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)

$$Y_i = \beta_1 + \beta_2 X_i + u_i \tag{2.4.2}$$

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_i, 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

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

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

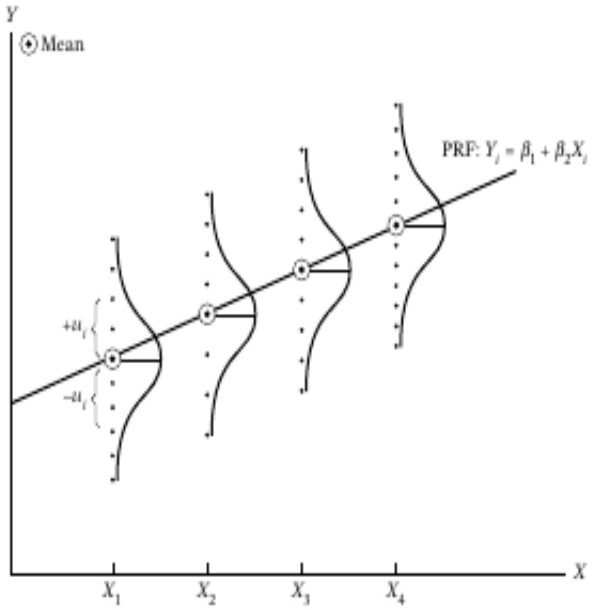

FIGURE 3.3 Conditional distribution of the disturbances u_i.

[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 $E(Y_i \mid X_i) = \beta_1 + \beta_2 X_i$. (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

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

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.

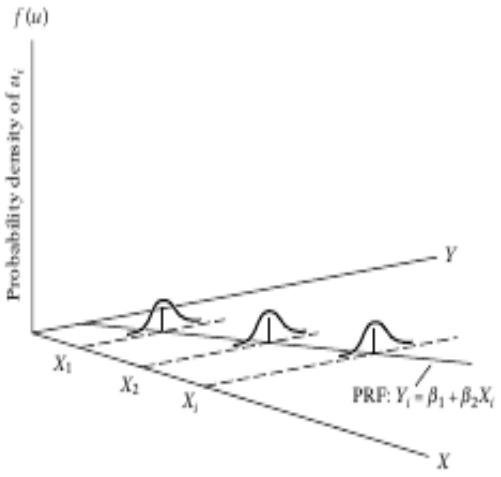

FIGURE 3.4 Homoscedasticity.

[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.

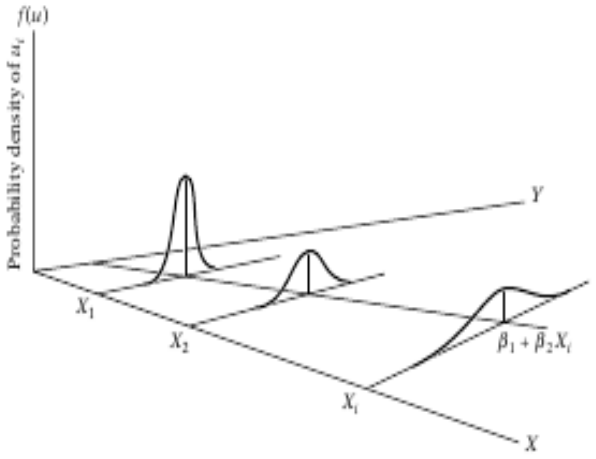

FIGURE 3.5 Heteroscedasticity.

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.

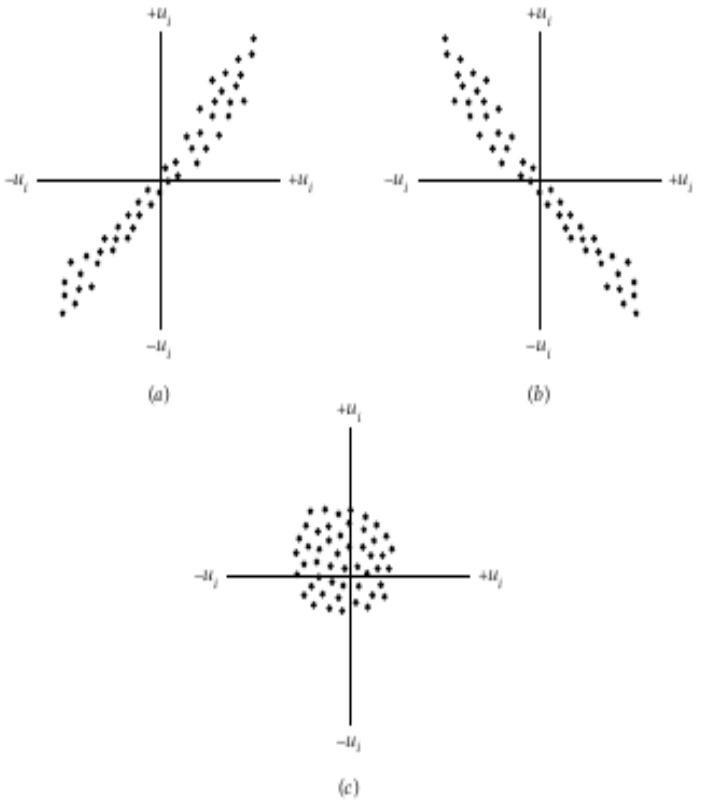

FIGURE 3.6 Patterns of correlation among the disturbances. (*a*) positive serial correlation; (*b*) negative serial correlation; (*c*) zero correlation.

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

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

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.

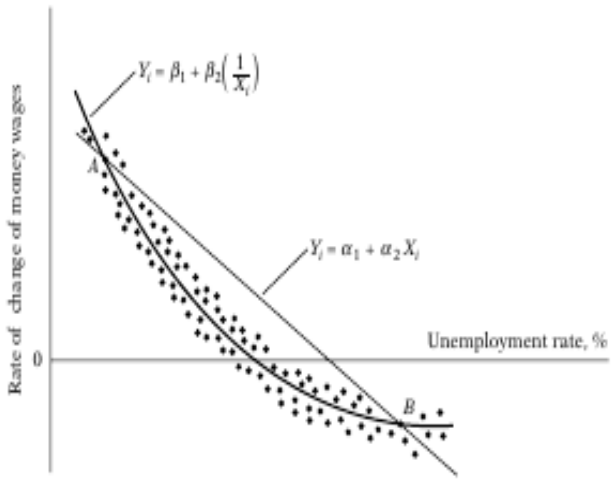

FIGURE 3.7 Linear and nonlinear Phillips curves.

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 $\text{var}\,(\hat{\beta}_2)$ 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.