7

Multiple Regression: Theory and Application

Most economic relations, and the processes they describe, involve more than one determinant of some particular dependent variable. For example, the earnings function examined in Chapter 6 presumes that education is the only identifiable variable that affects a person's earnings. Surely many more variables are directly relevant, and in this chapter we will go on to examine the roles of experience, demographic characteristics, and other variables.

This chapter covers both the theory and application of multiple regression, which involves more than one explanatory variable in a single regression equation. Most of the ideas regarding simple regression carry over, so there are relatively few new concepts to learn.

7.1 Two Explanatory Variables

As a first step in multiple regression, we consider an economic process in which the variable Y is determined by two given variables, X_1 and X_2. Much of what needs to be said about the theory and estimation of the corresponding model is a direct extension of the case of simple regression.

Our thinking is that n observations are subjected to this process, one at a time. For each observation separately, the values of X_1 and X_2 are fed in, and the value of Y is determined. The n values of X_1, X_2, and Y are all observable, and they can be collected in a data set. Graphically, the observations can be

plotted in a three-dimensional scatter diagram, in which the explanatory variables X_1 and X_2 are measured along the two axes defining the base and the dependent variable Y is measured vertically. Each observation is graphed as a single point, and together the observations resemble a cloud.

Two aspects of this process should be noted. First, X_1 and X_2 are the only identifiable and observable variables that affect Y. By explicit exclusion, other variables that might be measured for each observation are understood to play no direct role in the determination of Y. Second, the values of X_1 and X_2 are taken as given. These are determined outside the process under consideration, and we make no effort to understand why they take on whatever values they do.

We move toward statistical analysis by making a more concrete specification of the process by which Y is determined. We theorize that the value of variable Y for each observation is determined by the ***linear multiple regression model***

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i \tag{7.1}$$

Variable Y is the dependent variable (or regressand), and X_1 and X_2 are the explanatory variables (or regressors). The disturbance u is considered to be a random term that represents pure chance factors in the determination of Y. When a particular observation is referred to, a second subscript is used for the explanatory variables.

Based on (7.1), we can decompose each Y_i value into a systematic component

$$E[Y_i] = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} \tag{7.2}$$

and a random component, u_i. The systematic part of the relation between Y, X_1, and X_2 is given by an equation of the form

$$E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \tag{7.3}$$

which we call the true regression. This describes a plane graphed in the three-dimensional scatter diagram, but it is also known as a linear equation. With Y measured in the vertical direction, the observations (points) lie above or below the true regression, each at a vertical distance u_i.

Given this theoretical statement of how observable data are generated, our econometric task is to estimate the coefficients β_0, β_1, and β_2. As with simple regression, our technique is based on the method of ordinary least squares. When we fit a plane to the three-dimensional scatter of data points, it has an equation of the form

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 \tag{7.4}$$

This plane is called the estimated regression. It decomposes each actual Y_i value into its fitted (or predicted) value

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} \tag{7.5}$$

and its residual

$$e_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} \tag{7.6}$$

The OLS technique calculates $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ so as to make the sum of squared residuals as small as possible, and a derivation similar to that given in the appendix to Chapter 5 leads to estimators for the coefficients (i.e., formulas for calculating them). To simplify the presentation, we adopt the notation here that a lowercase letter stands for the deviation of a variable from its mean. Thus

$$y = Y - \bar{Y}, \quad x_1 = X_1 - \bar{X}_1, \quad \text{and} \quad x_2 = X_2 - \bar{X}_2 \tag{7.7}$$

The estimators are

$$\hat{\beta}_2 = \frac{\left(\sum x_2 y\right)\left(\sum x_1^2\right) - \left(\sum x_1 y\right)\left(\sum x_1 x_2\right)}{\left(\sum x_1^2\right)\left(\sum x_2^2\right) - \left(\sum x_1 x_2\right)^2} \tag{7.8}$$

$$\hat{\beta}_1 = \frac{\left(\sum x_1 y\right)\left(\sum x_2^2\right) - \left(\sum x_2 y\right)\left(\sum x_1 x_2\right)}{\left(\sum x_1^2\right)\left(\sum x_2^2\right) - \left(\sum x_1 x_2\right)^2} \tag{7.9}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2 \tag{7.10}$$

Notice that the estimators for all three coefficients involve all the values for all the variables. For example, $\hat{\beta}_2$ depends on the values of X_1 as well as those of Y and X_2. Hence these estimators are not equivalent to the coefficient estimators from the simple regression model applied twice. In other words, the multiple regression coefficients cannot be obtained by estimating two simple regressions, one of Y on X_1 and another of Y on X_2. [The exception to this is the special case when the correlation (and covariance) between X_1 and X_2 is zero; this would imply that $\Sigma\, x_1 x_2 = 0$.]

Our interpretations of the coefficients of the estimated regression differ in an important way from the case of simple regression. First, if we compare one possible observation with another, they may differ with regard to the values of X_1 and X_2 and also with regard to the predicted value $\hat{Y}$. These differences are linked by

$$\Delta\hat{Y} = \hat{\beta}_1\, \Delta X_1 + \hat{\beta}_2\, \Delta X_2 \tag{7.11}$$

This is derived by subtracting an equation like (7.5) specified for one point from the corresponding equation specified for another. This equation determines how changes in the values of the explanatory variables affect the predicted value of dependent variable. Graphically, this equation compares two points on the es-

timated regression plane and shows how differences in the three dimensions are related.

Now, if only one explanatory variable changes in value while the other remains the same, then

$$\Delta\hat{Y} = \hat{\beta}_1 \Delta X_1 \qquad \text{when} \quad \Delta X_2 = 0 \tag{7.12}$$

and

$$\Delta\hat{Y} = \hat{\beta}_2 \Delta X_2 \qquad \text{when} \quad \Delta X_1 = 0 \tag{7.13}$$

These provide the basis for interpreting the slope coefficients: $\hat{\beta}_1$ gives the impact on the predicted value of Y of a unit increase in X_1, holding constant the value of X_2. Note that this phrasing corresponds closely to the economic concept of ***ceteris paribus***. Similarly, $\hat{\beta}_2$ gives the impact on $\hat{Y}$ of a unit increase in X_2, holding X_1 constant. These interpretations will be illustrated and enhanced in later examples and discussion.

As with simple regression, the *SER* and R^2 serve as measures of goodness of fit, and their interpretations are the same. The general definitions of these measures are given in the next section.

The Earnings Function

As an example, we extend our analysis of the earnings function explored in Chapter 6. Suppose that labor market theory suggests that in addition to formal education, experience working in the labor force has a direct effect on workers' earnings. This may be because experience represents on-the-job training and thereby increases a person's productivity and wage, or it may be because of some other considerations. If this theory is correct and if the relation is linear, it is appropriate to formulate the multiple regression model

$$EARNS_i = \beta_0 + \beta_1 ED_i + \beta_2 EXP_i + u_i \tag{7.14}$$

Turning to estimation, the 100 observations in the cross-section data set in Chapter 2 are used to obtain the estimated regression

$$\begin{aligned} \widehat{EARNS}_i &= -6.179 + 0.978 ED_i + 0.124 EXP_i \\ R^2 &= .315 \qquad SER = 4.288 \end{aligned} \tag{7.15}$$

Holding constant the level of education, each year of experience is estimated to increase expected earnings by \$124. The *ceteris paribus* qualification "holding constant the level of education" is an interpretation based on (7.13); it does not mean that we or the computer hold some values specially fixed during estimation or make comparisons only among observations with common values for *ED*. Considering the 30-year age difference between the youngest and the oldest men in the sample, the estimated coefficient implies a (0.124)(30) = 3.720 thousand dollar annual earnings difference. Thus experience might be judged to be a moderately important factor in the determination of earnings.

The estimated impact of education on earnings is increased substantially from the finding in the simple regression (0.797 to 0.978). Of course, the true impact of education has not changed—we did nothing capable of altering that—but the change in specification has affected our estimated impact. We return to this later.

We can use the estimated model to predict earnings for given values of *ED* and *EXP* in the usual way. For example, the predicted earnings (in 1963) for a college graduate with five years of experience is

$$\widehat{EARNS}_i = -6.179 + (0.978)(16) + (0.124)(5) = 10.089 \quad (7.16)$$

thousand dollars. In addition, the method of constructing *EXP* provides auxiliary information that permits us to answer a question like this: taking into account that for a specific individual an increase in *ED* of one year means a decrease in *EXP* of one year, what is the total economic effect on earnings of going to school for one more year? That is, we are seeking the joint impact of $\Delta ED = 1$ and $\Delta EXP = -1$. Based on (7.11),

$$\widehat{\Delta EARNS} = (0.978)(1) + (0.124)(-1) = 0.854 \quad (7.17)$$

thousand dollars per year. Note that the estimated intercept plays no role in a calculation like this.

The R^2 in the multiple regression is .315, which is somewhat higher than the .285 in the simple regression. Does the fact that R^2 increases by only .030 mean that the impact of experience is marginal, compared with that of education? No, not necessarily. *ED* could seem to get most of the credit simply because we considered it first. The R^2 of a simple regression of *EARNS* on *EXP* in these data could be fairly high, although in fact it is not in this case.

