

## 4 THE LEAST SQUARES ESTIMATOR

### 4.1 INTRODUCTION

Chapter 3 treated fitting the linear regression to the data by least squares as a purely algebraic exercise. In this chapter, we will examine in detail least squares as an **estimator** of the model parameters of the linear regression model (defined in the following Table 4.1). We begin in Section 4.2 by returning to the question raised but not answered in Footnote 1, Chapter 3, that is, why should we use least squares? We will then analyze the estimator in detail. There are other candidates for estimating  $\beta$ . For example, we might use the coefficients that minimize the sum of absolute values of the residuals. The question of which estimator to choose is based on the **statistical properties** of the candidates, such as unbiasedness, consistency, efficiency, and their sampling distributions. Section 4.3 considers **finite-sample properties** such as unbiasedness. The finite-sample properties of the least squares estimator are independent of the sample size. The linear model is one of relatively few settings in which definite statements can be made about the exact finite sample properties of any estimator. In most cases, the only known properties are those that apply to large samples. Here, we can only approximate finite-sample behavior by using what we know about large-sample properties. Thus, in Section 4.4, we will examine the large-sample, or **asymptotic properties** of the least squares estimator of the regression model.

Discussions of the properties of an estimator are largely concerned with **point estimation** that is, in how to use the sample information as effectively as possible to produce the best single estimate of the model parameters. **Interval estimation**, considered in Section 4.5, is concerned with computing estimates that make explicit the uncertainty inherent in using randomly sampled data to estimate population quantities. We will consider some applications of interval estimation of parameters and some functions of parameters in Section 4.5. One of the most familiar applications of interval estimation is in using the model to predict the dependent variable, and to provide a plausible range of uncertainty for that prediction. Section 4.6 considers prediction and forecasting using the estimated regression model.

The analysis assumes that the data in hand correspond to the assumptions of the model. In Section 4.7, we consider several practical problems that arise in analyzing nonexperimental data. Assumption A2, full rank of  $X$ , is taken as a given. As we noted in Section 2.3.2, when this assumption is not met, the model is not estimable, regardless of the sample size. **Multicollinearity**, the near failure of this assumption in real world data is examined in Sections 4.7.1 to 4.7.3. Missing data have the potential to derail the entire analysis. The benign case in which missing values are simply manageable random gaps in the data set is considered in Section 4.7.4. The more complicated case of nonrandomly missing data is discussed in Chapter 18. Finally, the problem of badly measured data is examined in Section 4.7.5.

<sup>1</sup> This discussion will use our results on asymptotic distributions. It may be helpful to review Appendix D before proceeding to this material.

## 44 PART I ♦ The Linear Regression Model

TABLE 4.1 Assumptions of the Classical Linear Regression Model

- A1. Linearity:**  $y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \dots + \beta_K x_{iK} + \varepsilon_i$ . *italic*  
**A2. Full rank:** The  $n \times K$  sample data matrix,  $\mathbf{X}$  has full column rank.  
**A3. Exogeneity of the independent variables:**  $E[\varepsilon_i | x_{j1}, x_{j2}, \dots, x_{jK}] = 0$ ,  $i, j = 1, \dots, n$ .  
 There is no correlation between the disturbances and the independent variables.  
**A4. Homoscedasticity and nonautocorrelation:** Each disturbance,  $\varepsilon_i$ , has the same finite variance,  $\sigma^2$ , and is uncorrelated with every other disturbance,  $\varepsilon_j$  conditioned on  $x$ .  
**A5. Stochastic or nonstochastic data:**  $(x_{i1}, x_{i2}, \dots, x_{iK})$   $i = 1, \dots, n$ .  
**A6. Normal distribution:** The disturbances are normally distributed.

practical terms, is a minor consideration. Indeed, nearly all that we do with the regression model departs from this assumption fairly quickly. It serves only as a useful departure point. The issue is considered in Section 4.5. The normality of the disturbances assumed in A6 is crucial in obtaining the **sampling distributions** of several useful statistics that are used in the analysis of the linear model. We note that in the course of our analysis of the linear model as we proceed through the text, most of these assumptions will be discarded.

The **finite-sample properties** of the least squares estimator are independent of the sample size. But the classical regression model with normally distributed disturbances and independent observations is a special case that does not include many of the most common applications, such as panel data and most time-series models. Section 4.9 will generalize the classical regression model by relaxing these two important assumptions.

