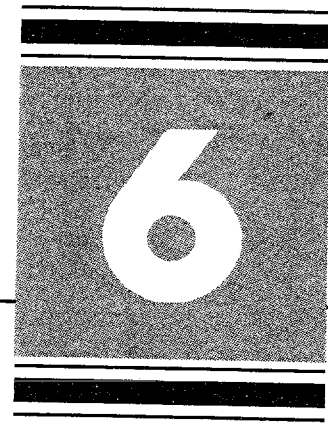

Simple Regression: Application

The simple regression model developed in Chapter 5 can be used to estimate the structural (behavioral) relation between two variables whenever we believe that the model accurately describes the process by which Y is determined. Although this model is too simple to describe most economic relations, it has many useful applications. In all these, the task of estimating the unknown coefficients is the same as in the general case, and the technique needs no further elaboration. What remains is to interpret the estimated model, and the first two sections of this chapter focus on that problem.

When we believe that the true process is more complex than the basic model, there are several paths open to us. If the true relation involves two variables in a nonlinear fashion, it is sometimes possible to transform the relation into a linear one and then apply the basic techniques already developed. Much of this chapter is devoted to exploring these possibilities. If more than one variable plays a systematic role in the determination of Y, we are led to multiple regression (Chapter 7).

6.1 Interpretation of the Coefficients

As presented in Chapter 5, the simple regression model

$$Y_i = \beta_0 + \beta_1 X_i + u_i \tag{6.1}$$

is a theoretical statement of the process by which the value of Y for each observation is determined on the basis of the given value of X for that observation. The systematic part of the relation is specified by the true regression line

$$E[Y] = \beta_0 + \beta_1 X \tag{6.2}$$

where $E[Y]$ denotes the expected value of Y associated with any particular value of X. This is graphed in Figure 6.1, which illustrates that $E[Y]$ is a linear function of X.

When we have data on Y and X, an estimated regression line

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X \tag{6.3}$$

can be determined by the method of ordinary least squares. This line is the empirical counterpart of the true regression line, and it serves as an estimated model of the systematic relation between Y and X. We should keep in mind, however, that even in the best of circumstances this model is only an estimate of the true relation—in the sense that $\hat{\beta}_0$ and $\hat{\beta}_1$ are only estimates of β_0 and β_1. Figure 6.1 illustrates an estimated regression line that (hypothetically) is based on data produced by the process described by the true regression line in the same figure. As explained in Section 5.3, the differences between the estimated and true regressions arise from the particular pattern of disturbances in the data set.

After estimation, our interest moves on to interpreting and using the estimated model. For this we blend together economic reasoning and some mathematical analysis. Since (6.3) is the equation of a straight line, the mathematics is easy.

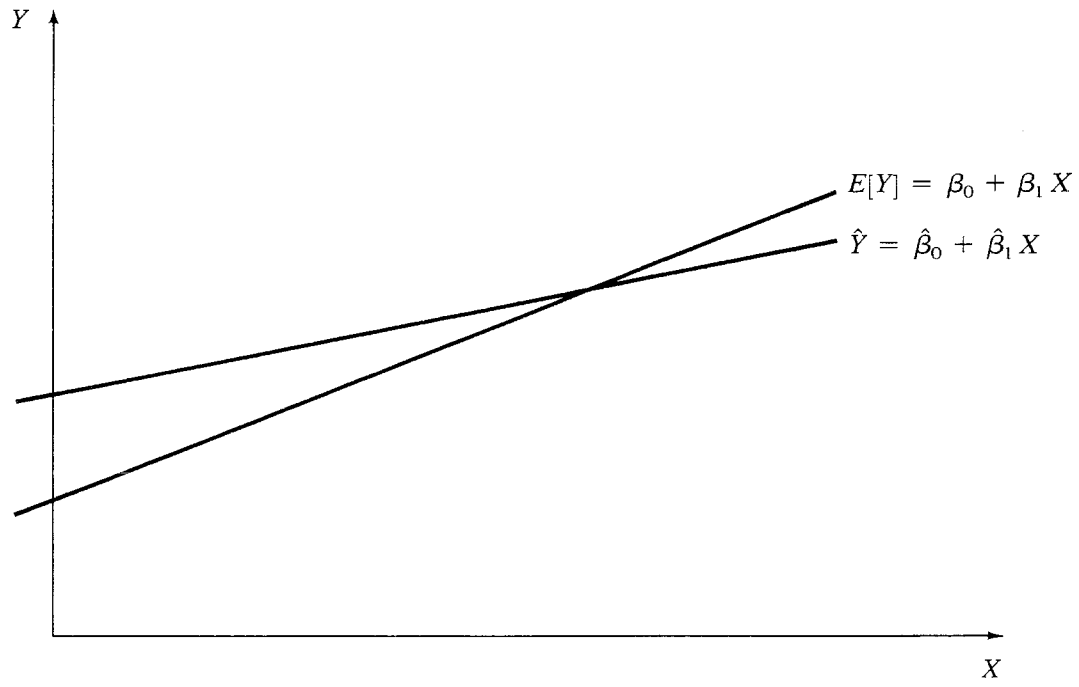

FIGURE 6.1 The systematic part of the simple regression model is specified by the true regression line, whose intercept is β_0 and whose slope is β_1. An estimated regression line fit to data produced by this process has an intercept of $\hat{\beta}_0$ and a slope of $\hat{\beta}_1$.

The parameter $\hat{\beta}_0$ in the estimated regression is called the ***intercept*** because it is the intercept of the estimated regression line with the vertical axis drawn through $X = 0$. Strictly speaking, the intercept gives the predicted value of Y for an observation with $X = 0$. In some cases this interpretation is of special interest. However, in many cases such an interpretation does not make economic sense. For example, with an aggregate consumption function it would be absurd to use a real-world model to predict what consumption would be if income were equal to zero: in all probability, either everyone would starve or chaos would reign as inventories of food and products were consumed.

The parameter $\hat{\beta}_1$ is called the ***slope coefficient.*** Based on the simple analytics of a straight line (see the appendix to Chapter 1), we know that the change in predicted Y that would be associated with any particular change in X is given by

$$\Delta\hat{Y} = \hat{\beta}_1 \Delta X \tag{6.4}$$

In this context, a "change in X" that is economically interesting might be the difference between the values of X for two observations or it might be a hypothetical change in the value of X for a particular observation. In either case, we are considering moving from one point to another along the estimated regression line.

Based on (6.4), we can see that if X changes by 1, $\hat{Y}$ will change by $\hat{\beta}_1$. That is,

$$\text{if} \quad \Delta X = 1, \quad \text{then} \quad \Delta\hat{Y} = \hat{\beta}_1 \tag{6.5}$$

This simple relation is the basis of our usual interpretation of a slope coefficient: $\hat{\beta}_1$ is the change in $\hat{Y}$ that results from a unit change in X. Meaning this, we say that $\hat{\beta}_1$ is the ***impact*** of X on $\hat{Y}$. (Note that a "unit change in X" means $\Delta X = 1$; this is different from a change in the units of measurement for X.)

When we believe that the form of the model correctly specifies the true process relating Y to X, which is a belief that we usually hold, we can recast our interpretations in terms of the true parameters. For example, since $\hat{\beta}_1$ is an estimate of the true β_1, we might say that $\hat{\beta}_1$ is an estimate of the impact of X on $E[Y]$.

A semantic problem arises in discussing the impact of the explanatory variable in a regression. It is tempting to use the word "significant" when the impact is big or noteworthy in an economic sense. Unfortunately, this use of "significant" in statistical discussion has been preempted by a technical meaning that is different from ordinary usage. As set out in Chapter 12, the meaning of "significant" in a basic hypothesis test is simply "not zero," and to use the word to discuss the importance of the impact of one variable on another can lead only to confusion.

