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

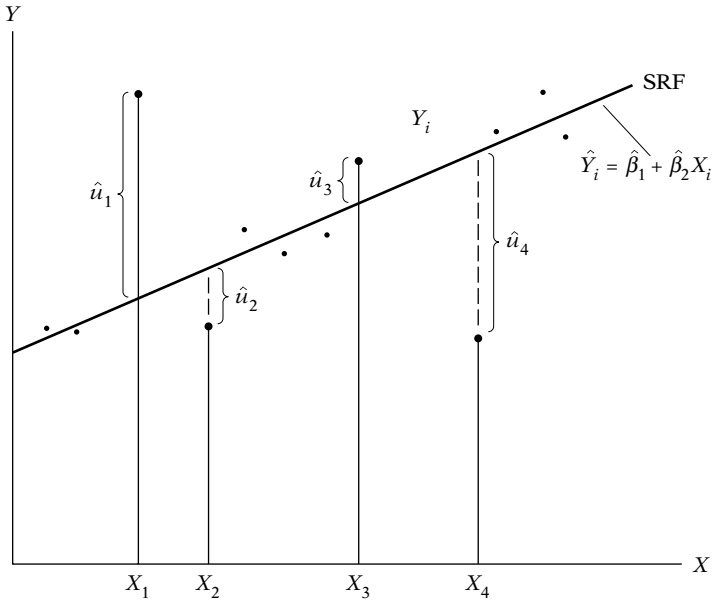


FIGURE 3.1 Least-squares criterion.

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_t (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:

$$\sum Y_i = n\hat{\beta}_1 + \hat{\beta}_2 \sum X_i \tag{3.1.4}$$

$$\sum Y_i X_i = \hat{\beta}_1 \sum X_i + \hat{\beta}_2 \sum X_i^2 \tag{3.1.5}$$

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

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

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.*

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

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: $\sum x_i y_i = \sum x_i (Y_i - \bar{Y}) = \sum x_i Y_i - \bar{Y}\sum x_i = \sum x_i Y_i - \bar{Y}\sum (X_i - \bar{X}) = \sum x_i Y_i$, 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.

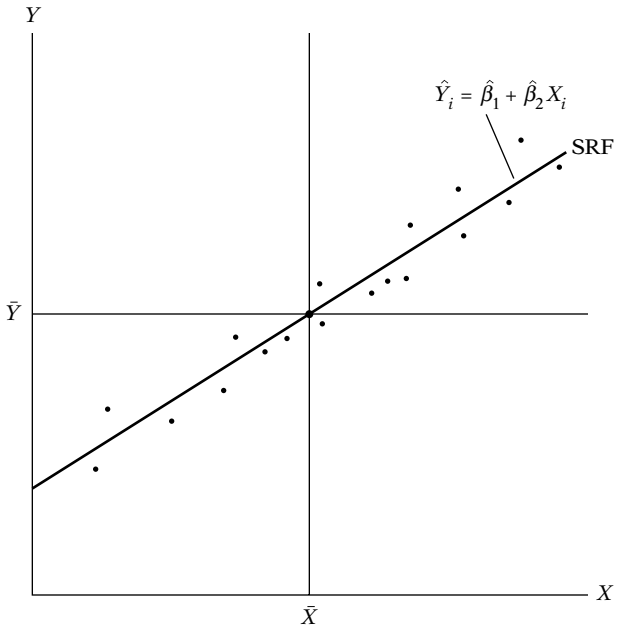


FIGURE 3.2 Diagram showing that the sample regression line passes through the sample mean values of Y and X.

[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.

3.2 THE CLASSICAL LINEAR REGRESSION MODEL: THE ASSUMPTIONS UNDERLYING THE METHOD OF LEAST SQUARES

If our objective is to estimate β_1 and β_2 only, the method of OLS discussed in the preceding section will suffice. But recall from Chapter 2 that in regression analysis our objective is not only to obtain $\hat{\beta}_1$ and $\hat{\beta}_2$ but also to draw inferences about the true β_1 and β_2. For example, we would like to know how close $\hat{\beta}_1$ and $\hat{\beta}_2$ are to their counterparts in the population or how close $\hat{Y}_i$ is to the true $E(Y \mid X_i)$. To that end, we must not only specify the functional form of the model, as in (2.4.2), but also make certain assumptions about

the manner in which Y_i are generated. To see why this requirement is needed, look at the PRF: $Y_i = \beta_1 + \beta_2 X_i + u_i$. It shows that Y_i depends on both X_i and u_i. Therefore, unless we are specific about how X_i and u_i are created or generated, there is no way we can make any statistical inference about the Y_i and also, as we shall see, about β_1 and β_2. Thus, the assumptions made about the X_i variable(s) and the error term are extremely critical to the valid interpretation of the regression estimates.

The Gaussian, standard, or classical linear regression model (CLRM), which is the cornerstone of most econometric theory, makes 10 assumptions.[7] We first discuss these assumptions in the context of the two-variable regression model; and in Chapter 7 we extend them to multiple regression models, that is, models in which there is more than one regressor.

Assumption 1: Linear regression model. The regression model is **linear in the parameters,** as shown in (2.4.2)

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

We already discussed model (2.4.2) in Chapter 2. Since linear-in-parameter regression models are the starting point of the CLRM, we will maintain this assumption throughout this book. Keep in mind that the regressand Y and the regressor X themselves may be nonlinear, as discussed in Chapter 2.[8]

Assumption 2: *X* values are fixed in repeated sampling. Values taken by the regressor X are considered fixed in repeated samples. More technically, X is assumed to be *nonstochastic.*

This assumption is implicit in our discussion of the PRF in Chapter 2. But it is very important to understand the concept of "fixed values in repeated sampling," which can be explained in terms of our example given in Table 2.1. Consider the various Y populations corresponding to the levels of income shown in that table. Keeping the value of income X fixed, say, at level \$80, we draw at random a family and observe its weekly family consumption expenditure Y as, say, \$60. Still keeping X at \$80, we draw at random another family and observe its Y value as \$75. In each of these drawings (i.e., repeated sampling), the value of X is fixed at \$80. We can repeat this process for all the X values shown in Table 2.1. As a matter of fact, the sample data shown in Tables 2.4 and 2.5 were drawn in this fashion.

What all this means is that our regression analysis is **conditional regression analysis,** that is, conditional on the given values of the regressor(s) X.

[7]It is classical in the sense that it was developed first by Gauss in 1821 and since then has served as a norm or a standard against which may be compared the regression models that do not satisfy the Gaussian assumptions.

[8]However, a brief discussion of nonlinear-in-the-parameter regression models is given in Chap. 14.

Assumption 3: Zero mean value of disturbance u_i. Given the value of X, the mean, or expected, value of the random disturbance term u_i is zero. Technically, the conditional mean value of u_i is zero. Symbolically, we have

$$E(u_i \mid X_i) = 0 \qquad \textbf{(3.2.1)}$$

Assumption 3 states that the mean value of u_i, conditional upon the given X_i, is zero. Geometrically, this assumption can be pictured as in Figure 3.3, which shows a few values of the variable X and the Y populations associated with each of them. As shown, each Y population corresponding to a given X is distributed around its mean value (shown by the circled points on the PRF) with some Y values above the mean and some below it. The distances above and below the mean values are nothing but the u_i, and what (3.2.1) requires is that the average or mean value of these deviations corresponding to any given X should be zero.[9]

This assumption should not be difficult to comprehend in view of the discussion in Section 2.4 [see Eq. (2.4.5)]. All that this assumption says is that the factors not explicitly included in the model, and therefore subsumed in u_i, do not systematically affect the mean value of Y; so to speak, the positive u_i

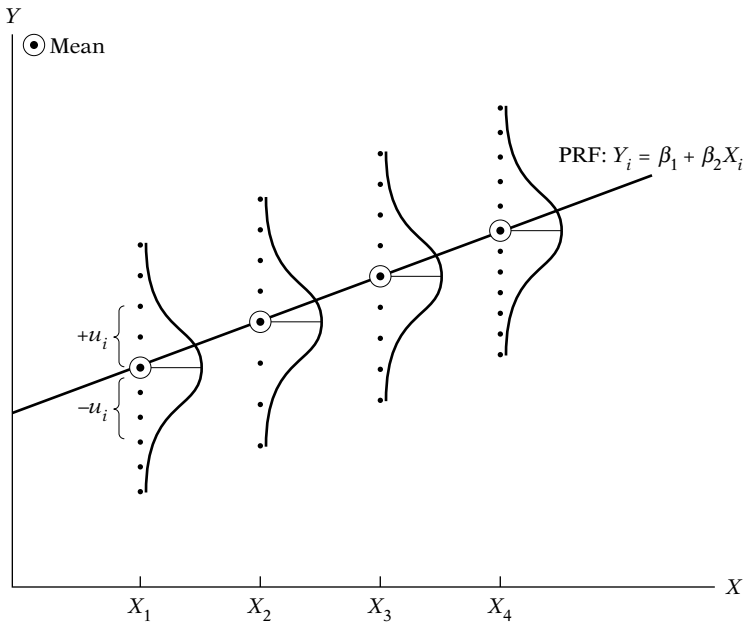


FIGURE 3.3 Conditional distribution of the disturbances u_i.

[9]For illustration, we are assuming merely that the u's are distributed symmetrically as shown in Figure 3.3. But shortly we will assume that the u's are distributed normally.

values cancel out the negative u_i values so that their average or mean effect on Y is zero.[10]

In passing, note that the assumption $E(u_i \mid X_i) = 0$ implies that $E(Y_i \mid X_i) = \beta_i + \beta_2 X_i$. (Why?) Therefore, the two assumptions are equivalent.

Assumption 4: Homoscedasticity or equal variance of u_i. Given the value of X, the variance of u_i is the same for all observations. That is, the conditional variances of u_i are identical. Symbolically, we have

$$\begin{aligned} \textbf{var}\,(u_i \mid X_i) &= E[u_i - E(u_i \mid X_i)]^2 \\ &= E\left(u_i^2 \mid X_i\right) \text{ because of Assumption 3} \\ &= \sigma^2 \end{aligned} \qquad \textbf{(3.2.2)}$$

where **var** stands for variance.

Eq. (3.2.2) states that the variance of u_i for each X_i (i.e., the conditional variance of u_i) is some positive constant number equal to σ^2. Technically, (3.2.2) represents the assumption of **homoscedasticity,** or *equal* (homo) *spread* (*scedasticity*) or *equal variance*. The word comes from the Greek verb *skedanime,* which means to disperse or scatter. Stated differently, (3.2.2) means that the Y populations corresponding to various X values have the same variance. Put simply, the variation around the regression line (which is the line of average relationship between Y and X) is the same across the X values; it neither increases or decreases as X varies. Diagrammatically, the situation is as depicted in Figure 3.4.

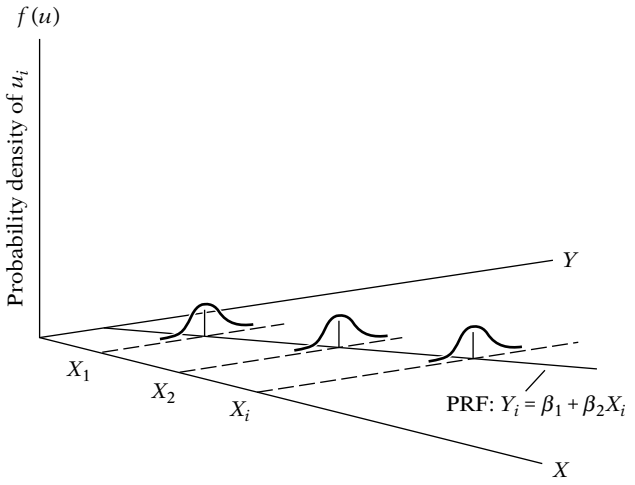


FIGURE 3.4 Homoscedasticity.

[10]For a more technical reason why Assumption 3 is necessary see E. Malinvaud, *Statistical Methods of Econometrics,* Rand McNally, Chicago, 1966, p. 75. See also exercise 3.3.

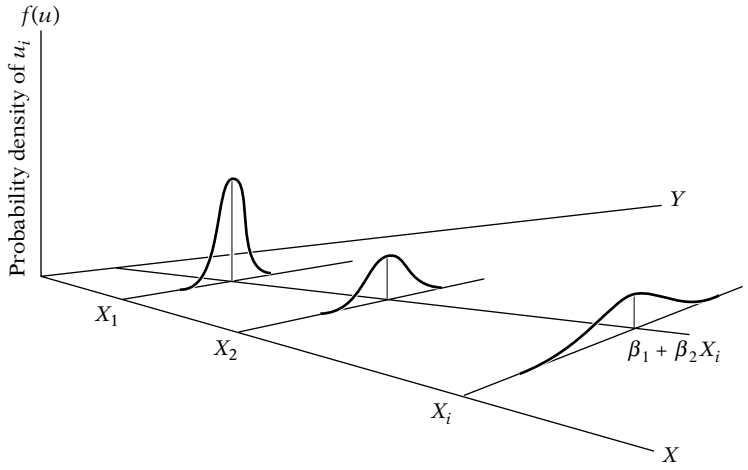


FIGURE 3.5 Heteroscedasticity.

In contrast, consider Figure 3.5, where the conditional variance of the Y population varies with X. This situation is known appropriately as **heteroscedasticity,** or *unequal spread,* or *variance.* Symbolically, in this situation (3.2.2) can be written as

$$\text{var}\,(u_i \mid X_i) = \sigma_i^2 \qquad \textbf{(3.2.3)}$$

Notice the subscript on σ^2 in Eq. (3.2.3), which indicates that the variance of the Y population is no longer constant.

To make the difference between the two situations clear, let Y represent weekly consumption expenditure and X weekly income. Figures 3.4 and 3.5 show that as income increases the average consumption expenditure also increases. But in Figure 3.4 the variance of consumption expenditure remains the same at all levels of income, whereas in Figure 3.5 it increases with increase in income. In other words, richer families on the average consume more than poorer families, but there is also more variability in the consumption expenditure of the former.