The linear model is one of relatively few settings in which any definite statements can be made about the exact finite sample properties of any estimator. In most cases, the only known properties of the estimators are those that apply to large samples. We can only approximate finite-sample behavior by using what we know about large-sample properties. This chapter will also examine the **asymptotic properties** of the parameter estimators in the classical regression model.<sup>1</sup>

## 4.2 MOTIVATING LEAST SQUARES

Ease of computation is one reason that least squares is so popular. However, there are several other justifications for this technique. First, least squares is a natural approach to estimation, which makes explicit use of the structure of the model as laid out in the assumptions. Second, even if the true model is not a linear regression, the regression line fit by least squares is an optimal linear predictor for the dependent variable. Thus, it enjoys a sort of robustness that other estimators do not. Finally, under the very specific assumptions of the classical model, by one reasonable criterion, least squares will be the most efficient use of the data. We will consider each of these in turn.

## 4.2.1 THE POPULATION ORTHOGONALITY CONDITIONS

Let  $\mathbf{x}$  denote the vector of independent variables in the population regression model and for the moment, based on assumption A5, the data may be stochastic or nonstochastic.

<sup>1</sup>Most of this discussion will use our results on asymptotic distributions. It may be helpful to review Appendix D before proceeding to this material.

## CHAPTER 4 ♦ Statistical Properties of the Least Squares Estimator 45

Assumption A3 states that the disturbances in the population are stochastically orthogonal to the independent variables in the model; that is,  $E[\varepsilon | \mathbf{x}] = 0$ . It follows that  $\text{Cov}[\mathbf{x}, \varepsilon] = \mathbf{0}$ . Since (by the law of iterated expectations—Theorem B.1)  $E_{\mathbf{x}}\{E[\varepsilon | \mathbf{x}]\} = E[\varepsilon] = 0$ , we may write this as

$$E_{\mathbf{x}} E_{\varepsilon}[\mathbf{x}\varepsilon] = E_{\mathbf{x}} E_{\varepsilon}[\mathbf{x}(y - \mathbf{x}'\beta)] = \mathbf{0}$$

or

$$E_{\mathbf{x}} E_y[\mathbf{x}y] = E_{\mathbf{x}}[\mathbf{x}\mathbf{x}']\beta. \quad (4-1)$$

(The right-hand side is not a function of  $y$  so the expectation is taken only over  $\mathbf{x}$ .) Now, recall the least squares normal equations,  $\mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{X}\mathbf{b}$ . Divide this by  $n$  and write it as a summation to obtain

$$\left(\frac{1}{n} \sum_{i=1}^n \mathbf{x}_i y_i\right) = \left(\frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i'\right) \mathbf{b}. \quad (4-2)$$

Equation (4-1) is a population relationship. Equation (4-2) is a sample analog. Assuming the conditions underlying the laws of large numbers presented in Appendix D are met, the sums on the left-hand and right-hand sides of (4-2) are estimators of their counterparts in (4-1). Thus, by using least squares, we are mimicking in the sample the relationship in the population. We'll return to this approach to estimation in Chapters 12 and 13 under the subject of GMM estimation.

## 4.2.2 MINIMUM MEAN SQUARED ERROR PREDICTOR

As an alternative approach, consider the problem of finding an **optimal linear predictor** for  $y$ . Once again, ignore Assumption A6 and, in addition, drop Assumption A1 that the conditional mean function,  $E[y | \mathbf{x}]$  is linear. For the criterion, we will use the mean squared error rule, so we seek the minimum mean squared error linear predictor of  $y$ , which we'll denote  $\mathbf{x}'\boldsymbol{\gamma}$ . The expected squared error of this predictor is

$$\text{MSE} = E_y E_{\mathbf{x}}[y - \mathbf{x}'\boldsymbol{\gamma}]^2.$$

This can be written as

$$\text{MSE} = E_{y,\mathbf{x}}\{y - E[y | \mathbf{x}]\}^2 + E_{y,\mathbf{x}}\{E[y | \mathbf{x}] - \mathbf{x}'\boldsymbol{\gamma}\}^2.$$

We seek the  $\boldsymbol{\gamma}$  that minimizes this expectation. The first term is not a function of  $\boldsymbol{\gamma}$ , so only the second term needs to be minimized. Note that this term is not a function of  $y$ , so the outer expectation is actually superfluous. But, we will need it shortly, so we will carry it for the present. The necessary condition is

$$\begin{aligned} \frac{\partial E_y E_{\mathbf{x}}\{[E(y | \mathbf{x}) - \mathbf{x}'\boldsymbol{\gamma}]^2\}}{\partial \boldsymbol{\gamma}} &= E_y E_{\mathbf{x}} \left\{ \frac{\partial [E(y | \mathbf{x}) - \mathbf{x}'\boldsymbol{\gamma}]^2}{\partial \boldsymbol{\gamma}} \right\} \\ &= -2 E_y E_{\mathbf{x}}\{\mathbf{x}[E(y | \mathbf{x}) - \mathbf{x}'\boldsymbol{\gamma}]\} = \mathbf{0}. \end{aligned}$$