Finally, since $\hat{\beta}_1$ estimates the impact of X on Y, it is tempting to jump to the conclusion that the larger is the $\hat{\beta}_1$ value in a regression the more important is X in explaining Y. This conclusion is valid when thinking about possible

values for the coefficient in a particular regression, but it is invalid when thinking about a comparison of coefficients in different regressions. Indeed, as shown in Section 5.5, the magnitude of a coefficient is affected by the units of measurement of the variables in the data. Therefore, the magnitude by itself does not indicate the importance of the coefficient.

6.2 The Earnings Function and the Consumption Function

In this section we present two examples of regular simple regression that serve as small case studies of econometric research. In both cases we start with a theory or belief about the form of the relation. The actual calculations are not shown, because they are standard applications of the methods of Chapter 5 and because we nearly always rely on a computer to carry them out anyway. The emphasis is on interpreting and using the estimated model.

The Earnings Function

We start with a simple theory that labor markets serve to determine persons' earnings according to their educational attainment as specified in

$$EARNS_i = \beta_0 + \beta_1 ED_i + u_i \tag{6.6}$$

using mnemonic variable names instead of Y and X. The theory might be based on the idea that education enhances productivity, which is rewarded in labor markets with greater earnings. Alternatively, it might be based on the ideas that educational attainment mainly certifies the existence of potential abilities, and that earnings are based on this certification. In either case, (6.6) is an appropriate model.

Using the 100 observations in our cross-section data set, in which *EARNS* measures earnings in thousands of dollars and *ED* measures years of schooling, and applying (5.12) and (5.13) to calculate the OLS coefficients, we find that

$$\begin{aligned} \widehat{EARNS}_i &= -1.315 + 0.797 ED_i \\ R^2 &= .285 \qquad SER = 4.361 \end{aligned} \tag{6.7}$$

The estimated coefficient on *ED* is $\hat{\beta}_1^* = 0.797$, which means that the estimated labor market reward for an extra year of schooling is 0.797 thousand dollars greater annual earnings. Put more simply, we have estimated that the impact of education on earnings is \$797 greater annual earnings for each additional year of schooling. Is this an important effect? Certainly it is not trivial. Given the relative ease of acquiring another year of education and taking into account the magnitude of this impact in comparison with the variation in earnings that exists among the observations, most economists (and educators) would say that education has a fairly important effect on earnings in this model. (We point

out that the fact that even though 0.797 looks like a small number, this alone tells us nothing about the importance of the coefficient.)

The intercept $\hat{\beta}_1^* = -1.315$ tells us that the predicted earnings of a person with no schooling is negative \$1315. Since labor markets do not offer negative earnings, this unrealistic finding needs careful attention. One possibility is that a worker with no schooling would indeed have negative productivity in a job, and so he would not be hired. In this case the estimated model is in accord with the true behavior of labor markets, but the naive interpretation of the intercept is misleading because it attempts to apply the model for a value of *ED* that is not appropriate. Another possibility is that workers with no schooling do indeed have positive earnings in labor markets, but that the particular outcomes for the disturbances in our data have led to a negative estimate for the intercept even though the true value is positive. Yet another possibility is that our model (6.6) is incorrect—the true form might be nonlinear, with positive earnings for workers with no education—and that the attempt to fit a straight line to the data has led to an unrealistic result. It turns out that in our data the minimum value for *ED* is 2 years, for which the estimated model predicts positive earnings. Thus we might proceed with some faith in the model and its estimation. (However, the issue unresolved, and we note that our attempt to make a careful interpretation of an estimated coefficient has led us to recognize the need for more research.)

The R^2 indicates that only about 30 percent of the variation in earnings among the observations is explained by the level of educational attainment. This may seem low, but it is similar to R^2 values found in other cross-section studies of earnings.

The relatively poor fit is also evidenced by the *SER*, which is 4.361 thousand dollars. To interpret this, the magnitude must be compared to values of *EARNS*, which is the dependent variable. In the data, *EARNS* ranges from 0.750 to 30.000 and has a mean of 7.911. Taking the mean value as a standard for comparison, the typical error of fit indeed seems quite large.

Using (6.7) we can predict that the earnings in 1963 of a college graduate ($ED = 16$) who was a male head of family aged 25–54 would have been

$$\widehat{EARNS} = -1.315 + (0.797)(16) = 11.437 \tag{6.8}$$

thousand dollars. The typical error associated with an out-of-sample prediction such as this turns out to be even larger than the standard error of the regression, *SER*. (In Chapter 13 we see how these typical prediction errors are calculated.)

It should be noted that our data pertain only to male heads of families in the 25–54 age range. Making predictions or assessing the impact of education on the basis of this estimated earnings function is appropriate only with reference to this particular group. We do not expect that it will adequately predict women's earnings, for example. Also, the dollar magnitudes are based on 1963 labor market conditions.

The Consumption Function

As discussed in Chapter 1, an aggregate consumption function based on simplified Keynesian ideas can be specified as

$$CON_i = \beta_0 + \beta_1 DPI_i + u_i \tag{6.9}$$

where CON_i is aggregate personal consumption expenditure in year i, and DPI_i is aggregate disposable personal income in the same year. β_1 is interpreted as the marginal propensity to consume, because (6.9) implies that if *DPI* increases by 1 dollar, then $E[CON]$ will increase by β_1 dollars (assuming that both variables are measured in the same units).

Using the 25 observations for the years 1956–1980 in our time-series data set, in which *CON* and *DPI* are both measured in billions of 1972 dollars, and calculating the ordinary least squares estimates, we find that

$$\widehat{CON}_i = 0.568 + 0.907 DPI_i \qquad R^2 = .997 \quad SER = 8.935 \tag{6.10}$$

The estimated marginal propensity to consume ($\hat{\beta}_1^*$) is 0.907, which is consistent with Keynes' conjecture. The fit is extraordinarily good: the R^2 of .997 means that nearly 100 percent of the variation in *CON* over this period is explained by the regression (i.e., is explained by variation in *DPI*). Although it seems that we might have discovered some fundamental economic law, judgment must be reserved on this question. Very high R^2 values are common in time-series studies because most variables tend to increase over time, and therefore high correlations will exist among them even if cause-and-effect relations are absent or weak. Also, other specifications of the process determining consumption behavior may be preferred in economic research.

We can use (6.10) to predict how high *CON* will be when *DPI* is 1.2 trillion dollars:

$$\widehat{CON} = 0.568 + (0.907)(1200) = 1089 \tag{6.11}$$

billion dollars. Also, if *DPI* were to decrease by 20 billion dollars (from whatever level it might be at), the estimated model predicts a change in consumption of

$$\Delta\widehat{CON} = (0.907)(-20) = -18.14 \tag{6.12}$$

billion dollars. Note that the estimated intercept plays no role in a calculation like this.

6.3 Alternative Model Specifications

As already noted, the appropriateness of using the regular simple regression model is contingent on its being an accurate description of the particular process

being studied. Most important, the model requires that for each observation the expected value of Y be a linear function of the value of X. In the earnings function and consumption function, this was taken to be an appropriate specification of the relations.

In other cases, however, theory and evidence may lead us to believe that the relation between two variables is definitely not linear, or that it is not contemporaneous. Thus the regular model will not be an accurate description of the process, and using it will be inappropriate. (This statement might be tempered with the notion that if the relation is *approximately* described by the regular model, its use can be considered appropriate.)

