3

TWO-VARIABLE REGRESSION MODEL: THE PROBLEM OF ESTIMATION

As noted in Chapter 2, our first task is to estimate the population regression function (PRF) on the basis of the sample regression function (SRF) as accurately as possible. In **Appendix A** we have discussed two generally used methods of estimation: (1) **ordinary least squares (OLS)** and (2) **maximum likelihood (ML).** By and large, it is the method of OLS that is used extensively in regression analysis primarily because it is intuitively appealing and mathematically much simpler than the method of maximum likelihood. Besides, as we will show later, in the linear regression context the two methods generally give similar results.

3.1 THE METHOD OF ORDINARY LEAST SQUARES

The method of ordinary least squares is attributed to Carl Friedrich Gauss, a German mathematician. Under certain assumptions (discussed in Section 3.2), the method of least squares has some very attractive statistical properties that have made it one of the most powerful and popular methods of regression analysis. To understand this method, we first explain the least-squares principle.

Recall the two-variable PRF:

$$Y_i = \beta_1 + \beta_2 X_i + u_i \qquad \textbf{(2.4.2)}$$

However, as we noted in Chapter 2, the PRF is not directly observable. We estimate it from the SRF:

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i \tag{2.6.2}$$

$$= \hat{Y}_i + \hat{u}_i \tag{2.6.3}$$

where $\hat{Y}_i$ is the estimated (conditional mean) value of Y_i.

But how is the SRF itself determined? To see this, let us proceed as follows. First, express (2.6.3) as

$$\begin{aligned} \hat{u}_i &= Y_i - \hat{Y}_i \\ &= Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i \end{aligned} \tag{3.1.1}$$

which shows that the $\hat{u}_i$ (the residuals) are simply the differences between the actual and estimated Y values.

Now given n pairs of observations on Y and X, we would like to determine the SRF in such a manner that it is as close as possible to the actual Y. To this end, we may adopt the following criterion: Choose the SRF in such a way that the sum of the residuals $\sum \hat{u}_i = \sum (Y_i - \hat{Y}_i)$ is as small as possible. Although intuitively appealing, this is not a very good criterion, as can be seen in the hypothetical scattergram shown in Figure 3.1.

If we adopt the criterion of minimizing $\sum \hat{u}_i$, Figure 3.1 shows that the residuals $\hat{u}_2$ and $\hat{u}_3$ as well as the residuals $\hat{u}_1$ and $\hat{u}_4$ receive the same weight in the sum $(\hat{u}_1 + \hat{u}_2 + \hat{u}_3 + \hat{u}_4)$, although the first two residuals are much closer to the SRF than the latter two. In other words, all the residuals receive

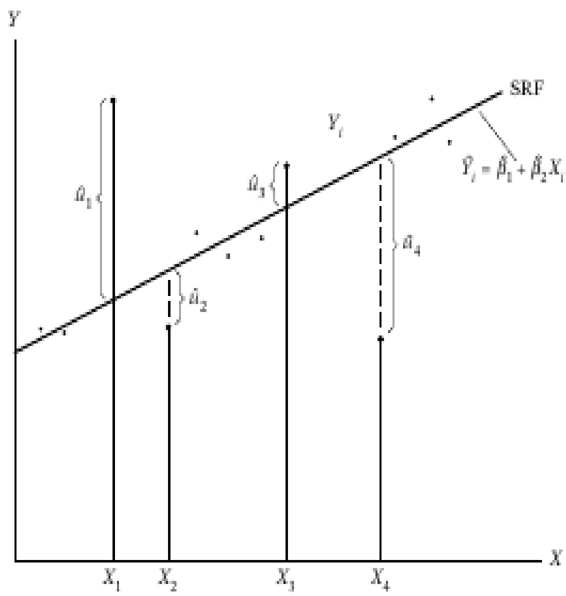

FIGURE 3.1 Least-squares criterion.