7.2 The General Case

The ***general linear multiple regression model*** is a particular specification of some economic process in which the values of a dependent variable (or regressand) are determined by several explanatory variables (or regressors). In general we may say that Y depends on k explanatory variables:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i \quad (7.18)$$

where the names of the variables are $X_1, X_2, \ldots, X_k$. A typical variable is denoted by X_j, and a typical coefficient by β_j. When a particular observation is referred to, a second subscript is used, so the ith observation on the jth variable is X_{ji}.

As with simple regression, our notion is that this model accurately reflects the way some process works. Indeed, this notion is more acceptable with multiple regression because this technique allows us to take into account all the important variables that help determine the value of the dependent variable.

Since most economic processes involve multiple causes of a single effect, this feature is especially important.

If it happens that $k = 1$, this model reduces to that of simple regression

$$Y_i = \beta_0 + \beta_1 X_{1i} + u_i \tag{7.19}$$

in which X_1 is the same as X in (5.1). We have treated the case of simple regression separately because it readily permits graphical interpretation and because the algebra of the derivations is relatively simple. If $k = 2$, the model reduces to the case with two explanatory variables, treated in the preceding section.

The ordinary least squares (OLS) technique for estimating the coefficients $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ is an extension of that for the simpler cases, and the estimates of these parameters are denoted by $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$. For any observation the true value of Y_i can be decomposed into the fitted value and the residual:

$$Y_i = \hat{Y}_i + e_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \cdots + \hat{\beta}_k X_{ki} + e_i \tag{7.20}$$

Recall that with simple regression ($k = 1$) the observations can be plotted in a two-dimensional graph, and the estimated regression is the line that provides the best possible fit to the scattered points. When $k = 2$, the observations can be plotted in a three-dimensional graph, and the estimated regression is the plane that provides the best possible fit to the scattered points. By extension, in the general case the observations can be thought of as plotted in a $(k + 1)$-dimensional graph, and the estimated regression is a hyperplane fit through the scattered points.

The OLS technique calculates the $\hat{\beta}_j$ so as to make the sum of squared residuals as small as possible:

$$\text{OLS criterion:} \quad \text{minimize} \quad SSR = \sum_i e_i^2 \tag{7.21}$$

Suffice it to say that we end up with a set of estimators analogous to those for the simpler cases. All the data are used together to solve simultaneously for the $k + 1$ coefficients; it is not the case that we just estimate k simple regressions. A computer can carry out the necessary calculations, and we rely on this facility.

The discussion in Section 5.3 regarding the comparison of the estimated regression coefficients with the corresponding true regression coefficients applies fully to multiple regression. Although we might wish that each estimated coefficient were equal to the corresponding true coefficient, so that $\hat{\beta}_j^* = \beta_j$, this is unlikely ever to occur. (As before, the asterisk indicates the computed values in a set of data.) The difference between $\hat{\beta}_j^*$ and β_j arises from the particular pattern of values taken on by the disturbances. (Just how the value of $\hat{\beta}_j$ might be related to β_j is the subject of sampling theory, which is discussed in Chapter 11.) Overall, we recognize that the estimated regression will be different from the true regression, but we hope that the differences are not too great.

To interpret and apply the estimated regression model we need to see how changes in the regressors affect the predicted value of the regressand. The estimated regression is given by

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_k X_k \tag{7.22}$$

For any given set of changes in the explanatory variables,

$$\Delta\hat{Y} = \hat{\beta}_1 \Delta X_1 + \hat{\beta}_2 \Delta X_2 + \cdots + \hat{\beta}_k \Delta X_k \tag{7.23}$$

provides the method for calculating the effect on the predicted value of Y of specified changes in the X_j's.

If only one explanatory variable changes in value while all the others remain the same, then

$$\Delta\hat{Y} = \hat{\beta}_j \Delta X_j \quad \text{, holding all other regressors constant} \tag{7.24}$$

This provides the basis for interpreting the value of any single coefficient: $\hat{\beta}_j$ gives the impact on the predicted value of Y of a unit increase in X_j, holding constant the values of all the other variables. The latter qualification is important, and it corresponds closely to the economic concept of ***ceteris paribus.*** This concept does not require or imply that there are no relations among the explanatory variables, but it ignores them in assessing the effect of each variable.

If a change in X_j does cause a change in other explanatory variables, we recognize that the total economic effect on $\hat{Y}$ includes the *ceteris paribus* effect $\hat{\beta}_j$ plus the effects of the consequent changes in the other variables, through (7.23). Additional information about how the explanatory variables affect each other would be needed to make this calculation, but usually it is not available. Sometimes such information is developed in multiequation models, which are beyond the scope of our present concern. In working with single-equation models, the only effects revealed are the *ceteris paribus* effects given by the individual regression coefficients, and these are what we are interested in.

Some of the properties of the estimated regression, which is usually called a line even though it is not one, are the same as for simple regression. First, it turns out that $\Sigma\, e_i = 0$; that is, the positive and negative residuals cancel out in summation, so the average error is zero. Second, the point $\overline{Y}, \overline{X}_1, \overline{X}_2, \ldots, \overline{X}_k$ lies on the fitted line; that is, the regression goes through the point of means. Third, the correlation between the residuals and any explanatory variable is zero.

The standard error of regression is given by

$$SER = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n - k - 1}} \tag{7.25}$$

and it measures the typical error of fit. Notice that the denominator is equal to n minus the number of coefficients (including $\hat{\beta}_0$) that are estimated; this count is the number of degrees of freedom for the estimation.

How well the estimated regression fits the data is also measured by the coefficient of determination, R^2, which is calculated as in the case of simple regression

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2} \tag{7.26}$$

Since each e_i is the part of Y_i that *is not* explained by the regression, R^2 is again interpreted as the proportion of the variation in Y that *is* explained by the regression (i.e., that is explained by the variation in the X_j's). A perfect fit, in which each $e_i = 0$, yields $R^2 = 1$, and if the fit is not very good at all, R^2 is close to zero.

In practical econometrics we often estimate more than one regression involving the same dependent variable. For example, so far we have estimated two earnings functions: one involving *ED* as the only explanatory variable, and a second involving both *ED* and *EXP* as explanatory variables. In comparing regressions like these it is useful to have a statistic or an indicator that tells us which regression fits better. Our natural candidates are the *SER* and R^2.

The R^2 is a more popular measure of fit for any single regression. However, for comparing regressions like our earnings functions, it always gives the same answer: the regression with additional variables included fits better. This is because the addition of an explanatory variable to an original regression model cannot raise the sum of squared residuals, *SSR*. (Since OLS is acting to minimize this sum, it need not allow an additional specified variable to increase the *SSR*; it could effectively ignore the new variable rather than let it worsen the *SSR*.) This sum, which is Σe_i^2, appears in the numerator of the ratio in (7.26). Thus for a given set of data on a dependent variable, Y, the addition of an explanatory variable to a regression model cannot decrease R^2, and in practice it always increases it at least a bit.

However, the increase in R^2 is obtained at a statistical "cost": the inclusion of another variable. An indicator of whether the new equation "really" fits better should assess whether the decrease in *SSR* achieved by including a new regressor is substantial enough to outweigh the cost of doing so. A statistic that does this is known as the ***adjusted*** or ***corrected*** R^2, which is denoted by $\bar{R}^2$ and defined as

$$\bar{R}^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2/(n - k - 1)}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2/(n - 1)} \tag{7.27}$$

The symbol $\bar{R}^2$ is conventionally read as "R bar squared." (The overbar notation does not signify a mean here.)

To examine what happens to $\bar{R}^2$ as another variable is added to a regression, we need only look at the numerator of the ratio in (7.27) because the denominator stays fixed. The numerator itself is a ratio. If adding a variable causes Σe_i^2 to decrease proportionately more than $n - k - 1$ decreases, R^2 will increase; if adding the variable causes Σe_i^2 to decrease only slightly, by proportionately less than the decrease in $n - k - 1$, $\bar{R}^2$ will decrease. Note that the expression $n - k - 1$ is the number of degrees of freedom, and its decrease is the link to the "statistical cost" of adding another variable.

For computational purposes, it is possible to determine $\bar{R}^2$ from R^2 and readily available parameters:

$$\bar{R}^2 = R^2 - \frac{k}{n - k - 1}(1 - R^2) \tag{7.28}$$

From this we see that $\bar{R}^2$ is less than R^2, except if $R^2 = 1$. Unfortunately, $\bar{R}^2$ does not have as straightforward an interpretation as R^2 does, and sometimes it can be negative.

Whether $\bar{R}^2$ would be higher or lower in comparable regressions can be determined by examining the *SER*s. To see this, note that the numerator of the ratio in (7.27) is the square of *SER*. Thus a decrease in the *SER* occurs whenever $\bar{R}^2$ increases, and an increase in the *SER* occurs whenever $\bar{R}^2$ decreases. In other words, the *SER* functions exactly the same as $\bar{R}^2$ as an indicator of whether the addition of an explanatory variable "really" improves the fit. For this reason, we make little use of the $\bar{R}^2$.