In this section and the next we see how some structural relations that do not conform to the regular model can be respecified in such a way that OLS can be applied. The key to this is realizing that the Y and X in the regular model (6.1) need not be the interesting variables themselves, but that they can be variables that are constructed from the interesting variables. To reduce confusion, it is useful to refer to Y as the ***regressand*** rather than the dependent variable and to X as the ***regressor*** rather than the explanatory (or independent) variable.

Ratios of Variables

Part of the controversy surrounding the Keynesian consumption theory focused on the average propensity to consume (APC), which is defined as

$$APC_i = \frac{CON_i}{DPI_i} \tag{6.13}$$

One tradition and body of evidence viewed the consumption–income ratio as an economic constant that did not change over time. By contrast, Keynes conjectured that the APC would decline over time.

One way to look at the data and provide evidence on this question begins by assuming that the average propensity to consume is a simple linear function of time. The variables CON and DPI from the time-series data set are used to construct a new variable, APC, defined by (6.13). A regression model is specified and then estimated using APC as the regressand and the time trend T as the regressor. With $n = 25$, the results are

$$\begin{aligned} \widehat{APC_i} &= 0.911 - 0.000213T_i \\ R^2 &= .021 \qquad SER = 0.011 \end{aligned} \tag{6.14}$$

Sometimes regressions like this one are reported as

$$\widehat{\left(\frac{CON_i}{DPI_i}\right)} = 0.911 - 0.000213T_i \tag{6.15}$$

to emphasize that the regressand is constructed as a ratio of two variables.

These results show that the average propensity to consume decreases over time, seemingly in conformity with the Keynesian view. Each year, the *APC* is estimated to decrease by about 0.000213. Is this a lot or a little? In the middle year of the sample (1968, with $T = 13$) the predicted *APC* is $0.911 - (0.000213)(13) = 0.908$, and the annual decrease (0.000213) is very small compared with this. The annual decrease seems quite unimportant. Hence some persons might be inclined to say that for all practical purposes the average propensity to consume is constant.

The Reciprocal Specification

Often economic theory predicts that the systematic relation between two variables is nonlinear. If the nonlinearity is judged to be not too severe, it might be reasonable to proceed with the regular simple regression model as an approximation. However, this is not usually advisable because it inhibits our investigating the nonlinear features. An alternative approach involves finding a nonlinear mathematical form to specify a relation that is appropriate for the economic process and that is transformable into a linear relation.

One possibility is the ***reciprocal*** relation

$$Y = \beta_0 + \frac{\beta_1}{X} \tag{6.16}$$

The geometry of the reciprocal relation is illustrated in the left side of Figure 6.2. Y may be positive or negative. We focus only on cases in which all the X values are positive; although the reciprocal relation is defined for negative X values, it is rarely used in these cases. If β_1 is negative, the slope of the relation between Y and X is positive and it becomes flatter as X increases. Y never rises above the value $Y = \beta_0$, and in fact it never quite reaches it. If β_1 is zero, Y is constant. If β_1 is positive, the slope of the relation between Y and X is negative and becomes flatter as X increases. Y never reaches or falls below the value $Y = \beta_0$.

Now consider a new variable, *XINV*, that is equal to the reciprocal (i.e., inverse) of X: $XINV = 1/X$. It follows from (6.16) that Y is a linear function of *XINV*:

$$Y = \beta_0 + \beta_1 XINV \tag{6.17}$$

The relation between Y and *XINV* is illustrated on the right side of Figure 6.2, with three cases depending on the value of β_1. We see that the specific form of the nonlinear relation (6.16) implies a linear relation between Y and *XINV*.

The potential for applying this bit of mathematical analysis to econometric regression modeling should be clear. If we have a theory or belief that the systematic part of the relation between Y and X is reciprocal, like (6.16), this theory can be reexpressed to state that Y is linearly related to *XINV*. Further,

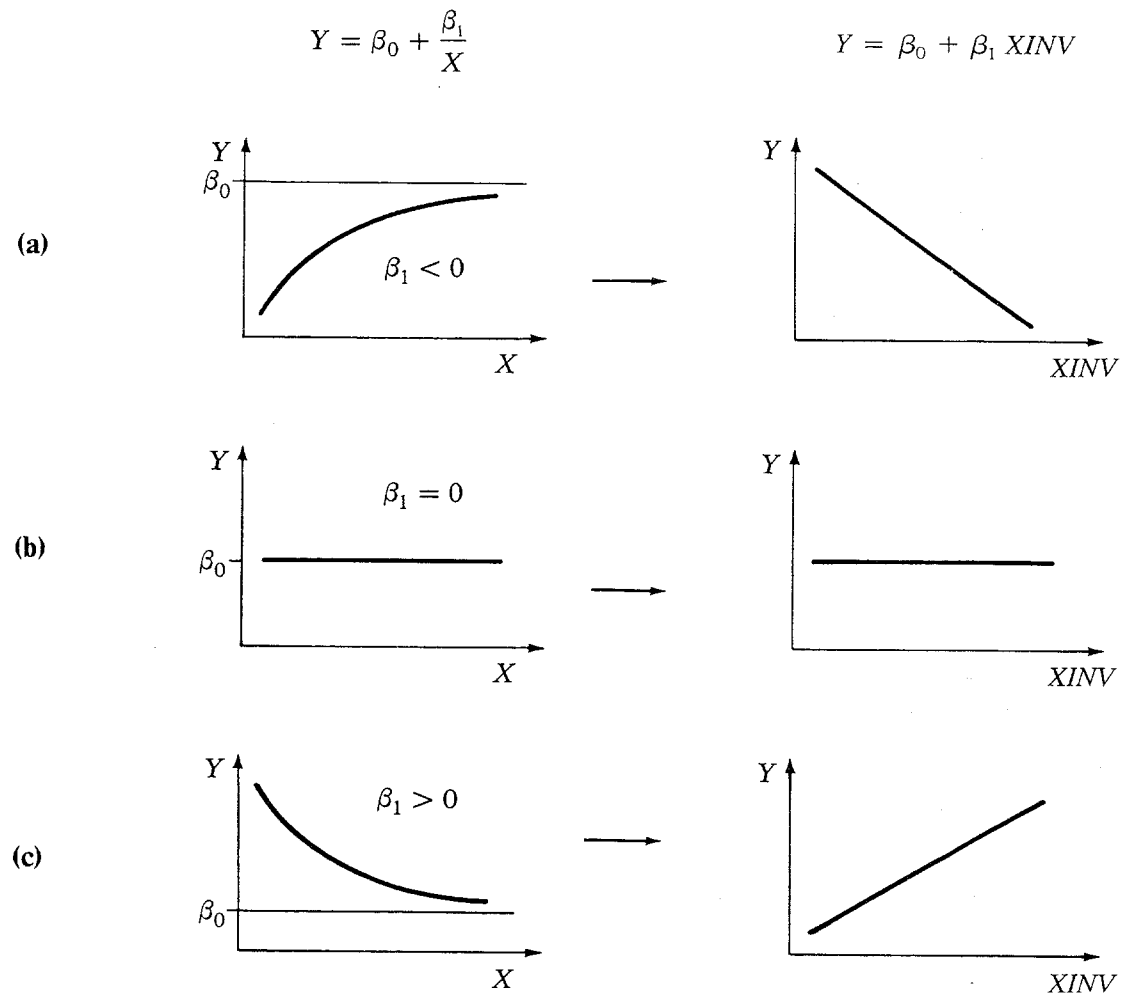

FIGURE 6.2 The geometry of the reciprocal relation depends on the sign of β_1. Except when $\beta_1 = 0$, Y is a nonlinear function of X; as X increases, Y increases ($\beta_1 < 0$) or decreases ($\beta_1 > 0$) and approaches the asymptotic limit β_0. Although Y is a nonlinear function of X, it is a linear function of the inverse of X—which is denoted by *XINV*.

the parameters β_0 and β_1 of the linear relation (between Y and *XINV*) are the same as the parameters of the reciprocal relation (between Y and X). If we add a disturbance term to (6.17), we have the specification of a regression model.

