Eq. (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 Eq. (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 Eq. (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 7 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.

[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.

ASSUMPTION 1 **Linear Regression Model:** The regression model is **linear in the parameters,** though it may or may not be linear in the variables. That is the regression model as shown in Eq. (2.4.2):

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

As will be discussed in Chapter 7, this model can be extended to include more explanatory variables.

We have 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 for most of this book.[8] Keep in mind that the regressand Y and the regressor X may be nonlinear, as discussed in Chapter 2.

ASSUMPTION 2 **Fixed X Values or X Values Independent of the Error Term:** Values taken by the regressor X may be considered fixed in repeated samples (the case of fixed regressor) or they may be sampled along with the dependent variable Y (the case of stochastic regressor). In the latter case, it is assumed that the X variable(s) and the error term are independent, that is, cov $(X_i, u_i) = 0$.

This can be explained in terms of our example given in Table 2.1 (page 35). Consider the various Y populations corresponding to the levels of income shown in the table. Keeping the value of income X fixed, say, at level \$80, we draw at random a family and observe its weekly family consumption Y as, say, \$60. Still keeping X at \$80, we draw at random another family and observe its Y value at \$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.

Why do we assume that the X values are nonstochastic? Given that, in most social sciences, data usually are collected randomly on both the Y and X variables, it seems natural to assume the opposite—that the X variable, like the Y variable, is also random or stochastic. But initially we assume that the X variable(s) is nonstochastic for the following reasons:

First, this is done initially to simplify the analysis and to introduce the reader to the complexities of regression analysis gradually. *Second,* in experimental situations it may not be unrealistic to assume that the X values are fixed. For example, a farmer may divide his land into several parcels and apply different amounts of fertilizer to these parcels to see its effect on crop yield. Likewise, a department store may decide to offer different rates of discount on a product to see its effect on consumers. Sometimes we may want to fix the X values for a specific purpose. Suppose we are trying to find out the average weekly earnings of workers (Y) with various levels of education (X), as in the case of the data given in Table 2.6. In this case, the X variable can be considered fixed or nonrandom. *Third,* as we show in Chapter 13, even if the X variables are stochastic, the statistical results of linear regression based

[8] However, a brief discussion of nonlinear-in-parameter regression models is given in Chapter 14 for the benefit of more advanced students.

on the case of fixed regressors are also valid when the X's are random, provided that some conditions are met. One condition is that regressor X and the error term u_i are independent. As James Davidson notes, ". . . this model [i.e., stochastic regressors] 'mimics' the fixed regressor model, and . . . many of the statistical properties of least squares in the fixed regressor model continue to hold."[9]

For all these reasons, we will first discuss the (fixed-regressor) CLRM in considerable detail. However, in Chapter 13 we will discuss the case of stochastic regressors in some detail and point out the occasions where we need to consider the stochastic regressor models. Incidentally, note that if the X variable(s) is stochastic, the resulting model is called the **neo-classical linear regression model (NLRM),**[10] in contrast to the CLRM, where the X's are treated as fixed or nonrandom. For discussion purposes, we will call the former the **stochastic regressor model** and the latter the **fixed regressor model.**

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. Symbolically, we have

$$E(u_i|X_i) = 0 \tag{3.2.1}$$

Or, if X is nonstochastic,

$$E(u_i) = 0$$

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

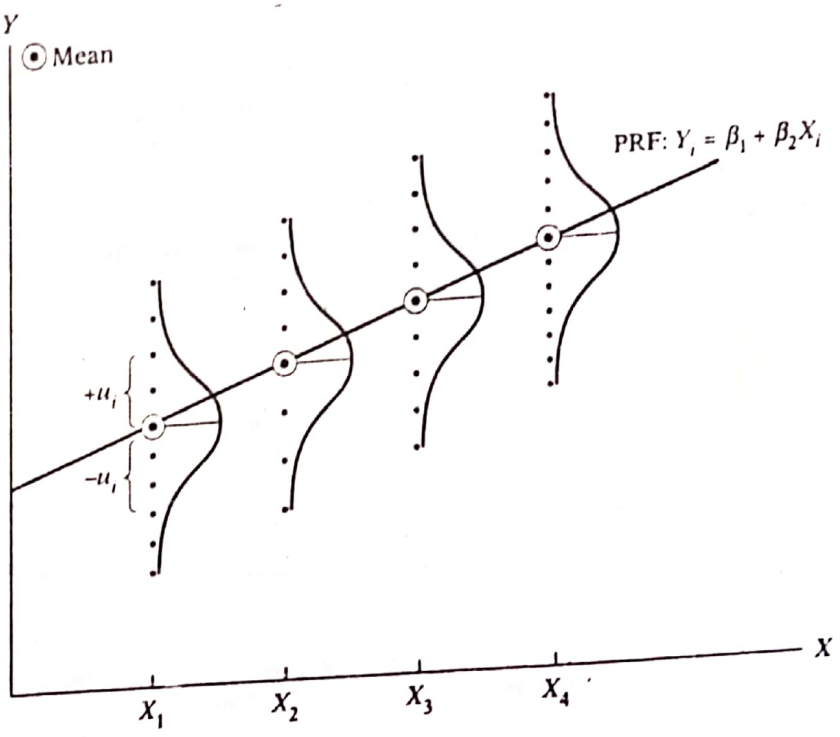

FIGURE 3.3 Conditional distribution of the disturbances u_i.