To understand the rationale behind this assumption, refer to Figure 3.5. As this figure shows, $\text{var}\,(u \mid X_1) < \text{var}\,(u \mid X_2), \ldots, < \text{var}\,(u \mid X_i)$. Therefore, the likelihood is that the Y observations coming from the population with $X = X_1$ would be closer to the PRF than those coming from populations corresponding to $X = X_2$, $X = X_3$, and so on. In short, not all Y values corresponding to the various X's will be equally reliable, reliability being judged by how closely or distantly the Y values are distributed around their means, that is, the points on the PRF. If this is in fact the case, would we not prefer to sample from those Y populations that are closer to their mean than those that are widely spread? But doing so might restrict the variation we obtain across X values.

By invoking Assumption 4, we are saying that at this stage all Y values corresponding to the various X's are equally important. In Chapter 11 we shall see what happens if this is not the case, that is, where there is heteroscedasticity.

In passing, note that Assumption 4 implies that the conditional variances of Y_i are also homoscedastic. That is,

$$\text{var}\,(Y_i \mid X_i) = \sigma^2 \tag{3.2.4}$$

Of course, the *unconditional variance* of Y is σ_Y^2. Later we will see the importance of distinguishing between conditional and unconditional variances of Y (see Appendix A for details of conditional and unconditional variances).

Assumption 5: No autocorrelation between the disturbances. Given any two X values, X_i and X_j $(i \neq j)$, the correlation between any two u_i and u_j $(i \neq j)$ is zero. Symbolically,

$$\begin{aligned} \textbf{cov}\,(u_i, u_j \mid X_i, X_j) &= E\{[u_i - E(u_i)] \mid X_i\}\{[u_j - E(u_j)] \mid X_j\} \\ &= E(u_i \mid X_i)(u_j \mid X_j) \qquad \text{(why?)} \\ &= 0 \end{aligned} \tag{3.2.5}$$

where i and j are two different observations and where **cov** means **covariance.**

In words, (3.2.5) postulates that the disturbances u_i and u_j are uncorrelated. Technically, this is the assumption of **no serial correlation,** or **no autocorrelation.** This means that, given X_i, the deviations of any two Y values from their mean value do not exhibit patterns such as those shown in Figure 3.6*a* and *b*. In Figure 3.6*a*, we see that the u's are **positively correlated,** a positive u followed by a positive u or a negative u followed by a negative u. In Figure 3.6*b*, the u's are **negatively correlated,** a positive u followed by a negative u and vice versa.

If the disturbances (deviations) follow systematic patterns, such as those shown in Figure 3.6*a* and *b,* there is auto- or serial correlation, and what Assumption 5 requires is that such correlations be absent. Figure 3.6*c* shows that there is no systematic pattern to the u's, thus indicating zero correlation.

The full import of this assumption will be explained thoroughly in Chapter 12. But intuitively one can explain this assumption as follows. Suppose in our PRF $(Y_t = \beta_1 + \beta_2 X_t + u_t)$ that u_t and u_{t-1} are positively correlated. Then Y_t depends not only on X_t but also on u_{t-1} for u_{t-1} to some extent determines u_t. At this stage of the development of the subject matter, by invoking Assumption 5, we are saying that we will consider the systematic effect, if any, of X_t on Y_t and not worry about the other influences that might act on Y as a result of the possible intercorrelations among the u's. But, as noted in Chapter 12, we will see how intercorrelations among the disturbances can be brought into the analysis and with what consequences.

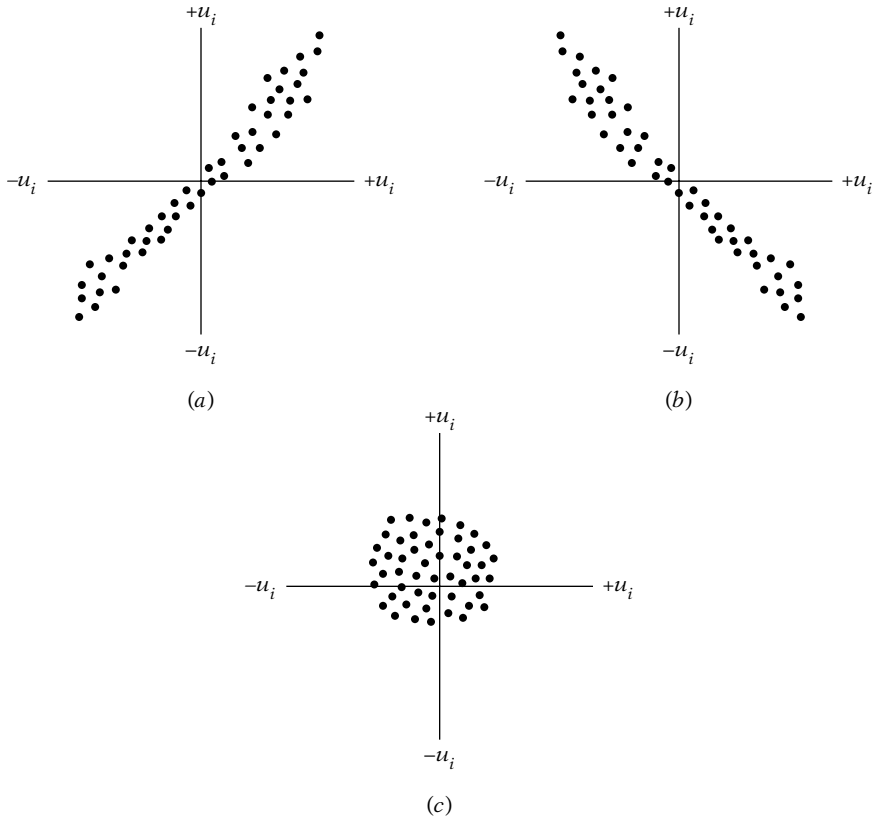


FIGURE 3.6 Patterns of correlation among the disturbances. (*a*) positive serial correlation; (*b*) negative serial correlation; (*c*) zero correlation.

Assumption 6: Zero covariance between u_i and X_i, or $E(u_i X_i) = 0$. Formally,

$$\begin{aligned}
\textbf{cov}\ (u_i, X_i) &= E[u_i - E(u_i)][X_i - E(X_i)] \\
&= E[u_i(X_i - E(X_i))] \qquad \text{since } E(u_i) = 0 \\
&= E(u_i X_i) - E(X_i)E(u_i) \qquad \text{since } E(X_i) \text{ is nonstochastic} \\
&= E(u_i X_i) \qquad \text{since } E(u_i) = 0 \\
&= 0 \qquad \text{by assumption}
\end{aligned} \tag{3.2.6}$$

Assumption 6 states that the disturbance u and explanatory variable X are uncorrelated. The rationale for this assumption is as follows: When we expressed the PRF as in (2.4.2), we assumed that X and u (which may represent the influence of all the omitted variables) have separate (and additive) influence on Y. But if X and u are correlated, it is not possible to assess their individual effects on Y. Thus, if X and u are positively correlated, X increases

when u increases and it decreases when u decreases. Similarly, if X and u are negatively correlated, X increases when u decreases and it decreases when u increases. In either case, it is difficult to isolate the influence of X and u on Y.

Assumption 6 is automatically fulfilled if X variable is nonrandom or nonstochastic and Assumption 3 holds, for in that case, $\text{cov}\,(u_i, X_i) = [X_i - E(X_i)]E[u_i - E(u_i)] = 0$. (Why?) But since we have assumed that our X variable not only is nonstochastic but also assumes fixed values in repeated samples,[11] Assumption 6 is not very critical for us; it is stated here merely to point out that the regression theory presented in the sequel holds true even if the X's are stochastic or random, provided they are independent or at least uncorrelated with the disturbances u_i.[12] (We shall examine the consequences of relaxing Assumption 6 in Part II.)

Assumption 7: The number of observations *n* must be greater than the number of parameters to be estimated. Alternatively, the number of observations *n* must be greater than the number of explanatory variables.

This assumption is not so innocuous as it seems. In the hypothetical example of Table 3.1, imagine that we had only the first pair of observations on Y and X (4 and 1). From this single observation there is no way to estimate the two unknowns, β_1 and β_2. We need at least two pairs of observations to estimate the two unknowns. In a later chapter we will see the critical importance of this assumption.

Assumption 8: Variability in *X* values. The *X* values in a given sample must not all be the same. Technically, var (*X*) must be a finite positive number.[13]

This assumption too is not so innocuous as it looks. Look at Eq. (3.1.6). If all the X values are identical, then $X_i = \bar{X}$ (Why?) and the denominator of that equation will be zero, making it impossible to estimate β_2 and therefore β_1. Intuitively, we readily see why this assumption is important. Looking at

[11] Recall that in obtaining the samples shown in Tables 2.4 and 2.5, we kept the same X values.

[12] As we will discuss in Part II, if the X's are stochastic but distributed independently of u_i, the properties of least estimators discussed shortly continue to hold, but if the stochastic X's are merely uncorrelated with u_i, the properties of OLS estimators hold true only if the sample size is very large. At this stage, however, there is no need to get bogged down with this theoretical point.

[13] The sample variance of X is

$$\text{var}\,(X) = \frac{\sum(X_i - \bar{X})^2}{n-1}$$

where n is sample size.

our family consumption expenditure example in Chapter 2, if there is very little variation in family income, we will not be able to explain much of the variation in the consumption expenditure. The reader should keep in mind that variation in both Y and X is essential to use regression analysis as a research tool. In short, the variables must vary!

Assumption 9: The regression model is correctly specified. Alternatively, there is no **specification bias or error** in the model used in empirical analysis.

As we discussed in the Introduction, the classical econometric methodology assumes implicitly, if not explicitly, that the model used to test an economic theory is "correctly specified." This assumption can be explained informally as follows. An econometric investigation begins with the specification of the econometric model underlying the phenomenon of interest. Some important questions that arise in the specification of the model include the following: (1) What variables should be included in the model? (2) What is the functional form of the model? Is it linear in the parameters, the variables, or both? (3) What are the probabilistic assumptions made about the Y_i, the X_i, and the u_i entering the model?

These are extremely important questions, for, as we will show in Chapter 13, by omitting important variables from the model, or by choosing the wrong functional form, or by making wrong stochastic assumptions about the variables of the model, the validity of interpreting the estimated regression will be highly questionable. To get an intuitive feeling about this, refer to the Phillips curve shown in Figure 1.3. Suppose we choose the following two models to depict the underlying relationship between the rate of change of money wages and the unemployment rate:

$$Y_i = \alpha_1 + \alpha_2 X_i + u_i \tag{3.2.7}$$

$$Y_i = \beta_1 + \beta_2 \left(\frac{1}{X_i}\right) + u_i \tag{3.2.8}$$

where Y_i = the rate of change of money wages, and X_i = the unemployment rate.

The regression model (3.2.7) is linear both in the parameters and the variables, whereas (3.2.8) is linear in the parameters (hence a linear regression model by our definition) but nonlinear in the variable X. Now consider Figure 3.7.

If model (3.2.8) is the "correct" or the "true" model, fitting the model (3.2.7) to the scatterpoints shown in Figure 3.7 will give us wrong predictions: Between points A and B, for any given X_i the model (3.2.7) is going to overestimate the true mean value of Y, whereas to the left of A (or to the right of B) it is going to underestimate (or overestimate, in absolute terms) the true mean value of Y.

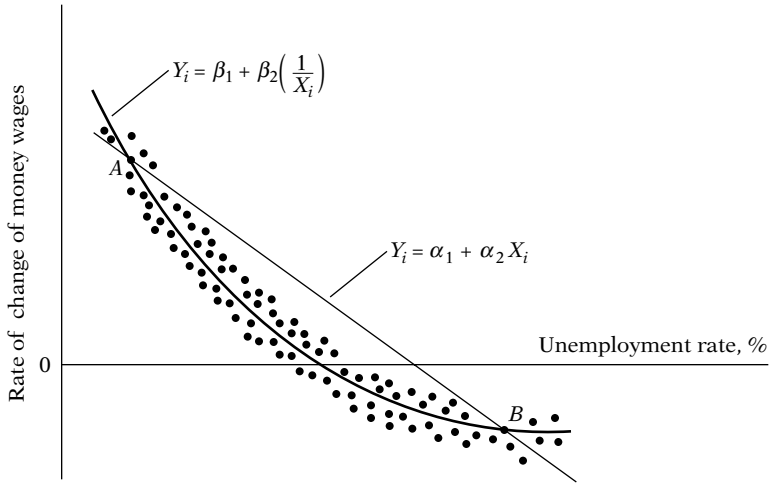


FIGURE 3.7 Linear and nonlinear Phillips curves.

The preceding example is an instance of what is called a **specification bias** or a **specification error;** here the bias consists in choosing the wrong functional form. We will see other types of specification errors in Chapter 13.