For example, before the combined high inflation and unemployment of the 1970s, the systematic part of the Phillips curve relation between the rate of inflation and the unemployment rate was considered to be nonlinear and resemble the left side of Figure 6.2c. Hence a reciprocal relation between inflation and unemployment was considered to be an appropriate form, and we use 15 annual observations (1956–1970) to estimate the model.

The regressand is simply *RINF1*. The regressor is a new variable, equal to the inverse of *UPCT*:

$$UINV_i = \frac{1}{UPCT_i} \tag{6.18}$$

The estimated regression is

$$\widehat{RINF1}_i = -1.984 + 22.234UINV_i \qquad (6.19)$$
$$R^2 = .549 \qquad SER = 0.956$$

Commonly, results such as these are reported simply with $1/UPCT$ in place of $UINV$, but we wish to stress that the actual regressor is a transformed variable.

The results of regressions involving transformed variables require care in interpretation. Since $\hat{\beta}_1^*$ is positive, the actual estimated regression resembles the right side of Figure 6.2c. However, the economic interpretation is best carried out in terms of the implied relation between predicted *RINF1* and *UPCT*, which resembles the left side of Figure 6.2c. Looking first at the estimated intercept, we see that as *UPCT* increases, *RINF1* decreases and approaches a lower limit of -1.984 percent. The sign of the estimated slope tells us the general shape (here, like the left side of Figure 6.2c), but the magnitude is difficult to interpret. To understand the meaning of this coefficient, it is useful to make predictions for the rate of inflation at various levels of *UPCT*. For example, if $UPCT = 6$ percent, the predicted *RINF1* is $-1.984 + (22.234)(1/6) = 1.72$ percent; if $UPCT = 3$ percent, the predicted *RINF1* is $-1.984 + (22.234)(1/3) = 5.43$ percent. The estimated regression is consistent with earlier findings on the Phillips curve.

The restriction of the data to observations from the period 1956–1970 reflects the fundamental econometric requirement that all the data used to estimate a model must have been generated from the same economic process. We believe that after 1970 expectations of continuing inflation began to get built into the behavior of the economy in a way that they had not been in the earlier period. In other words, the behavioral pattern changed. This means that although the model (6.19) may be quite appropriate for the earlier period, it cannot be used to make predictions after 1970. For this later period, a more complicated model is required.

Lagged Variables and First Differences

The regular simple regression model is specified so that Y_i is related to X_i. In a time-series context, this means that the values of Y and X are to be measured in the same time period. However, economic behavior is dynamic, and an effect may occur substantially later than its cause. For example, a firm's investment decision may be made at one point in time but the machines might not be produced and delivered until a year or more later. Similarly, people may budget their consumption expenditures on the basis of last year's income rather than its current level. In these cases the explanatory variable is said to determine the dependent variable with a ***lag***.

If there is a one-period lag between cause and effect, it is natural to formulate a simple regression model as

$$Y_i = \beta_0 + \beta_1 X_{i-1} + u_i \tag{6.20}$$

where X_{i-1} is the value of X one period before i. In other words, the current (ith period) value of Y depends on the one-period lagged value of X and the current disturbance. This formulation appears to complicate the estimation of the coefficients, because the paired Y_i, X_{i-1} values are no longer a row in a rectangular data matrix and the OLS estimators are no longer appropriately defined.

However, the creation of a new regressor straightens out these difficulties. Table 6.1 illustrates the procedure. The table shows some data on two variables Y and X. A new regressor is created in such a way that each of its values is equal to the previous period's value of X. Appropriately, this new regressor is called $XLAG$. In period i, the value $XLAG_i$ is equal to X_{i-1}, and

$$Y_i = \beta_0 + \beta_1 XLAG_i + u_i \tag{6.21}$$

has the same meaning as (6.20). In this specification the earlier difficulties disappear, and the coefficients can be estimated in the ordinary way. It should be noted that in the construction of $XLAG$, no value could be assigned to $XLAG_1$ because X_0 is not in the data set. Hence in estimating (6.21) we must ignore the first observation in the data, and use only 2 through n.

For example, using observations 2 through 25 of the time-series data set, an aggregate consumption function embodying the theory that DPI affects CON with a one-period lag is estimated as

$$\begin{aligned} \widehat{CON}_i &= 10.913 + 0.923 DPILAG_i \\ R^2 &= .993 \qquad SER = 14.953 \end{aligned} \tag{6.22}$$

The results differ slightly from the contemporaneous model (6.10). The standard error of regression, which measures the typical error of fit, is about 50 percent greater here, but the R^2 is only slightly lower. Although the original specification provides a better fit, we do not have a statistical basis yet for choosing between

TABLE 6.1 Construction of the Lag Regressor

i	Y	X	$XLAG$
1	405.4	446.2	—
2	413.8	455.5	446.2
3	418.0	460.7	455.5
⋮			
$i-1$	Y_{i-1}	X_{i-1}	$XLAG_{i-1}$
i	Y_i	X_i	$XLAG_i$
⋮			

the two. This lag specification is so common and easy to understand that reports of equations like (6.22) often are written with DPI_{i-1} rather than $DPILAG_i$ on the right-hand side, because no confusion is likely to occur.

Another common specification in time-series modeling involves letting the regressand or regressor, or both, be the ***first difference*** of a variable. The first difference is a constructed variable whose value in any period is equal to the value of the original variable in that period minus its value in the previous period. For example, starting with data on *GNP*, the new variable ΔGNP is defined by

$$\Delta GNP_i = GNP_i - GNP_{i-1} \tag{6.23}$$

In other words, the first difference for *GNP* is equal to *GNP* minus the lagged value of *GNP*, for each and every observation. (Note that if we have n observations on *GNP*, the values of ΔGNP are defined only for observations 2 through n.)

The accelerator theory of aggregate investment behavior is based on the idea that changes in the level of *GNP* are the main determinant of the level of investment spending by business. This leads to a simple regression model

$$INV_i = \beta_0 + \beta_1 \Delta GNP_i + u_i \tag{6.24}$$

where *INV* is real gross private domestic investment. Using observations 2 through 25 from the time-series data set in Chapter 2, the estimated model is

$$\begin{aligned} \widehat{INV}_i &= 136.6 + 0.691\ \Delta GNP_i \\ R^2 &= .175 \qquad SER = 40.4 \end{aligned} \tag{6.25}$$

The positive intercept gives the estimated amount of investment (136.6 billion dollars) that would occur if *GNP* were not growing (i.e., if $\Delta GNP = 0$); this might reflect investment to replace depreciated assets. The estimated slope is substantially smaller than predicted by simple accelerator theory, which suggests that (6.24) might not be an appropriate model of investment behavior.

6.4 Logarithmic Functional Forms*

As seen in Section 6.3, in some cases a transformation of a nonlinear relation leads to an equivalent linear relation involving newly created variables. The benefit of this is that a linear regression model can be used to estimate the parameters. The focus of our interest and interpretation, however, remains with the original relation.

