5

Simple Regression: Theory

As previewed in Chapter 1, regression analysis is a technique for estimating the values of the structural coefficients in a model of an economic process. In this chapter and the next we discuss simple regression, which applies to models involving just two variables. In Chapter 7 we discuss multiple regression, which applies to models that are more complex.

The discussion in this chapter is limited to the theory and mechanics of estimating a simple regression model of regular form. In the next chapter we apply this theory in a variety of cases, and we see how the estimates and measures developed here are put into practice.

5.1 Specification of the Model

The econometric use of simple regression starts from the theory or presumption that there is some true relation between two variables. In economics a relation is usually based on behavior, as in the case of a consumption function, but sometimes the relation is based on technology, as in the case of a production function. For simplicity we refer to either type as a behavioral relation.

In technical discussion we think of such a relation as being an ***economic process*** that can be described very concretely in mathematical terms. As an analogy, it helps to think of some engineering process, with an input and an output. The input and output are observable, but the operations of the process itself are not.

Suppose that we have data for n observations on two variables, Y and X, that we believe are related in such a way that we would say that Y is determined by X. Our thinking is that there is some process occurring in which a value of X is fed in and a value for Y is produced. If n values of X are fed in one at a time, n corresponding values for Y are determined one at a time. The n values of X and Y are all observable, and they can be collected as a set of data.

Two aspects of this process should be noted. First, X is the only identifiable and observable variable that affects Y; it is the only variable that feeds into the process. By explicit exclusion from the discussion, 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 are taken as given. The economic process under consideration does not affect those values and it plays no role in their determination. In addition, we make no effort to understand why X takes on whatever values it does. Our interest is in how the value of Y is related to the value of X for any observation.

We move toward statistical analysis by making a more concrete specification of the process by which Y is determined; this specification is known as the ***simple regression model.*** We theorize that the value of variable Y for each observation is determined by the equation.

$$Y_i = \beta_0 + \beta_1 X_i + u_i \tag{5.1}$$

In this regression model, Y is called the ***dependent variable,*** and X is called the ***explanatory variable*** (sometimes X is called the ***independent variable***). In the model, β_0 and β_1 are parameters that have fixed values throughout (β is lowercase "beta," the Greek "b"); they are called the ***coefficients*** of the regression model. The term u_i is called the ***disturbance.***

The disturbance u_i is considered to be an unobservable random term that does not depend on the value of X_i. The disturbances are meant to represent pure chance factors in the determination of Y. Among these factors there may be one that is best described as luck. Also, we might believe that Y is affected by a host of minor factors that we cannot identify and whose combined impact is indistinguishable from pure chance. Finally, we recognize possible measurement error in Y (but we presume that there is no measurement error in X). For any particular observation, u_i can be positive or negative, small or large. Overall, the average disturbance is anticipated to be close to zero, but only by coincidence would it be exactly zero in any given set of data.

For example, consider the process by which families determine their annual saving. Economic theory suggests that saving depends mostly on income. Letting Y be family saving and X be family income, (5.1) is a simple model of family saving behavior in which different families are the observations. The model does not explain the level of income of each family; it takes this as given. The parameter β_0 is the saving (probably negative) of families with zero income, and β_1 is the increase in saving resulting from a unit increase in family income (i.e., β_1 is the marginal propensity to save).

Based on the simple regression specification (5.1), we can decompose each Y_i value into a systematic component $\beta_0 + \beta_1 X_i$ and a random component u_i. The value of the systematic component is called the ***expected value*** of Y for each observation:

$$E[Y_i] = \beta_0 + \beta_1 X_i \tag{5.2}$$

This concept will be given a precise statistical meaning in Chapter 11, but for now it seems fairly intuitive: given (5.1), it is the value that Y_i would take on if the disturbance were equal to zero. This decomposition of each Y_i value into its systematic and random components can be compactly rewritten as

$$Y_i = E[Y_i] + u_i \tag{5.3}$$

The model specified in (5.1) is represented graphically in Figure 5.1. The systematic part of relation between Y and X is graphed as the line

$$E[Y] = \beta_0 + \beta_1 X \tag{5.4}$$

which is called the ***true regression line.*** We consider only a specific set of n observations whose X values are somehow given and whose Y values are determined according to (5.1). The resulting data points are plotted.

Following (5.3), the value Y_i for a single observation can be decomposed vertically into the distance up to a point on the true regression line and the distance from that point to the observation. The decomposition is shown explicitly for two observations. For the first, which has the values X_1 and Y_1 for

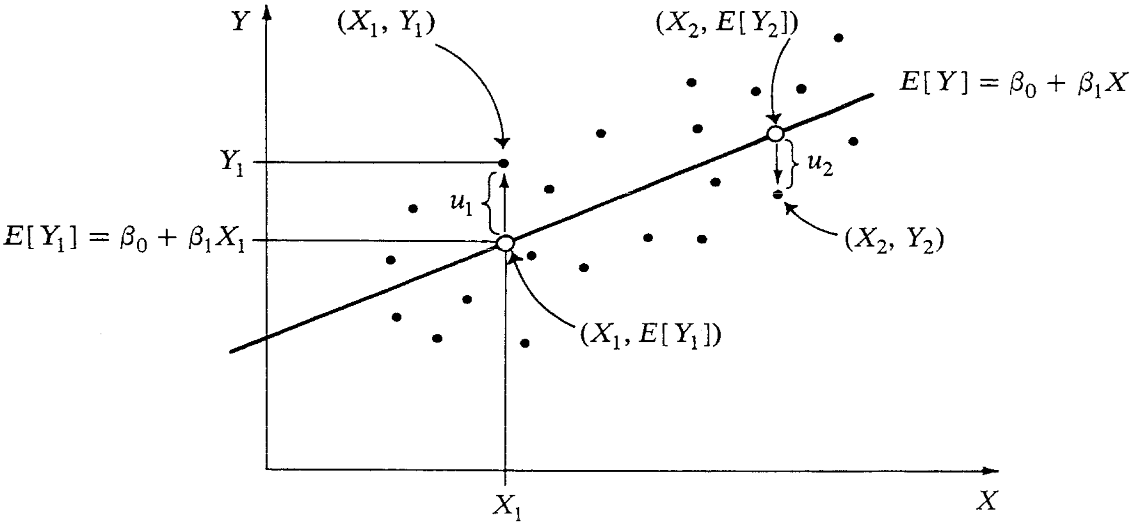

FIGURE 5.1 For each observation, the value Y_i can be decomposed into two parts: the systematic part $E[Y_i]$ (which equals $\beta_0 + \beta_1 X_i$), and the disturbance u_i. The systematic part is given by the height of the true regression line above the observation's X value, and the disturbance is given by the distance from the point on the regression line to the data point. The disturbance is positive for the first observation and negative for the second.

the two variables, the plotted point is shown and the X and Y values are traced to the axes. Directly above the value X_1 on the X axis we find a point on the true regression line; its vertical value is $E[Y_1]$, and this is traced over to the Y axis. The vertical distance from $E[Y_1]$ to Y_1 is u_1, which is shown also. The situation for the second observation (X_2, Y_2) is the same except that the data point lies below the true regression line, so the disturbance u_2 is negative.