[9]James Davidson, *Econometric Theory*, Blackwell Publishers, U.K., 2000, p. 10.

[10]A term due to Arthur S. Goldberger, *A Course in Econometrics*, Harvard University Press, Cambridge, MA, 1991, p. 264.

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. Equation 3.2.1 requires that the average or mean value of these deviations corresponding to any given X should be zero.

This assumption should not be difficult to comprehend in view of the discussion in Section 2.4 (see Eq. [2.4.5]). Assumption 3 simply says that the factors not explicitly included in the model, and therefore subsumed in u_i, do not systematically affect the mean value of Y; in other words, the positive u_i values cancel out the negative u_i values so that their average or mean effect on Y is zero.[11]

In passing, note that the assumption $E(u_i|X_i) = 0$ implies that $E(Y_i|X_i) = \beta_1 + \beta_2 X_i$. (Why?) Therefore, the two assumption are equivalent.

It is important to point out that Assumption 3 implies that there is no **specification bias** or **specification error** in the model used in empirical analysis. In other words, the regression model is correctly specified. Leaving out important explanatory variables, including unnecessary variables, or choosing the wrong functional form of the relationship between the Y and X variables are some examples of specification error. We will discuss this topic in considerable detail in Chapter 13.

Note also that if the conditional mean of one random variable given another random variable is zero, the covariance between the two variables is zero and hence the two variables are uncorrelated. Assumption 3 therefore implies that X_i and u_i are uncorrelated.[12]

The reason for assuming that the disturbance term u and the explanatory variable(s) X are uncorrelated is simple. When we expressed the PRF as in Eq. (2.4.2), we assumed that X and u (which represent the influence of all omitted variables) have separate (and additive) influences 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 decreases when u decreases. Similarly, if X and u are negatively correlated, X increases when u decreases and decreases when u increases. In situations like this it is quite possible that the error term actually includes some variables that should have been included as additional regressors in the model. This is why Assumption 3 is another way of stating that there is no specification error in the chosen regression model.

ASSUMPTION 4 **Homoscedasticity or Constant Variance of u_i:** The variance of the error, or disturbance, term is the same regardless of the value of X. Symbolically,

$$\begin{aligned}\text{var}(u_i) &= E[u_i - E(u_i|X_i)]^2 \\ &= E(u_i^2|X_i), \text{ because of Assumption 3} \\ &= E(u_i^2), \text{ if } X_i \text{ are nonstochastic} \\ &= \sigma^2\end{aligned} \tag{3.2.2}$$

where var stands for variance.

[11]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.

[12]The converse, however, is not true because correlation is a measure of linear association only. That is, even if X_i and u_i are uncorrelated, the conditional mean of u_i given X_i may not be zero. However, if X_i and u_i are correlated, $E(u_i|X_i)$ must be nonzero, violating Assumption 3. We owe this point to Stock and Watson. See James H. Stock and Mark W. Watson, *Introduction to Econometrics*, Addison-Wesley, Boston, 2003, pp. 104–105.

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, $\text{var}(u|X_1) < \text{var}(u|X_2), \ldots, < \text{var}(u|X_i)$. 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,

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

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. In short, the observations are sampled independently. Symbolically,

$$\text{cov}(u_i, u_j | X_i, X_j) = 0 \tag{3.2.5}$$
$$\text{cov}(u_i, u_j) = 0, \text{ if } X \text{ is nonstochastic}$$

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

In words, Equation 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 Figures 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 Figures 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.

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 7 **The Nature of X Variables:** The X values in a given sample must not all be the same. Technically, var (X) must be a positive number. Furthermore, there can be no outliers in the values of the X variable, that is, values that are very large in relation to the rest of the observations.

The assumption that there is variability in the X values is also not as 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 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!

The requirement that there are no outliers in the X values is to avoid the regression results being dominated by such outliers. If there are a few X values that are, say, 20 times the average of the X values, the estimated regression lines with or without such observations might be vastly different. Very often such outliers are the result of human errors of arithmetic or mixing samples from different populations. In Chapter 13 we will discuss this topic further.

Our discussion of the assumptions underlying the classical linear regression model is now complete. It is important to note that all of 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 $E(u_i|X_i) = 0$. 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 are not only autocorrelated but are also heteroscedastic (see Chapter 12).

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."[13]

One may not subscribe to this viewpoint fully, but recall that in any scientific study we make certain assumptions because they facilitate the development of the subject matter in gradual steps, not because they are necessarily realistic in the sense that they replicate

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