A special class of nonlinear relations become linear when they are transformed with logarithms. The wide range of nonlinearities that can be captured

*This section is relatively difficult and can be skipped without loss of continuity.

and the associated ease of interpretation make these specifications very popular with applied econometricians. These relations and their transformations are explored in this section, and the appendix to this chapter contains the derivations of some of the interpretations given here.

The Log-Linear Specification

Suppose that we think the exact relation between Y and X is

$$Y = e^{\beta_0} X^{\beta_1} \tag{6.26}$$

where e^{β_0} stands for any positive constant. If we take the natural logarithm of both sides of the equation, we obtain

$$\ln Y = \beta_0 + \beta_1 \ln X \tag{6.27}$$

This is known as the ***log-linear*** relation between Y and X because it is linear in the logarithms of the original variables. Since (6.27) involves the logarithm of Y and X, the relation is applicable only if all the values of Y and X are positive: none can be zero, none can be negative.

The log-linear relation is particularly useful because of the variety of graphical shapes that it can represent. Figure 6.3 shows the geometry of (6.26) and its logarithmic transformation (6.27) when β_1 takes on different values. If β_1 is negative, the relation between Y and X is downward sloping and its slope becomes flatter as X increases. If β_1 is zero, then Y is just a constant. If β_1 is between 0 and 1, then the relation between Y and X extends out from the origin and slopes upward, but the slope becomes flatter as X increases. If β_1 is equal to 1, the relation between Y and X passes through the origin and is linear; Y is proportional to X. If β_1 is greater than 1, the relation between Y and X extends out from the origin and slopes upward, but the slope becomes steeper as X increases. In all cases, β_0 is a factor that affects all possible observations equiproportionally.

Perhaps the most attractive feature of this model is that β_1 can be directly interpreted as the ***elasticity*** of Y with respect to X. In economics, this elasticity is equal to the proportional change in Y divided by the proportional change in X resulting from a movement along the relation between Y and X. As shown in the appendix to this chapter, the specification underlying (6.26) and (6.27) is such that the elasticity is the same everywhere along the whole function when the considered changes in Y and X are small. That is, the point elasticity is constant, and it is exactly equal to β_1:

$$\beta_1 = \frac{dY/Y}{dX/X} \tag{6.28}$$

where dY and dX can be thought of as small changes (Δ's) in Y and X, respectively. Hence the log-linear relation is sometimes called the ***constant-elasticity*** relation.

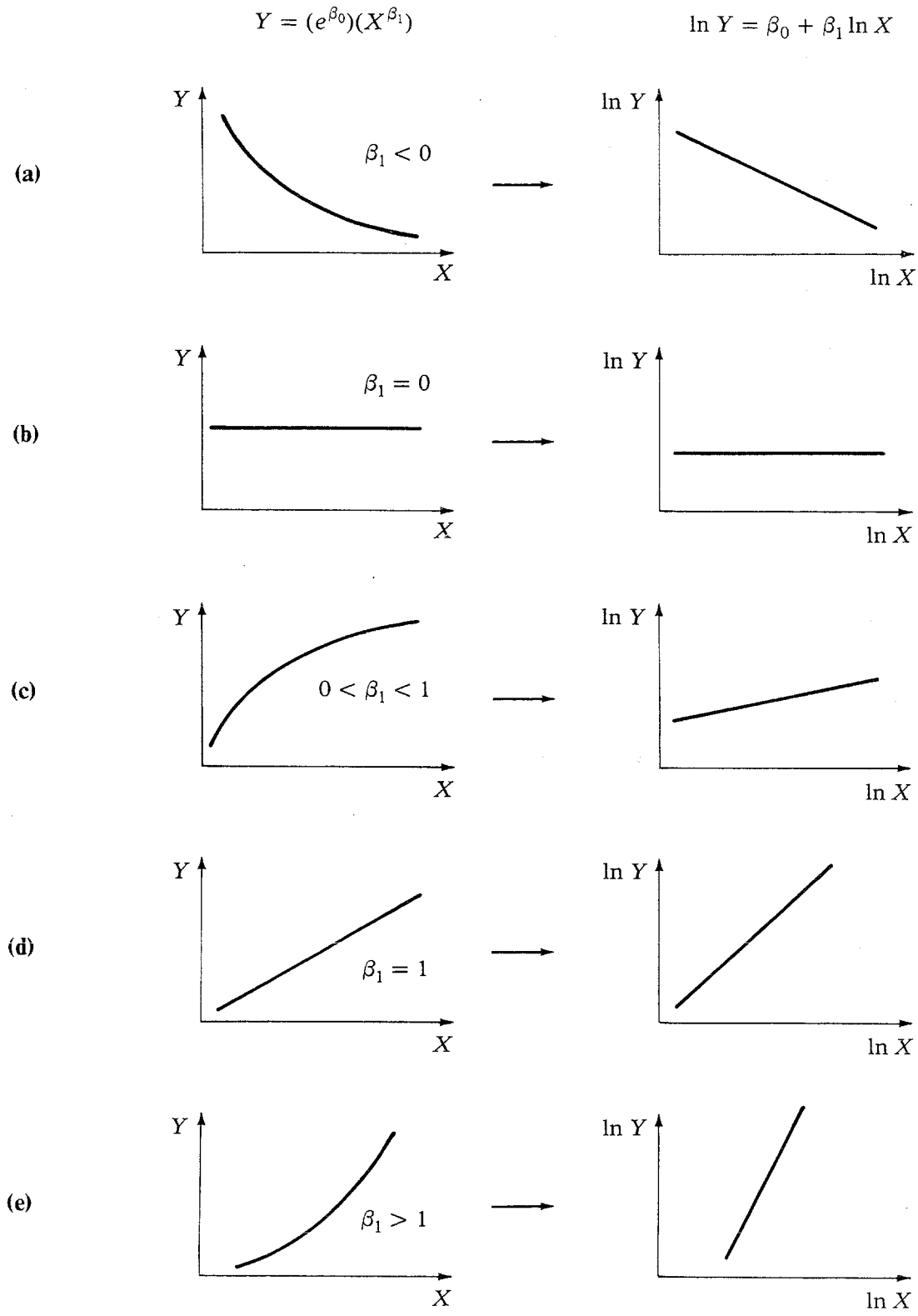

FIGURE 6.3 The geometry of the log-linear relation depends on the sign of β_1. When Y decreases as X increases [case (a), with $\beta_1 < 0$], it is concave upward. When Y increases with X (with $\beta_1 > 0$), the concavity may be upward or downward, depending on the magnitude of β_1. Although Y is a nonlinear function of X, $\ln Y$ is a linear function of $\ln X$; the slope of that line is the same β_1 as in the original formulation. The parameter β_1 is the elasticity of Y with respect to X.

When the elasticity of the relation is known, it provides the basis for calculating the effect on Y of changes in X. Let "pc of Y" be the proportionate change in Y, which might otherwise be denoted by $\Delta Y/Y$. The definition of elasticity implies that

$$\text{pc of } Y \approx (\beta_1)(\text{pc of } X) \tag{6.29}$$

For example, if the elasticity (β_1) equals 1.2, a 5 percent change in X will lead to a 6 percent change in Y. [The "pc" should be entered into (6.29) as a regular proportion—not in percentage points; that is, it should be entered like .05, not 5. However, it is conventional to verbalize a regular proportion like .05 as "5 percent." It might be noted that entering the "pc" in percentage points will not lead to an error in (6.29), but doing so in similar formulas will.]

So far we have focused on an exact theoretical relation between variables Y and X. It is easy to see that the log-linear relation can be taken as the basis of a simple regression model

$$\ln Y_i = \beta_0 + \beta_1 \ln X_i + u_i \tag{6.30}$$

where the regressand is ln Y and the regressor is ln X. This econometric model is appropriate if the process determining Y is such that the expected value of ln Y is a linear function of ln X.

