themselves. The classical linear model assumptions can be pretty restrictive for time series applications, but they are a natural starting point. We have applied them to both static regression and finite distributed lag models.

Logarithms and dummy variables are used regularly in time series applications and in event studies. We also discussed index numbers and time series measured in terms of nominal and real dollars.

Trends and seasonality can be easily handled in a multiple regression framework by including time and seasonal dummy variables in our regression equations. We presented problems with the usual *R*-squared as a goodness-of-fit measure and suggested some simple alternatives based on detrending or deseasonalizing.

KEY TERMS

Autocorrelation
Base Period
Base Value
Contemporaneously Exogenous
Deseasonalizing
Detrending
Event Study
Exponential Trend
Finite Distributed Lag (FDL) Model
Growth Rate
Impact Multiplier
Impact Propensity
Index Number
Lag Distribution
Linear Time Trend
Long-Run Elasticity
Long-Run Multiplier
Long-Run Propensity (LRP)
Seasonal Dummy Variables
Seasonality
Seasonally Adjusted
Serial Correlation
Short-Run Elasticity
Spurious Regression
Static Model
Stochastic Process
Strictly Exogenous
Time Series Process
Time Trend

PROBLEMS

10.1 Decide if you agree or disagree with each of the following statements and give a brief explanation of your decision:

(i) Like cross-sectional observations, we can assume that most time series observations are independently distributed.

(ii) The OLS estimator in a time series regression is unbiased under the first three Gauss-Markov assumptions.

(iii) A trending variable cannot be used as the dependent variable in multiple regression analysis.

(iv) Seasonality is not an issue when using annual time series observations.

10.2 Let $gGDP_t$ denote the annual percentage change in gross domestic product and let int_t denote a short-term interest rate. Suppose that $gGDP_t$ is related to interest rates by

$$gGDP_t = \alpha_0 + \delta_0 int_t + \delta_1 int_{t-1} + u_t,$$

where u_t is uncorrelated with int_t, int_{t-1}, and all other past values of interest rates. Suppose that the Federal Reserve follows the policy rule:

$$int_t = \gamma_0 + \gamma_1(gGDP_{t-1} - 3) + v_t,$$

where $\gamma_1 > 0$. (When last year's GDP growth is above 3%, the Fed increases interest rates to prevent an "overheated" economy.) If v_t is uncorrelated with all past values of int_t and u_t, argue that int_t must be correlated with u_{t-1}. (*Hint*: Lag the first equation for one time period and substitute for $gGDP_{t-1}$ in the second equation.) Which Gauss-Markov assumption does this violate?

10.3 Suppose y_t follows a second order FDL model:

$$y_t = \alpha_0 + \delta_0 z_t + \delta_1 z_{t-1} + \delta_2 z_{t-2} + u_t.$$

Let z^* denote the *equilibrium value* of z_t and let y^* be the equilibrium value of y_t, such that

$$y^* = \alpha_0 + \delta_0 z^* + \delta_1 z^* + \delta_2 z^*.$$

Show that the change in y^*, due to a change in z^*, equals the long-run propensity times the change in z^*:

$$\Delta y^* = LRP \cdot \Delta z^*.$$

This gives an alternative way of interpreting the LRP.

10.4 When the three event indicators *befile6*, *affile6*, and *afdec6* are dropped from equation (10.22), we obtain $R^2 = .281$ and $\bar{R}^2 = .264$. Are the event indicators jointly significant at the 10% level?

10.5 Suppose you have quarterly data on new housing starts, interest rates, and real per capita income. Specify a model for housing starts that accounts for possible trends and seasonality in the variables.

10.6 In Example 10.4, we saw that our estimates of the individual lag coefficients in a distributed lag model were very imprecise. One way to alleviate the multicollinearity

problem is to assume that the δ_j follow a relatively simple pattern. For concreteness, consider a model with four lags:

$$y_t = \alpha_0 + \delta_0 z_t + \delta_1 z_{t-1} + \delta_2 z_{t-2} + \delta_3 z_{t-3} + \delta_4 z_{t-4} + u_t.$$

Now, let us assume that the δ_j follow a quadratic in the lag, j:

$$\delta_j = \gamma_0 + \gamma_1 j + \gamma_2 j^2,$$

for parameters γ_0, γ_1, and γ_2. This is an example of a *polynomial distributed lag (PDL) model.*

(i) Plug the formula for each δ_j into the distributed lag model and write the model in terms of the parameters γ_h, for $h = 0,1,2$.

(ii) Explain the regression you would run to estimate the γ_h.