In summary, the simple regression model illustrated in Figure 5.1 is a specification of the theoretical process that we use to describe the relation between two observable variables. For each observation the value X_i is determined outside the process. Given this X_i, the value Y_i is determined by (5.1). It will be convenient to say that the process produces the data on Y and X, even though the X values are determined outside.

Our theory is that the simple regression model actually reflects the way some economic behavior works. Surely this is a very simple model of how the values of a variable are determined: most economic variables systematically depend on more than one other variable, and that is why multiple regression is usually more appropriate. However, there are many instances in which simple regression is quite reasonable, and it is a useful tool of econometrics. Also, the presumption that the relation is strictly linear is not so constraining as it might seem. In some cases, we might well accept a linear model if we believe that the true relation is approximately linear. More important, we see in Chapter 6 that some truly nonlinear relations can be transformed into linear ones. In these cases we treat the transformed relation just as we treat the model specified in (5.1).

5.2 Estimation of the Model

When we have a set of data on Y and X that we presume was generated by a process described by (5.1), the values of β_0 and β_1 are unknown. Our interest focuses on estimating the values of these parameters, based on the data we have.

Our job now is pictured in Figure 5.2. Temporarily ignore the line drawn there. The data we have are plotted, and these points represent all that we can observe. To say, as above, that we presume or theorize that the data were generated by the process (5.1) means that we presume that there is some unobservable true regression line underlying the data. Figure 5.1, which illustrates the theory, explicitly shows a true regression line. Careful comparison of Figures 5.1 and 5.2 shows that the data points are exactly the same in both. In other words, Figure 5.1 presents our theory of how the data in Figure 5.2 were generated. Hence the true regression line in Figure 5.1 underlies the data in Figure 5.2. Of course, we do not see it there, because it is not observable. But if we accept the theory that the true regression line underlies the data, we can use the data to estimate the parameters of the line.

The task of estimating the parameters β_0 and β_1 of the unobservable true regression line is carried out by drawing or fitting an actual line through the

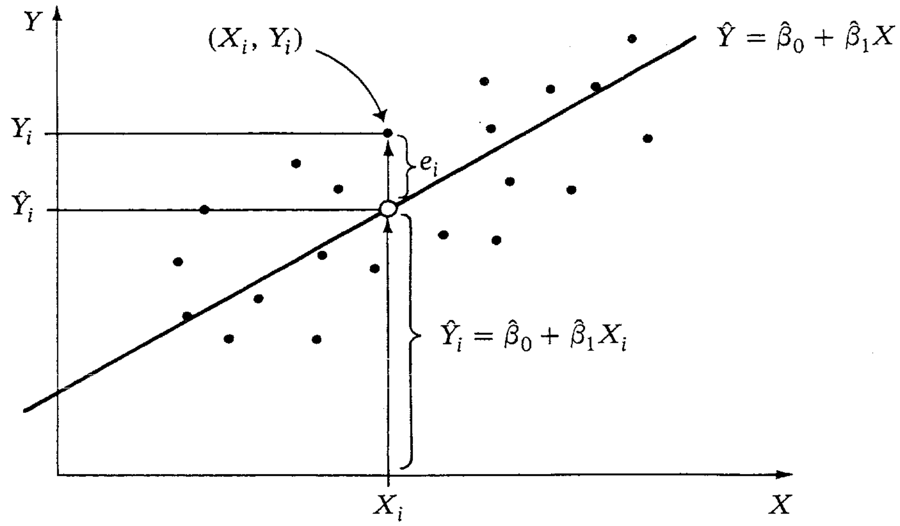
Y
(X_i, Y_i)
$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$
Y_i
e_i
$\hat{Y}_i$
$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
X_i
X

FIGURE 5.2 The estimated regression line is based on the data on Y and X, as plotted. It serves as an estimate of the true regression line in Figure 5.1, but it is generally different from it. For each observation, the value Y_i can be decomposed into two parts: the fitted value $\hat{Y}_i$ (which equals $\hat{\beta}_0 + \hat{\beta}_1 X_i$), and the residual e_i. The ordinary least squares criterion is to minimize the sum of the squares of these residuals.

data. The intercept of this actual line is denoted by $\hat{\beta}_0$ and it serves as our estimate of β_0, the intercept of the true regression line. Similarly, the slope of the actual line is denoted by $\hat{\beta}_1$ and it serves as our estimate of β_1. (The circumflex looks like a hat, so $\hat{\beta}_1$ is conventionally read as "beta-one-hat.")

Suppose that the $\hat{\beta}_0$ and $\hat{\beta}_1$ we end up with correspond to the line

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X \tag{5.5}$$

drawn through the data in Figure 5.2. This is called the ***estimated regression line*** or the ***fitted regression line.*** For any data point such as (X_i, Y_i), this line decomposes the total value of Y_i into two parts: $\hat{\beta}_0 + \hat{\beta}_1 X_i$, which is the height of the line above X_i, and e_i, which is the vertical distance from the line to the data point. The first part is the ***fitted*** or ***predicted value*** for Y:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \tag{5.6}$$

The second part, e_i, is called the ***residual,*** or the ***error*** of fit or prediction:

$$e_i = Y_i - \hat{Y}_i \tag{5.7}$$

Clearly,

$$Y_i = \hat{Y}_i + e_i \tag{5.8}$$

How can we determine the parameters $\hat{\beta}_0$ and $\hat{\beta}_1$ of the estimated regression line? It seems reasonable to try to find the line that best fits the scatter of data. There are a variety of alternative techniques that might be used, and we consider several before focusing on the one we adopt.

One possibility is just to draw a line that seems to fit pretty well and accept that. In some cases this might be satisfactory, but it is not very precise. Two different persons analyzing the same data would undoubtedly come up with different estimates of the parameters, and we would have no basis for deciding which were better.

Another possibility is to measure the perpendicular distance from each point to the fitted line and then develop a method that calculates values for the parameters so as to minimize some overall measure of these distances. The results might be called "perpendicular estimators."

Instead of dealing with the perpendicular distances, another possibility is to measure the vertical distances from each point to the fitted line. These distances are the residuals, defined by (5.7). We might try to find a line for which the sum of the residuals is zero. It turns out that many lines satisfy this criterion, and some of them definitely do not fit very well. Alternatively, we might try to minimize the sum of the absolute values of the residuals. This is reasonable but somewhat awkward.

The standard approach in much practical work, which is the approach we adopt, is called the method of ***ordinary least squares*** (OLS). With this method, the criterion for being the best fit is that the line must make the ***sum of the squared residuals*** (*SSR*) as small as possible:

$$\text{OLS criterion:} \quad \text{minimize } SSR = \sum_{i=1}^{n} e_i^2 \tag{5.9}$$

Based on this criterion, we can develop mathematical rules or formulas for calculating $\hat{\beta}_0$ and $\hat{\beta}_1$.