For example, consider the aggregate demand for money in the United States. Good economics leads us to specify the model in real terms, so we construct the variable M to be the real quantity of money (see Section 2.3), based on variables in our data set:

$$M_i = \frac{M1_i}{PGNP_i}(100) \tag{6.31}$$

Economic theory suggests that it is reasonable to specify the constant-elasticity form

$$\ln M_i = \beta_0 + \beta_1 \ln GNP_i + u_i \tag{6.32}$$

as the regression model. It is important to realize that the regressand and regressor of the model are transformed variables: they are the logarithms of M and GNP, respectively. For notational convenience we let the names of the new variables be LNM and $LNGNP$.

Using the 25 observations in the time-series data set, the estimated regression is

$$\begin{aligned} \widehat{LNM}_i &= 3.948 + 0.215 LNGNP_i \\ R^2 &= .780 \qquad SER = 0.0305 \end{aligned} \tag{6.33}$$

Since $0 < \hat{\beta}_1^* < 1$, we see that the implied relation between M and GNP resembles the left side of Figure 6.3c. The estimated income elasticity of the demand for money is 0.215; in other words, if GNP increases by 1 percent, we

predict that the demand for money will increase by 0.215 percent. Using (6.29) we see that a 5 percent increase in *GNP* leads to a 1.075 percent increase in predicted *M*. However, when we start to consider large changes in *GNP* or *M*, the calculation based on the point elasticity holds only approximately.

Now, suppose that we wish to predict the demand for money when $GNP = 1000$. First, we calculate $\ln(1000) = 6.908$. Then we determine the predicted value of $\ln M$: $3.948 + (0.215)(6.908) = 5.433$. Finally, we take the antilog of this number, yielding a predicted demand for money of 228.8 billion dollars. (A refinement of this procedure is suggested in Section 16.2.)

The constant elasticity model is probably the second most useful specification for a simple regression, after the regular linear form. This is because economic theory often leads us to characterize relations in elasticity terms, and the model allows for a simple approach to estimating "the elasticity." It should be noted that the presumption that the elasticity is constant throughout the relation is a strong one—but so is the corresponding assumption that the slope is constant in a regular linear model.

The Semilog Specification

Suppose that we think the exact relation between Y and X is

$$Y = e^{\beta_0} e^{\beta_1 X} \tag{6.34}$$

If we take the natural logarithm of both sides of the equation, we obtain

$$\ln Y = \beta_0 + \beta_1 X \tag{6.35}$$

This is known as the ***semilog*** relation because only part of it is specified in logarithmic form. (Another variant, which we do not consider, specifies Y as a function of $\ln X$.)

The geometry of (6.34) and (6.35) is illustrated in Figure 6.4. X may take on positive or negative values, but Y must be positive if $\ln Y$ is to be defined. If β_1 is negative, Y decreases as X increases and its slope becomes flatter. If $\beta_1 = 0$, Y is constant. If β_1 is positive, Y increases as X increases and its slope becomes steeper.

Part of the usefulness of the semilog relation derives from the ease of interpretation of β_1. Thinking of dY and dX as small changes resulting from movement along the relation between Y and X, it is shown in the appendix to this chapter that

$$\beta_1 = \frac{dY/Y}{dX} \tag{6.36}$$

That is, β_1 can be interpreted as the proportional change in Y that results from a unit change in X. This implies that

$$\text{pc of } Y \approx \beta_1 \, \Delta X \tag{6.37}$$

when the changes in X and Y are small.

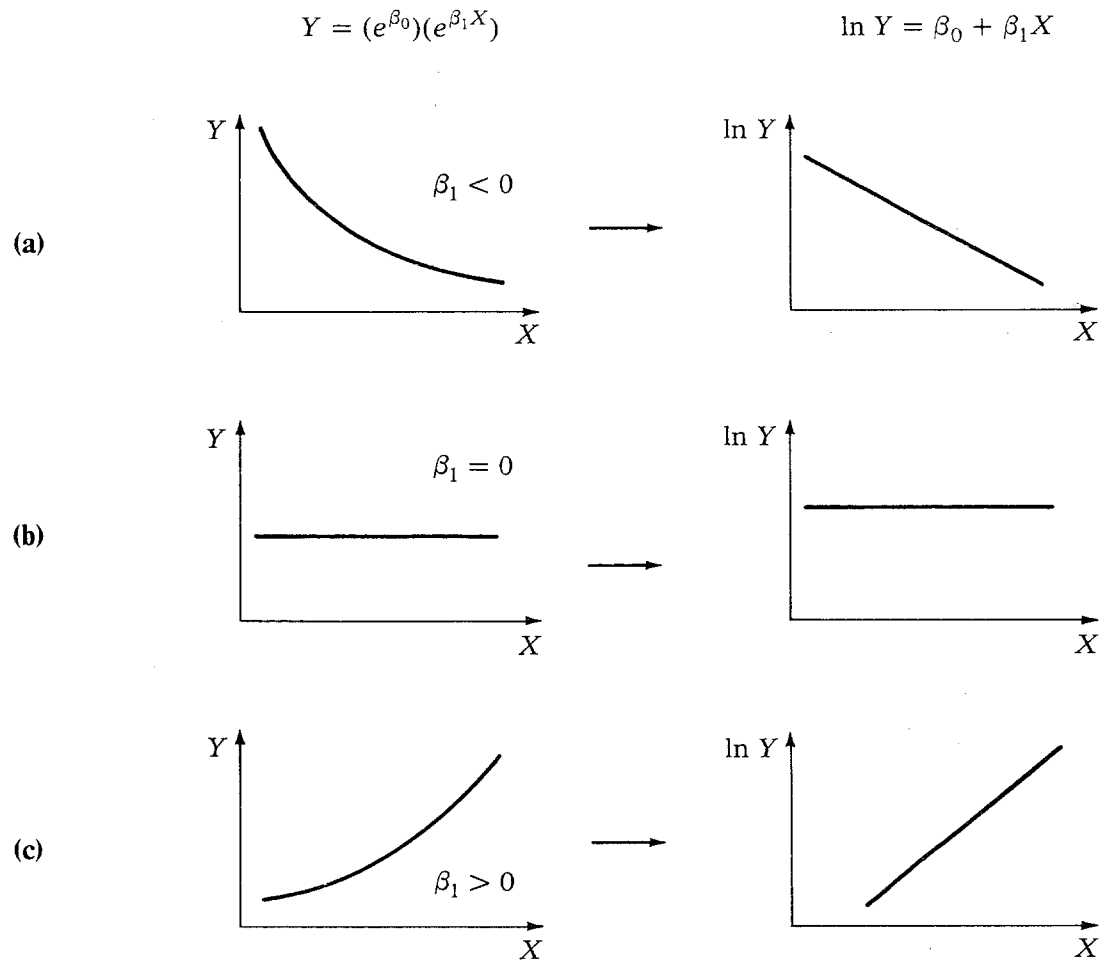

FIGURE 6.4 The geometry of the semilog relation depends on the sign of β_1. Except when $\beta_1 = 0$, Y is a nonlinear function of X. As X increases, Y decreases ($\beta_1 < 0$) or increases ($\beta_1 > 0$), and the relation is concave upward in both cases. Although Y is a nonlinear function of X, ln Y is a linear function of X; the slope of that line is the same β_1 as in the original formulation. The parameter β_1 can be interpreted as the proportional change in Y that results from a unit change in X.

