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.**

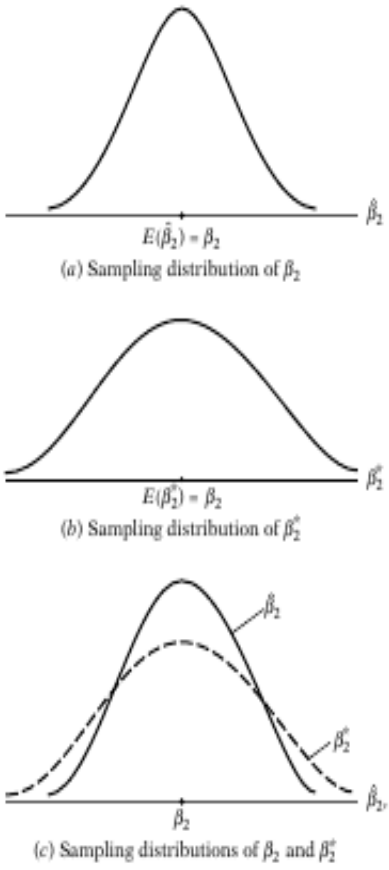
In the regression context it can be proved that the OLS estimators are BLUE. This is the gist of the famous Gauss–Markov theorem, which can be stated as follows:

> **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 $\hat{\beta}_2$

(*b*) Sampling distribution of β_2^*

(*c*) Sampling distributions of $\hat{\beta}_2$ 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.

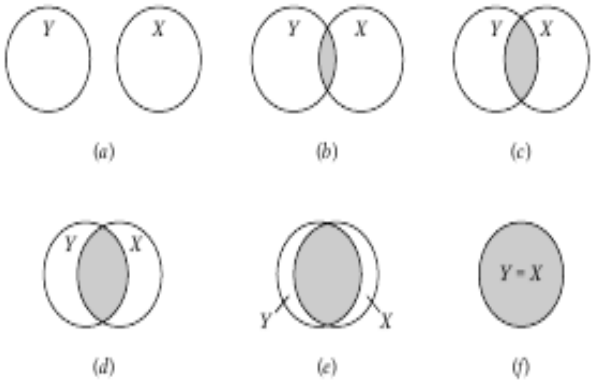


FIGURE 3.9 The Ballentine view of r^2: (a) $r^2 = 0$; (f) $r^2 = 1$.

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

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

or in the deviation form

$$y_i = \hat{y}_i + \hat{u}_i \tag{3.5.1}$$

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.

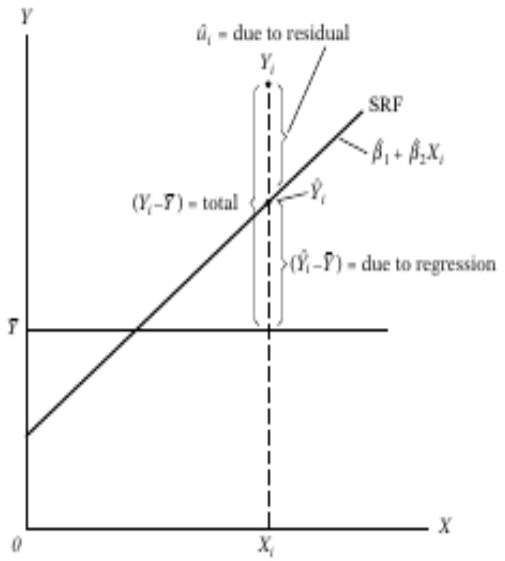


FIGURE 3.10 Breakdown of the variation of Y_i into two components.

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**.

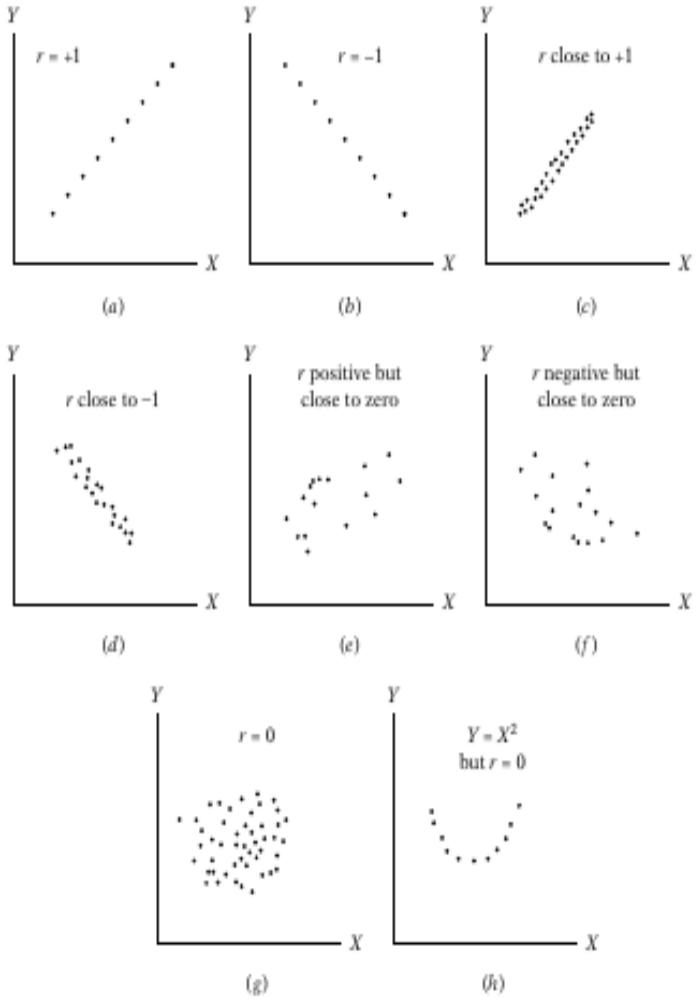


FIGURE 3.11 Correlation patterns (adapted from Henri Theil, *Introduction to Econometrics*, Prentice-Hall, Englewood Cliffs, N.J., 1978, p. 86).

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.