An implication of this is that our two basic measures of fit, *SER* and R^2, can sometimes give conflicting signals. Whenever a variable is added to the specification of a regression, R^2 will increase and our basic interpretation is that the "overall fit" is improved. However, when the improvement in the fit is relatively small, *SER* will increase and our interpretation is that the typical error of fit got worse.

The Consumption Function

The original Keynesian idea that aggregate consumption is determined primarily by income can be expanded to include other potentially important explanatory variables. We might think that the rate of interest available on savings and the rate of inflation in consumer prices are important. Economic theory does not specify strictly whether these effects would be positive or negative, so we approach the data with an exploratory frame of mind.

Using the 25 annual observations from our time-series data set, we estimate

$$\widehat{CON}_t = -2.370 + 0.910DPI_t + 0.500RAAA_t - 0.562RINF2_t \tag{7.29}$$

$$R^2 = .997 \qquad SER = 9.309$$

The estimated marginal propensty to consume, which gives the impact on *CON* of a unit increase in *DPI* while holding *RAAA* and *RINF2* constant, is 0.910. We see also that a one-percentage-point increase in the long-term interest rate is estimated to increase consumption (i.e., decrease saving) by 0.500 billion dollars, and a one-point increase in the rate of inflation decreases consumption by 0.562 billion dollars.

In comparison with the simple regression of *CON* on *DPI* (6.10), we see that the estimated marginal propensity to consume is changed only slightly. The R^2 is slightly higher (in the fifth decimal position, not shown) and also the *SER* is higher here than in (6.10). As noted above, this situation leads to a lower $\overline{R}^2$ in the present model.

We see that the addition of *RAAA* and *RINF2* to the simple form has not led to a model with appreciably more explanatory power. This, in itself, does not mean that the new variables do not belong in the model. If *DPI* were treated as a "new" variable, and compared with a simple regression of *CON* or *RAAA* or *RINF2*, then it might appear to have little additional explanatory power. (We see from Table 3.4 that the correlation between *CON* and *RAAA* is .952.) In Chapter 12 we present formal tests that also bear on this question.

7.3 Dummy Variables

In all the regression models considered so far, every variable has been a cardinal measure of some economic characteristic. For example, *CON* measures aggregate consumption in billions of constant dollars and *ED* measures the years of schooling completed by individual persons. These variables are included in a regression model in a natural way, so that changes in the numerical value of an explanatory variable have consistent numerical effects on the dependent variable.

Another type of data variable introduced in Chapter 2 carries information that is essentially categorical, such as a person's race, sex, or region of residence. This information can be used to classify or categorize observations or to separate them in some way, but the characteristic cannot be measured in any meaningful way. For example, the variable *REG* in the cross-section data set is equal to 1 if the worker lives in the Northeast, 2 if he lives in the North Central region, and so on. The variable *REG* indicates where the worker lives, but the values 1, 2, 3, and 4 do not result from measuring or counting anything. Hence the information contained in *REG* cannot be included directly into a regression model.

However, when the numerical outcome of an economic process depends in part on some categorical characteristic of the observation, this information must be brought into the regression specification somehow in order for the model to describe the process correctly. The technique for doing this involves constructing new regressors known as ***dummy variables*** and treating them exactly like other regressors in the multiple regression framework.

To start with a simple case, suppose that we are focusing on the relation between Y and X and that some such relation occurs both for men and women. One possibility is that the process determining Y is quite different for men and women. In this case, where there are essentially two separate processes occurring, we would separate the data according to sex and carry out separate statistical analyses. That is, we would have two separate models, one estimated with data for men and the other estimated with data for women.

Another possibility, which leads to the use of dummy variables, is that we think the process is such that the effect on Y of a change in X is the same for both sexes, but that there may be a systematic difference between men and women in the levels of Y associated with each particular value of X. These ideas are represented in Figure 7.1, which shows that the expected value of Y is a separate linear function of X for each sex, with equal slopes but different intercepts. Let the common slope value be β_1 and let the intercept for men be β_0. The intercept for women could be given a separate symbol, but instead we let β_2 be the difference between the women's and men's intercepts, so that the intercept for women is $\beta_0 + \beta_2$. Since the two functions have the same slope, β_2 also is the difference in $E[Y]$ between women and men having any particular value of X. As drawn in the figure, β_2 is positive so that the relation for women lies above that for men. In our general thinking the sign of β_2 is not specified, so that the relation for women may lie above or below that for men.

These three parameters can be estimated by constructing a dummy variable, S, that is equal to 0 for every observation that is a man and is equal to 1 for every observation that is a woman. We can formulate our specification so far as

$$E[Y] = \beta_0 + \beta_1 X + \beta_2 S \tag{7.30}$$

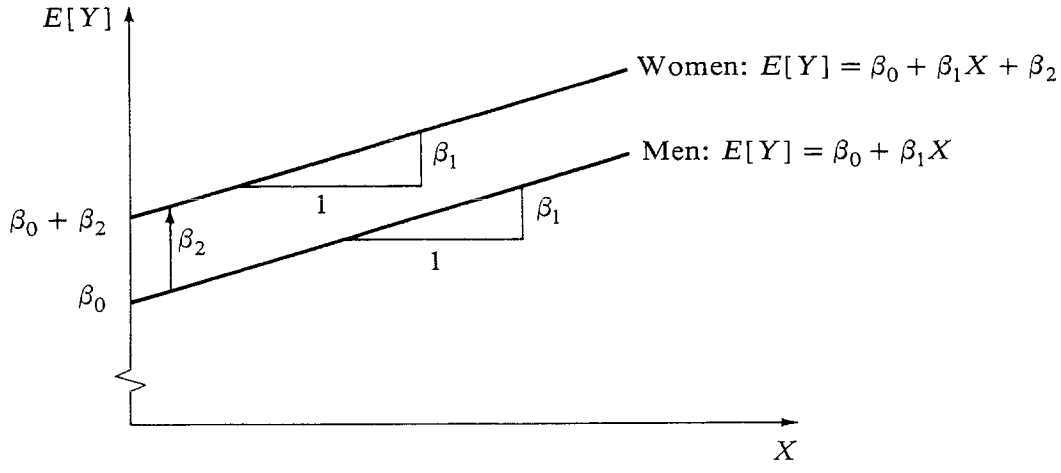

FIGURE 7.1 This illustrates an economic process in which the impact on Y of a change in X is the same for men and women, but in which men and women having the same value for X have different expected values for Y. Assuming linearity, the parameterization in the diagram leads naturally to the dummy variable formulation of a regression model.

Regardless of the value of S, the slope in the relation between $E[Y]$ and X is β_1. For men, with $S = 0$, the right-hand side of (7.30) equals $\beta_0 + \beta_1 X$, so the intercept is β_0. For women, with $S = 1$, the right-hand side of (7.30) can be rearranged as $(\beta_0 + \beta_2) + \beta_1 X$, so the intercept is $\beta_0 + \beta_2$. Adding a disturbance for econometric reality, we have a multiple regression model

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 S_i + u_i \tag{7.31}$$

which can be estimated by OLS.

The essential property of a dummy variable is that it identifies each of the observations as being in one of two groups. For all the observations in the first group the variable is set equal to zero, and for all the observations in the second it is set to 1. When the dummy variable is included as a regressor in a multiple regression, its coefficient represents the difference in the intercept between the second group and the first. Therefore, it also measures the impact on the expected value of the regressand of an observation's being in the second group rather than the first, holding all the other regressors constant. Sometimes the dummy is called a shift variable, because it simply causes a shift in the relation between $E[Y]$ and the other regressors; it does not otherwise alter that relation.

When the regression model is specified, the group given a dummy value of zero is called the ***excluded group,*** and the group given a dummy value of 1 is called the ***included group.*** With a single dummy variable in the regression model, the intercept for the excluded group is simply β_0 and the intercept for the included group is $\beta_0 + \beta_2$. That is, the intercept for the included group is equal to the intercept for the excluded group plus the coefficient on the dummy variable. Given two groups, either may be taken as the included one; this choice will affect the values of particular coefficients but not the overall interpretation of the regression.

For example, suppose theory suggests that earnings depends linearly on educational attainment but that there is a shiftlike difference between races. Our cross-section data set includes the variable *RACE,* which was coded in the interview as 1 for whites and 2 for blacks; other races were ignored in the data selection. This variable is unsatisfactory for our purposes. We construct instead a new regressor, *DRACE,* which is a dummy variable taking the value 0 for whites and 1 for blacks. Algebraically,

$$DRACE_i = RACE_i - 1 \tag{7.32}$$

The earnings function theory leads to a multiple regression, which is estimated as

$$\begin{aligned} \widehat{EARNS}_i &= -0.778 + 0.762ED_i - 1.926DRACE_i \\ R^2 &= .293 \qquad SER = 4.356 \end{aligned} \tag{7.33}$$

The coefficient $\hat{\beta}_1^* = 0.762$ estimates that the impact on earnings of an additional year of schooling is \$762 for both blacks and whites. The estimated impact

TABLE 7.1 First 10 Observations For Estimating Equation (7.33)

Obs.	*EARNS*	*ED*	*DRACE*
1	1.920	2	1
2	12.403	9	0
3	5.926	17	0
4	7.000	9	0
5	6.990	12	0
6	6.500	13	0
7	26.000	17	0
8	15.000	16	0
9	5.699	9	1
10	8.820	16	0

of race is that the earnings of blacks are \$1926 less than those of whites, holding constant the level of education.