Note that we have interchanged the operations of expectation and differentiation in the middle step, since the range of integration is not a function of  $\boldsymbol{\gamma}$ . Finally, we have

## 46 PART I ♦ The Linear Regression Model

the equivalent condition

$$E_y E_x [\mathbf{x} E(y | \mathbf{x})] = E_y E_x [\mathbf{x} \mathbf{x}'] y.$$

The left-hand side of this result is  $E_x E_y [\mathbf{x} E(y | \mathbf{x})] = \text{Cov}[\mathbf{x}, E(y | \mathbf{x})] + E[\mathbf{x}] E_x [E(y | \mathbf{x})] = \text{Cov}[\mathbf{x}, y] + E[\mathbf{x}] E[y] = E_x E_y [\mathbf{x} y]$ . (We have used Theorem B.2.) Therefore, the necessary condition for finding the minimum MSE predictor is

$$E_x E_y [\mathbf{x} y] = E_x E_y [\mathbf{x} \mathbf{x}'] y. \quad (4-3)$$

This is the same as (4-1), which takes us to the least squares condition once again. Assuming that these expectations exist, they would be estimated by the sums in (4-2), which means that regardless of the form of the conditional mean, least squares is an estimator of the coefficients of the minimum expected mean squared error linear predictor. We have yet to establish the conditions necessary for the if part of the theorem, but this is an opportune time to make it explicit:

**THEOREM 4.1 Minimum Mean Squared Error Predictor**

If the data generating mechanism generating  $(x_i, y_i)_{i=1, \dots, n}$  is such that the law of large numbers applies to the estimators in (4-2) of the matrices in (4-1), then the minimum expected squared error linear predictor of  $y_i$  is estimated by the least squares regression line.

## 4.2.3 MINIMUM VARIANCE LINEAR UNBIASED ESTIMATION

Finally, consider the problem of finding a **linear unbiased estimator**. If we seek the one that has smallest variance, we will be led once again to least squares. This proposition will be proved in Section 4.3.5.

The preceding does not assert that no other competing estimator would ever be preferable to least squares. We have restricted attention to linear estimators. The result immediately above precludes what might be an acceptably biased estimator. And, of course, the assumptions of the model might themselves not be valid. Although A5 and A6 are ultimately of minor consequence, the failure of any of the first four assumptions would make least squares much less attractive than we have suggested here.

**4.3 UNBIASED ESTIMATION**

The least squares estimator is unbiased in every sample. To show this, write

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}. \quad (4-4)$$

Now, take expectations, iterating over  $\mathbf{X}$ ;

$$E[\mathbf{b} | \mathbf{X}] = \boldsymbol{\beta} + E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} | \mathbf{X}].$$