Moving toward econometrics, it is easy to see that the exact relation (6.35) can be taken as the basis of a simple regression model

$$\ln Y_i = \beta_0 + \beta_1 X_i + u_i \tag{6.38}$$

where the regressand is ln Y and the regressor is simply X.

One application of the semilog specification is to the human capital theory of earnings determination. In one derivation, the theory states that the systematic relation between earnings and educational attainment is

$$\ln EARNS = \beta_0 + \beta_1 ED \tag{6.39}$$

Adding a disturbance term yields a semilog regression model. The estimate of this earnings function, based on the 100 observations in our cross-section data set, is

$$\widehat{LNEARNS}_i = 0.673 + 0.107ED_i \qquad (6.40)$$
$$R^2 = .405 \qquad SER = 0.446$$

Since $\hat{\beta}_1^* > 0$, we see that the implied relation between *EARNS* and *ED* resembles the left side of Figure 6.4c. On the basis of this estimation we predict that each additional year of schooling increases earnings approximately by the proportion 0.107, or 10.7 percent. In other words, the new level of earnings will be 1.107 times the original level. Using (6.37), a three-year increase in schooling would lead earnings to increase approximately by the proportion 0.321, or 32.1 percent.

It is interesting to compare this with the result of (6.7), which estimates that the impact of each additional year of schooling is to increase annual earnings by \$797. Which model is more appropriate for the earnings function? On the basis of general principles, we might surmise that the semilog model is appropriate when theory suggests that equal-sized increases in X lead to equal proportionate increases in Y for different observations, whereas the linear model is appropriate when theory suggests that equal-sized increases in X lead to equal absolute increases in Y for different observations. In practice, one way to compare (6.7) with (6.40) is to graph the residuals against the explanatory variable, *ED* (a computer can do this easily). If the residuals appear to be unrelated to *ED* in one case, but related to *ED* in a $\cup$ or $\cap$ shape in the other, then the case with the unrelated residuals may be judged better because it is in greater conformity with regression theory. Note, however, that there will be no linear relation between the residuals and *ED* because OLS always makes this correlation equal to zero.

Another common application of the semilog specification is to estimate the trend rate of growth or shrinkage in a time-series variable. Consider a simple model of growth

$$Y_i = Y_0(1 + r)^i \qquad (6.41)$$

where i indicates the number of years since period 0. Since our time variable (T) conforms to our observation numbering, we can write

$$Y_i = Y_0(1 + r)^{T_i} \qquad (6.42)$$

Taking logarithms yields

$$\ln Y_i = \ln Y_0 + T_i \ln (1 + r) \qquad (6.43)$$

Now, $\ln Y_0$ and $\ln (1 + r)$ are both constants, which we can rename β_0 and β_1, respectively. Adding a disturbance for econometric reality, (6.43) becomes

$$\ln Y_i = \beta_0 + \beta_1 T_i + u_i \qquad (6.44)$$

which is suitable for OLS estimation using the transformed variable $\ln Y$ as the regressand. Since $\beta_1 = \ln (1 + r) \approx r$, $\hat{\beta}_1$ is approximately the estimated annual rate of growth [see (2.12)].

For example, choosing *GNP* as a variable of interest, we find that

$$\widehat{LNGNP}_i = 6.456 + 0.0354T_i \qquad (6.45)$$
$$R^2 = .988 \qquad SER = 0.0297$$

for the 25 observations in the time-series data set. The annual rate of growth of *GNP* is estimated to be 3.54 percent based on the assumption behind model (6.44) that the systematic rate of growth was constant over this period. The disturbance term allows for random fluctuation of *GNP* from its trend each year.

Problems

Section 6.1

★ **6.1** If the expected value of Y for the ith observation is 35, how could it be that the actual value is 33?

6.2 Suppose that an appropriately estimated regression is $\hat{Y}_i = 3 + 4X_i$.
(a) Determine the change in $\hat{Y}$ associated with $\Delta X = 2$.
(b) Comparing two particular observations in the data, it turns out that $\Delta Y = 9$ while $\Delta X = 2$. What accounts for the difference between this ΔY and your answer to part (a)?

6.3 In regressions of saving on income reported in Section 5.5, the slope coefficient was 86.3 in one case and 0.0863 in another. In which case is income more important in explaining saving? Why?

Section 6.2

6.4 Based on the regression reported as Equation (6.7):
(a) Determine the predicted value of *EARNS* corresponding to $ED = 12$.
(b) Determine the change in predicted *EARNS* associated with a change from $ED = 12$ to $ED = 16$.
(c) Thinking of $\Delta ED = 1$, is the effect on earnings of the senior year in college the same as for the sophomore year?

★ **6.5** Suppose that Equation (6.6) is the correct specification of the relation between *EARNS* and *ED*. If the regression (6.7) explains 28.5 percent of the variation in earnings in the sample, what accounts for the rest of the variation?

★ **6.6** Suppose that in 1980 all earnings levels had been inflated by a factor of 100 percent compared with the levels of 1963. If the relation between *EARNS* and *ED* otherwise remained unchanged, what would be the impact of an additional year of schooling on predicted earnings in 1980?

6.7 Based on Equation (6.10), how much of an increase in *DPI* is required to increase predicted *CON* by 1 billion dollars?

6.8 Interpret the estimated intercept in Equation (6.10).

Section 6.3

6.9 Suppose that instead of T in Equation (6.14) we had used the actual year number, t, as the regressor (i.e., $t_1 = 1956$, $t_2 = 1957$, etc.). What would be the estimated regression of APC on t?

⋆ **6.10** From Table 2.3, find the actual values of CON and DPI for 1974. What is the actual value for the APC? What is the predicted value of the APC based on Equation (6.14)?

6.11 Assuming that DPI increases steadily over time, what does Equation (6.10) imply about what happens to the average propensity to consume over time?

6.12 Suppose that in normal times the yield on bonds increases with the time to maturity, but that it increases at a decreasing rate and that it never goes above some natural level. Explain how a regression model can best be used to estimate the relation between yield and time to maturity.

6.13 Based on the estimated Phillips curve (6.19), what level of unemployment would have been required to bring the predicted rate of inflation down to zero?

⋆ **6.14** Based on Equation (6.22), what is the predicted value of consumption for 1974? What is the predicted value based on Equation (6.10)?

6.15 Based on Equation (6.25), determine the predicted level of investment for 1976.

Section 6.4

6.16 With reference to Equation (6.26), what happens to the value of all possible observations on Y if β_0 is increased by a constant amount?

6.17 Based on the slope coefficient in Equation (6.33), what would be the proportional impact on predicted M of the actual increase in GNP that occurred between 1967 and 1968? What was the actual proportional change in M? (See Table 2.3.)

⋆ **6.18** Suppose that $\ln Y = 1.0 + 0.25 \ln X$. Compute the predicted values of Y for these four values of X: 100, 500, 1000, 1500.

6.19 Plot the four Y, X points determined in Problem 6.18. How does the shape of this graph compare with Figure 6.3?

⋆ **6.20** Suppose that X increases from 500 to 1000. Based on your calculations for Problem 6.18, what is the percentage increase in predicted Y? How does this compare with the estimated elasticity?

6.21 Based on Equation (6.40), what is the proportional change in predicted earnings that would result from gaining a college education rather than stopping after completing high school?

⋆ **6.22** Consider the specification of a demand curve: $Q = AP^b$. Interpret the meaning of b. Based on examination of Figure 6.3 and knowledge of

economics, is b likely to be positive or negative? How could you use simple regression to estimate the value of b?

6.23 Suppose that the labor force has been growing at a steady rate over time. Explain how a regression model can be used to estimate the rate of growth.