Unfortunately, in practice one rarely knows the correct variables to include in the model or the correct functional form of the model or the correct probabilistic assumptions about the variables entering the model for the theory underlying the particular investigation (e.g., the Phillips-type money wage change–unemployment rate tradeoff) may not be strong or robust enough to answer all these questions. Therefore, in practice, the econometrician has to use some judgment in choosing the number of variables entering the model and the functional form of the model and has to make some assumptions about the stochastic nature of the variables included in the model. To some extent, there is some trial and error involved in choosing the "right" model for empirical analysis.[14]

If judgment is required in selecting a model, what is the need for Assumption 9? Without going into details here (see Chapter 13), this assumption is there to remind us that our regression analysis and therefore the results based on that analysis are conditional upon the chosen model and to warn us that we should give very careful thought in formulating econometric

[14]But one should avoid what is known as **"data mining,"** that is, trying every possible model with the hope that at least one will fit the data well. That is why it is essential that there be some economic reasoning underlying the chosen model and that any modifications in the model should have some economic justification. A purely ad hoc model may be difficult to justify on theoretical or a priori grounds. In short, theory should be the basis of estimation. But we will have more to say about data mining in Chap. 13, for there are some who argue that in some situations data mining can serve a useful purpose.

models, especially when there may be several competing theories trying to explain an economic phenomenon, such as the inflation rate, or the demand for money, or the determination of the appropriate or equilibrium value of a stock or a bond. *Thus, econometric model-building, as we shall discover, is more often an art rather than a science.*

Our discussion of the assumptions underlying the classical linear regression model is now completed. It is important to note that all these assumptions pertain to the PRF only and not the SRF. But it is interesting to observe that the method of least squares discussed previously has some properties that are similar to the assumptions we have made about the PRF. For example, the finding that $\sum \hat{u}_i = 0$, and, therefore, $\bar{\hat{u}} = 0$, is akin to the assumption that $E(u_i \mid X_i) = 0$. Likewise, the finding that $\sum \hat{u}_i X_i = 0$ is similar to the assumption that $\text{cov}(u_i, X_i) = 0$. It is comforting to note that the method of least squares thus tries to "duplicate" some of the assumptions we have imposed on the PRF.

Of course, the SRF does not duplicate all the assumptions of the CLRM. As we will show later, although $\text{cov}(u_i, u_j) = 0 \, (i \neq j)$ by assumption, it is *not* true that the *sample* $\text{cov}(\hat{u}_i, \hat{u}_j) = 0 \, (i \neq j)$. As a matter of fact, we will show later that the residuals not only are autocorrelated but also are heteroscedastic (see Chapter 12).

When we go beyond the two-variable model and consider multiple regression models, that is, models containing several regressors, we add the following assumption.

> **Assumption 10: There is no perfect multicollinearity.** That is, there are *no perfect linear relationships among the explanatory variables.*

We will discuss this assumption in Chapter 7, where we discuss multiple regression models.

A Word about These Assumptions

The million-dollar question is: How realistic are all these assumptions? The "reality of assumptions" is an age-old question in the philosophy of science. Some argue that it does not matter whether the assumptions are realistic. What matters are the predictions based on those assumptions. Notable among the "irrelevance-of-assumptions thesis" is Milton Friedman. To him, unreality of assumptions is a positive advantage: "to be important . . . a hypothesis must be descriptively false in its assumptions."[15]

One may not subscribe to this viewpoint fully, but recall that in any scientific study we make certain assumptions because they facilitate the

[15]Milton Friedman, *Essays in Positive Economics,* University of Chicago Press, Chicago, 1953, p. 14.

development of the subject matter in gradual steps, not because they are necessarily realistic in the sense that they replicate reality exactly. As one author notes, ". . . if simplicity is a desirable criterion of good theory, all good theories idealize and oversimplify outrageously."[16]

What we plan to do is first study the properties of the CLRM thoroughly, and then in later chapters examine in depth what happens if one or more of the assumptions of CLRM are not fulfilled. At the end of this chapter, we provide in Table 3.4 a guide to where one can find out what happens to the CLRM if a particular assumption is not satisfied.

As a colleague pointed out to me, when we review research done by others, we need to consider whether the assumptions made by the researcher are appropriate to the data and problem. All too often, published research is based on implicit assumptions about problem and data that are likely not correct and that produce estimates based on these assumptions. Clearly, the knowledgeable reader should, realizing these problems, adopt a skeptical attitude toward the research. The assumptions listed in Table 3.4 therefore provide a checklist for guiding our research and for evaluating the research of others.

With this backdrop, we are now ready to study the CLRM. In particular, we want to find out the **statistical properties** of OLS compared with the purely **numerical properties** discussed earlier. The statistical properties of OLS are based on the assumptions of CLRM already discussed and are enshrined in the famous **Gauss–Markov theorem.** But before we turn to this theorem, which provides the theoretical justification for the popularity of OLS, we first need to consider the **precision** or **standard errors** of the least-squares estimates.

3.3 PRECISION OR STANDARD ERRORS OF LEAST-SQUARES ESTIMATES

From Eqs. (3.1.6) and (3.1.7), it is evident that least-squares estimates are a function of the sample data. But since the data are likely to change from sample to sample, the estimates will change ipso facto. Therefore, what is needed is some measure of "reliability" or **precision** of the estimators $\hat{\beta}_1$ and $\hat{\beta}_2$. In statistics the precision of an estimate is measured by its standard error (se).[17] Given the Gaussian assumptions, it is shown in Appendix 3A, Section 3A.3 that the standard errors of the OLS estimates can be obtained

[16]Mark Blaug, *The Methodology of Economics: Or How Economists Explain,* 2d ed., Cambridge University Press, New York, 1992, p. 92.

[17]The **standard error** is nothing but the standard deviation of the sampling distribution of the estimator, and the sampling distribution of an estimator is simply a probability or frequency distribution of the estimator, that is, a distribution of the set of values of the estimator obtained from all possible samples of the same size from a given population. Sampling distributions are used to draw inferences about the values of the population parameters on the basis of the values of the estimators calculated from one or more samples. (For details, see **App. A**.)

as follows:

$$\text{var}\,(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_i^2} \tag{3.3.1}$$

$$\text{se}\,(\hat{\beta}_2) = \frac{\sigma}{\sqrt{\sum x_i^2}} \tag{3.3.2}$$

$$\text{var}\,(\hat{\beta}_1) = \frac{\sum X_i^2}{n \sum x_i^2}\sigma^2 \tag{3.3.3}$$

$$\text{se}\,(\hat{\beta}_1) = \sqrt{\frac{\sum X_i^2}{n \sum x_i^2}}\sigma \tag{3.3.4}$$

where var = variance and se = standard error and where σ^2 is the constant or homoscedastic variance of u_i of Assumption 4.

All the quantities entering into the preceding equations except σ^2 can be estimated from the data. As shown in Appendix 3A, Section 3A.5, σ^2 itself is estimated by the following formula:

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-2} \tag{3.3.5}$$

where $\hat{\sigma}^2$ is the OLS estimator of the true but unknown σ^2 and where the expression $n - 2$ is known as the **number of degrees of freedom (df),** $\sum \hat{u}_i^2$ being the sum of the residuals squared or the **residual sum of squares (RSS).**[18]

Once $\sum \hat{u}_i^2$ is known, $\hat{\sigma}^2$ can be easily computed. $\sum \hat{u}_i^2$ itself can be computed either from (3.1.2) or from the following expression (see Section 3.5 for the proof):

$$\sum \hat{u}_i^2 = \sum y_i^2 - \hat{\beta}_2^2 \sum x_i^2 \tag{3.3.6}$$

Compared with Eq. (3.1.2), Eq. (3.3.6) is easy to use, for it does not require computing $\hat{u}_i$ for each observation although such a computation will be useful in its own right (as we shall see in Chapters 11 and 12).

Since

$$\hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2}$$

[18]The term **number of degrees of freedom** means the total number of observations in the sample ($= n$) less the number of independent (linear) constraints or restrictions put on them. In other words, it is the number of independent observations out of a total of n observations. For example, before the RSS (3.1.2) can be computed, $\hat{\beta}_1$ and $\hat{\beta}_2$ must first be obtained. These two estimates therefore put two restrictions on the RSS. Therefore, there are $n - 2$, not n, independent observations to compute the RSS. Following this logic, in the three-variable regression RSS will have $n - 3$ df, and for the k-variable model it will have $n - k$ df. **The general rule is this:** df = (n − number of parameters estimated).

an alternative expression for computing $\sum \hat{u}_i^2$ is

$$\sum \hat{u}_i^2 = \sum y_i^2 - \frac{\left(\sum x_i y_i\right)^2}{\sum x_i^2} \tag{3.3.7}$$

In passing, note that the positive square root of $\hat{\sigma}^2$

$$\hat{\sigma} = \sqrt{\frac{\sum \hat{u}_i^2}{n-2}} \tag{3.3.8}$$

is known as the **standard error of estimate** or **the standard error of the regression (se).** It is simply the standard deviation of the Y values about the estimated regression line and is often used as a summary measure of the "goodness of fit" of the estimated regression line, a topic discussed in Section 3.5.

Earlier we noted that, given X_i, σ^2 represents the (conditional) variance of both u_i and Y_i. Therefore, the standard error of the estimate can also be called the (conditional) standard deviation of u_i and Y_i. Of course, as usual, σ_Y^2 and σ_Y represent, respectively, the unconditional variance and unconditional standard deviation of Y.

Note the following features of the variances (and therefore the standard errors) of $\hat{\beta}_1$ and $\hat{\beta}_2$.

1. The variance of $\hat{\beta}_2$ is directly proportional to σ^2 but inversely proportional to $\sum x_i^2$. That is, given σ^2, the larger the variation in the X values, the smaller the variance of $\hat{\beta}_2$ and hence the greater the precision with which β_2 can be estimated. In short, given σ^2, if there is substantial variation in the X values (recall Assumption 8), β_2 can be measured more accurately than when the X_i do not vary substantially. Also, given $\sum x_i^2$, the larger the variance of σ^2, the larger the variance of β_2. Note that as the sample size n increases, the number of terms in the sum, $\sum x_i^2$, will increase. As n increases, the precision with which β_2 can be estimated also increases. (Why?)

2. The variance of $\hat{\beta}_1$ is directly proportional to σ^2 and $\sum X_i^2$ but inversely proportional to $\sum x_i^2$ and the sample size n.

3. Since $\hat{\beta}_1$ and $\hat{\beta}_2$ are estimators, they will not only vary from sample to sample but in a given sample they are likely to be dependent on each other, this dependence being measured by the covariance between them. It is shown in Appendix 3A, Section 3A.4 that

$$\begin{aligned} \operatorname{cov}(\hat{\beta}_1, \hat{\beta}_2) &= -\bar{X} \operatorname{var}(\hat{\beta}_2) \\ &= -\bar{X}\left(\frac{\sigma^2}{\sum x_i^2}\right) \end{aligned} \tag{3.3.9}$$

Since var $(\hat{\beta}_2)$ is always positive, as is the variance of any variable, the nature of the covariance between $\hat{\beta}_1$ and $\hat{\beta}_2$ depends on the sign of $\bar{X}$. If $\bar{X}$ is positive, then as the formula shows, the covariance will be negative. Thus, if the slope coefficient β_2 is *overestimated* (i.e., the slope is too steep), the intercept coefficient β_1 will be *underestimated* (i.e., the intercept will be too small). Later on (especially in the chapter on multicollinearity, Chapter 10), we will see the utility of studying the covariances between the estimated regression coefficients.

How do the variances and standard errors of the estimated regression coefficients enable one to judge the reliability of these estimates? This is a problem in statistical inference, and it will be pursued in Chapters 4 and 5.

3.4 PROPERTIES OF LEAST-SQUARES ESTIMATORS: THE GAUSS–MARKOV THEOREM[19]

As noted earlier, given the assumptions of the classical linear regression model, the least-squares estimates possess some ideal or optimum properties. These properties are contained in the well-known **Gauss–Markov theorem.** To understand this theorem, we need to consider the **best linear unbiasedness property** of an estimator.[20] As explained in Appendix A, an estimator, say the OLS estimator $\hat{\beta}_2$, is said to be a best linear unbiased estimator (BLUE) of β_2 if the following hold:

1. It is **linear,** that is, a linear function of a random variable, such as the dependent variable Y in the regression model.

2. It is **unbiased,** that is, its average or expected value, $E(\hat{\beta}_2)$, is equal to the true value, β_2.

3. It has minimum variance in the class of all such linear unbiased estimators; an unbiased estimator with the least variance is known as an **efficient estimator.**

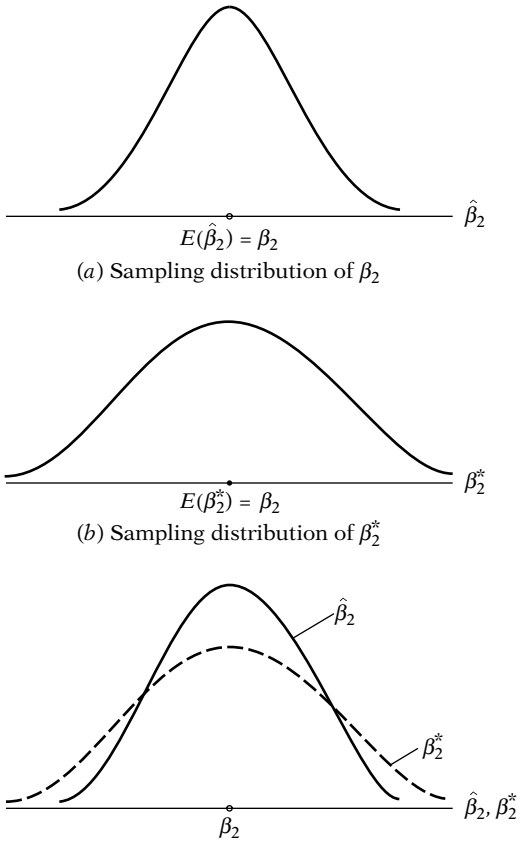
In the regression context it can be proved that the OLS estimators are BLUE. This is the gist of the famous Gauss–Markov theorem, which can be stated as follows:

> **Gauss–Markov Theorem:** Given the assumptions of the classical linear regression model, the least-squares estimators, in the class of unbiased linear estimators, have minimum variance, that is, they are BLUE.

The proof of this theorem is sketched in **Appendix 3A**, **Section 3A.6**. The full import of the Gauss–Markov theorem will become clearer as we move

[19]Although known as the *Gauss–Markov theorem,* the least-squares approach of Gauss antedates (1821) the minimum-variance approach of Markov (1900).

[20]The reader should refer to **App. A** for the importance of linear estimators as well as for a general discussion of the desirable properties of statistical estimators.