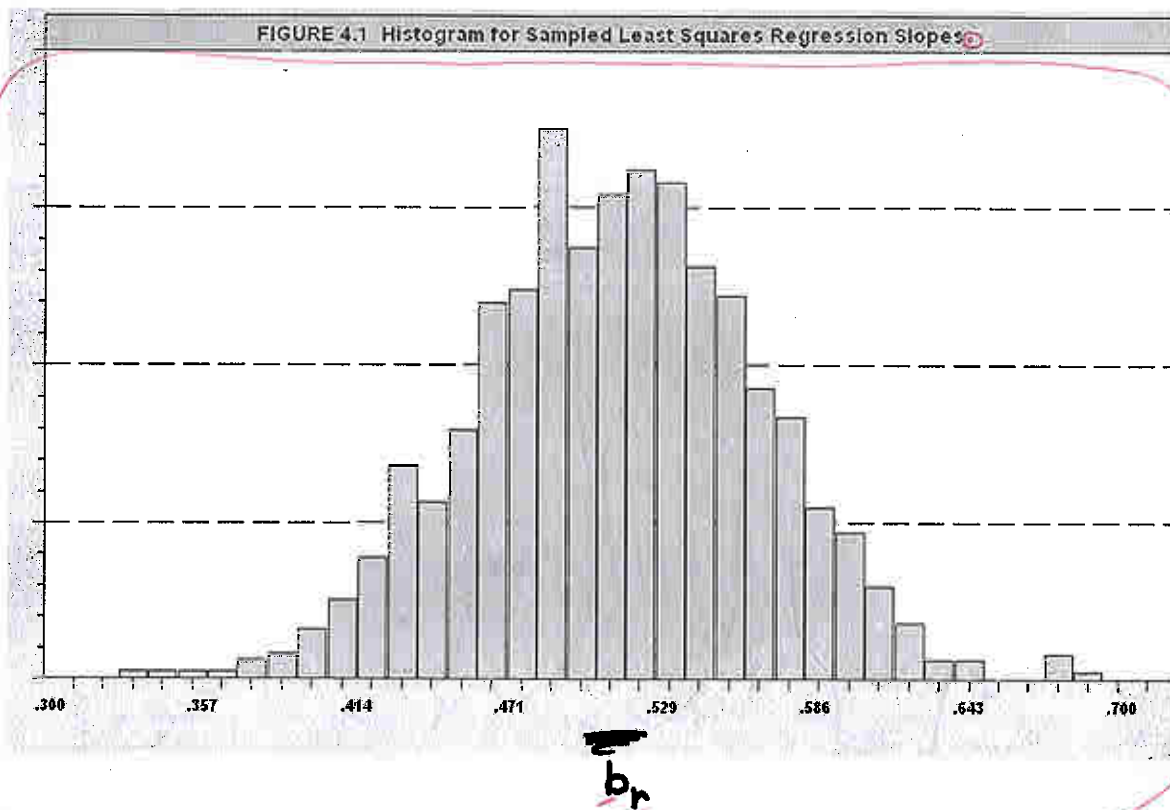


### 4.3 FINITE SAMPLE PROPERTIES OF LEAST SQUARES

An “estimator” is a strategy, or formula, for using the sample data that are drawn from a population. The “properties” of that estimator are a description of how that estimator can be expected to behave when it is applied to a sample of data. To consider an example, the concept of unbiasedness implies that “on average” an estimator (strategy) will correctly estimate the parameter in question; it will not be systematically too high or too low. It seems less than obvious how one could know this if they were only going to draw a single sample of data from the population and analyze that one sample. The argument adopted in classical econometrics is provided by the sampling properties of the estimation strategy. A conceptual experiment lies behind the description. One imagines “repeated sampling” from the population and characterizes the behavior of the “sample of samples.” The underlying statistical theory of the estimator provides the basis of the description. Example 4.1 illustrates.

#### Example 4.1 The Sampling Distribution of a Least Squares Estimator

The following sampling experiment shows the nature of a sampling distribution and the implication of unbiasedness. We drew two samples of 10,000 random draws on variables  $w_i$  and  $x_i$  from the standard normal population (mean zero, variance 1). We generated a set of  $\varepsilon_i$ s equal to  $0.5w_i$  and then  $y_i = 0.5 + 0.5x_i + \varepsilon_i$ . We take this to be our population. We then drew 1000 random samples of 100 observations on  $(y_i, x_i)$  from this population, and with each one, computed the least squares slope, using at replication  $r$ ,  $b_r = \left[ \sum_{j=1}^{100} (x_{jr} - \bar{x}_r) y_{jr} \right] / \left[ \sum_{j=1}^{100} (x_{jr} - \bar{x}_r)^2 \right]$ . The histogram in Figure 4.1 shows the result of the experiment. Note that the distribution of slopes has a mean roughly equal to the “true value” of 0.5, and it has a substantial variance, reflecting the fact that the regression slope, like any other statistic computed from the sample, is a random variable. The concept of unbiasedness relates to the central tendency of this distribution of values obtained in repeated sampling from the population. The shape of the histogram also suggests the normal distribution of the estimator that we will show theoretically in Section 4.3.8 (The experiment should be replicable with any regression program that provides a random number generator and a means of drawing a random sample of observations from a master data set.)