Appendix

⋆ **6.24** Based on the relation in Problem 6.18, follow the procedure in Equation (6.48) to determine the exact predicted percentage increase in Y associated with a 100 percent increase in X. Compare this result with that in Problem 6.20.

6.25 Based on Equation (6.40), what is the exact proportionate change for Problem 6.21? Compare this with the simpler calculation.

APPENDIX: The Coefficients in Logarithmic Models*

The slope coefficients in log-linear and semilog models have straightforward interpretations when ΔX and the accompanying ΔY are small. These are given in Section 6.4 and are derived here. For large changes these interpretations are poor approximations, and to determine the effect ΔY resulting from a given ΔX we must work through the mathematical specification of the model, as also shown here.

For the log-linear functional form, the point elasticity interpretation is derived using calculus:

$$\begin{aligned} \ln Y &= \beta_0 + \beta_1 \ln X \\ d(\ln Y) &= d\beta_0 + d(\beta_1 \ln X) \\ dY/Y &= 0 + \beta_1(dX/X) \\ \beta_1 &= \frac{dY/Y}{dX/X} = \frac{dY}{dX} \cdot \frac{X}{Y} \end{aligned} \tag{6.46}$$

The middle expression in the last line of (6.46) is the elasticity, and the derivation shows that this is equal to β_1, a constant. Roughly speaking, we may say that a 1 percent change in X leads to a β_1 percent change in Y.

For discrete changes in X and Y we consider the movement from p' to p'' along the relation between Y and X illustrated in Figure 6.5. Letting X_1 and Y_1 correspond to p', and X_2 and Y_2 correspond to p'', we find that

*This appendix is relatively difficult and can be skipped without loss of continuity.

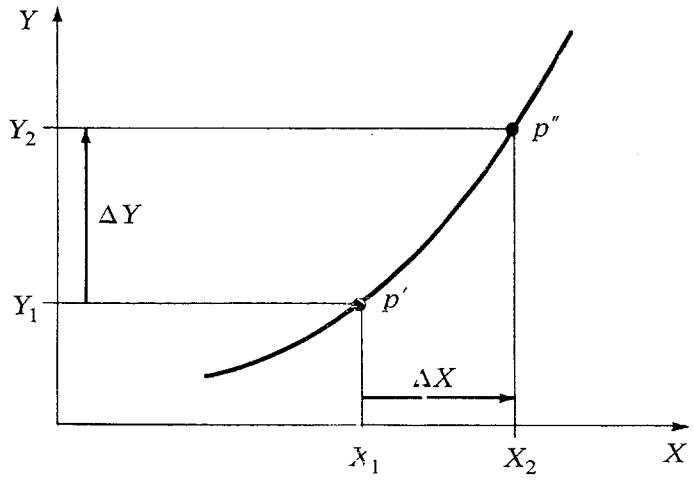

FIGURE 6.5 Moving from p' to p'' along a relation between Y and X is viewed as a change in Y (i.e., ΔY) resulting from a given change in X (i.e., ΔX). In the linear form (not illustrated here), Y and X are related in such a way that ΔY is a constant multiple of ΔX. In the log-linear form, Y and X are related in such a way that $\Delta Y/Y$ is a constant multiple of $\Delta X/X$, approximately; that is, the elasticity is constant, and therefore it does not vary with the level of X. In the semilog form, Y and X are related in such a way that $\Delta Y/Y$ is a constant multiple of ΔX, approximately; that is, the proportionate change in Y resulting from a given ΔX does not depend on the level of X.

$$\begin{aligned}
\ln Y_2 &= \beta_0 + \beta_1 \ln X_2 \\
\ln Y_1 &= \beta_0 + \beta_1 \ln X_1 \\
\ln Y_2 - \ln Y_1 &= \beta_0 - \beta_0 + \beta_1 \ln X_2 - \beta_1 \ln X_1 \\
\ln (Y_2/Y_1) &= \beta_1 \ln (X_2/X_1) \\
\beta_1 &= \frac{\ln (Y_2/Y_1)}{\ln (X_2/X_1)} = \frac{\ln (1 + \text{pc of } Y)}{\ln (1 + \text{pc of } X)} \\
&\approx \frac{\text{pc of } Y}{\text{pc of } X}
\end{aligned} \tag{6.47}$$

where "pc of Y" denotes the proportionate change in Y. The first two lines are statements of the relation holding at p'' and p', and the third line results from subtracting the second line from the first. The fourth line is obtained by applying the rules of logarithms. The fifth line isolates β_1 and shows what it is equal to in terms of proportionate changes of Y and X. When the proportionate changes are small, the final approximation is good; this yields the elasticity.

We can use this result to calculate the effect on Y resulting from any given change in X. For example, $\hat{\beta}_1^*$ in the demand for money equation (6.33) is 0.215. If *GNP* were to increase by 100 percent (pc = 1.00), as it does over the 1956–1980 sample period, the proportionate effect on M could be determined as follows. First (6.47) is rewritten as

$$\begin{aligned}
\ln (1 + \text{pc of } Y) &= \beta_1 \ln (1 + \text{pc of } X) = \beta_1 \ln (1 + 1) \\
&= (0.215)(0.693) = 0.149
\end{aligned} \tag{6.48}$$

Taking antilogs yields

$$1 + \text{pc of } Y = 1.161, \quad \text{so} \quad \text{pc of } Y = 0.161 \tag{6.49}$$

Hence the correct predicted proportionate change in Y is 16.1 percent, which is considerably smaller than the 21.5 percent change that would be predicted by simply applying the point elasticity.

For the semilog form,

$$\begin{aligned} \ln Y &= \beta_0 + \beta_1 X \\ d(\ln Y) &= d\beta_0 + d\beta_1 X = \beta_1\, dX \\ \beta_1 &= \frac{d(\ln Y)}{dX} = \frac{dY/Y}{dX} \end{aligned} \tag{6.50}$$

Roughly speaking, a unit change in X leads to a β_1 proportionate change in Y.

For discrete changes in X and Y we consider moving from p' to p'' in Figure 6.5 again:

$$\begin{aligned} \ln Y_2 &= \beta_0 + \beta_1 X_2 \\ \ln Y_1 &= \beta_0 + \beta_1 X_1 \\ \ln Y_2 - \ln Y_1 &= \beta_0 - \beta_0 + \beta_1 X_2 - \beta_1 X_1 \\ \ln (Y_2/Y_1) &= \beta_1\, \Delta X \\ \beta_1 &= \frac{\ln (1 + \text{pc of } Y)}{\Delta X} \approx \frac{\text{pc of } Y}{\Delta X} \end{aligned} \tag{6.51}$$

When the changes are small, the final approximation is good; this is our easy interpretation.

The effect on Y resulting from any given change in X may be derived as follows. Restating (6.51) leads to

$$\begin{aligned} \ln (1 + \text{pc of } Y) &= \beta_1\, \Delta X \\ 1 + \text{pc of } Y &= e^{\beta_1 \Delta X} \\ \text{pc of } Y &= e^{\beta_1 \Delta X} - 1 \end{aligned} \tag{6.52}$$

For example, in the earnings function (6.40) the estimated slope coefficient, 0.107, was interpreted as indicating that each additional year of schooling increases earnings by approximately the multiplicative factor 0.107, or 10.7 percent. The actual computed increase is $e^{0.107} - 1 = 0.113$, or 11.3 percent. A three-year increase in schooling increases earnings by the factor $e^{0.321} - 1 = 0.379$, or 37.9 percent. This is substantially larger than the 32.1 percent computed by the approximation (6.37).