Putting this more formally, suppose that we have n observations on Y and X, as illustrated in Figure 5.2. Any line drawn through the data will be of the form (5.5), and associated with particular $\hat{\beta}_0$ and $\hat{\beta}_1$ values are a set of residuals, e_i, that are determined by

$$e_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i \tag{5.10}$$

Consider the sum of the squared residuals around a fitted line:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 \tag{5.11}$$

For a given set of data the X_i and Y_i are specific numbers, and our interest is in finding the $\hat{\beta}_0$ and $\hat{\beta}_1$ values that minimize this expression.

A calculus derivation, given in the appendix to this chapter, leads us to the following:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})Y_i}{\sum_{i=1}^{n} (X_i - \overline{X})^2} \tag{5.12}$$

and

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{X} \qquad (5.13)$$

These are known as the OLS ***estimators*** of β_1 and β_0: they are the formulas for calculating the $\hat{\beta}_1$ and $\hat{\beta}_0$ of the estimated regression line. In practice we calculate $\hat{\beta}_1$ first, and then using that value we calculate $\hat{\beta}_0$. Whenever we use these estimators we can be confident of getting the best-fitting line.

In the derivation of the OLS estimators, the only special assumption made is that not all the X values are identical. If this were the case, the denominator in (5.12) would be zero and $\hat{\beta}_1$ would be undefined. Graphically, this would be a case in which all the data points lie along a vertical line.

Three properties of the least squares fit always hold true. First, the sum of the residuals is exactly zero: $\Sigma\, e_i = 0$. Thus the average error of fit is also zero: $\bar{e} = 0$. This property holds true even though the sum of the squared residuals, which has been minimized, is some positive amount. Interestingly, the least squares line is not the only line for which the associated sum of residuals is zero; one could construct a very poor fitting line for which this property also holds.

Second, the fitted regression line goes through the ***point of means,*** which is the point $(\overline{X}, \overline{Y})$. This property makes it easy to graph the line: one point on the estimated line is its intercept with the vertical axis $(0, \hat{\beta}_0)$ and now we know that a second point is $(\overline{X}, \overline{Y})$. We can graph the line through these two points. It should be noted that usually none of the observations in the data lie at the point of means, although this could happen by coincidence.

Third, there is zero correlation between the residuals and the explanatory variable. We already know that the average residual is zero. This third property assures us that for observations with X above $\overline{X}$, there is no tendency for the residuals to average differently from zero—and similarly for $X < \overline{X}$. To see what this means, think of all the data points as lying close to some straight line. The zero correlation between e and X assures us that the OLS fit cuts through the points rather than across them.

As an example of computation, consider the five observations on Y and X in Table 5.1. These are the first five observations on *SAVING* (Y) and *INCOME* (X) from our cross-section data set, with the values rounded to one decimal position for simplicity. As discussed above, the simple regression model (5.1) can be taken as a theoretical statement describing family saving behavior. The calculations are shown in the table, and the estimated regression is reported as

$$\hat{Y}_i = -0.0386 + 0.0863X_i \qquad (5.14)$$

The data are plotted in a scatter diagram in Figure 5.3, and the fitted regression line is drawn in.

TABLE 5.1 Calculation of Regression Estimates

(1) Y_i	(2) X_i	(3) $X_i - \bar{X}$	(4) $(X_i - \bar{X})^2$	(5) $(X_i - \bar{X})Y_i$
0.0	1.9	−5.04	25.4016	0.0
0.9	12.4	5.46	29.8116	4.914
0.4	6.4	−0.54	0.2916	−0.216
1.2	7.0	0.06	0.0036	0.072
0.3	7.0	0.06	0.0036	0.018
2.8	34.7	0.0	55.5120	4.788

$$\bar{X} = \Sigma X_i/n = 34.7/5 = 6.94$$

$$\bar{Y} = \Sigma Y_i/n = 2.8/5 = 0.56$$

$$\hat{\beta}_1 = [\Sigma (X_i - \bar{X})Y_i]/[\Sigma (X_i - \bar{X})^2] = 4.788/55.512 = 0.0863$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X} = 0.56 - (0.0863)(6.94) = -0.0386$$

5.3 Interpretation of the Regression

The econometric approach to analyzing data is based on the theory that there is some stable process of economic behavior that underlies the data we have. We use the data to learn about the process. In our work, the systematic part of the process is specified by the true regression, and calculating the best-fitting line through the data yields the estimated regression.

Consider how the actual estimated regression line compares with the theoretical true regression line. Suppose that we have a set of n observations on Y and X that were generated by an economic process correctly described by the simple regression model (5.1). The technique of ordinary least squares provides

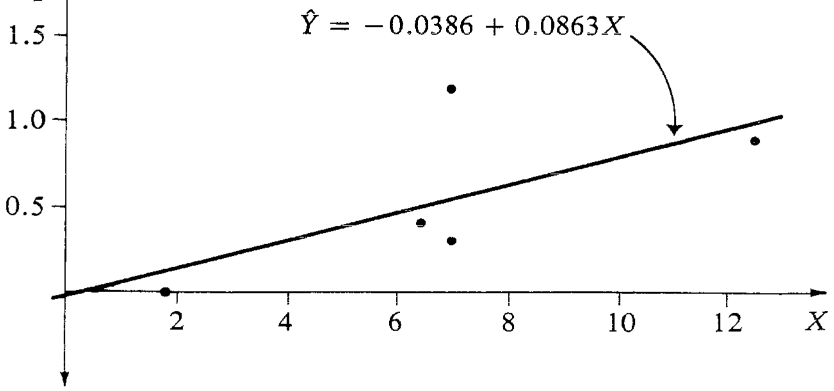

FIGURE 5.3 The scatter diagram displays the five observations on *SAVING* (Y) and *INCOME* (X) from Table 5.1, and the estimated regression line is drawn in. Compare this with Figure 3.4.

a method for calculating estimates of β_0 and β_1. Letting asterisks denote the actually computed values in a set of data, the estimates are $\hat{\beta}_0^*$ and $\hat{\beta}_1^*$.

How is $\hat{\beta}_0^*$ related to β_0 and how is $\hat{\beta}_1^*$ related to β_1? We might hope at first that $\hat{\beta}_0^* = \beta_0$ and $\hat{\beta}_1^* = \beta_1$, so that we would have discovered the true values, but this is unlikely ever to occur. We could hardly expect to learn the precise value of the coefficients in the process (5.1) on the basis of a limited set of data generated by the process.

This thinking is reflected in Figure 5.4. Suppose that we are dealing with three observations, having actual X values X_1, X_2, and X_3. We do not explore why the observations take on these values; even in theory we accept them as given. We theorize that the actual Y values for the observations are produced by the process (5.1). Accordingly, in Figure 5.4 we graph the true regression line, which is the systematic part of the process, and we also plot the actual data points resulting from the process. Given these data, the estimated regression line is determined by the OLS estimators, and it is drawn in the figure also.