The first 10 observations used for estimating (7.33) are shown in Table 7.1. The variables *EARNS* and *ED* are taken directly from Table 2.2. The variable *DRACE* is constructed according to (7.32). Like any dummy variable, *DRACE* takes on only the values 0 and 1.

In time-series regressions it may be that for some periods the relation between a dependent variable and a set of explanatory variables is shifted by a constant amount. For example, during war years the consumption function might shift down at all levels of income, because of the decrease in production of consumer goods and the special incentives given to saving. In this case, a model of the form

$$CON_i = \beta_0 + \beta_1 DPI_i + \beta_2 WAR_i + u_i \tag{7.34}$$

would be appropriate. The dummy variable *WAR* takes on the value 1 for wartime observations and zero for others; the coefficient β_2 is the shift in the consumption function, and we expect that $\hat{\beta}_2$ will be negative.

Another example is provided by the Phillips curve, which represents the trade-off between inflation and unemployment. It is conjectured that the relation shifted up in the middle of the 1960s because of the Viet Nam War and structural changes in the economy. To examine this conjecture, we add a dummy variable to the model earlier specified (6.19):

$$RINF1_i = \beta_0 + \beta_1 \left(\frac{1}{UPCT_i}\right) + \beta_2 D_i + u_i \tag{7.35}$$

where $D_i = 0$ for the observations 1956–1964 and $D_i = 1$ for 1965–1970. With 15 observations in total, we find that

$$\widehat{RINF1}_i = -0.803 + 15.030 \left(\frac{1}{UPCT_i}\right) + 0.894 D_i \tag{7.36}$$
$$R^2 = .600 \qquad SER = 0.936$$

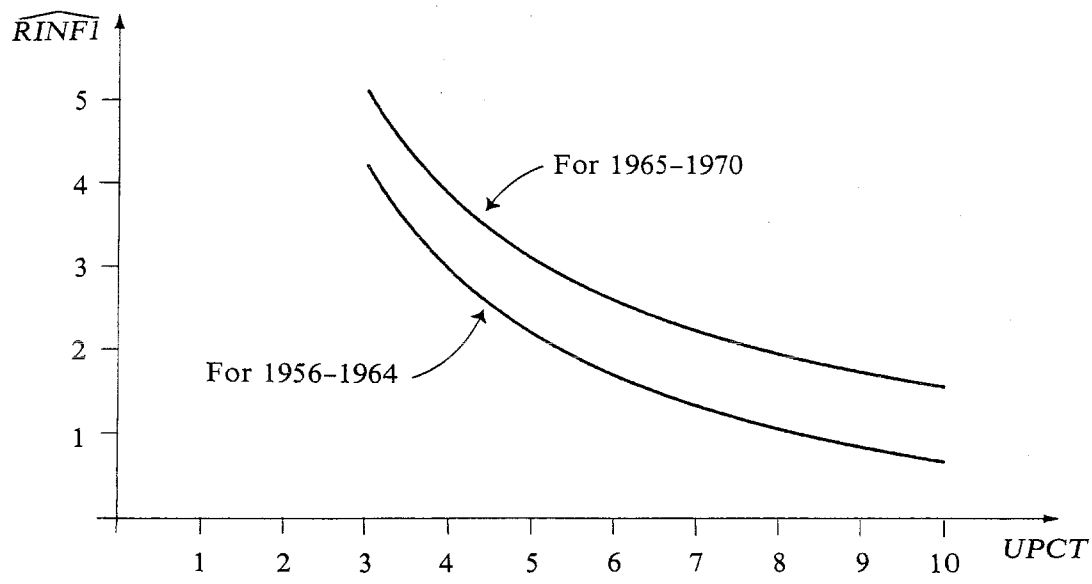

FIGURE 7.2 The Phillips curve shows the trade-off between inflation and unemployment. Using a dummy variable specification, Equation (7.36) finds evidence that is consistent with the conjecture that the curve shifted upward in the middle of the 1960s, as compared with its position earlier. The reciprocal specification of the effect of unemployment on inflation leads to the nonlinear relation shown here.

(Note that the first regressor is actually *UINV* as defined previously.) The coefficient on the dummy variable indicates that the Phillips curve shifted up by 0.894 percentage point in the latter part of the sample period as compared with where it was earlier. The two estimated Phillips curves are shown in Figure 7.2.

Using categorical information with the dummy variable technique is more complicated when there are more than two categories involved. For example, the variable *REG* in our cross-section data set carries information about the region of the country in which the person lives. The information is coded 1 for Northeast, 2 for North Central, 3 for South, and 4 for West. Suppose that we are focusing on a linear relation between Y and X, but want to take region into account. Would the specification

$$E[Y] = \beta_0 + \beta_1 X + \beta_2 REG \tag{7.37}$$

make sense? No. This specification does show a common impact of X on $E[Y]$ among the regions, but holding X constant it specifies that $E[Y]$ is β_2 greater in the North Central region than in the Northeast, $2\beta_2$ greater in the South, and $3\beta_2$ greater in the West. There is no reason why the actual differences should be so ordered or why they should be multiples of one another.

To bring in multiple-category information like region, we must construct a set of dummy variables. One way to think about doing this is to create a separate dummy variable for each category (region), taking on the value 1 if the observation belongs to that category and 0 if it does not. From *REG*, we can create four dummy variables: *DNEAST*, *DNCENT*, *DSOUTH*, and *DWEST*, to use

mnemonic names. As before, each dummy variable serves to identify each obervation as being in one of two groups (i.e., in the specified region or not). In formulating the regression model, one of the dummy variables must be excluded. Then, the coefficient on each of the included dummy variables represents the difference between the intercepts of that category and the excluded category. The intercept for the excluded category is simply β_0. Were all the categories' dummy variables included, we would have one more coefficient than we could interpret logically; such a redundancy is symptomatic of a major mistake in specification, which we review in Section 7.6.

Revising (7.37) we have

$$E[Y] = \beta_0 + \beta_1 X + \beta_2 DNCENT + \beta_3 DSOUTH + \beta_4 DWEST \tag{7.38}$$

If a person lives in the excluded Northeast, the value of each of the included dummy variables is zero and

$$E[Y] = \beta_0 + \beta_1 X \tag{7.39}$$

If a person lives in one of the other regions, the corresponding dummy variable is 1 but the other two are 0; thus

$$E[Y] = \beta_0 + \beta_1 X + \beta_j \qquad (j = 2, 3, \text{ or } 4) \tag{7.40}$$

The interpretation of each dummy variable coefficient (β_j) is the difference in $E[Y]$ of living in that region rather than the excluded Northeast region, holding X constant.

For example, we reconsider the simplest earnings function. Now taking region into account, the estimated regression is

$$\begin{aligned} \widehat{EARNS}_i = {} & -0.803 + 0.794 ED_i + 0.288 DNCENT_i \\ & -0.828 DSOUTH_i - 1.992 DWEST_i \\ R^2 = {} & .310 \qquad SER = 4.351 \end{aligned} \tag{7.41}$$

which is graphed in Figure 7.3. The coefficients on the dummy variables estimate shiftlike differences among the predicted earnings for persons living in different regions. It does not matter for the ultimate interpretation which of the region dummies is excluded, but it does affect the constant and dummy coefficients that are actually found.

The dummy variable technique can be extended to cases with more than one categorizing variable. For example, if theory specifies that earnings is a function of education, race, and region, then the regression model would include *DRACE* and the three dummy variables for region. For men in the sample who are white and live in the Northeast, the values of all the dummy variables are zero; this is the reference group to which other observations are compared.