(*a*) Sampling distribution of β_2

(*b*) Sampling distribution of β_2^*

(*c*) Sampling distributions of β_2 and β_2^*

FIGURE 3.8 Sampling distribution of OLS estimator $\hat{\beta}_2$ and alternative estimator β_2^*.

along. It is sufficient to note here that the theorem has theoretical as well as practical importance.[21]

What all this means can be explained with the aid of Figure 3.8.

In Figure 3.8(*a*) we have shown the **sampling distribution** of the OLS estimator $\hat{\beta}_2$, that is, the distribution of the values taken by $\hat{\beta}_2$ in repeated sampling experiments (recall Table 3.1). For convenience we have assumed $\hat{\beta}_2$ to be distributed symmetrically (but more on this in Chapter 4). As the figure shows, the mean of the $\hat{\beta}_2$ values, $E(\hat{\beta}_2)$, is equal to the true β_2. In this situation we say that $\hat{\beta}_2$ is an *unbiased estimator* of β_2. In Figure 3.8(*b*) we have shown the sampling distribution of β_2^*, an alternative estimator of β_2

[21]For example, it can be proved that any linear combination of the β's, such as $(\beta_1 - 2\beta_2)$, can be estimated by $(\hat{\beta}_1 - 2\hat{\beta}_2)$, and this estimator is BLUE. For details, see Henri Theil, *Introduction to Econometrics,* Prentice-Hall, Englewood Cliffs, N.J., 1978, pp. 401–402. Note a technical point about the Gauss–Markov theorem: It provides only the sufficient (but not necessary) condition for OLS to be efficient. I am indebted to Michael McAleer of the University of Western Australia for bringing this point to my attention.

obtained by using another (i.e., other than OLS) method. For convenience, assume that β_2^*, like $\hat{\beta}_2$, is unbiased, that is, its average or expected value is equal to β_2. Assume further that both $\hat{\beta}_2$ and β_2^* are linear estimators, that is, they are linear functions of Y. Which estimator, $\hat{\beta}_2$ or β_2^*, would you choose?

To answer this question, superimpose the two figures, as in Figure 3.8(c). It is obvious that although both $\hat{\beta}_2$ and β_2^* are unbiased the distribution of β_2^* is more diffused or widespread around the mean value than the distribution of $\hat{\beta}_2$. In other words, the variance of β_2^* is larger than the variance of $\hat{\beta}_2$. Now given two estimators that are both linear and unbiased, one would choose the estimator with the smaller variance because it is more likely to be close to β_2 than the alternative estimator. In short, one would choose the BLUE estimator.

The Gauss–Markov theorem is remarkable in that it makes no assumptions about the probability distribution of the random variable u_i, and therefore of Y_i (in the next chapter we will take this up). As long as the assumptions of CLRM are satisfied, the theorem holds. As a result, we need not look for another linear unbiased estimator, for we will not find such an estimator whose variance is smaller than the OLS estimator. Of course, if one or more of these assumptions do not hold, the theorem is invalid. For example, if we consider nonlinear-in-the-parameter regression models (which are discussed in Chapter 14), we may be able to obtain estimators that may perform better than the OLS estimators. Also, as we will show in the chapter on heteroscedasticity, if the assumption of homoscedastic variance is not fulfilled, the OLS estimators, although unbiased and consistent, are no longer minimum variance estimators even in the class of linear estimators.

The statistical properties that we have just discussed are known as **finite sample properties:** These properties hold regardless of the sample size on which the estimators are based. Later we will have occasions to consider the **asymptotic properties,** that is, properties that hold only if the sample size is very large (technically, infinite). A general discussion of finite-sample and large-sample properties of estimators is given in **Appendix A.**

3.5 THE COEFFICIENT OF DETERMINATION r^2: A MEASURE OF "GOODNESS OF FIT"

Thus far we were concerned with the problem of estimating regression coefficients, their standard errors, and some of their properties. We now consider the **goodness of fit** of the fitted regression line to a set of data; that is, we shall find out how "well" the sample regression line fits the data. From Figure 3.1 it is clear that if all the observations were to lie on the regression line, we would obtain a "perfect" fit, but this is rarely the case. Generally, there will be some positive $\hat{u}_i$ and some negative $\hat{u}_i$. What we hope for is that these residuals around the regression line are as small as possible. The **coefficient of determination** r^2 (two-variable case) or R^2 (multiple regression) is a summary measure that tells how well the sample regression line fits the data.

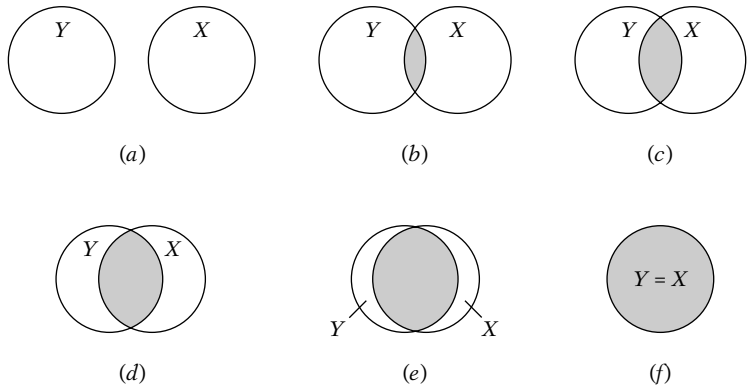


FIGURE 3.9 The Ballentine view of r^2: (*a*) $r^2 = 0$; (*f*) $r^2 = 1$.

Before we show how r^2 is computed, let us consider a heuristic explanation of r^2 in terms of a graphical device, known as the **Venn diagram,** or the **Ballentine,** as shown in Figure 3.9.[22]

In this figure the circle Y represents variation in the dependent variable Y and the circle X represents variation in the explanatory variable X.[23] The overlap of the two circles (the shaded area) indicates the extent to which the variation in Y is explained by the variation in X (say, via an OLS regression). The greater the extent of the overlap, the greater the variation in Y is explained by X. The r^2 is simply a numerical measure of this overlap. In the figure, as we move from left to right, the area of the overlap increases, that is, successively a greater proportion of the variation in Y is explained by X. In short, r^2 increases. When there is no overlap, r^2 is obviously zero, but when the overlap is complete, r^2 is 1, since 100 percent of the variation in Y is explained by X. As we shall show shortly, r^2 lies between 0 and 1.

To compute this r^2, we proceed as follows: Recall that

$$Y_i = \hat{Y}_i + \hat{u}_i \qquad \textbf{(2.6.3)}$$

or in the deviation form

$$y_i = \hat{y}_i + \hat{u}_i \qquad \textbf{(3.5.1)}$$

where use is made of (3.1.13) and (3.1.14). Squaring (3.5.1) on both sides

[22] See Peter Kennedy, "Ballentine: A Graphical Aid for Econometrics," *Australian Economics Papers,* vol. 20, 1981, pp. 414–416. The name Ballentine is derived from the emblem of the well-known Ballantine beer with its circles.

[23] The term *variation* and *variance* are different. Variation means the sum of squares of the deviations of a variable from its mean value. Variance is this sum of squares divided by the appropriate degrees of freedom. In short, variance = variation/df.

and summing over the sample, we obtain

$$\begin{aligned}\sum y_i^2 &= \sum \hat{y}_i^2 + \sum \hat{u}_i^2 + 2\sum \hat{y}_i \hat{u}_i \\ &= \sum \hat{y}_i^2 + \sum \hat{u}_i^2 \\ &= \hat{\beta}_2^2 \sum x_i^2 + \sum \hat{u}_i^2\end{aligned} \tag{3.5.2}$$

since $\sum \hat{y}_i \hat{u}_i = 0$ (why?) and $\hat{y}_i = \hat{\beta}_2 x_i$.

The various sums of squares appearing in (3.5.2) can be described as follows: $\sum y_i^2 = \sum (Y_i - \bar{Y})^2 =$ total variation of the actual Y values about their sample mean, which may be called the **total sum of squares (TSS).** $\sum \hat{y}_i^2 = \sum (\hat{Y}_i - \bar{\hat{Y}})^2 = \sum (\hat{Y}_i - \bar{Y})^2 = \hat{\beta}_2^2 \sum x_i^2 =$ variation of the estimated Y values about their mean ($\bar{\hat{Y}} = \bar{Y}$), which appropriately may be called the sum of squares due to regression [i.e., due to the explanatory variable(s)], or explained by regression, or simply the **explained sum of squares (ESS).** $\sum \hat{u}_i^2 =$ residual or **unexplained** variation of the Y values about the regression line, or simply the **residual sum of squares (RSS).** Thus, (3.5.2) is

$$\text{TSS} = \text{ESS} + \text{RSS} \tag{3.5.3}$$

and shows that the total variation in the observed Y values about their mean value can be partitioned into two parts, one attributable to the regression line and the other to random forces because not all actual Y observations lie on the fitted line. Geometrically, we have Figure 3.10.

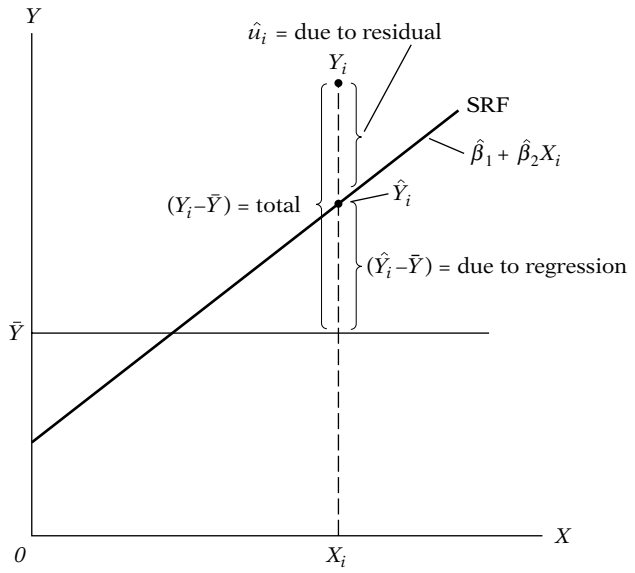


FIGURE 3.10 Breakdown of the variation of Y_i into two components.

Now dividing (3.5.3) by TSS on both sides, we obtain

$$\begin{aligned} 1 &= \frac{\text{ESS}}{\text{TSS}} + \frac{\text{RSS}}{\text{TSS}} \\ &= \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} + \frac{\sum \hat{u}_i^2}{\sum(Y_i - \bar{Y})^2} \end{aligned} \tag{3.5.4}$$

We now define r^2 as

$$r^2 = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = \frac{\text{ESS}}{\text{TSS}} \tag{3.5.5}$$

or, alternatively, as

$$\begin{aligned} r^2 &= 1 - \frac{\sum \hat{u}_i^2}{\sum(Y_i - \bar{Y})^2} \\ &= 1 - \frac{\text{RSS}}{\text{TSS}} \end{aligned} \tag{3.5.5a}$$

The quantity r^2 thus defined is known as the (sample) **coefficient of determination** and is the most commonly used measure of the goodness of fit of a regression line. Verbally, r^2 *measures the proportion or percentage of the total variation in Y explained by the regression model.*

Two properties of r^2 may be noted:

1. It is a nonnegative quantity. (Why?)

2. Its limits are $0 \leq r^2 \leq 1$. An r^2 of 1 means a perfect fit, that is, $\hat{Y}_i = Y_i$ for each i. On the other hand, an r^2 of zero means that there is no relationship between the regressand and the regressor whatsoever (i.e., $\hat{\beta}_2 = 0$). In this case, as (3.1.9) shows, $\hat{Y}_i = \hat{\beta}_1 = \bar{Y}$, that is, the best prediction of any Y value is simply its mean value. In this situation therefore the regression line will be horizontal to the X axis.

Although r^2 can be computed directly from its definition given in (3.5.5), it can be obtained more quickly from the following formula:

$$\begin{aligned} r^2 &= \frac{\text{ESS}}{\text{TSS}} \\ &= \frac{\sum \hat{y}_i^2}{\sum y_i^2} \\ &= \frac{\hat{\beta}_2^2 \sum x_i^2}{\sum y_i^2} \\ &= \hat{\beta}_2^2 \left(\frac{\sum x_i^2}{\sum y_i^2} \right) \end{aligned} \tag{3.5.6}$$

If we divide the numerator and the denominator of (3.5.6) by the sample size n (or $n-1$ if the sample size is small), we obtain

$$r^2 = \hat{\beta}_2^2\left(\frac{S_x^2}{S_y^2}\right) \tag{3.5.7}$$

where S_y^2 and S_x^2 are the sample variances of Y and X, respectively.

Since $\hat{\beta}_2 = \sum x_i y_i / \sum x_i^2$, Eq. (3.5.6) can also be expressed as

$$r^2 = \frac{\left(\sum x_i y_i\right)^2}{\sum x_i^2 \sum y_i^2} \tag{3.5.8}$$

an expression that may be computationally easy to obtain.

Given the definition of r^2, we can express ESS and RSS discussed earlier as follows:

$$\begin{aligned} \text{ESS} &= r^2 \cdot \text{TSS} \\ &= r^2 \sum y_i^2 \end{aligned} \tag{3.5.9}$$

$$\begin{aligned} \text{RSS} &= \text{TSS} - \text{ESS} \\ &= \text{TSS}(1 - \text{ESS/TSS}) \\ &= \sum y_i^2 \cdot (1 - r^2) \end{aligned} \tag{3.5.10}$$

Therefore, we can write

$$\begin{aligned} &\text{TSS} = \text{ESS} + \text{RSS} \\ &\sum y_i^2 = r^2 \sum y_i^2 + (1 - r^2) \sum y_i^2 \end{aligned} \tag{3.5.11}$$

an expression that we will find very useful later.