For the data in Figure 5.4, the estimated line differs from the theoretical true regression line because of the particular pattern taken on by the disturbances: u_1 is large and positive, while u_2 and u_3 are relatively small. This causes the estimated regression line to have a flatter slope than the true one, in this case. Note that while $\Sigma\, e_i = 0$ for the OLS fit, it is usually true that the disturbances do not sum exactly to zero because they reflect uncontrolled random factors.

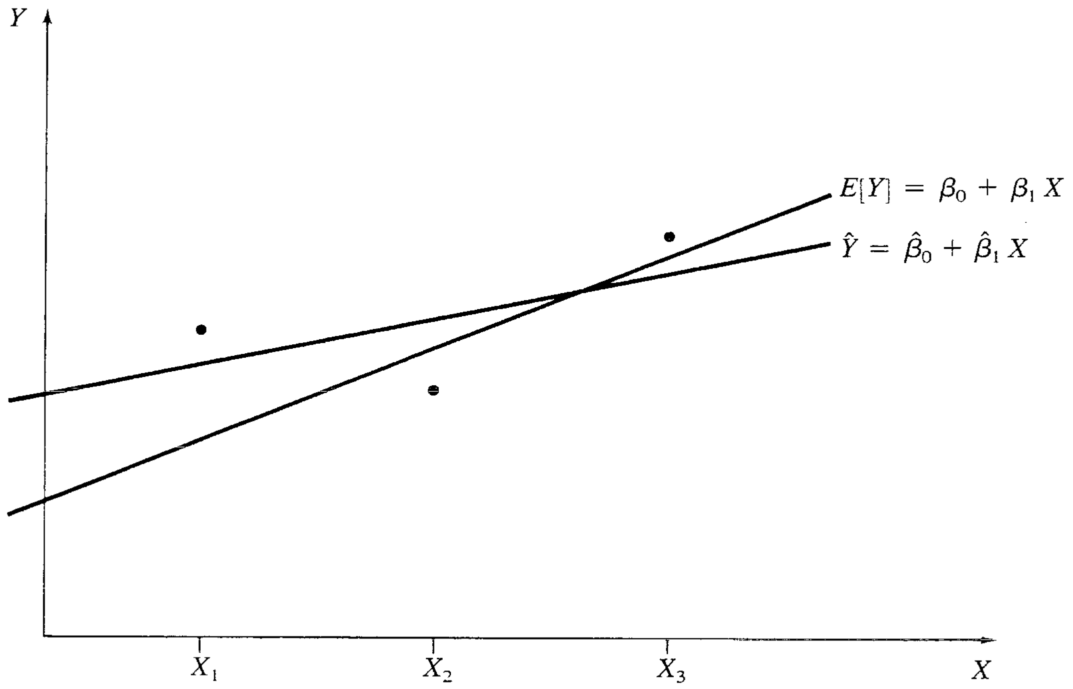
Y
$E[Y] = \beta_0 + \beta_1 X$
$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$
X_1
X_2
X_3
X

FIGURE 5.4 For the given set of X values, the Y values are determined around the true regression line as in Figure 5.1, with a large positive disturbance u_1 in this case. For these data on Y and X, the estimated regression line is also shown, as in Figure 5.2. The difference between the true regression and the estimated regression arises from the particular pattern of values taken on by the disturbances.

In general, the differences between an estimated regression line and the underlying true regression line arise from the particular pattern of values that the disturbances happen to take on. The differences between $\hat{\beta}_0^*$ and β_0, and between $\hat{\beta}_1^*$ and β_1, are estimation errors. Since the true values of β_0 and β_1 are unknown, we cannot determine the size of these estimation errors in any particular case. Of course, we always hope that the errors are small, but we can never be sure that they are. (In Part IV we develop methods of statistical inference that permit us to assess how large or small the estimation errors might be.)

What if we apply this method of estimating a regression line to data from some process that is not correctly described by the simple regression model? For example, suppose that the true process relating Y to X is

$$Y_i = X_i^{\beta_1} + u_i \tag{5.15}$$

We will explore specifications similar to this in Chapter 6, but the essential point here is that the systematic part of the relation between Y and X is not linear. If we apply our OLS methods directly to estimate a regular linear regression, our results clearly cannot provide us with a correct understanding of the true process determining Y. In general, the usefulness and validity of a regression model depends on its being an accurate specification of the process being examined.

To foreshadow the analysis that we carry out in various applications in Chapter 6, let us look at the estimated regression (5.14) of family saving on family income. Remember that as an example of computation we used only five observations, so that these numerical results can hardly be considered seriously.

Just as the theoretical true regression line underlying the data determines the systematic part of saving for a family with a given income, the estimated regression line can be used to make predictions of saving for a family with some specified income. A family whose income is 8 thousand dollars has a predicted saving of

$$\hat{Y}_i = -0.0386 + (0.0863)(8.0) = 0.6518 \tag{5.16}$$

thousand dollars (i.e., about \$652). Notice that in predicting a level for the dependent variable we are finding the height of the estimated regression line corresponding to a specific level of X.

The estimated intercept and slope values are just that: estimates of the corresponding parameters of the behavioral relation. Special attention focuses on the slope here, because the economics of the behavior is that β_1 is the marginal propensity to save. Our estimate of this parameter is 0.0863, and in a serious study we would hope that this is not too far from the true value.

Looking at the estimated slope, we see that if a family's income were to increase by 1 thousand dollars (i.e., $\Delta X = 1$) its predicted saving would increase by 0.0863 thousand dollars (i.e., \$86.30). This is a basic interpretation that we can make without any computation. By contrast, if a family's income were to increase by \$2500 (i.e., $\Delta X = 2.5$), we use our knowledge of slopes (see the appendix to Chapter 1) to compute

$$\Delta\hat{Y} = 0.0863\Delta X = (0.0863)(2.5) = 0.21575 \qquad (5.17)$$

thousand dollars. That is, predicted saving increases by about \$216. Notice that the intercept plays no role in a computation like this.

Looking at the estimated intercept, we see that if a family's income were zero, its predicted saving would be -0.0386 thousand dollars (i.e., minus \$38.60). Although this may seem odd at first, it does make economic sense: many families with temporarily low incomes do have some accumulated wealth that they would spend if their income were zero, and decreases in wealth are measured as negative saving.

5.4 Measures of Goodness of Fit

The technique of ordinary least squares guarantees that the estimated regression line is the best-fitting line that can be drawn through the data, in the sense that it has the smallest possible sum of squared residuals. Although it is the best-fitting line, whether we would judge that it fits well or not so well depends on the data: if the data in the scatterplot are widely dispersed, no line can fit very well; if the data seem to lie close to some line, a good-fitting line can be found. In this section we develop two statistics that allow us to quantify how well the regression fits.