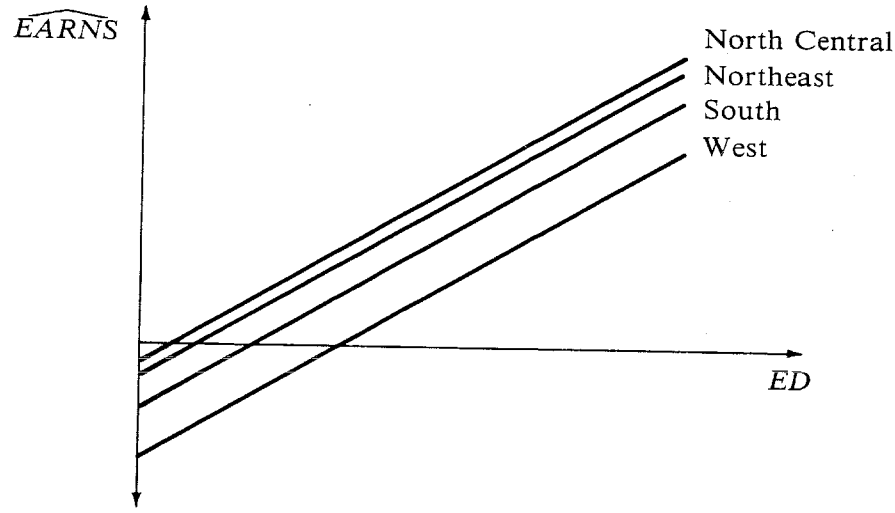

FIGURE 7.3 The simplest earnings function taking region into account uses three dummy variables to estimate the shiftlike differences among the regions. In Equation (7.41), the dummy for the Northeast region is excluded, and the other dummy-variable coefficients give the shift between each of the other regions and the Northeast. If a different region were excluded, the resulting graphical representation would be exactly the same as shown here, but the dummy variable coefficients and the regression intercept would be different because of the change of reference.

7.4 Polynomial Specifications

In Chapter 6 a variety of functional forms were introduced that allow the linear regression model to be used even when the basic behavioral relation is essentially nonlinear. The idea is that transformations of variables create new variables that can be examined in a regression framework. In the multiple regression extension of this idea, the same transformation can be applied to all the original variables, or different transformations can be applied to different variables. The important requirement is that a linear relation be specified between the ultimate regressand and the ultimate regressors.

In this section we look at another functional form that enhances the flexibility of the regression model. In the example provided, one of the original explanatory variables is left unchanged while the other is treated in the new way.

Thinking purely in mathematical terms, suppose that the exact relation between Y and X is like that in Figure 7.4a or b. These relations are clearly nonlinear, but they cannot be characterized by any of the functional forms examined in Chapter 6. The shape is that of a parabola, whose equation is written as

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 \tag{7.42}$$

If $\beta_2 > 0$, the parabola is concave upward, and if $\beta_2 < 0$, the parabola is concave downward. The slope of a parabola is given by

$$\text{slope of parabola} = \beta_1 + 2\beta_2 X \tag{7.43}$$

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2$$

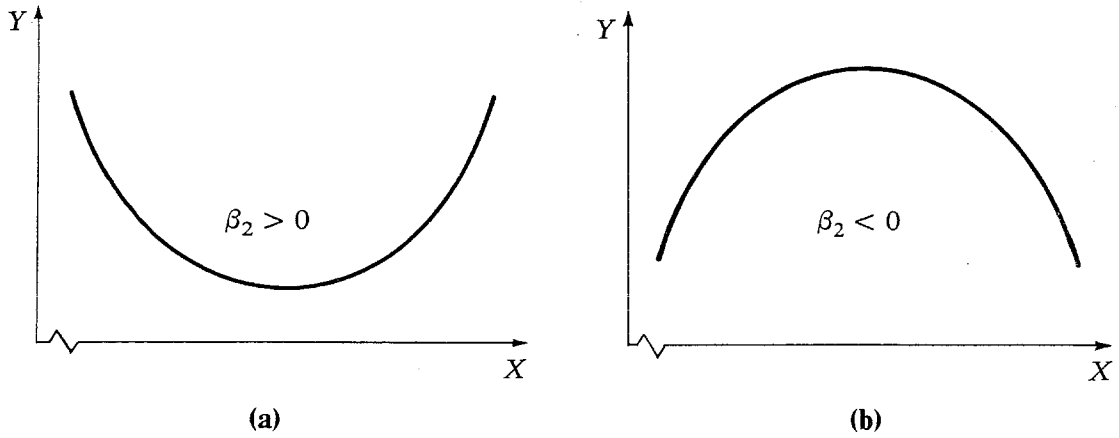

FIGURE 7.4 A parabolic relation between Y and X is concave upward if $\beta_2 > 0$ and concave downward if $\beta_2 < 0$. Although Y is a nonlinear function of X, Y also can be viewed as a linear function of X and X^2 together. This allows a multiple regression model to specify a parabolic form for the systematic relation between Y and X.

And as with any mathematical function, we know that when $\Delta X = 1$, the corresponding ΔY is approximately equal to the slope.

Returning to econometrics, if the systematic part of the relation between two variables, Y and X, is parabolic (or quadratic), we can use a linear multiple regression model of the form

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i \tag{7.44}$$

to describe the process. To do this, we let the regressor X_1 be equal to the original variable X and we let the regressor X_2 be equal to X^2. That is, we make two regressors out of a single original variable.

With this understanding, it is conventional to rewrite (7.44) as

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + u_i \tag{7.45}$$

Although the regression model specifies that Y is a linear function of X and X^2, our interest is in the implied parabolic relation between Y and X. We satisfy this interest by recognizing that X^2 must change if X does. Thus it does not make sense to say that β_1 alone measures the impact of X on Y. Rather, the impact on $E[Y]$ of a unit change in X is given approximately by the slope of the implicit quadratic equation: $\beta_1 + 2\beta_2 X$.

In studies of earnings functions it sometimes is suggested that the impact of experience diminishes and perhaps even becomes negative as the amount of experience increases. This is based on notions of physical and mental aging, diminishing returns, and optimal investment in human capital. Holding constant the level of education, the impact of experience on earnings is theorized to be

like a hill-shaped parabola. Letting *EXPSQ* be the name of the regressor that is equal to the square of *EXP*, we estimate an earnings function of the form

$$\widehat{EARNS}_i = -9.791 + 0.995ED_i + 0.471EXP_i - 0.00751EXPSQ_i \qquad (7.46)$$

$$R^2 = .329 \qquad SER = 4.267$$

In reporting regressions like this, the symbol EXP^2 is sometimes used to denote the second regressor.

To assess the impact of experience, we see first that the coefficient on *EXPSQ* is negative. Therefore, the estimated relation between *EARNS* and *EXP* (holding *ED* constant at any level) is a hill-shaped parabola, as in Figure 7.5. The slope of this relation is

$$\text{slope} = 0.471 + (2)(-0.00751)EXP \qquad (7.47)$$

Using this, we find that a man with five years of experience will have his earnings increased by about \$396 after gaining another year, but a man with 20 years of experience will have his earnings increased by about only \$171 after gaining another year. To find the level of experience that corresponds to the peak of earnings, we solve (7.47) for the *EXP* value associated with a zero slope. Here, peak earnings occurs after about 31 years of experience, and beyond that negative returns to experience set in. All this holds true regardless of what the level of education is.

To assess the effect on predicted earnings of any particular change in experience two approaches are reasonable. The first, which makes sense for small changes in experience, makes use of the slope at the original point to make an approximation:

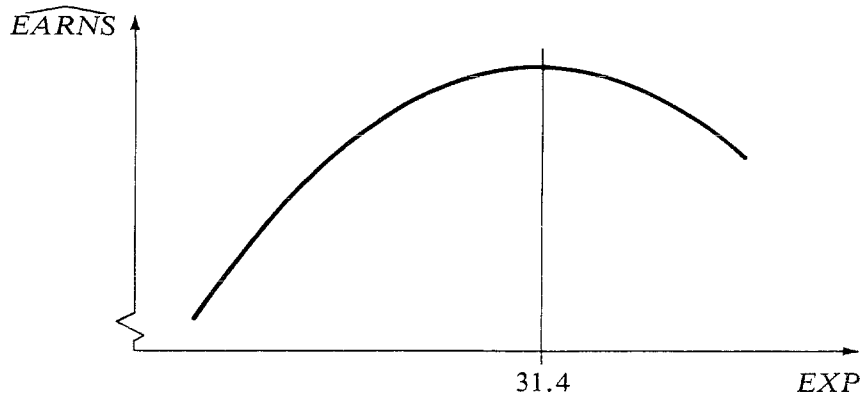

FIGURE 7.5 The earnings function (7.46) includes a parabolic (quadratic) specification for the partial relation between earnings and experience. Holding education constant at any level, the relation between predicted earnings and the level of experience is illustrated here. The slope, which is given by Equation (7.47), gives the impact on predicted earnings of a one-year increase in experience. For any given level of education, predicted earnings reach a peak at 31.4 years of experience.

$$\Delta EARNS \approx [0.471 + (2)(-0.00751)EXP_i]\ \Delta EXP \tag{7.48}$$

The second approach makes use of (7.23) to find the exact change along the estimated regression:

$$\Delta EARNS = 0.471\Delta EXP - 0.00751\Delta EXPSQ \tag{7.49}$$

In this computation, note that $\Delta EXPSQ$ is not equal to the square of ΔEXP.

It sometimes is theorized that the relation between Y and X is specified by a higher-order polynomial, such as

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 \tag{7.50}$$

Clearly, our technique can be extended to bring in the values of X, X^2, and X^3 as three separate regressors in order to make estimates of the unknown parameters. These applications are rare, however, in contrast to the more common quadratic specification.