(iii) The polynomial distributed lag model is a restricted version of the general model. How many restrictions are imposed? How would you test these? (*Hint*: Think F test.)

COMPUTER EXERCISES

10.7 In October 1979, the Federal Reserve changed its policy of targeting the money supply and instead began to focus directly on short-term interest rates. Using the data in INTDEF.RAW, define a dummy variable equal to one for years after 1979. Include this dummy in equation (10.15) to see if there is a shift in the interest rate equation after 1979. What do you conclude?

10.8 Use the data in BARIUM.RAW for this exercise.

(i) Add a linear time trend to equation (10.22). Are any variables, other than the trend, statistically significant?

(ii) In the equation estimated in part (i), test for joint significance of all variables except the time trend. What do you conclude?

(iii) Add monthly dummy variables to this equation and test for seasonality. Does including the monthly dummies change any other estimates or their standard errors in important ways?

10.9 Add the variable log(*prgnp*) to the minimum wage equation in (10.38). Is this variable significant? Interpret the coefficient. How does adding log(*prgnp*) affect the estimated minimum wage effect?

10.10 Use the data in FERTIL3.RAW to verify that the standard error for the LRP in equation (10.19) is about .030.

10.11 Use the data in EZANDERS.RAW for this exercise. The data are on monthly unemployment claims in Anderson Township in Indiana, from January 1980 through November 1988. In 1984, an enterprise zone (EZ) was located in Anderson (as well as other cities in Indiana). [See Papke (1994) for details.]

(i) Regress log(*uclms*) on a linear time trend and 11 monthly dummy variables. What was the overall trend in unemployment claims over this period? (Interpret the coefficient on the time trend.) Is there evidence of seasonality in unemployment claims?

(ii) Add *ez*, a dummy variable equal to one in the months Anderson had an EZ, to the regression in part (i). Does having the enterprise zone seem to decrease unemployment claims? By how much? [You should use formula (7.10) from Chapter 7.]

(iii) What assumptions do you need to make to attribute the effect in part (ii) to the creation of an EZ?

10.12 Use the data in FERTIL3.RAW for this exercise.

(i) Regress gfr_t on t and t^2 and save the residuals. This gives a detrended gfr_t, say $\ddot{gfr}_t$.

(ii) Regress $\ddot{gfr}_t$ on all of the variables in equation (10.35), including t and t^2. Compare the R-squared with that from (10.35). What do you conclude?

(iii) Reestimate equation (10.35) but add t^3 to the equation. Is this additional term statistically significant?

10.13 Use the data set CONSUMP.RAW for this exercise.

(i) Estimate a simple regression model relating the growth in real per capita consumption (of nondurables and services) to the growth in real per capita disposable income. Use the change in the logarithms in both cases. Report the results in the usual form. Interpret the equation and discuss statistical significance.

(ii) Add a lag of the growth in real per capita disposable income to the equation from part (i). What do you conclude about adjustment lags in consumption growth?

(iii) Add the real interest rate to the equation in part (i). Does it affect consumption growth?

10.14 Use the data in FERTIL3.RAW for this exercise.

(i) Add pe_{t-3} and pe_{t-4} to equation (10.19). Test for joint significance of these lags.

(ii) Find the estimated long-run propensity and its standard error in the model from part (i). Compare these with those obtained from equation (10.19).

(iii) Estimate the polynomial distributed lag model from Problem 10.6. Find the estimated LRP and compare this with what is obtained from the unrestricted model.

10.15 Use the data in VOLAT.RAW for this exercise. The variable *rsp500* is the monthly return on the Standard & Poors 500 stock market index, at an annual rate. (This includes price changes as well as dividends.) The variable *i3* is the return on three-month T-bills, and *pcip* is the percentage change in industrial production; these are also at an annual rate.

(i) Consider the equation

$$rsp500_t = \beta_0 + \beta_1 pcip_t + \beta_2 i3_t + u_t.$$

What signs do you think β_1 and β_2 should have?

(ii) Estimate the previous equation by OLS, reporting the results in standard form. Interpret the signs and magnitudes of the coefficients.

(iii) Which of the variables is statistically significant?

(iv) Does your finding from part (iii) imply that the return on the S&P 500 is predictable? Explain.

10.16 Consider the model estimated in (10.15); use the data in INTDEF.RAW.

(i) Find the correlation between *inf* and *def* over this sample period and comment.

(ii) Add a single lag of *inf* and *def* to the equation and report the results in the usual form.

(iii) Compare the estimated LRP for the effect of inflation from that in equation (10.15). Are they vastly different?

(iv) Are the two lags in the model jointly significant at the 5% level?