Starting from a set of data on Y and X, the estimated regression line yields a set of fitted values, or predictions, for the actual Y_i values. These are given by

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \qquad (5.18)$$

The associated errors of fit are given by the residuals

$$e_i = Y_i - \hat{Y}_i \qquad (5.19)$$

These residuals serve as the basis for our two measures of goodness of fit.

The Standard Error of the Regression, *SER*

The n residuals constitute a data variable, e, that can be described by the methods of Chapter 3. As noted above, it is a property of OLS estimation that the mean residual is always zero: $\bar{e} = 0$. The standard deviation of the residuals (S_e) will be some positive number, and its usual interpretation will still be valid: S_e measures the typical deviation of e from its mean, without regard to sign.

Since $\bar{e} = 0$, each deviation $(e_i - \bar{e})$ is the same as the value of the variable (e_i) itself. Hence the standard deviation can be interpreted here as the typical value of the variable, without regard to sign. Thus S_e could serve as an interesting measure of how well the regression fits the data: it answers the question, "What is the typical error of fit?" To satisfy some purposes discussed in Chapter 11, the actual measure we adopt is a slight modification of S_e.

The ***standard error of regression*** (*SER*), which is sometimes called the "standard error of estimate," is defined by

$$SER = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n - 2}} \tag{5.20}$$

and it gives the typical error of fit. (Notice that if $n - 2$ were replaced by $n - 1$, the expression on the right-hand side would equal S_e.) As with the standard deviation, the denominator in this expression is identified as the number of degrees of freedom.

To illustrate how the *SER* is computed, we continue with the previous example dealing with family saving behavior. Table 5.2 starts with the same data on Y and X as Table 5.1. For each observation, separately, the fitted or predicted value $\hat{Y}_i$ is determined from the estimated regression (5.14) by substituting in the X_i value. Then, again for each observation separately, the residual (5.19) is calculated and squared. The sum of these squares is $\Sigma\ e_i^2$. The *SER* is then calculated directly according to (5.20).

The units of measurement of the *SER* are always the same as those of the dependent variable Y because each residual is equal to the actual value Y_i minus the fitted value $\hat{Y}_i$. In this case the *SER* is 0.416 thousand dollars, because family saving (Y) is measured in those units. We interpret this value as the typical error of fit for the regression of family saving on family income.

To assess its magnitude, the *SER* is usually compared with some value of the dependent variable. If we are making one prediction, a comparison of the *SER* with the $\hat{Y}_i$ value suggests how much in error the prediction might be. (A probabilistic analysis of prediction errors is given in Chapter 13.) If we are

TABLE 5.2 Calculation of R^2 and *SER*

(1) Y_i	(2) X_i	(3) $\hat{Y}_i$	(4) e_i	(5) e_i^2	(6) $Y_i - \overline{Y}$	(7) $(Y_i - \overline{Y})^2$
0.0	1.9	0.126	−0.125	0.0157	−0.56	0.3136
0.9	12.4	1.031	−0.131	0.0171	0.34	0.1156
0.4	6.4	0.513	−0.113	0.0129	−0.16	0.0256
1.2	7.0	0.565	0.635	0.4030	0.64	0.4096
0.3	7.0	0.565	−0.265	0.0703	−0.26	0.0676
			0.0	0.5190	0.0	0.9320

$$R^2 = 1 - \Sigma\ e_i^2/[\Sigma\ (Y_i - \overline{Y})^2] = 1 - 0.5190/0.9320 = 0.443$$

$$SER = \sqrt{\Sigma\ e_i^2/(n - 2)} = \sqrt{0.5190/3} = 0.416$$

considering the overall goodness of fit of the regression, a comparison of the *SER* with the mean of Y is useful. For example, in the saving regression, the *SER* is \$416 and $\overline{Y}$ is \$560; the typical error of fit is relatively large, indicating a fairly poor fit. However, simple comparisons with the mean are sometimes misleading; if the same residuals had been obtained with families having greater income and saving levels (and thus greater $\overline{Y}$), the *SER* would not seem so "relatively large."

As evident in Table 5.2 and Figure 5.3, one of the residuals is much larger than the others in absolute value. This, combined with the small size of the sample, leads the *SER* to be larger than four of the five residuals. (The mean absolute residual is 0.254 here, which is substantially smaller than the *SER*.) Because of this, we might hesitate before accepting the conclusion that the fit is "fairly poor."

In practical applications of simple regression, we sometimes end up estimating two or more regressions having the same dependent variable. For example, if we have two sets of data on the same variables, we might estimate two separate regressions of Y on X. Or if we have three variables in single data set, we might estimate one regression of Y on X and another regression of Y on Z. In comparing the two regressions, it makes sense to say that the one with the smaller *SER* has the better fit.

By contrast, it generally does not make sense to compare the *SER*s of two regressions when they have different dependent variables. This can lead to illogical comparisons. For example, if we have three variables in a single data set, we might estimate one regression of Y on X and another regression of Z on X. In each regression the computed *SER* measures the typical error of fit. But if Y is measured in dollars and Z is measured in percentage points, we clearly cannot compare the *SER*s to judge which regression has the better fit.

The Coefficient of Determination, R^2

The second measure we develop to quantify how well the estimated regression line fits the data yields a pure, dimensionless number. Like the magnitude of the correlation, this measure varies between zero and 1, with a higher value indicating a better fit. We go through several steps of a formal development because these are useful for understanding the interpretation we use.

Making reference to Figure 5.5, it should be clear that for any observation i,

$$(Y_i - \overline{Y}) = (\hat{Y}_i - \overline{Y}) + (Y_i - \hat{Y}_i) = (\hat{Y}_i - \overline{Y}) + e_i \tag{5.21}$$

This is a decomposition, showing that for each observation the total deviation of Y_i from the mean is equal to the deviation of the regression's predicted value from the mean plus the regression's error of fit. In other words, the ***total deviation*** of Y_i from $\overline{Y}$ is equal to the ***deviation explained by the regression*** plus the ***unexplained deviation.***

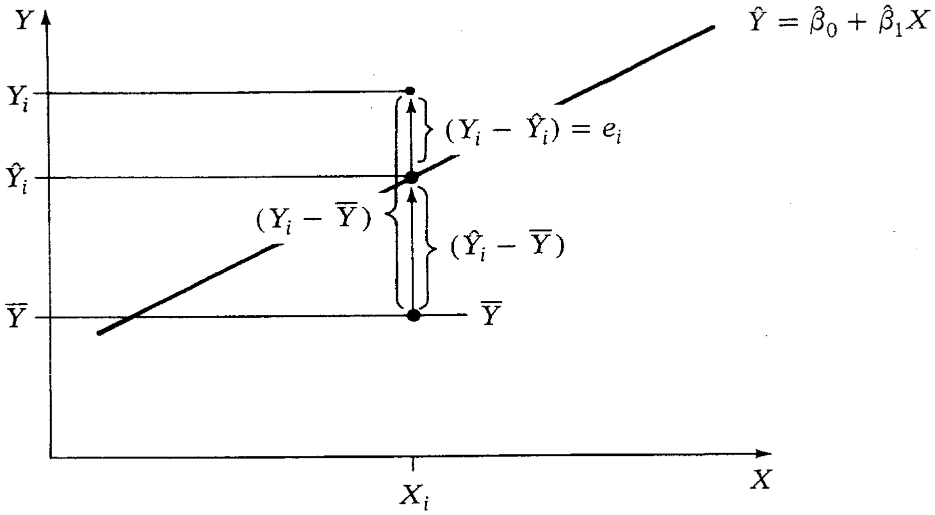
$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$
Y
Y_i
$\hat{Y}_i$
$\overline{Y}$
$(Y_i - \hat{Y}_i) = e_i$
$(Y_i - \overline{Y})$
$(\hat{Y}_i - \overline{Y})$
$\overline{Y}$
X_i
X