7.5 Logarithmic Specifications*

If the regressand in a multiple regression is the logarithm of a regular variable, then the regressors might be any combination of logarithmic, regular, and dummy variable forms. The model might be pure log-linear, pure semilog, or a hybrid of types.

The Demand for Money

Although the demand for money depends importantly on the level of income because money holdings are used to finance the transactions that generate income, it probably also depends on the rate of interest because money is held as an asset, as part of wealth. In addition, even the money balances held for transactions purposes may be interest sensitive. For both reasons, the interest impact is theorized to be negative. Thus a multiple regression in which the amount of money demanded is related to both the level of income and the rate of interest seems more appropriate than the simple model proposed in Chapter 6.

The most common specification in money demand regressions is the pure log-linear form, because it yields constant elasticity estimates. Using the 25 annual observations from 1956 through 1980, we find

$$\widehat{LNM_i} = 3.759 + 0.246LNGNP_i - 0.0205LNRTB_i \tag{7.51}$$
$$R^2 = .785 \qquad SER = 0.0309$$

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

where M is the real quantity of money [see (6.31)], GNP is real national income, and RTB is the interest rate on Treasury bills. As theory predicts, the income elasticity is positive and the interest elasticity is negative. Comparing the multiple with the simple regression (6.33), we see that the estimated income elasticity is somewhat greater (0.246 versus 0.215) as is the R^2 (.785 versus .780).

How important is the interest rate relative to the level of income? We may directly compare the regression coefficients, because they are elasticities: a 1 percent increase in income leads to a 0.246 percent increase in the predicted demand for money, while a 1 percent increase in the interest rate leads to a 0.0205 percent decrease. Thus the interest rate appears to be much less important than income, but this comparison is misleading. In the short run the interest rate is proportionately much more variable than income: year-to-year changes of 25 percent (not percentage points!) or more often occur in the interest rate, while changes of only 5 percent or more in income occur with roughly the same frequency. Comparing these two hypothetical changes in RTB and GNP, we find the resulting impact of the interest rate to be about half as much as that of the level of income. Thus the rate of interest should be viewed as having a moderately important impact on the demand for money.

The Earnings Function

In applied labor market research, the preferred form for earnings functions specifies the regressand to be the logarithm of earnings. When education is the only explanatory variable, we have a pure semilog form as in (6.39). As more variables are taken into account the form of the function may become mixed.

For example, we consider a model in which the logarithm of earnings depends on the level of education, the amount of experience (entered quadratically), the logarithm of the number of months worked (to yield an elasticity), and the person's race and region of residence. Based on our cross-section data set, the estimated regression is

$$\begin{aligned}\widehat{LNEARNS}_i = {} & -2.031 + 0.106ED_i + 0.0501EXP_i \\ & -0.000930EXPSQ_i + 0.908LNMONTHS_i \\ & -0.239DRACE_i - 0.00468DNCENT_i \\ & -0.193DSOUTH_i - 0.162DWEST_i \end{aligned} \tag{7.52}$$

$$R^2 = .511 \qquad SER = 0.420$$

The impact of each coefficient is interpreted in the same way as in simpler formulations.

An additional year of schooling increases the level of earnings by approximately 10.6 percent, which is quite close to the finding in the simple semilog model (6.40). The level of earnings increases with the amount of experience up

to a peak of about 27 years of experience and decreases thereafter. The elasticity of *EARNS* with respect to *MONTHS* worked is 0.908; if the elasticity were 1.0, then earnings would be proportional to months worked, which might be expected on the basis of simple reasoning.

The dummy variable coefficients show the impact on the regressand of being in the included category. We find that *LNEARNS* is 0.239 lower for blacks than for whites. Being black rather than white corresponds to a unit change in *DRACE* (i.e., $\Delta DRACE = 1$), so the effect of being black is to change earnings by approximately -23.9 percent. In other words, earnings are about 23.9 percent less for blacks than for whites. Note that this estimate of the racial difference in earnings pertains to a comparison in which all the other explanatory factors (education, experience, months worked, and region) are held constant; it is not an estimate of the difference between the average earnings of blacks and whites. The coefficients on the regional dummies show that earnings are lower in all these regions than in the Northeast, holding constant the other explanatory factors. Earnings are approximately 1/2 of 1 percent lower in the North Central region, 19 percent lower in the South, and 16 percent lower in the West.

The R^2 value indicates that more than half of the variation in observed *LNEARNS* has been explained by the regression, which is fairly good for this type of regression.

7.6 Specification Questions

In our development and use of multiple regression, we have followed a consistent approach in all cases. We start from the presumption that there is some stable process determining the values of some variable. Next we set up a multiple regression model that correctly describes the process in mathematical terms. Finally, we use data to estimate the unknown parameters and interpret or apply the estimated model according to our needs.

One of the difficulties with this approach is that we must correctly describe the process in terms of a regression model. This is a very demanding requirement. To meet it, or at least to come close to meeting it, we need to know more about how to specify regression models. In this section we limit ourselves to questions relating to the selection of variables for a linear model.

The Causal Nexus

In thinking about how several variables together affect or determine another one, it is natural to make a distinction between direct and indirect effects. Figure 7.6 sketches the causal linkages in a hypothetical case, with arrows indicating the paths and directions of causation. Variable X_1 is directly affected by X_3 and unlabeled variables. Variable X_2 is directly affected by X_1 and an unlabeled variable, and it is indirectly affected by X_3 and the other unlabeled variables.

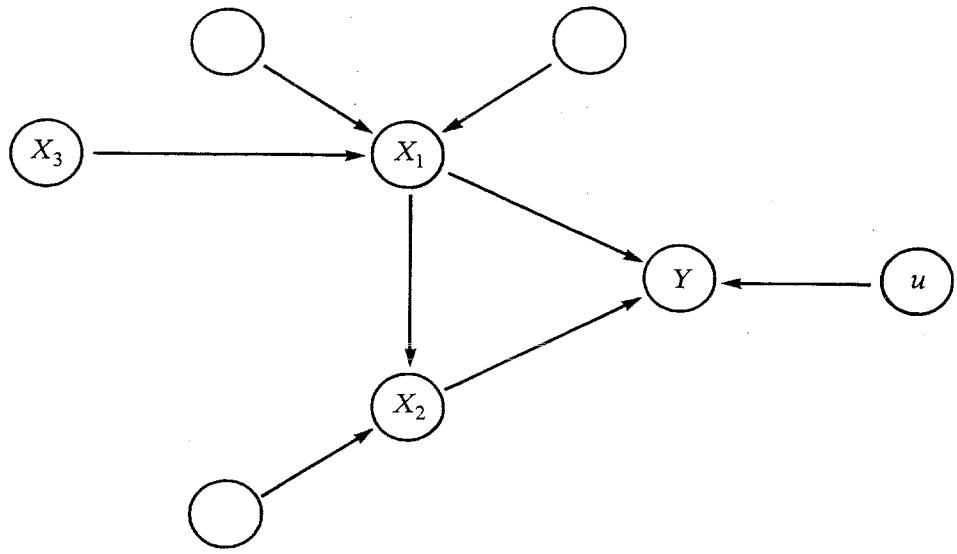

FIGURE 7.6 The causal nexus determining Y in a hypothetical case is illustrated in this schematic diagram. Arrows indicate the paths and directions of causation, and unlabeled circles contain other variables. This behavioral process is properly specified by Equation (7.53), assuming linearity. A regression model that mistakenly excludes X_2 or one that mistakenly includes X_3 misspecifies this process, but these two types of misspecification have very different consequences.

Variable Y is directly affected by X_1 and X_2, and it is indirectly affected by X_1, X_3, and the unlabeled variables. Note that X_1 has both direct and indirect effects on Y. Also, note that the disturbance u has a direct effect on Y and no linkage at all to any of the other variables.

For example, in thinking about the aggregate demand for money in macroeconomics, theory suggests that it depends directly on the level of GNP and on some interest rate and indirectly on the Federal Reserve discount rate. Through bank behavior and the action of financial markets, the discount rate affects the general level of the interest rate. Through expenditure decisions, the level of the interest rate affects GNP. And through the behavior of firms and individuals, the interest rate and the level of GNP affect the demand for money. In a commonsense way, this logic is represented in Figure 7.6, with Y being the demand for money, X_1 being the interest rate, X_2 being the level of GNP, and X_3 being the Federal Reserve discount rate.

Suppose that the correct form of the model determining Y is linear. Which variables should be included? The answer here is that only X_1 and X_2 should be included as explanatory variables, so that the proper model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i \tag{7.53}$$

That is, in the specification of a multiple regression model, all the variables that have direct effects should be included, and variables that have only indirect effects (or no effects at all) should not be included.

The validity of this general rule depends on specific notions of what "direct" and "indirect" mean. For our purposes, a variable has a direct effect on Y only if in our general thinking about the process determining Y it is true that a change

in the variable would lead to a change in Y while all other variables affecting Y are held constant. In other words, if a variable truly has a *ceteris paribus* effect on Y, then it has a direct effect. This leads to the general rule for specifying the regression model, because each coefficient is only a *ceteris paribus* effect.