equal importance no matter how close or how widely scattered the individual observations are from the SRF. A consequence of this is that it is quite possible that the algebraic sum of the $\hat{u}_i$ is small (even zero) although the $\hat{u}_i$ are widely scattered about the SRF. To see this, let $\hat{u}_1, \hat{u}_2, \hat{u}_3$, and $\hat{u}_4$ in Figure 3.1 assume the values of 10, −2, +2, and −10, respectively. The algebraic sum of these residuals is zero although $\hat{u}_1$ and $\hat{u}_4$ are scattered more widely around the SRF than $\hat{u}_2$ and $\hat{u}_3$. We can avoid this problem if we adopt the *least-squares criterion*, which states that the SRF can be fixed in such a way that

$$\begin{aligned}\sum \hat{u}_i^2 &= \sum (Y_i - \hat{Y}_i)^2 \\ &= \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2\end{aligned} \tag{3.1.2}$$

is as small as possible, where $\hat{u}_i^2$ are the squared residuals. By squaring $\hat{u}_i$, this method gives more weight to residuals such as $\hat{u}_1$ and $\hat{u}_4$ in Figure 3.1 than the residuals $\hat{u}_2$ and $\hat{u}_3$. As noted previously, under the minimum $\sum \hat{u}_i$ criterion, the sum can be small even though the $\hat{u}_i$ are widely spread about the SRF. But this is not possible under the least-squares procedure, for the larger the $\hat{u}_i$ (in absolute value), the larger the $\sum \hat{u}_i^2$. A further justification for the least-squares method lies in the fact that the estimators obtained by it have some very desirable statistical properties, as we shall see shortly.

It is obvious from (3.1.2) that

$$\sum \hat{u}_i^2 = f(\hat{\beta}_1, \hat{\beta}_2) \tag{3.1.3}$$

that is, the sum of the squared residuals is some function of the estimators $\hat{\beta}_1$ and $\hat{\beta}_2$. For any given set of data, choosing different values for $\hat{\beta}_1$ and $\hat{\beta}_2$ will give different $\hat{u}$'s and hence different values of $\sum \hat{u}_i^2$. To see this clearly, consider the hypothetical data on Y and X given in the first two columns of Table 3.1. Let us now conduct two experiments. In experiment 1,

TABLE 3.1 EXPERIMENTAL DETERMINATION OF THE SRF

Y_i (1)	X_i (2)	$\hat{Y}_{1i}$ (3)	$\hat{u}_{1i}$ (4)	$\hat{u}_{1i}^2$ (5)	$\hat{Y}_{2i}$ (6)	$\hat{u}_{2i}$ (7)	$\hat{u}_{2i}^2$ (8)
4	1	2.929	1.071	1.147	4	0	0
5	4	7.000	−2.000	4.000	7	−2	4
7	5	8.357	−1.357	1.841	8	−1	1
12	6	9.714	2.286	5.226	9	3	9
Sum: 28	16		0.0	12.214		0	14

Notes: $\hat{Y}_{1i} = 1.572 + 1.357X_i$ (i.e., $\hat{\beta}_1 = 1.572$ and $\hat{\beta}_2 = 1.357$)
$\hat{Y}_{2i} = 3.0 + 1.0X_i$ (i.e., $\hat{\beta}_1 = 3$ and $\hat{\beta}_2 = 1.0$)
$\hat{u}_{1i} = (Y_i - \hat{Y}_{1i})$
$\hat{u}_{2i} = (Y_i - \hat{Y}_{2i})$