FIGURE 5.5 For any observation, the total deviation of Y_i from the mean, $(Y_i - \overline{Y})$, can be decomposed into two parts: the unexplained deviation $(Y_i - \hat{Y}_i)$, which is just the residual e_i, and the explained deviation $(\hat{Y}_i - \overline{Y})$. The coefficient of determination, R^2, is defined as the ratio of an overall measure of the explained deviations to an overall measure of the total deviations. The standard error of the regression, SER, is a measure of the typical unexplained deviation—that is, it is a measure of the typical error of fit, e_i.

If we square the leftmost and rightmost portions of (5.21), we get

$$(Y_i - \overline{Y})^2 = (\hat{Y}_i - \overline{Y})^2 + e_i^2 + 2e_i(\hat{Y}_i - \overline{Y}) \tag{5.22}$$

and if we add up the n equations like (5.22) that hold for each i, we get

$$\sum (Y_i - \overline{Y})^2 = \sum (\hat{Y}_i - \overline{Y})^2 + \sum e_i^2 + 2 \sum e_i(\hat{Y}_i - \overline{Y}) \tag{5.23}$$

It can be shown that the last term on the right-hand side of (5.23) is equal to zero, so that what remains is

$$\sum (Y_i - \overline{Y})^2 = \sum (\hat{Y}_i - \overline{Y})^2 + \sum e_i^2 \tag{5.24}$$

All the terms in this equation are positive or zero because they are sums of squares. The expression on the left-hand side is called the ***total variation*** of Y, because it is the sum of squares of the total deviations identified above. Similarly, the first expression on the right-hand side is called the ***explained variation*** and the second is called the ***unexplained variation.*** Thus, as a consequence of the original decomposition illustrated in Figure 5.5, we can say that the total variation of Y is equal to the variation that is explained by the regression plus the unexplained variation.

Rearranging and dividing by $\Sigma\ (Y_i - \overline{Y})^2$, we get

$$1 - \frac{\sum e_i^2}{\sum (Y_i - \overline{Y})^2} = \frac{\sum (\hat{Y}_i - \overline{Y})^2}{\sum (Y_i - \overline{Y})^2} \tag{5.25}$$

The right-hand side of this equation is the ratio of the explained variation to the total variation. The numerator and the denominator are positive, and since the numerator is a component of the denominator the ratio can take on values only between zero and 1. This ratio could serve as a measure of goodness of fit, and we adopt it for that purpose. The left-hand side of (5.25) is often easier to compute. In part for that reason, we use the left-hand side in the definition but give it the interpretation that applies directly to the right-hand side.

The ***coefficient of determination,*** which is usually denoted by R^2 and read as "R-squared," is defined by

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2} \tag{5.26}$$

Based on the analysis of (5.25), the general form of our interpretation of R^2 is that it measures "the proportion of the total variation of Y that is explained by the regression." Since the regression model explains, or predicts, the values of Y on the basis of the given values of X, we might say instead that R^2 measures "the proportion of the total variation of Y that is explained by X."

The computation of R^2 is illustrated in Table 5.2. In addition to calculating $\Sigma\, e_i^2$, which was done for the *SER,* we need to calculate the total variation of Y, $\Sigma\,(Y_i - \bar{Y})^2$, as an intermediate step. Finding $R^2 = .443$, we say that the regression explains about 44 percent of the variation of Y.

In the appendix to this chapter it is proved that R^2 is equal to the square of r, the correlation between X and Y. Thus these two measures serve in the same way to quantify how well the estimated regression line fits the data. In econometric analysis we usually talk in terms of R^2 rather than r, partly because of its useful interpretation in terms of explaining the total variation of Y.

The R^2 is a zero-to-1 dimensionless measure of goodness of fit, and knowing the numerical value in any case helps us conjure up an image of the scatterplot in our minds. An R^2 of .443 is usually associated with a scatterplot in which the data are fairly dispersed around the estimated regression. In our example, we might qualify this description by noting that one residual is much larger than the others.

Our interpretation of the magnitude of R^2 also depends on the nature of the economic process being analyzed. The R^2 is often high in time-series work because Y and X often have a common trend. By contrast, the R^2 tends to be lower in cross-section work because there is no trend and because of the substantial natural variation in individual behavior. If $R^2 = .443$ were reported for a macroeconomic time-series saving function, it would be judged quite low: experienced researchers expect a regression of aggregate saving on income to have an R^2 of .95 or higher. However, an R^2 of .443 for a large-sample, cross-section saving function would be quite high compared with similar studies, and the regression would be judged to have a relatively good fit.

The R^2 can be used in a limited way to compare the fits of two or more regressions. In comparing regressions, it is natural to note which one is the most successful in explaining the variation in its dependent variable. However, these comparisons must be made with care, especially when the regressions have different dependent variables. A better fit does not necessarily mean a better regression.

For any given set of data and the corresponding estimated regression, R^2 and *SER* are useful measures of goodness of fit. It is common to report both along with the estimated regression in a display like

$$\begin{aligned} \hat{Y}_i &= -0.0386 + 0.0863X_i \\ R^2 &= .443 \qquad SER = 0.416 \end{aligned} \tag{5.27}$$

The R^2 tells us that about 44 percent of the variation in family saving is explained by the regression. The *SER* tells us that the typical error of fit for saving is 0.416 thousand dollars (i.e., \$416). In Chapter 11 we develop additional calculations in regression estimation that will be added to this reporting style.

5.5 The Effects of Linear Transformation

The units in which we measure variables are arbitrarily chosen. For example, *GNP* for 1980 might be measured as 1,480,700,000,000 dollars, 1,480.7 billion dollars, or 1.4807 trillion dollars. What impact does this choice have on our regression results? We can show that the economic sense of the fitted regression does not depend on the units of measurement, even though the actual regression coefficients do.

Suppose that we start with data on Y and X. The most common form of units adjustment involves multiplying a variable by a constant. We view this as transforming the original variables, Y and X, into new variables, y and x, according to

$$y_i = b_y Y_i \quad \text{and} \quad x_i = b_x X_i \tag{5.28}$$

For example, if Y is the unemployment rate expressed as a proportion of the labor force and if b_y equals 100, y is the unemployment rate expressed in percentage points.