A quantity closely related to but conceptually very much different from r^2 is the **coefficient of correlation,** which, as noted in Chapter 1, is a measure of the degree of association between two variables. It can be computed either from

$$r = \pm\sqrt{r^2} \tag{3.5.12}$$

or from its definition

$$\begin{aligned} r &= \frac{\sum x_i y_i}{\sqrt{\left(\sum x_i^2\right)\left(\sum y_i^2\right)}} \\ &= \frac{n\sum X_i Y_i - \left(\sum X_i\right)\left(\sum Y_i\right)}{\sqrt{\left[n\sum X_i^2 - \left(\sum X_i\right)^2\right]\left[n\sum Y_i^2 - \left(\sum Y_i\right)^2\right]}} \end{aligned} \tag{3.5.13}$$

which is known as the **sample correlation coefficient.**[24]

[24]The population correlation coefficient, denoted by ρ, is defined in **App. A**.

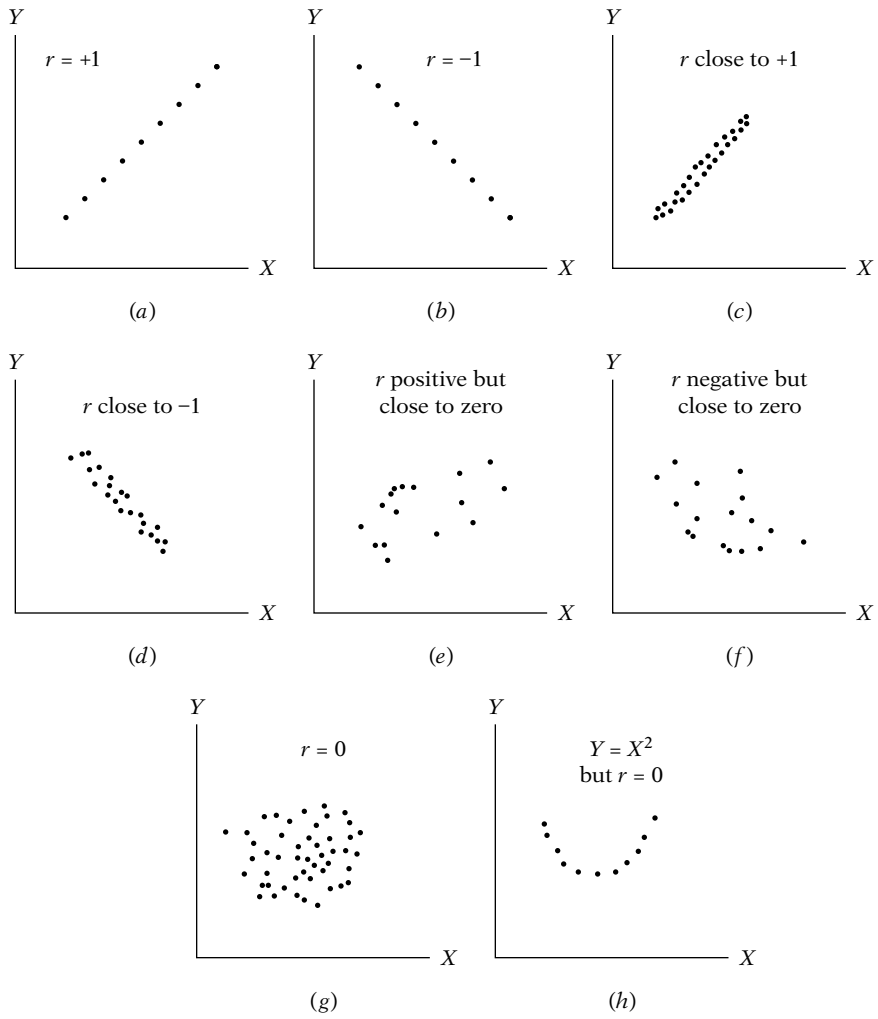


FIGURE 3.11 Correlation patterns (adapted from Henri Theil, *Introduction to Econometrics,* Prentice-Hall, Englewood Cliffs, N.J., 1978, p. 86).

Some of the properties of r are as follows (see Figure 3.11):

1. It can be positive or negative, the sign depending on the sign of the term in the numerator of (3.5.13), which measures the sample *covariation* of two variables.

2. It lies between the limits of -1 and $+1$; that is, $-1 \leq r \leq 1$.

3. It is symmetrical in nature; that is, the coefficient of correlation between X and $Y(r_{XY})$ is the same as that between Y and $X(r_{YX})$.

4. It is independent of the origin and scale; that is, if we define $X_i^* = aX_i + C$ and $Y_i^* = bY_i + d$, where $a > 0$, $b > 0$, and c and d are constants,

then r between X^* and Y^* is the same as that between the original variables X and Y.

5. If X and Y are statistically independent (see **Appendix A** for the definition), the correlation coefficient between them is zero; but if $r = 0$, it does not mean that two variables are independent. In other words, **zero correlation does not necessarily imply independence.** [See Figure 3.11(h).]

6. It is a measure of *linear association* or *linear dependence* only; it has no meaning for describing nonlinear relations. Thus in Figure 3.11(h), $Y = X^2$ is an exact relationship yet r is zero. (Why?)

7. Although it is a measure of linear association between two variables, it does not necessarily imply any cause-and-effect relationship, as noted in Chapter 1.

In the regression context, r^2 is a more meaningful measure than r, for the former tells us the proportion of variation in the dependent variable explained by the explanatory variable(s) and therefore provides an overall measure of the extent to which the variation in one variable determines the variation in the other. The latter does not have such value.[25] Moreover, as we shall see, the interpretation of r ($= R$) in a multiple regression model is of dubious value. However, we will have more to say about r^2 in Chapter 7.

In passing, note that the r^2 defined previously *can also be computed as the squared coefficient of correlation between actual Y_i and the estimated Y_i*, namely, $\hat{Y}_i$. That is, using (3.5.13), we can write

$$r^2 = \frac{\left[\sum (Y_i - \bar{Y})(\hat{Y}_i - \bar{Y})\right]^2}{\sum (Y_i - \bar{Y})^2 \sum (\hat{Y}_i - \bar{Y})^2}$$

That is,

$$r^2 = \frac{\left(\sum y_i \hat{y}_i\right)^2}{\left(\sum y_i^2\right)\left(\sum \hat{y}_i^2\right)} \tag{3.5.14}$$

where Y_i = actual Y, $\hat{Y}_i$ = estimated Y, and $\bar{Y} = \bar{\hat{Y}}$ = the mean of Y. For proof, see exercise 3.15. Expression (3.5.14) justifies the description of r^2 as a measure of goodness of fit, for it tells how close the estimated Y values are to their actual values.

3.6 A NUMERICAL EXAMPLE

We illustrate the econometric theory developed so far by considering the Keynesian consumption function discussed in the Introduction. Recall that Keynes stated that "The fundamental psychological law . . . is that men

[25] In regression modeling the underlying theory will indicate the direction of causality between Y and X, which, in the context of single-equation models, is generally from X to Y.

TABLE 3.2 HYPOTHETICAL DATA ON WEEKLY FAMILY CONSUMPTION EXPENDITURE *Y* AND WEEKLY FAMILY INCOME *X*

Y, $	*X*, $
70	80
65	100
90	120
95	140
110	160
115	180
120	200
140	220
155	240
150	260

[women] are disposed, as a rule and on average, to increase their consumption as their income increases, but not by as much as the increase in their income," that is, the marginal propensity to consume (MPC) is greater than zero but less than one. Although Keynes did not specify the exact functional form of the relationship between consumption and income, for simplicity assume that the relationship is linear as in (2.4.2). As a test of the Keynesian consumption function, we use the sample data of Table 2.4, which for convenience is reproduced as Table 3.2. The raw data required to obtain the estimates of the regression coefficients, their standard errors, etc., are given in Table 3.3. From these raw data, the following calculations are obtained, and the reader is advised to check them.

$$\begin{aligned} \hat{\beta}_1 &= 24.4545 \quad \text{var}(\hat{\beta}_1) = 41.1370 \quad \text{and} \quad \text{se}(\hat{\beta}_1) = 6.4138 \\ \hat{\beta}_2 &= \ \ 0.5091 \quad \text{var}(\hat{\beta}_2) = \ \ 0.0013 \quad \text{and} \quad \text{se}(\hat{\beta}_2) = 0.0357 \\ &\text{cov}(\hat{\beta}_1, \hat{\beta}_2) = -0.2172 \quad \hat{\sigma}^2 = 42.1591 \\ &r^2 = 0.9621 \quad r = 0.9809 \quad \text{df} = 8 \end{aligned} \tag{3.6.1}$$

The estimated regression line therefore is

$$\hat{Y}_i = 24.4545 + 0.5091 X_i \tag{3.6.2}$$

which is shown geometrically as Figure 3.12.

Following Chapter 2, the SRF [Eq. (3.6.2)] and the associated regression line are interpreted as follows: Each point on the regression line gives an *estimate* of the expected or mean value of Y corresponding to the chosen X value; that is, $\hat{Y}_i$ is an estimate of $E(Y \mid X_i)$. The value of $\hat{\beta}_2 = 0.5091$, which measures the slope of the line, shows that, within the sample range of X between \$80 and \$260 per week, as X increases, say, by \$1, the estimated increase in the mean or average weekly consumption expenditure amounts to about 51 cents. The value of $\hat{\beta}_1 = 24.4545$, which is the intercept of the

TABLE 3.3 RAW DATA BASED ON TABLE 3.2

Y_i (1)	X_i (2)	Y_iX_i (3)	X_i^2 (4)	$x_i = X_i - \bar{X}$ (5)	$y_i = Y_i - \bar{Y}$ (6)	x_i^2 (7)	x_iy_i (8)	$\hat{Y}_i$ (9)	$\hat{u}_i = Yi - \hat{Y}_i$ (10)	$\hat{Y}_i\hat{u}_i$ (11)
70	80	5600	6400	−90	−41	8100	3690	65.1818	4.8181	314.0524
65	100	6500	10000	−70	−46	4900	3220	75.3636	−10.3636	−781.0382
90	120	10800	14400	−50	−21	2500	1050	85.5454	4.4545	381.0620
95	140	13300	19600	−30	−16	900	480	95.7272	−0.7272	−69.6128
110	160	17600	25600	−10	−1	100	10	105.9090	4.0909	433.2631
115	180	20700	32400	10	4	100	40	116.0909	−1.0909	−126.6434
120	200	24000	40000	30	9	900	270	125.2727	−6.2727	−792.0708
140	220	30800	48400	50	29	2500	1450	136.4545	3.5454	483.7858
155	240	37200	57600	70	44	4900	3080	145.6363	8.3636	1226.4073
150	260	39000	67600	90	39	8100	3510	156.8181	−6.8181	−1069.2014
Sum 1110	1700	205500	322000	0	0	33000	16800	1109.9995 ≈ 1110.0	0	0.0040 ≈ 0.0
Mean 111	170	nc	nc	0	0	nc	nc	110	0	0

$$\hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2} = 16{,}800/33{,}000 = 0.5091$$

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2\bar{X} = 111 - 0.5091(170) = 24.4545$$

Notes: ≈ symbolizes "approximately equal to"; nc means "not computed."

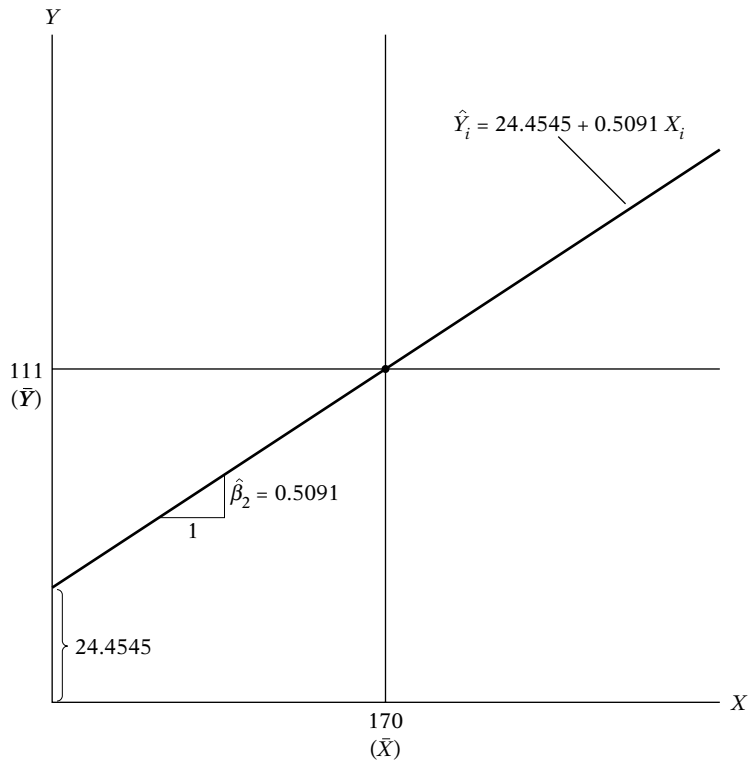


FIGURE 3.12 Sample regression line based on the data of Table 3.2.

line, indicates the average level of weekly consumption expenditure when weekly income is zero. However, this is a mechanical interpretation of the intercept term. In regression analysis such literal interpretation of the intercept term may not be always meaningful, although in the present example it can be argued that a family without any income (because of unemployment, layoff, etc.) might maintain some minimum level of consumption expenditure either by borrowing or dissaving. But in general one has to use common sense in interpreting the intercept term, for very often the sample range of X values may not include zero as one of the observed values.

Perhaps it is best to interpret the intercept term as the mean or average effect on Y of all the variables omitted from the regression model. The value of r^2 of 0.9621 means that about 96 percent of the variation in the weekly consumption expenditure is explained by income. Since r^2 can at most be 1, the observed r^2 suggests that the sample regression line fits the data very well.[26] The coefficient of correlation of 0.9809 shows that the two variables, consumption expenditure and income, are highly positively correlated. The estimated standard errors of the regression coefficients will be interpreted in Chapter 5.