let $\hat{\beta}_1 = 1.572$ and $\hat{\beta}_2 = 1.357$ (let us not worry right now about how we got these values; say, it is just a guess).[1] Using these $\hat{\beta}$ values and the X values given in column (2) of Table 3.1, we can easily compute the estimated Y_i given in column (3) of the table as $\hat{Y}_{1i}$ (the subscript 1 is to denote the first experiment). Now let us conduct another experiment, but this time using the values of $\hat{\beta}_1 = 3$ and $\hat{\beta}_2 = 1$. The estimated values of Y_i from this experiment are given as $\hat{Y}_{2i}$ in column (6) of Table 3.1. Since the $\hat{\beta}$ values in the two experiments are different, we get different values for the estimated residuals, as shown in the table; $\hat{u}_{1i}$ are the residuals from the first experiment and $\hat{u}_{2i}$ from the second experiment. The squares of these residuals are given in columns (5) and (8). Obviously, as expected from (3.1.3), these residual sums of squares are different since they are based on different sets of $\hat{\beta}$ values.

Now which sets of $\hat{\beta}$ values should we choose? Since the $\hat{\beta}$ values of the first experiment give us a lower $\sum \hat{u}_i^2$ (= 12.214) than that obtained from the $\hat{\beta}$ values of the second experiment (= 14), we might say that the $\hat{\beta}$'s of the first experiment are the "best" values. But how do we know? For, if we had infinite time and infinite patience, we could have conducted many more such experiments, choosing different sets of $\hat{\beta}$'s each time and comparing the resulting $\sum \hat{u}_i^2$ and then choosing that set of $\hat{\beta}$ values that gives us the least possible value of $\sum \hat{u}_i^2$ assuming of course that we have considered all the conceivable values of β_1 and β_2. But since time, and certainly patience, are generally in short supply, we need to consider some shortcuts to this trial-and-error process. Fortunately, the method of least squares provides us such a shortcut. The principle or the method of least squares chooses $\hat{\beta}_1$ and $\hat{\beta}_2$ in such a manner that, for a given sample or set of data, $\sum \hat{u}_i^2$ is as small as possible. In other words, for a given sample, the method of least squares provides us with unique estimates of β_1 and β_2 that give the smallest possible value of $\sum \hat{u}_i^2$. How is this accomplished? This is a straight-forward exercise in differential calculus. As shown in Appendix 3A, Section 3A.1, the process of differentiation yields the following equations for estimating β_1 and β_2:

$$\sum Y_i = n\hat{\beta}_1 + \hat{\beta}_2 \sum X_i \qquad \textbf{(3.1.4)}$$

$$\sum Y_i X_i = \hat{\beta}_1 \sum X_i + \hat{\beta}_2 \sum X_i^2 \qquad \textbf{(3.1.5)}$$

where n is the sample size. These simultaneous equations are known as the **normal equations.**

[1] For the curious, these values are obtained by the method of least squares, discussed shortly. See Eqs. (3.1.6) and (3.1.7).

Solving the normal equations simultaneously, we obtain

$$\begin{aligned}\hat{\beta}_2 &= \frac{n\sum X_iY_i - \sum X_i \sum Y_i}{n\sum X_i^2 - \left(\sum X_i\right)^2} \\ &= \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - X)^2} \\ &= \frac{\sum x_i y_i}{\sum x_i^2}\end{aligned} \tag{3.1.6}$$

where $\bar{X}$ and $\bar{Y}$ are the sample means of X and Y and where we define $x_i = (X_i - \bar{X})$ and $y_i = (Y_i - \bar{Y})$. *Henceforth we adopt the convention of letting the lowercase letters denote deviations from mean values.*

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum X_i^2 \sum Y_i - \sum X_i \sum X_iY_i}{n\sum X_i^2 - \left(\sum X_i\right)^2} \\ &= \bar{Y} - \hat{\beta}_2\bar{X}\end{aligned} \tag{3.1.7}$$

The last step in (3.1.7) can be obtained directly from (3.1.4) by simple algebraic manipulations.