Using the original data, let the estimated regression of Y on X be denoted by

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \tag{5.29}$$

And using the transformed data, let the estimated regression of y on x be denoted by

$$\hat{y}_i = \hat{\gamma}_0 + \hat{\gamma}_1 x_i \tag{5.30}$$

The choice of notation here makes it clear that the estimated intercepts may be different in the two regressions and that the estimated slopes may be different also. (Note that γ is lowercase "gamma," the Greek "g.")

The question is: what are the relations between $\hat{\gamma}_0$ and $\hat{\beta}_0$ and between

$\hat{\gamma}_1$ and $\hat{\beta}_1$? In other words, how do the new coefficients compare with the old ones? To answer this, we start with the formulas for the OLS estimators giving $\hat{\gamma}_0$ and $\hat{\gamma}_1$ in terms of y and x, then substitute for y and x what they are in terms of Y and X, and finally rearrange until a useful statement is found. Our results from Chapter 3 regarding transformations of variables are important here. We see that

$$\begin{aligned}\hat{\gamma}_1 &= \frac{\sum (x_i - \bar{x})y_i}{\sum (x_i - \bar{x})^2} \\ &= \frac{\sum (b_x X_i - b_x\bar{X})b_y Y}{\sum (b_x X_i - b_x\bar{X})^2} \\ &= \frac{b_x b_y \sum (X_i - \bar{X})Y_i}{(b_x)^2 \sum (X_i - \bar{X})^2} \\ &= \frac{b_y}{b_x}\hat{\beta}_1\end{aligned} \tag{5.31}$$

and

$$\begin{aligned}\hat{\gamma}_0 &= \bar{y} - \hat{\gamma}_1\bar{x} \\ &= b_y\bar{Y} - \left(\frac{b_y}{b_x}\hat{\beta}_1\right)(b_x\bar{X}) \\ &= b_y(\bar{Y} - \hat{\beta}_1\bar{X}) \\ &= b_y\hat{\beta}_0\end{aligned} \tag{5.32}$$

Collecting these results, we see that when Y and X are transformed by multiplicative constants as specified in (5.28), the new coefficient estimates are related to the original ones by (5.31) and (5.32).

For example, the saving (Y) and income (X) data used in this chapter are measured in thousands of dollars. If both variables were transformed to express the amounts in dollars, the new variables would be

$$y_i = 1000Y_i \quad \text{and} \quad x_i = 1000X_i \tag{5.33}$$

From (5.27), (5.31), and (5.32), we see quickly that the new regression of saving on income would be

$$\hat{y}_i = -38.6 + 0.0863x_i \tag{5.34}$$

That is, the slope is unchanged but the new intercept is 1000 times greater than the old one. By contrast, if saving were transformed to dollars while income remained in the original form, the regression of saving on income would be

$$\hat{y}_i = -38.6 + 86.3X_i \tag{5.35}$$

The R^2 in these regressions would be the same as in (5.27), but in this case the original *SER* would be multiplied by a factor of 1000.

The question of units choice can be generalized somewhat to the following problem. Suppose that we have data on Y and X and have estimated the regression

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \tag{5.36}$$

Suppose now that we create a new variable, y, from Y according to the linear transformation

$$y_i = a_y + b_y Y_i \tag{5.37}$$

and a new variable, x, according to

$$x_i = a_x + b_x X_i \tag{5.38}$$

(Note that a_y and b_y need not be related to a_x and b_x at all; the notation is used just to economize on symbols.) If we estimate an OLS regression of y on x, we get

$$\hat{y}_i = \hat{\gamma}_0 + \hat{\gamma}_1 x_i \tag{5.39}$$

and we ask: How do the new coefficients compare with the original ones? Derivations similar to those above lead to

$$\hat{\gamma}_1 = \frac{b_y}{b_x}\hat{\beta}_1 \tag{5.40}$$

and

$$\hat{\gamma}_0 = a_y + b_y\hat{\beta}_0 - \frac{a_x b_y}{b_x}\hat{\beta}_1 \tag{5.41}$$

For example, suppose that after-tax income (x) is related to before-tax income (X), both expressed in thousands of dollars, by

$$x_i = 2 + 0.75X_i \tag{5.42}$$

Comparing this to (5.38), we have $a_x = 2$ and $b_x = 0.75$. Suppose now that we are interested in the relation between saving and after-tax income. The saving concept is unchanged: $y_i = Y_i$, so $a_y = 0$ and $b_y = 1$. Without estimating another regression, we can determine from (5.27) that the regression of saving on after-tax income is

$$\begin{aligned}\hat{y}_i &= \left[0 + (1)(-0.0386) - \frac{(2)(1)}{0.75}(0.0863)\right] \\ &\quad + \left[\frac{1}{0.75}(0.0863)\right] x_i \\ &= -0.269 + 0.115x_i\end{aligned} \tag{5.43}$$

The R^2 of the original and new regressions would be the same; in this case the *SER* is also unchanged because the units of measurement for saving, the dependent variable, are unchanged.

Understanding the effects of units adjustment and linear transformation is useful for us in two ways. First, given an estimated regression, we are able to restate it to make more sensible reports without actually computing new estimates. For example, it generally does not make sense to report a regression of saving on income with one variable measured in dollars and the other in thousands of dollars.

Second, this analysis makes us aware that the magnitude of estimated regression coefficients depend in part on the units of measurement that happen to have been used in the data. Without knowing what the units are in any particular regression, finding a coefficient equal to 3400.0 is no more interesting or important than finding 0.00034 as the estimated value.

Problems

Section 5.1

5.1 In a simple linear regression model of market demand involving the quantity demanded and the price of a product, which variable is the dependent one and which is the explanatory one? Draw a figure showing the true regression line and plot some of the observations that might be observed.

⋆ **5.2** In the simple regression model illustrated in Figure 5.1, is it possible that all the actual Y_i values would lie above the true regression line? Explain.

5.3 Suppose that the quantity demanded in a market depends on the price of the product and the income of consumers. Would a simple regression model explaining the quantity demanded be appropriate? Could income be considered part of the disturbance?

5.4 Suppose that the true regression line is $E[Y] = 2 + 3X$. Determine the value of the disturbance for an observation having (X_i, Y_i) values (3, 8). Determine the disturbance for an observation (6, 21).

Section 5.2

5.5 In an OLS regression estimation, is it possible that all the actual Y_i values would lie above the estimated regression line? Explain.

⋆ **5.6** Construct a diagram to show a case for which $\Sigma\, e_i = 0$ around a line that is clearly not the best-fitting line.

5.7 In the following table Y stands for *EARNS* and X stands for *ED* from the cross-section data set, and the data are for observations 26 through

30 (with *EARNS* rounded). Plot Y and X in a scatter diagram, with Y on the vertical axis.