3.7 ILLUSTRATIVE EXAMPLES

EXAMPLE 3.1

CONSUMPTION–INCOME RELATIONSHIP IN THE UNITED STATES, 1982–1996

Let us return to the consumption income data given in Table I.1 of the Introduction. We have already shown the data in Figure I.3 along with the estimated regression line (I.3.3). Now we provide the underlying OLS regression results. (The results were obtained from the statistical package *Eviews 3*.) *Note:* Y = personal consumption expenditure (PCE) and X = gross domestic product (GDP), all measured in 1992 billions of dollars. In this example, our data are *time series* data.

$$\hat{Y}_i = -184.0780 + 0.7064X_i \qquad \textbf{(3.7.1)}$$

$$\text{var}\,(\hat{\beta}_1) = 2140.1707 \qquad \text{se}\,(\hat{\beta}_1) = 46.2619$$

$$\text{var}\,(\hat{\beta}_2) = 0.000061 \qquad \text{se}\,(\hat{\beta}_2) = 0.007827$$

$$r^2 = 0.998406 \qquad \hat{\sigma}^2 = 411.4913$$

Equation (3.7.1) is the aggregate (i.e., for the economy as a whole) Keynesian consumption function. As this equation shows, the **marginal propensity to consume (MPC)** is about 0.71, suggesting that if income goes up by a dollar, the average personal consumption expenditure (PCE) goes up by about 71 cents. From Keynesian theory, the MPC is less than 1. The intercept value of about −184 tells us that if income were zero, the PCE would be about −184 billion dollars. Of course, such a mechanical interpretation of the intercept term does not make economic sense in the present instance because the zero income value is out of the range of values we are working with and does not represent a likely outcome (see Table I.1). As we will see on many an occasion, very often the intercept term may not make much economic sense. Therefore, in practice the intercept term may not be very meaningful, although on occasions it can be very meaningful, as we will see in some illustrative examples. The more meaningful value is the slope coefficient, MPC in the present case.

The r^2 value of 0.9984 means approximately 99 percent of the variation in the PCE is explained by variation in the GDP. Since r^2 at most can be 1, we can say that the regression line in (3.7.1), which is shown in Figure I.3, fits our data extremely well; as you can see from that figure the actual data points are very tightly clustered around the estimated regression line. As we will see throughout this book, in regressions involving time series data one generally obtains high r^2 values. In the chapter on autocorrelation, we will see the reasons behind this phenomenon.

[26] A formal test of the significance of r^2 will be presented in Chap. 8.

EXAMPLE 3.2

FOOD EXPENDITURE IN INDIA

Refer to the data given in Table 2.8 of exercise 2.15. The data relate to a sample of 55 rural households in India. The regressand in this example is expenditure on food and the regressor is total expenditure, a proxy for income, both figures in rupees. The data in this example are thus *cross-sectional* data.

On the basis of the given data, we obtained the following regression:

$$\widehat{\text{FoodExp}}_i = 94.2087 + 0.4368\ \text{TotalExp}_i \qquad \textbf{(3.7.2)}$$

$$\text{var}\,(\hat{\beta}_1) = 2560.9401 \qquad \text{se}\,(\hat{\beta}_1) = 50.8563$$

$$\text{var}\,(\hat{\beta}_2) = 0.0061 \qquad \text{se}\,(\hat{\beta}_2) = 0.0783$$

$$r^2 = 0.3698 \qquad \hat{\sigma}^2 = 4469.6913$$

From (3.7.2) we see that if total expenditure increases by 1 rupee, on average, expenditure on food goes up by about 44 paise (1 rupee = 100 paise). If total expenditure were zero, the average expenditure on food would be about 94 rupees. Again, such a mechanical interpretation of the intercept may not be meaningful. However, in this example one could argue that even if total expenditure is zero (e.g., because of loss of a job), people may still maintain some minimum level of food expenditure by borrowing money or by dissaving.

The r^2 value of about 0.37 means that only 37 percent of the variation in food expenditure is explained by the total expenditure. This might seem a rather low value, but as we will see throughout this text, in cross-sectional data, typically one obtains low r^2 values, possibly because of the diversity of the units in the sample. We will discuss this topic further in the chapter on heteroscedasticity (see Chapter 11).

EXAMPLE 3.3

THE RELATIONSHIP BETWEEN EARNINGS AND EDUCATION

In Table 2.6 we looked at the data relating average hourly earnings and education, as measured by years of schooling. Using that data, if we regress[27] average hourly earnings (Y) on education (X), we obtain the following results.

$$\hat{Y}_i = -0.0144 + 0.7241\ X_i \qquad \textbf{(3.7.3)}$$

$$\text{var}\,(\hat{\beta}_1) = 0.7649 \quad se\,(\hat{\beta}_1) = 0.8746$$

$$\text{var}\,(\hat{\beta}_2) = 0.00483 \quad se\,(\hat{\beta}_2) = 0.0695$$

$$r^2 = 0.9077 \qquad \hat{\sigma}^2 = 0.8816$$

As the regression results show, there is a positive association between education and earnings, an unsurprising finding. For every additional year of schooling, the average hourly earnings go up by about 72 cents an hour. The intercept term is positive but it may have no economic meaning. The r^2 value suggests that about 89 percent of the variation in average hourly earnings is explained by education. For cross-sectional data, such a high r^2 is rather unusual.

3.8 A NOTE ON MONTE CARLO EXPERIMENTS

In this chapter we showed that under the assumptions of CLRM the least-squares estimators have certain desirable statistical features summarized in the BLUE property. In the appendix to this chapter we prove this property

[27] Every field of study has its jargon. The expression "regress Y on X" simply means treat Y as the regressand and X as the regressor.

more formally. But in practice how does one know that the BLUE property holds? For example, how does one find out if the OLS estimators are unbiased? The answer is provided by the so-called **Monte Carlo** experiments, which are essentially computer simulation, or sampling, experiments.

To introduce the basic ideas, consider our two-variable PRF:

$$Y_i = \beta_1 + \beta_2 X_i + u_i \tag{3.8.1}$$

A Monte Carlo experiment proceeds as follows:

1. Suppose the true values of the parameters are as follows: $\beta_1 = 20$ and $\beta_2 = 0.6$.

2. You choose the sample size, say $n = 25$.

3. You fix the values of X for each observation. In all you will have 25 X values.

4. Suppose you go to a random number table, choose 25 values, and call them u_i (these days most statistical packages have built-in random number generators).[28]

5. Since you *know* β_1, β_2, X_i, and u_i, using (3.8.1) you obtain 25 Y_i values.

6. Now using the 25 Y_i values thus generated, you regress these on the 25 X values chosen in step 3, obtaining $\hat{\beta}_1$ and $\hat{\beta}_2$, the least-squares estimators.

7. Suppose you repeat this experiment 99 times, each time using the same β_1, β_2, and X values. Of course, the u_i values will vary from experiment to experiment. Therefore, in all you have 100 experiments, thus generating 100 values each of β_1 and β_2. (In practice, many such experiments are conducted, sometimes 1000 to 2000.)

8. You take the averages of these 100 estimates and call them $\bar{\hat{\beta}}_1$ and $\bar{\hat{\beta}}_2$.

9. If these average values are about the same as the true values of β_1 and β_2 assumed in step 1, this Monte Carlo experiment "establishes" that the least-squares estimators are indeed unbiased. Recall that under CLRM $E(\hat{\beta}_1) = \beta_1$ and $E(\hat{\beta}_2) = \beta_2$.

These steps characterize the general nature of the Monte Carlo experiments. Such experiments are often used to study the statistical properties of various methods of estimating population parameters. They are particularly useful to study the behavior of estimators in small, or finite, samples. These experiments are also an excellent means of driving home the concept of **repeated sampling** that is the basis of most of classical statistical inference, as we shall see in Chapter 5. We shall provide several examples of Monte Carlo experiments by way of exercises for classroom assignment. (See exercise 3.27.)

[28] In practice it is assumed that u_i follows a certain probability distribution, say, normal, with certain parameters (e.g., the mean and variance). Once the values of the parameters are specified, one can easily generate the u_i using statistical packages.

3.9 SUMMARY AND CONCLUSIONS

The important topics and concepts developed in this chapter can be summarized as follows.

1. The basic framework of regression analysis is the **CLRM.**

2. The CLRM is based on a set of assumptions.

3. Based on these assumptions, the least-squares estimators take on certain properties summarized in the Gauss–Markov theorem, which states that in the class of linear unbiased estimators, the least-squares estimators have minimum variance. In short, they are BLUE.

4. The *precision* of OLS estimators is measured by their **standard errors.** In Chapters 4 and 5 we shall see how the standard errors enable one to draw inferences on the population parameters, the β coefficients.

5. The overall goodness of fit of the regression model is measured by the **coefficient of determination,** r^2. It tells what proportion of the variation in the dependent variable, or regressand, is explained by the explanatory variable, or regressor. This r^2 lies between 0 and 1; the closer it is to 1, the better is the fit.

6. A concept related to the coefficient of determination is the **coefficient of correlation,** r. It is a measure of *linear association* between two variables and it lies between -1 and $+1$.

7. The CLRM is a theoretical construct or abstraction because it is based on a set of assumptions that may be stringent or "unrealistic." But such abstraction is often necessary in the initial stages of studying any field of knowledge. Once the CLRM is mastered, one can find out what happens if one or more of its assumptions are not satisfied. The first part of this book is devoted to studying the CLRM. The other parts of the book consider the refinements of the CLRM. Table 3.4 gives the road map ahead.

TABLE 3.4 WHAT HAPPENS IF THE ASSUMPTIONS OF CLRM ARE VIOLATED?

Assumption number	Type of violation	Where to study?
1	Nonlinearity in parameters	Chapter 14
2	Stochastic regressor(s)	Introduction to Part II
3	Nonzero mean of u_i	Introduction to Part II
4	Heteroscedasticity	Chapter 11
5	Autocorrelated disturbances	Chapter 12
6	Nonzero covariance between disturbances and regressor	Introduction to Part II and Part IV
7	Sample observations less than the number of regressors	Chapter 10
8	Insufficient variability in regressors	Chapter 10
9	Specification bias	Chapters 13, 14
10	Multicollinearity	Chapter 10
11*	Nonnormality of disturbances	Introduction to Part II

**Note:* The assumption that the disturbances u_i are normally distributed is not a part of the CLRM. But more on this in Chapter 4.

EXERCISES

Questions

3.1. Given the assumptions in column 1 of the table, show that the assumptions in column 2 are equivalent to them.

ASSUMPTIONS OF THE CLASSICAL MODEL

(1)	(2)
$E(u_i \mid X_i) = 0$	$E(Y_i \mid X_i) = \beta_2 + \beta_2 X$
cov $(u_i, u_j) = 0$ $i \neq j$	cov $(Y_i, Y_j) = 0$ $i \neq j$
var $(u_i \mid X_i) = \sigma^2$	var $(Y_i \mid X_i) = \sigma^2$

3.2. Show that the estimates $\hat{\beta}_1 = 1.572$ and $\hat{\beta}_2 = 1.357$ used in the first experiment of Table 3.1 are in fact the OLS estimators.

3.3. According to Malinvaud (see footnote 10), the assumption that $E(u_i \mid X_i) = 0$ is quite important. To see this, consider the PRF: $Y = \beta_1 + \beta_2 X_i + u_i$. Now consider two situations: (i) $\beta_1 = 0$, $\beta_2 = 1$, and $E(u_i) = 0$; and (ii) $\beta_1 = 1$, $\beta_2 = 0$, and $E(u_i) = (X_i - 1)$. Now take the expectation of the PRF conditional upon X in the two preceding cases and see if you agree with Malinvaud about the significance of the assumption $E(u_i \mid X_i) = 0$.

3.4. Consider the sample regression

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

Imposing the restrictions (i) $\sum \hat{u}_i = 0$ and (ii) $\sum \hat{u}_i X_i = 0$, obtain the estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ and show that they are identical with the least-squares estimators given in (3.1.6) and (3.1.7). This method of obtaining estimators is called the **analogy principle.** Give an intuitive justification for imposing restrictions (i) and (ii). (*Hint:* Recall the CLRM assumptions about u_i.) In passing, note that the analogy principle of estimating unknown parameters is also known as the **method of moments** in which sample moments (e.g., sample mean) are used to estimate population moments (e.g., the population mean). As noted in **Appendix A**, a **moment** is a summary statistic of a probability distribution, such as the expected value and variance.

3.5. Show that r^2 defined in (3.5.5) ranges between 0 and 1. You may use the Cauchy–Schwarz inequality, which states that for any random variables X and Y the following relationship holds true:

$$[E(XY)]^2 \leq E(X^2)E(Y^2)$$

3.6. Let $\hat{\beta}_{YX}$ and $\hat{\beta}_{XY}$ represent the slopes in the regression of Y on X and X on Y, respectively. Show that

$$\hat{\beta}_{YX}\hat{\beta}_{XY} = r^2$$

where r is the coefficient of correlation between X and Y.

3.7. Suppose in exercise 3.6 that $\hat{\beta}_{YX}\hat{\beta}_{XY} = 1$. Does it matter then if we regress Y on X or X on Y? Explain carefully.

3.8. Spearman's rank correlation coefficient r_s is defined as follows:

$$r_s = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$$

where d = difference in the ranks assigned to the same individual or phenomenon and n = number of individuals or phenomena ranked. Derive r_s from r defined in (3.5.13). *Hint:* Rank the X and Y values from 1 to n. Note that the sum of X and Y ranks is $n(n + 1)/2$ each and therefore their means are $(n + 1)/2$.

3.9. Consider the following formulations of the two-variable PRF:

$$\text{Model I:} \quad Y_i = \beta_1 + \beta_2 X_i + u_i$$

$$\text{Model II:} \quad Y_i = \alpha_1 + \alpha_2 (X_i - \bar{X}) + u_i$$

a. Find the estimators of β_1 and α_1. Are they identical? Are their variances identical?

b. Find the estimators of β_2 and α_2. Are they identical? Are their variances identical?

c. What is the advantage, if any, of model II over model I?