In the causal nexus illustrated in Figure 7.6, variable X_3 has only an indirect effect. We recognize its effect as working this way: a change in the value of X_3 affects X_1, and any change in X_1 affects Y. There is no other path along which X_3 affects Y. Thus if X_1 somehow remains constant there is no way for X_3 to affect Y. Since X_3 has no *ceteris paribus* effect, it does not have a direct effect and it should not be included in the regression model. In the demand for money example, neither firms nor individuals care about the Federal Reserve discount rate. It affects the demand for money only through the general interest rate. If the general interest rate somehow remains constant, changes in the discount rate will not affect the demand for money.

The distinction between direct and indirect effects means that we must be especially careful in interpreting regression coefficients for variables like X_1, which has both kinds. The proper interpretation of β_1 is that it gives the impact on Y of a change in X_1, holding constant the value of X_2. This is the direct effect. The indirect effect of X_1 works through consequent changes in X_2. If we knew the amount of change in X_2 that would be associated with the initial change in X_1, we would have a basis for calculating the magnitude of the indirect effect. However, a regression model does not provide any information about the linkages between or among its explanatory variables, so generally it does not provide enough information to determine the magnitude of an indirect effect like that of X_1. Thus a regression model always tells us about the partial effects of the explanatory variables, but it is silent on the question of total effects. In applied research it is often true that there are some linkages between or among explanatory variables, and therefore careful interpretation is usually needed.

In thinking about what determines some economic outcome, sometimes it is natural to consider a variable that has only an indirect effect, and sometimes this indirect effect may be of special interest to us. For example, we may want to know what effect the Federal Reserve discount rate has on the demand for money. If the variable, like X_3, is not included in the model, how can we learn about its effect? What is needed is a second regression model that describes how X_1 is determined by X_3 and other variables. This model and (7.53) taken together form a recursive system of equations that can be used to analyze the effect X_3 has on Y. Such a system is a special case of ***simultaneous-equation models,*** which are discussed in Chapter 17. Until then our attention will be focused on single-equation models, which we realize are necessarily limited in scope.

Consequences of Misspecification

Sometimes it is difficult to know whether a variable under consideration for inclusion in a regression model has a direct effect, only an indirect effect, or no effect at all. Hence it is difficult to decide which variables should be included

in the model, and it is likely that some models we see or create will be misspecified.

Let us return to the causal nexus illustrated in Figure 7.6. The process is correctly described by (7.53), as explained above. Variables X_1 and X_2 are known as ***relevant variables*** because they should be included in the regression, and all other variables (such as X_3) are known as ***irrelevant variables*** because they should be excluded. Suppose that we have data on all the necessary variables. The estimated form of the correct model is

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} \tag{7.54}$$

In carrying out our research we might make either of two mistakes, unfortunately: we might exclude a relevant variable from the regression, or we might include an irrelevant one. What are the consequences of these mistakes?

Suppose first that we leave out a relevant variable, say X_2, from the equation and estimate just the simple regression

$$\hat{Y}_i = \hat{\gamma}_0 + \hat{\gamma}_1 X_{1i} \tag{7.55}$$

(γ is lowercase gamma, the Greek "g"). Here $\hat{\gamma}_1$ denotes the estimated coefficient on X_1. In general, the value of $\hat{\gamma}_1$ in (7.55) will be different from the value of $\hat{\beta}_1$ in (7.54) when the regressions are estimated from the same set of data. This is because the formulas used are different: the simple regression slope coefficient is calculated by (5.12), and it depends only on the values of Y and X_1; the multiple regression slope coefficient on X_1 is calculated by (7.9), and it depends on the values of X_2 as well as on the values of Y and X_1.

In general we believe that the $\hat{\beta}_1$ in the estimated multiple regression provides the best estimate of the true *ceteris paribus* effect of X_1 on Y, which is denoted by β_1 in the true regression model (7.53). This idea is explored further in Chapter 13. Since $\hat{\gamma}_1$ is different from $\hat{\beta}_1$, it makes sense to say that it should be considered a not-so-good estimate of the true effect. Mathematical analysis of the estimating formulas shows that $\hat{\gamma}_1$ differs from $\hat{\beta}_1$ by factors that represent the relation between X_1 and X_2 and the relation between Y and X_2. Loosely speaking, the calculation of $\hat{\gamma}_1$ captures both the direct effect of X_1 on Y and some part of the direct effect of X_2 on Y. The latter effect is captured because of the correlation between X_2 and X_1 in the data. Hence $\hat{\gamma}_1$ is systematically distorted from the true value β_1, which is only the direct effect of X_1 on Y.

For example, consider the earnings functions (7.15) and (6.7). In the estimated multiple regression the coefficient on *ED* is 0.978. In the estimated simple regression the coefficient on *ED* is 0.797, which is substantially smaller. This difference is consistent with what would be expected if the multiple regression were the true model. To see this we note that *ED* and *EXP* are negatively correlated ($r = -.57$) in the data. This means that higher-than-average *ED* values tend to be accompanied by lower-than-average *EXP* values among the observations. Since experience is estimated to have a positive direct effect [$\hat{\beta}_2 = 0.124$ in (7.15)], observations with higher-than-average values of *ED*

tend to have their earnings diminished by their lower-than-average values of *EXP*. Thus it would be expected that $\hat{\gamma}_1$ would be smaller than $\hat{\beta}_1$.

Suppose now that we make the mistake of including an irrelevant variable, say X_3, in the regression specification and that we end up with

$$\hat{Y}_i = \hat{\delta}_0 + \hat{\delta}_1 X_{1i} + \hat{\delta}_2 X_{2i} + \hat{\delta}_3 X_{3i} \tag{7.56}$$

(δ is lowercase delta, the Greek "d"). In comparison with (7.54), the inclusion of the irrelevant variable affects the estimates of the other coefficients in the sense that $\hat{\delta}_2 \neq \hat{\beta}_2$, $\hat{\delta}_1 \neq \hat{\beta}_1$, and $\hat{\delta}_0 \neq \hat{\beta}_0$. However, the effects on these estimates are rather random and usually mild. The coefficient $\hat{\delta}_3$ serves to estimate the true partial impact of X_3 on Y, which is zero (because X_3 is irrelevant). However, because of the randomness introduced into the data by the disturbances, the actual estimated value is not likely to be zero. Hence inclusion of the irrelevant variable can lead to misinterpretation of the true economic process.

In deciding whether or not to include a variable in a regression specification, the consequences of these two types of mistakes must be compared. In most cases the introduction of additional randomness into the estimation process is less serious than the introduction of systematic distortion. Hence, including an irrelevant variable is usually considered to be less of a problem than excluding a relevant one. If one has good theoretical reasons for including a variable, it is best to do so. However, this should not be taken as a suggestion to hunt for variables with the hope that some might turn out to look good. In practice, most researchers estimate more than one specification of the process they are studying and then try to determine which of them is best. This judgment must be based on a blending of economic and econometric analysis, including considerations covered in Chapters 11 through 13.

Finally, it is often the case that our interest is in determining just the effect of one variable on another. For example, we might want to estimate the effect of education on earnings but we might be unconcerned with the role of experience. It is tempting to estimate just a simple regression of earnings on education and look at the slope coefficient. However, this is not what we are interested in. The analysis above shows that the estimate of the slope in the simple regression is not a good estimate of the *ceteris paribus* effect of education on earnings. Also, it is not a good estimate of what was identified in the preceding section as the total effect of education on earnings, because to calculate that we would need a simultaneous-equation model.

To generalize this, even when we are interested in the effect of a particular variable, it is necessary to specify and estimate a regression model that fully reflects the behavior of the economic process at work. Sometimes the variables that we are not interested in are called ***statistical controls.*** For example, we might say that the multiple regression (7.15) estimates the effect of education on earnings, controlling for experience. Similarly, the coefficient on *DRACE* in (7.33) estimates the effect of race (i.e., the effect of being black rather than white) on earnings, controlling for education. What all this amounts to saying

is that the only effects that can be estimated in a single-equation, multiple regression model are the *ceteris paribus* (direct) effects and that the proper way to estimate these effects is with a correctly specified model.

Multicollinearity

In the data used to estimate a multiple regression model, it is usually the case that there is some correlation or (more technically) some degree of linear dependence among the explanatory variables. For example, in Figure 7.6 a case is illustrated in which there is a direct relation between X_1 and X_2. In other cases there may be a correlation between explanatory variables even when they are not connected by a behavioral relation. The general model and our method of estimation accept this situation as valid; indeed, the absence of any relations among the explanatory variables is a very special case that we rarely encounter in econometrics. The main consequence of this situation, so far, has been that we must be careful to interpret regression coefficients as direct effects and to recognize that there may be indirect effects as well.