Y	X
12.0	12
3.6	8
9.6	10
3.7	3
6.5	12

5.8 Based on the data of Problem 5.7, estimate the coefficients of the OLS regression of Y on X and graph the estimated regression line through the scatter diagram of Problem 5.7. (Use a table format to organize your calculations.)

5.9 For a data set of three observations whose (X_i, Y_i) values are (10, 5), (8, 7), and (12, 9), estimate the regression of Y on X. Calculate the sum of the residuals and the covariance between e and X. Verify that the regression goes through the point of means.

5.10 Calculate the sum of squared residuals for the estimated regression in Problem 5.9. Now, add 0.5 to the intercept to get another line through the data, and calculate the sum of squared residuals for this line. Which of the two calculated *SSR*s is smaller? Why?

Section 5.3

★ **5.11** Draw a scatter diagram of the first five observations on *EARNS* and *ED* from the cross-section data set, and roughly draw in the best-fitting line. How does comparing this to the graph from Problem 5.8 illustrate the existence of estimation errors?

5.12 Based on the estimated regression (5.14), determine the predicted saving of a family whose income is 25 thousand dollars. For what level of income would predicted saving be zero?

Section 5.4

★ **5.13** Based on Problem 5.8, compute the values of R^2 and *SER*.

5.14 OLS finds the best-fitting line, while R^2 measures the goodness of fit. Does this mean that R^2 will always be high if the OLS technique is used?

5.15 Prove that the third term in Equation (5.23) is equal to zero. [*Hint:* At one stage use the fact that $\Sigma\, e_i X_i = 0$, which stems from Equation (5.48).]

5.16 Compute the values of R^2 and *SER* for the regression estimated in Problem 5.9.

5.17 In Figure 5.3 and Table 5.2, the observation with the large residual could be called an "outlier." Recompute the estimated regression, the

R^2, and the *SER* after deleting this outlier (i.e., using only four observations) and compare the results to those gathered in Equation (5.27).

Section 5.5

⋆ **5.18** Based on Problems 5.7 and 5.8, suppose that *EARNS* is transformed into dollars (from thousands of dollars). What would be the new estimated regression of earnings on *ED*?

5.19 Based on the saving function presented in Equation (5.27), what would be the estimated regression of saving on income if income were transformed to dollars but saving remained in the original form?

5.20 Derive the relations presented as Equations (5.40) and (5.41).

Appendix

5.21 Prove that Equations (5.54) and (5.53) are equivalent. [*Hint:* Start from the numerator in (5.53).]

5.22 Show why Equation (5.48) is equivalent to stating that e and X are uncorrelated.

⋆ **5.23** Consider a regression model in which β_0 is specified to be zero: $Y_i = \beta_1 X_i + u_i$. Derive the OLS estimator for β_1.

APPENDIX*

Derivation of the OLS Estimators

The derivation of the OLS estimators is a standard calculus minimization problem in which the objective function to be minimized is the sum of squared residuals [see (5.9)]. Given a set of data, all the values like Y_i and X_i are fixed numbers. Our task is to find the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that make *SSR* as small as possible. In this context, $\hat{\beta}_0$ and $\hat{\beta}_1$ are variables (arguments of the function) while the Y's and X's are constants.

The sum of squared residuals is given by

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 \qquad (5.44)$$

To find the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize this expression, we take the partial derivatives of (5.44) with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$ and set them equal to zero:

$$\frac{\partial}{\partial \hat{\beta}_0}\left[\sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2\right] = 0 \qquad (5.45)$$

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

and

$$\frac{\partial}{\partial\hat{\beta}_1}\left[\sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2\right] = 0 \tag{5.46}$$

Evaluating these partial derivatives gives us

$$-2\sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0 \tag{5.47}$$

$$-2\sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)X_i = 0 \tag{5.48}$$

Next, we divide each equation by -2, leaving each side equal to zero, and then rearrange terms to get

$$\sum Y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum X_i \tag{5.49}$$

$$\sum Y_i X_i = \hat{\beta}_0 \sum X_i + \hat{\beta}_1 \sum X_i^2 \tag{5.50}$$

This pair of equations is a set of two simultaneous equations in which $\hat{\beta}_0$ and $\hat{\beta}_1$ are unknown and all the X_i and Y_i values are known.

A convenient way to solve these equations is to solve (5.49) for

$$\hat{\beta}_0 = \left\{\sum Y_i - \hat{\beta}_1 \sum X_i\right\} /n \tag{5.51}$$

and substitute this for $\hat{\beta}_0$ in (5.50). That equation then can be solved to yield

$$\hat{\beta}_1 = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - \left(\sum X_i\right)^2} \tag{5.52}$$

For theoretical and computational convenience, we note that this expression can be arranged in various ways, including

$$\hat{\beta}_1 = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2} \tag{5.53}$$

and

$$\hat{\beta}_1 = \frac{\sum (X_i - \overline{X})Y_i}{\sum (X_i - \overline{X})^2} \tag{5.54}$$

which is the same as (5.12). Now (5.51) can be manipulated to yield

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} \tag{5.55}$$

which is the same as (5.13). We could find instead an expression that gives $\hat{\beta}_0$ solely in terms of X and Y values, but this is not useful.

The property that $\Sigma\, e_i = 0$ follows from (5.47) and that e and X are uncorrelated follows from (5.48).

Equivalence of R^2 and $(r)^2$

The value of R^2 is exactly equal to the square of r, the correlation between X and Y. To see this, first note that

$$\begin{aligned}\sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2 &= \sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 X_i - \hat{\beta}_0 - \hat{\beta}_1 \overline{X})^2 \\ &= \sum_{i=1}^{n} [\hat{\beta}_1 (X_i - \overline{X})]^2 \qquad (5.56) \\ &= (\hat{\beta}_1)^2 \sum_{i=1}^{n} (X_i - \overline{X})^2\end{aligned}$$

Now,

$$\begin{aligned}R^2 &= \frac{\sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{n} (Y_i - \overline{Y})^2} \\ &= \frac{(\hat{\beta}_1)^2 \sum (X_i - \overline{X})^2}{\sum (Y_i - \overline{Y})^2} = (\hat{\beta}_1)^2 \frac{\sum (X_i - \overline{X})^2}{\sum (Y_i - \overline{Y})^2} \\ &= \left[\frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2}\right]^2 \frac{\sum (X_i - \overline{X})^2}{\sum (Y_i - \overline{Y})^2} \qquad (5.57) \\ &= \frac{\left[\sum (X_i - \overline{X})(Y_i - \overline{Y})\right]^2}{\sum (X_i - \overline{X})^2 \sum (Y_i - \overline{Y})^2} \\ &= \left[\frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum (X_i - \overline{X})^2}\sqrt{\sum (Y_i - \overline{Y})^2}}\right]^2 \\ &= (r)^2\end{aligned}$$

Since we showed earlier that the maximum value of R^2 is 1 and the minimum is 0, the equivalence of R^2 and $(r)^2$ proves that the maximum value of r is 1 and the minimum is -1.