3.10. Suppose you run the following regression:

$$y_i = \hat{\beta}_1 + \hat{\beta}_2 x_i + \hat{u}_i$$

where, as usual, y_i and x_i are deviations from their respective mean values. What will be the value of $\hat{\beta}_1$? Why? Will $\hat{\beta}_2$ be the same as that obtained from Eq. (3.1.6)? Why?

3.11. Let r_1 = coefficient of correlation between n pairs of values (Y_i, X_i) and r_2 = coefficient of correlation between n pairs of values $(aX_i + b,\ cY_i + d)$, where a, b, c, and d are constants. Show that $r_1 = r_2$ and hence *establish the principle that the coefficient of correlation is invariant with respect to the change of scale and the change of origin.*

Hint: Apply the definition of r given in (3.5.13).

Note: The operations aX_i, $X_i + b$, and $aX_i + b$ are known, respectively, as the *change of scale, change of origin,* and *change of both scale and origin.*

3.12. If r, the coefficient of correlation between n pairs of values (X_i, Y_i), is positive, then determine whether each of the following statements is true or false:

a. r between $(-X_i, -Y_i)$ is also positive.

b. r between $(-X_i, Y_i)$ and that between $(X_i, -Y_i)$ can be either positive or negative.

c. Both the slope coefficients β_{yx} and β_{xy} are positive, where β_{yx} = slope coefficient in the regression of Y on X and β_{xy} = slope coefficient in the regression of X on Y.

3.13. If X_1, X_2, and X_3 are uncorrelated variables each having the same standard deviation, show that the coefficient of correlation between $X_1 + X_2$ and $X_2 + X_3$ is equal to $\frac{1}{2}$. Why is the correlation coefficient not zero?

3.14. In the regression $Y_i = \beta_1 + \beta_2 X_i + u_i$ suppose we *multiply* each X value by a constant, say, 2. Will it change the residuals and fitted values of Y? Explain. What if we *add* a constant value, say, 2, to each X value?

3.15. Show that (3.5.14) in fact measures the coefficient of determination. *Hint:* Apply the definition of r given in (3.5.13) and recall that $\sum y_i \hat{y}_i = \sum(\hat{y}_i + \hat{u}_i)\hat{y}_i = \sum \hat{y}_i^2$, and remember (3.5.6).

3.16. Explain *with reason* whether the following statements are true, false, or uncertain:

- **a.** Since the correlation between two variables, Y and X, can range from -1 to $+1$, this also means that cov (Y, X) also lies between these limits.
- **b.** If the correlation between two variables is zero, it means that there is no relationship between the two variables whatsoever.
- **c.** If you regress Y_i on $\hat{Y}_i$ (i.e., actual Y on estimated Y), the intercept and slope values will be 0 and 1, respectively.

3.17. *Regression without any regressor.* Suppose you are given the model: $Y_i = \beta_1 + u_i$. Use OLS to find the estimator of β_1. What is its variance and the RSS? Does the estimated β_1 make intuitive sense? Now consider the two-variable model $Y_i = \beta_1 + \beta_2 X_i + u_i$. Is it worth adding X_i to the model? If not, why bother with regression analysis?

Problems

3.18. In Table 3.5, you are given the ranks of 10 students in midterm and final examinations in statistics. Compute Spearman's coefficient of rank correlation and interpret it.

3.19. *The relationship between nominal exchange rate and relative prices.* From the annual observations from 1980 to 1994, the following regression results were obtained, where Y = exchange rate of the German mark to the U.S. dollar (GM/$) and X = ratio of the U.S. consumer price index to the German consumer price index; that is, X represents the relative prices in the two countries:

$$\hat{Y}_t = 6.682 - 4.318 X_t \qquad r^2 = 0.528$$
$$\text{se} = (1.22)(1.333)$$

- **a.** Interpret this regression. How would you interpret r^2?
- **b.** Does the negative value of X_t make economic sense? What is the underlying economic theory?
- **c.** Suppose we were to redefine X as the ratio of German CPI to the U.S. CPI. Would that change the sign of X? And why?

3.20. Table 3.6 gives data on indexes of output per hour (X) and real compensation per hour (Y) for the business and nonfarm business sectors of the U.S. economy for 1959–1997. The base year of the indexes is 1982 = 100 and the indexes are seasonally adjusted.

TABLE 3.5

	Student									
Rank	A	B	C	D	E	F	G	H	I	J
Midterm	1	3	7	10	9	5	4	8	2	6
Final	3	2	8	7	9	6	5	10	1	4

TABLE 3.6 PRODUCTIVITY AND RELATED DATA, BUSINESS SECTOR, 1959–98
[Index numbers, 1992 = 100; quarterly data seasonally adjusted]

	Output per hour of all persons[1]		Compensation per hour[2]	
Year or quarter	Business sector	Nonfarm business sector	Business sector	Nonfarm business sector
1959	50.5	54.2	13.1	13.7
1960	51.4	54.8	13.7	14.3
1961	53.2	56.6	14.2	14.8
1962	55.7	59.2	14.8	15.4
1963	57.9	61.2	15.4	15.9
1964	60.6	63.8	16.2	16.7
1965	62.7	65.8	16.8	17.2
1966	65.2	68.0	17.9	18.2
1967	66.6	69.2	18.9	19.3
1968	68.9	71.6	20.5	20.8
1969	69.2	71.7	21.9	22.2
1970	70.6	72.7	23.6	23.8
1971	73.6	75.7	25.1	25.4
1972	76.0	78.3	26.7	27.0
1973	78.4	80.7	29.0	29.2
1974	77.1	79.4	31.8	32.1
1975	79.8	81.6	35.1	35.3
1976	82.5	84.5	38.2	38.4
1977	84.0	85.8	41.2	41.5
1978	84.9	87.0	44.9	45.2
1979	84.5	86.3	49.2	49.5
1980	84.2	86.0	54.5	54.8
1981	85.8	87.0	59.6	60.2
1982	85.3	88.3	64.1	64.6
1983	88.0	89.9	66.8	67.3
1984	90.2	91.4	69.7	70.2
1985	91.7	92.3	73.1	73.4
1986	94.1	94.7	76.8	77.2
1987	94.0	94.5	79.8	80.1
1988	94.7	95.3	83.6	83.7
1989	95.5	95.8	85.9	86.0
1990	96.1	96.3	90.8	90.7
1991	96.7	97.0	95.1	95.1
1992	100.0	100.0	100.0	100.0
1993	100.1	100.1	102.5	102.2
1994	100.7	100.6	104.4	104.2
1995	101.0	101.2	106.8	106.7
1996	103.7	103.7	110.7	110.4
1997	105.4	105.1	114.9	114.5

[1]Output refers to real gross domestic product in the sector.
[2]Wages and salaries of employees plus employers' contributions for social insurance and private benefit plans. Also includes an estimate of wages, salaries, and supplemental payments for the self-employed.
Source: Economic Report of the President, 1999, Table B-49, p. 384.

a. Plot Y against X for the two sectors separately.
b. What is the economic theory behind the relationship between the two variables? Does the scattergram support the theory?
c. Estimate the OLS regression of Y on X. Save the results for a further look after we study Chapter 5.

3.21. From a sample of 10 observations, the following results were obtained:

$$\sum Y_i = 1110 \qquad \sum X_i = 1700 \qquad \sum X_i Y_i = 205{,}500$$

$$\sum X_i^2 = 322{,}000 \qquad \sum Y_i^2 = 132{,}100$$

with coefficient of correlation $r = 0.9758$. But on rechecking these calculations it was found that two pairs of observations were recorded:

Y	X		Y	X
90	120	instead of	80	110
140	220		150	210

What will be the effect of this error on r? Obtain the correct r.

3.22. Table 3.7 gives data on gold prices, the Consumer Price Index (CPI), and the New York Stock Exchange (NYSE) Index for the United States for the period 1977–1991. The NYSE Index includes most of the stocks listed on the NYSE, some 1500 plus.

TABLE 3.7

Year	Price of gold at New York, $ per troy ounce	Consumer Price Index (CPI), 1982–84 = 100	New York Stock Exchange (NYSE) Index, Dec. 31, 1965 = 100
1977	147.98	60.6	53.69
1978	193.44	65.2	53.70
1979	307.62	72.6	58.32
1980	612.51	82.4	68.10
1981	459.61	90.9	74.02
1982	376.01	96.5	68.93
1983	423.83	99.6	92.63
1984	360.29	103.9	92.46
1985	317.30	107.6	108.90
1986	367.87	109.6	136.00
1987	446.50	113.6	161.70
1988	436.93	118.3	149.91
1989	381.28	124.0	180.02
1990	384.08	130.7	183.46
1991	362.04	136.2	206.33

Source: Data on CPI and NYSE Index are from the *Economic Report of the President,* January 1993, Tables B-59 and B-91, respectively. Data on gold prices are from U.S. Department of Commerce, Bureau of Economic Analysis, *Business Statistics, 1963–1991,* p. 68.

a. Plot in the same scattergram gold prices, CPI, and the NYSE Index.
b. An investment is supposed to be a hedge against inflation if its price and/or rate of return at least keeps pace with inflation. To test this hypothesis, suppose you decide to fit the following model, assuming the scatterplot in **a** suggests that this is appropriate:

$$\text{Gold price}_t = \beta_1 + \beta_2\, \text{CPI}_t + u_t$$
$$\text{NYSE index}_t = \beta_1 + \beta_2\, \text{CPI}_t + u_t$$

3.23. Table 3.8 gives data on gross domestic product (GDP) for the United States for the years 1959–1997.
a. Plot the GDP data in current and constant (i.e., 1992) dollars against time.
b. Letting Y denote GDP and X time (measured chronologically starting with 1 for 1959, 2 for 1960, through 39 for 1997), see if the following model fits the GDP data:

$$Y_t = \beta_1 + \beta_2\, X_t + u_t$$

Estimate this model for both current and constant-dollar GDP.
c. How would you interpret β_2?
d. If there is a difference between β_2 estimated for current-dollar GDP and that estimated for constant-dollar GDP, what explains the difference?

TABLE 3.8 NOMINAL AND REAL GDP, UNITED STATES, 1959–1997

Year	NGDP	RGDP	Year	NGDP	RGDP
1959	507.2000	2210.200	1979	2557.500	4630.600
1960	526.6000	2262.900	1980	2784.200	4615.000
1961	544.8000	2314.300	1981	3115.900	4720.700
1962	585.2000	2454.800	1982	3242.100	4620.300
1963	617.4000	2559.400	1983	3514.500	4803.700
1964	663.0000	2708.400	1984	3902.400	5140.100
1965	719.1000	2881.100	1985	4180.700	5323.500
1966	787.7000	3069.200	1986	4422.200	5487.700
1967	833.6000	3147.200	1987	4692.300	5649.500
1968	910.6000	3293.900	1988	5049.600	5865.200
1969	982.2000	3393.600	1989	5438.700	6062.000
1970	1035.600	3397.600	1990	5743.800	6136.300
1971	1125.400	3510.000	1991	5916.700	6079.400
1972	1237.300	3702.300	1992	6244.400	6244.400
1973	1382.600	3916.300	1993	6558.100	6389.600
1974	1496.900	3891.200	1994	6947.000	6610.700
1975	1630.600	3873.900	1995	7269.600	6761.700
1976	1819.000	4082.900	1996	7661.600	6994.800
1977	2026.900	4273.600	1997	8110.900	7269.800
1978	2291.400	4503.000			

Note: NGDP = nominal GDP (current dollars in billions).
RGDP = real GDP (1992 billions of dollars).
Source: Economic Report of the President, 1999, Tables B-1 and B-2, pp. 326–328.

e. From your results what can you say about the nature of inflation in the United States over the sample period?

3.24. Using the data given in Table I.1 of the Introduction, verify Eq. (3.7.1).

3.25. For the S.A.T. example given in exercise 2.16 do the following:

a. Plot the female verbal score against the male verbal score.

b. If the scatterplot suggests that a linear relationship between the two seems appropriate, obtain the regression of female verbal score on male verbal score.

c. If there is a relationship between the two verbal scores, is the relationship *causal?*

3.26. Repeat exercise 3.24, replacing math scores for verbal scores.

3.27. Monte Carlo study *classroom assignment:* Refer to the 10 X values given in Table 3.2. Let $\beta_1 = 25$ and $\beta_2 = 0.5$. Assume $u_i \approx N(0, 9)$, that is, u_i are normally distributed with mean 0 and variance 9. Generate 100 samples using these values, obtaining 100 estimates of β_1 and β_2. Graph these estimates. What conclusions can you draw from the Monte Carlo study? *Note:* Most statistical packages now can generate random variables from most well-known probability distributions. Ask your instructor for help, in case you have difficulty generating such variables.

APPENDIX 3A

3A.1 DERIVATION OF LEAST-SQUARES ESTIMATES

Differentiating (3.1.2) partially with respect to $\hat{\beta}_1$ and $\hat{\beta}_2$, we obtain

$$\frac{\partial(\sum \hat{u}_i^2)}{\partial \hat{\beta}_1} = -2\sum(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) = -2\sum \hat{u}_i \tag{1}$$

$$\frac{\partial(\sum \hat{u}_i^2)}{\partial \hat{\beta}_2} = -2\sum(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)X_i = -2\sum \hat{u}_i X_i \tag{2}$$

Setting these equations to zero, after algebraic simplification and manipulation, gives the estimators given in Eqs. (3.1.6) and (3.1.7).