Incidentally, note that, by making use of simple algebraic identities, formula (3.1.6) for estimating β_2 can be alternatively expressed as

$$\begin{aligned}\hat{\beta}_2 &= \frac{\sum x_i y_i}{\sum x_i^2} \\ &= \frac{\sum x_i Y_i}{\sum X_i^2 - n\bar{X}^2} \\ &= \frac{\sum X_i y_i}{\sum X_i^2 - n\bar{X}^2}\end{aligned} \tag{3.1.8}$$ [2]

The estimators obtained previously are known as the **least-squares estimators,** for they are derived from the least-squares principle. Note the following **numerical properties** of estimators obtained by the method of OLS: "Numerical properties are those that hold as a consequence of the use

[2] *Note 1:* $\sum x_i^2 = \sum (X_i - \bar{X})^2 = \sum X_i^2 - 2\sum X_i\bar{X} + \sum \bar{X}^2 = \sum X_i^2 - 2\bar{X}\sum X_i + \sum \bar{X}^2$, since $\bar{X}$ is a constant. Further noting that $\sum X_i = n\bar{X}$ and $\sum \bar{X}^2 = n\bar{X}^2$ since $\bar{X}$ is a constant, we finally get $\sum x_i^2 = \sum X_i^2 - n\bar{X}^2$.

Note 2: $\sum x_i y_i = \sum x_i(Y_i - \bar{Y}) = \sum x_iY_i - \bar{Y}\sum x_i = \sum x_iY_i - \bar{Y}\sum (X_i - \bar{X}) = \sum x_iY_i$, since $\bar{Y}$ is a constant and since the sum of deviations of a variable from its mean value [e.g., $\sum (X_i - \bar{X})$] is always zero. Likewise, $\sum y_i = \sum (Y_i - \bar{Y}) = 0$.

of ordinary least squares, regardless of how the data were generated."[3] Shortly, we will also consider the **statistical properties** of OLS estimators, that is, properties "that hold only under certain assumptions about the way the data were generated."[4] (See the classical linear regression model in Section 3.2.)

I. The OLS estimators are expressed solely in terms of the observable (i.e., sample) quantities (i.e., X and Y). Therefore, they can be easily computed.

II. They are **point estimators;** that is, given the sample, each estimator will provide only a single (point) value of the relevant population parameter. (In Chapter 5 we will consider the so-called **interval estimators,** which provide a range of possible values for the unknown population parameters.)

III. Once the OLS estimates are obtained from the sample data, the sample regression line (Figure 3.1) can be easily obtained. The regression line thus obtained has the following properties:

1. It passes through the sample means of Y and X. This fact is obvious from (3.1.7), for the latter can be written as $\bar{Y} = \hat{\beta}_1 + \hat{\beta}_2\bar{X}$, which is shown diagrammatically in Figure 3.2.

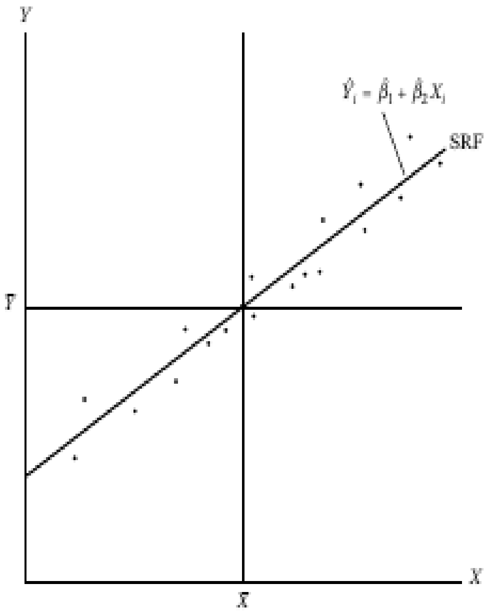

FIGURE 3.2 Diagram showing that the sample regression line passes through the sample mean values of Y and X.