In addition, when two explanatory variables have a very high correlation or when there are some other special relations among the explanatory variables, the situation has some unfortunate consequences for statistical inference. We will examine these in Chapter 13. Loosely speaking, it becomes very difficult to disentangle the separate effects of the explanatory variables on the dependent variable. For example, suppose that aggregate consumption depends on aggregate income, the Treasury bill rate, and the interest rate on consumer debt. Because of the behavior of financial markets, there is likely to be a high correlation between the two interest rates over time. One might guess that it would be difficult to determine the effects of each interest rate separately with any great precision because the two variables might be nearly linear transformations of each other. A common consequence of this is that if we happen to add a few new observations to the data set, or drop a few from it, the new regression coefficients may be very different from the original ones.

This situation is known as ***multicollinearity***. From the point of view of specifying the model, it does not indicate any mistake. Rather, multicollinearity arises from the nature of the data, and usually we have to accept it as part of reality. Multicollinearity is common in time-series regressions, because several of the explanatory variables may increase over time and therefore be highly correlated.

However, consider the possibility that there is a perfect correlation ($r = 1$) between a pair of explanatory variables, or (more generally) that there is a perfect linear dependence among the explanatory variables. Technically, this is a limiting case of multicollinearity, and indeed it is called ***perfect multicollinearity***. In this situation, the OLS method no longer can produce estimates. The point of difficulty may be seen in the case with two explanatory variables: the denom-

inators in (7.9) and (7.8), which define the estimators for $\hat{\beta}_1$ and $\hat{\beta}_2$, become equal to zero.

Although this would seem to complicate matters for us immensely, it turns out not to be much of a problem. In contrast to regular multicollinearity, which is a situation that occurs naturally in data, perfect multicollinearity nearly always is the result of making a mistake in the specification of the model. The remedy is simple: respecify the model appropriately. To understand this, we consider several cases in which perfect linear dependence can arise in a regression model. These mistakes share the characteristic that they include in the specification some variable that is not really needed or that does not bring new information to the model; in this sense, the mistaken specification is redundant.

A very special case of perfect linear dependence occurs if an explanatory variable, X_j, is a constant. In this case, the coefficient β_j plays the same role as the intercept β_0 in the regression specification: β_0 and the term $\beta_j X_j$ are constants that are just added in during the determination of Y. There is no unique way for any statistical technique to assign some of the constancy to β_0 and the rest to $\beta_j X_j$.

Perfect linear dependence also occurs if one variable is simply a multiple of another. For example, suppose that we are trying to explain the exports of cars from Japan to the United States and that we include both the price of these cars in Japan (measured in yen) and the price in the United States (measured in dollars) among the explanatory variables. If all our observations are from a period of fixed exchange rates during which all the dollar prices were the same multiple of the yen prices, then the two price variables measure exactly the same set of economic facts. It does not make sense to include them both, and because of the linear dependence we could not.

Perfect linear dependence also occurs if some set of the explanatory variables satisfy an additive identity. For example, suppose that we are interested in estimating the marginal propensities to consume (mpc's) out of labor income, property income, and total income. We might think of regressing consumption on these three income variables in one equation. However, since total income equals labor income plus property income, it must be that the mpc out of total income equals the sum of the two type-specific mpc's. Trying to estimate three mpc's is redundant and therefore not necessary; since it involves a linear dependence among the explanatory variables, it is also impossible.

A final case of perfect linear dependence occurs if dummy variables for all the groups of a categorical variable are included in the regression. For example, in our treatment of region in Section 7.3 we distinguished four groups but included only three in the specified earnings function. This was adequate to specify the theory behind the model, because each dummy variable coefficient specified the difference between the estimated intercept for that group and the intercept (β_0) for the excluded group. Including the fourth dummy variable would be redundant and would introduce a linear dependence.

As might be realized from these cases, it is quite possible to specify a model

with perfect multicollinearity if the work is done with insufficient thought. Usually, a computer program will detect the situation and give some kind of error message. However, because of either imprecision in the data or design of the computational algorithm, it is possible that a computer program might not detect the situation and it would produce some calculations. In this case, the user would think he has an estimated regression when in fact he has nonsense.

Problems

Section 7.1

7.1 Formulate a multiple regression model showing how the quantity demanded of a certain product depends on both the price of the product and the income of consumers. What are the anticipated signs of the coefficients?

7.2 What is the graphical interpretation of the demand model estimated from the specification in Problem 7.1?

★ **7.3** Continuing Problem 7.2, if income is fixed at a certain amount, what is the graphical interpretation of the relation between predicted demand and price? How does this graph illustrate the *ceteris paribus* concept?

7.4 Based on Equation (7.15), what is the impact on predicted earnings of gaining a college education ($ED = 16$) rather than stopping after completing high school ($ED = 12$)? Assume that EXP is held constant.

★ **7.5** Consider two men of age 35. Suppose that the first has four more years of schooling than the second and therefore has four fewer years of working experience. Based on Equation (7.15), who has greater predicted earnings? By how much?

7.6 Show that if the covariance between X_1 and X_2 is zero, the estimator given in Equation (7.8) is identical to the slope estimator in the simple regression of Y on X_2.

Section 7.2

7.7 Based on Equation (7.29), what is the impact on predicted consumption of an increase in the interest rate from 5 percent to 7 percent?

7.8 Suppose that inflationary forces increase both the interest rate and the rate of inflation by three percentage points. Based on Equation (7.29), determine the effect of these changes on predicted consumption.

7.9 Suppose that a one-percentage-point increase in the rate of interest causes DPI to decrease by 200 million dollars. Based on Equation (7.29), determine the total effect of this change in the rate of interest.

7.10 Determine the values of $\overline{R}^2$ for Equations (7.29) and (6.10).

7.11 Derive the relation in Equation (7.28).

Section 7.3

⋆ **7.12** Based on Equation (7.33), what is the predicted level of earnings for a black man with 16 years of schooling? For a white man with 12 years of schooling?

7.13 Graph the estimated regression (7.33), clearly labeling all its features.

7.14 Based on Equation (7.41), determine the predicted levels of earnings for high school graduates ($ED = 12$) in each of the four regions of the country.

⋆ **7.15** Suppose that you want to estimate the impact of education and marital status on the earnings of women. If the data show three marital status categories (single, married, and divorced), how would you set up a regression model?

7.16 Based on Equation (7.41) and Figure 7.3, determine the coefficients of the estimated regression of *EARNS* on *ED* and the regional dummies if the West is the excluded region.

7.17 Does the estimated coefficient on *DRACE* in Equation (7.33) give the total effect of race on earnings? Explain.

7.18 Based on Equation (7.36), determine the predicted rate of inflation in 1960 and compare this with the actual rate. Now, try to predict the rate of inflation for 1980, and compare whatever prediction you make with the actual rate. Explain.

Section 7.4

7.19 Based on Equation (7.46), what is the effect on the predicted earnings of a person with 25 years of experience gaining one more year? What about a person with 35 years of experience?

7.20 Based on Equation (7.46), determine the effect on predicted earnings of a five-year increase in experience for a worker already having 20 years of experience. Do this first using an approximation based on the slope of the implied relation and then using an exact calculation in the estimated regression.

7.21 Verify that the peak in earnings in Figure 7.5 occurs at about 31.4 years of experience.

7.22 Suppose that we have data on a factory's average cost of production and the amount of output in different periods. Specify a regression model that could estimate a U-shaped average cost curve.

Section 7.5

7.23 Suppose that the quantity demanded of a certain product depends on its price and consumers' income. Formulate a constant-elasticity

regression model for estimating this demand relation. What are the anticipated signs of the estimated elasticities?

* **7.24** In a cross-section context, suppose that output depends on labor and capital inputs. Formulate a regression model for estimating the output elasticities of labor and capital.

* **7.25** In a time-series context, suppose that "disembodied technical progress" leads the output yielded by all combinations of capital and labor inputs to grow at a fixed rate per year. Assuming constant output elasticities for labor and capital, formulate a regression model for estimating the rate of disembodied technical progress.

7.26 Modify the model formulated in Problem 7.25 to take account of the effect of "energy restrictions" that prevailed during three years. Explain clearly the econometric assumption underlying this modification.

7.27 In the earnings function estimated as Equation (7.52), would it make sense to have all the regressors in logarithmic form? Explain.

Section 7.6

7.28 Supposing that the rate of inflation affects the interest rate, and making any other economic assumptions that seem appropriate, illustrate the causal nexus determining aggregate consumption as estimated in Equation (7.29). Where could the Federal Reserve discount rate enter?

7.29 Suppose that a properly specified earnings function includes *ED*, *EXP*, and *DRACE* as explanatory variables. Illustrate the causal nexus determining earnings and explain the linkages.

7.30 Compare the estimated effect of education in Equations (7.33) and (6.7) from the point of view of possible misspecification.

7.31 Suppose that we wish to estimate the effect of being a union member on workers' earnings, using cross-section data. Specify a regression model that would be appropriate for estimating this effect. What theory or assumptions are required to make it appropriate?

7.32 Consider the Phillips curve model in Equation (7.36). Could this be estimated using data just for 1965 through 1970? Explain.

7.33 Suppose that the fourth dummy variable *DNEAST* were added to the regression specification in Equation (7.41) and that a computer provided "estimates." Try to interpret all the coefficients.