3A.2 LINEARITY AND UNBIASEDNESS PROPERTIES OF LEAST-SQUARES ESTIMATORS

From (3.1.8) we have

$$\hat{\beta}_2 = \frac{\sum x_i Y_i}{\sum x_i^2} = \sum k_i Y_i \tag{3}$$

where

$$k_i = \frac{x_i}{(\sum x_i^2)}$$

which shows that $\hat{\beta}_2$ is a **linear estimator** because it is a linear function of Y; actually it is a weighted average of Y_i with k_i serving as the weights. It can similarly be shown that $\hat{\beta}_1$ too is a linear estimator.

Incidentally, note these properties of the weights k_i:

1. Since the X_i are assumed to be nonstochastic, the k_i are nonstochastic too.

2. $\sum k_i = 0$.

3. $\sum k_i^2 = 1/\sum x_i^2$.

4. $\sum k_i x_i = \sum k_i X_i = 1$. These properties can be directly verified from the definition of k_i.

For example,

$$\begin{aligned}\sum k_i = \sum\left(\frac{x_i}{\sum x_i^2}\right) &= \frac{1}{\sum x_i^2}\sum x_i, && \text{since for a given sample } \textstyle\sum x_i^2 \text{ is known}\\ &= 0, && \text{since } \textstyle\sum x_i, \text{ the sum of deviations from the mean value, is always zero}\end{aligned}$$

Now substitute the PRF $Y_i = \beta_1 + \beta_2 X_i + u_i$ into (3) to obtain

$$\begin{aligned}\hat{\beta}_2 &= \sum k_i(\beta_1 + \beta_2 X_i + u_i)\\ &= \beta_1\sum k_i + \beta_2\sum k_i X_i + \sum k_i u_i\\ &= \beta_2 + \sum k_i u_i\end{aligned} \tag{4}$$

where use is made of the properties of k_i noted earlier.

Now taking expectation of (4) on both sides and noting that k_i, being nonstochastic, can be treated as constants, we obtain

$$\begin{aligned}E(\hat{\beta}_2) &= \beta_2 + \sum k_i E(u_i)\\ &= \beta_2\end{aligned} \tag{5}$$

since $E(u_i) = 0$ by assumption. Therefore, $\hat{\beta}_2$ is an unbiased estimator of β_2. Likewise, it can be proved that $\hat{\beta}_1$ is also an unbiased estimator of β_1.

3A.3 VARIANCES AND STANDARD ERRORS OF LEAST-SQUARES ESTIMATORS

Now by the definition of variance, we can write

$$\begin{aligned}\text{var}\,(\hat{\beta}_2) &= E[\hat{\beta}_2 - E(\hat{\beta}_2)]^2\\ &= E(\hat{\beta}_2 - \beta_2)^2 && \text{since } E(\hat{\beta}_2) = \beta_2\\ &= E\left(\sum k_i u_i\right)^2 && \text{using Eq. (4) above}\\ &= E\left(k_1^2u_1^2 + k_2^2u_2^2 + \cdots + k_n^2u_n^2 + 2k_1k_2u_1u_2 + \cdots + 2k_{n-1}k_nu_{n-1}u_n\right)\end{aligned} \tag{6}$$

Since by assumption, $E(u_i^2) = \sigma^2$ for each i and $E(u_i u_j) = 0$, $i \neq j$, it follows that

$$\begin{aligned} \operatorname{var}(\hat{\beta}_2) &= \sigma^2 \sum k_i^2 \\ &= \frac{\sigma^2}{\sum x_i^2} \qquad \text{(using the definition of } k_i^2\text{)} \\ &= \text{Eq. (3.3.1)} \end{aligned} \tag{7}$$

The variance of $\hat{\beta}_1$ can be obtained following the same line of reasoning already given. Once the variances of $\hat{\beta}_1$ and $\hat{\beta}_2$ are obtained, their positive square roots give the corresponding standard errors.

3A.4 COVARIANCE BETWEEN $\hat{\beta}_1$ AND $\hat{\beta}_2$

By definition,

$$\begin{aligned} \operatorname{cov}(\hat{\beta}_1, \hat{\beta}_2) &= E\{[\hat{\beta}_1 - E(\hat{\beta}_1)][\hat{\beta}_2 - E(\hat{\beta}_2)]\} \\ &= E(\hat{\beta}_1 - \beta_1)(\hat{\beta}_2 - \beta_2) \qquad \text{(Why?)} \\ &= -\bar{X} E(\hat{\beta}_2 - \beta_2)^2 \\ &= -\bar{X} \operatorname{var}(\hat{\beta}_2) \\ &= \text{Eq. (3.3.9)} \end{aligned} \tag{8}$$

where use is made of the fact that $\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}$ and $E(\hat{\beta}_1) = \bar{Y} - \beta_2 \bar{X}$, giving $\hat{\beta}_1 - E(\hat{\beta}_1) = -\bar{X}(\hat{\beta}_2 - \beta_2)$. *Note:* $\operatorname{var}(\hat{\beta}_2)$ is given in (3.3.1).

3A.5 THE LEAST-SQUARES ESTIMATOR OF σ^2

Recall that

$$Y_i = \beta_1 + \beta_2 X_i + u_i \tag{9}$$

Therefore,

$$\bar{Y} = \beta_1 + \beta_2 \bar{X} + \bar{u} \tag{10}$$

Subtracting (10) from (9) gives

$$y_i = \beta_2 x_i + (u_i - \bar{u}) \tag{11}$$

Also recall that

$$\hat{u}_i = y_i - \hat{\beta}_2 x_i \tag{12}$$

Therefore, substituting (11) into (12) yields

$$\hat{u}_i = \beta_2 x_i + (u_i - \bar{u}) - \hat{\beta}_2 x_i \tag{13}$$

Collecting terms, squaring, and summing on both sides, we obtain

$$\sum \hat{u}_i^2 = (\hat{\beta}_2 - \beta_2)^2 \sum x_i^2 + \sum (u_i - \bar{u})^2 - 2(\hat{\beta}_2 - \beta_2) \sum x_i (u_i - \bar{u}) \quad \textbf{(14)}$$

Taking expectations on both sides gives

$$\begin{aligned} E\left(\sum \hat{u}_i^2\right) &= \sum x_i^2 E(\hat{\beta}_2 - \beta_2)^2 + E\left[\sum (u_i - \bar{u})^2\right] - 2E\left[(\hat{\beta}_2 - \beta_2) \sum x_i (u_i - \bar{u})\right] \\ &= \sum x_i^2 \operatorname{var}(\hat{\beta}_2) + (n-1) \operatorname{var}(u_i) - 2E\left[\sum k_i u_i (x_i u_i)\right] \\ &= \sigma^2 + (n-1)\sigma^2 - 2E\left[\sum k_i x_i u_i^2\right] \\ &= \sigma^2 + (n-1)\sigma^2 - 2\sigma^2 \\ &= (n-2)\sigma^2 \end{aligned} \quad \textbf{(15)}$$

where, in the last but one step, use is made of the definition of k_i given in Eq. (3) and the relation given in Eq. (4). Also note that

$$\begin{aligned} E\sum (u_i - \bar{u})^2 &= E\left[\sum u_i^2 - n\bar{u}^2\right] \\ &= E\left[\sum u_i^2 - n\left(\frac{\sum u_i}{n}\right)^2\right] \\ &= E\left[\sum u_i^2 - \frac{1}{n}\sum (u_i^2)\right] \\ &= n\sigma^2 - \frac{n}{n}\sigma^2 = (n-1)\sigma^2 \end{aligned}$$

where use is made of the fact that the u_i are uncorrelated and the variance of each u_i is σ^2.

Thus, we obtain

$$E\left(\sum \hat{u}_i^2\right) = (n-2)\sigma^2 \quad \textbf{(16)}$$

Therefore, if we define

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-2} \quad \textbf{(17)}$$

its expected value is

$$E(\hat{\sigma}^2) = \frac{1}{n-2} E\left(\sum \hat{u}_i^2\right) = \sigma^2 \qquad \text{using (16)} \quad \textbf{(18)}$$

which shows that $\hat{\sigma}^2$ is an unbiased estimator of true σ^2.

3A.6 MINIMUM-VARIANCE PROPERTY OF LEAST-SQUARES ESTIMATORS

It was shown in Appendix 3A, Section 3A.2, that the least-squares estimator $\hat{\beta}_2$ is linear as well as unbiased (this holds true of $\hat{\beta}_1$ too). To show that these estimators are also minimum variance in the class of all linear unbiased estimators, consider the least-squares estimator $\hat{\beta}_2$:

$$\hat{\beta}_2 = \sum k_i Y_i$$

where

$$k_i = \frac{X_i - \bar{X}}{\sum (X_i - \bar{X})^2} = \frac{x_i}{\sum x_i^2} \qquad \text{(see Appendix 3A.2)} \tag{19}$$

which shows that $\hat{\beta}_2$ is a weighted average of the Y's, with k_i serving as the weights.

Let us define an alternative linear estimator of β_2 as follows:

$$\beta_2^* = \sum w_i Y_i \tag{20}$$

where w_i are also weights, not necessarily equal to k_i. Now

$$\begin{aligned} E(\beta_2^*) &= \sum w_i E(Y_i) \\ &= \sum w_i (\beta_1 + \beta_2 X_i) \\ &= \beta_1 \sum w_i + \beta_2 \sum w_i X_i \end{aligned} \tag{21}$$

Therefore, for β_2^* to be unbiased, we must have

$$\sum w_i = 0 \tag{22}$$

and

$$\sum w_i X_i = 1 \tag{23}$$

Also, we may write

$$\begin{aligned} \operatorname{var}(\beta_2^*) &= \operatorname{var} \sum w_i Y_i \\ &= \sum w_i^2 \operatorname{var} Y_i \qquad [\textit{Note:} \operatorname{var} Y_i = \operatorname{var} u_i = \sigma^2] \\ &= \sigma^2 \sum w_i^2 \qquad [\textit{Note:} \operatorname{cov}(Y_i, Y_j) = 0 \, (i \neq j)] \\ &= \sigma^2 \sum \left(w_i - \frac{x_i}{\sum x_i^2} + \frac{x_i}{\sum x_i^2} \right)^2 \qquad \text{(Note the mathematical trick)} \\ &= \sigma^2 \sum \left(w_i - \frac{x_i}{\sum x_i^2} \right)^2 + \sigma^2 \frac{\sum x_i^2}{\left(\sum x_i^2 \right)^2} + 2\sigma^2 \sum \left(w_i - \frac{x_i}{\sum x_i^2} \right) \left(\frac{x_i}{\sum x_i^2} \right) \\ &= \sigma^2 \sum \left(w_i - \frac{x_i}{\sum x_i^2} \right)^2 + \sigma^2 \left(\frac{1}{\sum x_i^2} \right) \end{aligned} \tag{24}$$

because the last term in the next to the last step drops out. (Why?)

Since the last term in (24) is constant, the variance of (β_2^*) can be minimized only by manipulating the first term. If we let

$$w_i = \frac{x_i}{\sum x_i^2}$$

Eq. (24) reduces to

$$\begin{aligned} \text{var}\,(\beta_2^*) &= \frac{\sigma^2}{\sum x_i^2} \\ &= \text{var}\,(\hat{\beta}_2) \end{aligned} \tag{25}$$

In words, with weights $w_i = k_i$, which are the least-squares weights, the variance of the linear estimator β_2^* is equal to the variance of the least-squares estimator $\hat{\beta}_2$; otherwise $\text{var}\,(\beta_2^*) > \text{var}\,(\hat{\beta}_2)$. To put it differently, if there is a minimum-variance linear unbiased estimator of β_2, it must be the least-squares estimator. Similarly it can be shown that $\hat{\beta}_1$ is a minimum-variance linear unbiased estimator of β_1.

3A.7 CONSISTENCY OF LEAST-SQUARES ESTIMATORS

We have shown that, in the framework of the classical linear regression model, the least-squares estimators are unbiased (and efficient) in any sample size, small or large. But sometimes, as discussed in **Appendix A,** an estimator may not satisfy one or more desirable statistical properties in small samples. But as the sample size increases indefinitely, the estimators possess several desirable statistical properties. These properties are known as the **large sample,** or **asymptotic, properties.** In this appendix, we will discuss one large sample property, namely, the property of **consistency,** which is discussed more fully in **Appendix A.** For the two-variable model we have already shown that the OLS estimator $\hat{\beta}_2$ is an unbiased estimator of the true β_2. Now we show that $\hat{\beta}_2$ is also a consistent estimator of β_2. As shown in **Appendix A,** a sufficient condition for consistency is that $\hat{\beta}_2$ is unbiased and that its variance tends to zero as the sample size n tends to infinity.

Since we have already proved the unbiasedness property, we need only show that the variance of $\hat{\beta}_2$ tends to zero as n increases indefinitely. We know that

$$\text{var}\,(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_i^2} = \frac{\sigma^2/n}{\sum x_i^2/n} \tag{26}$$

By dividing the numerator and denominator by n, we do not change the equality.

Now

$$\underbrace{\lim \text{var}\,(\hat{\beta}_2)}_{n \to \infty} = \underbrace{\lim \left(\frac{\sigma^2/n}{\sum x_i^2/n} \right)}_{n \to \infty} = 0 \tag{27}$$

where use is made of the facts that (1) the limit of a ratio quantity is the limit of the quantity in the numerator to the limit of the quantity in the denominator (refer to any calculus book); (2) as n tends to infinity, σ^2/n tends to zero because σ^2 is a finite number; and $[(\sum x_i^2)/n] \neq 0$ because the variance of X has a finite limit because of Assumption 8 of CLRM.

The upshot of the preceding discussion is that the OLS estimator $\hat{\beta}_2$ is a consistent estimator of true β_2. In like fashion, we can establish that $\hat{\beta}_1$ is also a consistent estimator. Thus, in repeated (small) samples, the OLS estimators are unbiased and as the sample size increases indefinitely the OLS estimators are consistent. As we shall see later, even if some of the assumptions of CLRM are not satisfied, we may be able to obtain consistent estimators of the regression coefficients in several situations.