[3]Russell Davidson and James G. MacKinnon, *Estimation and Inference in Econometrics,* Oxford University Press, New York, 1993, p. 3.
[4]*Ibid.*

2. The mean value of the estimated $Y = \hat{Y}_i$ is equal to the mean value of the actual Y for

$$\begin{aligned}\hat{Y}_i &= \hat{\beta}_1 + \hat{\beta}_2 X_i \\ &= (\bar{Y} - \hat{\beta}_2 \bar{X}) + \hat{\beta}_2 X_i \\ &= \bar{Y} + \hat{\beta}_2 (X_i - \bar{X})\end{aligned} \tag{3.1.9}$$

Summing both sides of this last equality over the sample values and dividing through by the sample size n gives

$$\bar{\hat{Y}} = \bar{Y} \tag{3.1.10}$$[5]

where use is made of the fact that $\sum(X_i - \bar{X}) = 0$. (Why?)

3. The mean value of the residuals $\hat{u}_i$ is zero. From Appendix 3A, Section 3A.1, the first equation is

$$-2\sum(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) = 0$$

But since $\hat{u}_i = Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i$, the preceding equation reduces to $-2\sum \hat{u}_i = 0$, whence $\bar{\hat{u}} = 0$.[6]

As a result of the preceding property, the sample regression

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i \tag{2.6.2}$$

can be expressed in an alternative form where both Y and X are expressed as deviations from their mean values. To see this, sum (2.6.2) on both sides to give

$$\begin{aligned}\sum Y_i &= n\hat{\beta}_1 + \hat{\beta}_2 \sum X_i + \sum \hat{u}_i \\ &= n\hat{\beta}_1 + \hat{\beta}_2 \sum X_i \qquad \text{since } \sum \hat{u}_i = 0\end{aligned} \tag{3.1.11}$$

Dividing Eq. (3.1.11) through by n, we obtain

$$\bar{Y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{X} \tag{3.1.12}$$

which is the same as (3.1.7). Subtracting Eq. (3.1.12) from (2.6.2), we obtain

$$Y_i - \bar{Y} = \hat{\beta}_2 (X_i - \bar{X}) + \hat{u}_i$$

[5]Note that this result is true only when the regression model has the intercept term β_1 in it. As **App. 6A**, **Sec. 6A.1** shows, this result need not hold when β_1 is absent from the model.

[6]This result also requires that the intercept term β_1 be present in the model (see **App. 6A**, **Sec. 6A.1**).

or

$$y_i = \hat{\beta}_2 x_i + \hat{u}_i \tag{3.1.13}$$

where y_i and x_i, following our convention, are deviations from their respective (sample) mean values.

Equation (3.1.13) is known as the **deviation form.** Notice that the intercept term $\hat{\beta}_1$ is no longer present in it. But the intercept term can always be estimated by (3.1.7), that is, from the fact that the sample regression line passes through the sample means of Y and X. An advantage of the deviation form is that it often simplifies computing formulas.

In passing, note that in the deviation form, the SRF can be written as

$$\hat{y}_i = \hat{\beta}_2 x_i \tag{3.1.14}$$

whereas in the original units of measurement it was $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$, as shown in (2.6.1).

4. The residuals $\hat{u}_i$ are uncorrelated with the predicted Y_i. This statement can be verified as follows: using the deviation form, we can write

$$\begin{aligned}\sum \hat{y}_i \hat{u}_i &= \hat{\beta}_2 \sum x_i \hat{u}_i \\ &= \hat{\beta}_2 \sum x_i (y_i - \hat{\beta}_2 x_i) \\ &= \hat{\beta}_2 \sum x_i y_i - \hat{\beta}_2^2 \sum x_i^2 \\ &= \hat{\beta}_2^2 \sum x_i^2 - \hat{\beta}_2^2 \sum x_i^2 \\ &= 0\end{aligned} \tag{3.1.15}$$

where use is made of the fact that $\hat{\beta}_2 = \sum x_i y_i / \sum x_i^2$.

5. The residuals $\hat{u}_i$ are uncorrelated with X_i; that is, $\sum \hat{u}_i X_i = 0$. This fact follows from Eq. (2) in Appendix 3A, Section 3A.1.