## 46 PART I ♦ The Linear Regression Model

the equivalent condition

$$E_y E_x [\mathbf{x} E(y | \mathbf{x})] = E_y E_x [\mathbf{x} \mathbf{x}'] y.$$

The left-hand side of this result is  $E_x E_y [\mathbf{x} E(y | \mathbf{x})] = \text{Cov}[\mathbf{x}, E(y | \mathbf{x})] + E[\mathbf{x}] E_x [E(y | \mathbf{x})] = \text{Cov}[\mathbf{x}, y] + E[\mathbf{x}] E[y] = E_x E_y [\mathbf{x} y]$ . (We have used Theorem B.2.) Therefore, the necessary condition for finding the minimum MSE predictor is

$$E_x E_y [\mathbf{x} y] = E_x E_y [\mathbf{x} \mathbf{x}'] y. \quad (4-3)$$

This is the same as (4-1), which takes us to the least squares condition once again. Assuming that these expectations exist, they would be estimated by the sums in (4-2), which means that regardless of the form of the conditional mean, least squares is an estimator of the coefficients of the minimum expected mean squared error linear predictor. We have yet to establish the conditions necessary for the if part of the theorem, but this is an opportune time to make it explicit:

**THEOREM 4.1 Minimum Mean Squared Error Predictor**

*If the data generating mechanism generating  $(x_i, y_i)_{i=1, \dots, n}$  is such that the law of large numbers applies to the estimators in (4-2) of the matrices in (4-1), then the minimum expected squared error linear predictor of  $y_i$  is estimated by the least squares regression line.*

**4.2.3 MINIMUM VARIANCE LINEAR UNBIASED ESTIMATION**

Finally, consider the problem of finding a **linear unbiased estimator**. If we seek the one that has smallest variance, we will be led once again to least squares. This proposition will be proved in Section 4.4.

The preceding does not assert that no other competing estimator would ever be preferable to least squares. We have restricted attention to linear estimators. The result immediately above precludes what might be an acceptably biased estimator. And, of course, the assumptions of the model might themselves not be valid. Although A5 and A6 are ultimately of minor consequence, the failure of any of the first four assumptions would make least squares much less attractive than we have suggested here.

## 4.3.1

**UNBIASED ESTIMATION**

The least squares estimator is unbiased in every sample. To show this, write

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\boldsymbol{\varepsilon}. \quad (4-4)$$

Now, take expectations, iterating over  $\mathbf{X}$ ;

$$E[\mathbf{b} | \mathbf{X}] = \boldsymbol{\beta} + E[(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\boldsymbol{\varepsilon} | \mathbf{X}].$$

## CHAPTER 4 ♦ Statistical Properties of the Least Squares Estimator 47

By Assumption A3, the second term is 0, so

$$E[b | \mathbf{X}] = \beta.$$

(4-5)

Therefore,

$$E[b] = E_X\{E[b | \mathbf{X}]\} = E_X[\beta] = \beta.$$

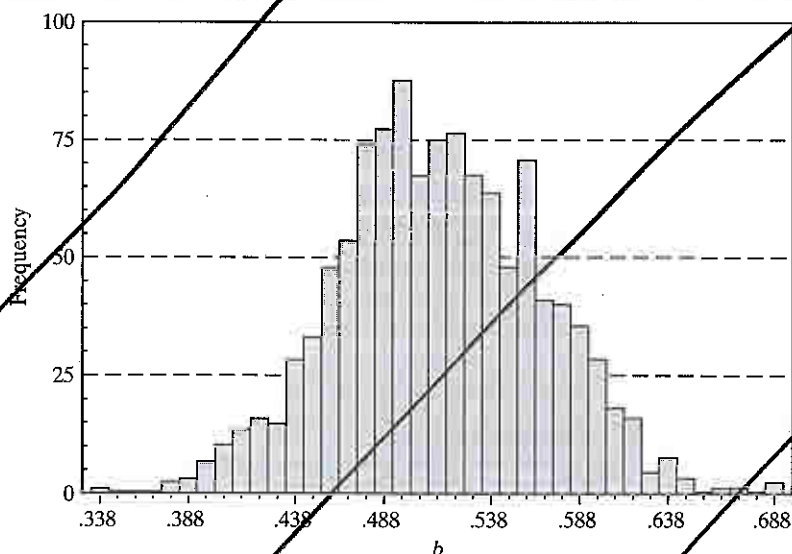
(4-6)

The interpretation of this result is that for any particular set of observations,  $\mathbf{X}$ , the least squares estimator has expectation  $\beta$ . Therefore, when we average this over the possible values of  $\mathbf{X}$  we find the unconditional mean is  $\beta$  as well.

**Example 4.1 The Sampling Distribution of a Least Squares Estimator**

The following sampling experiment, which can be replicated with any regression program that provides a random number generator and a means of drawing a random sample of observations from a master data set, shows the nature of a sampling distribution and the implication of unbiasedness. We drew two samples of 10,000 random draws on  $w_i$  and  $x_i$  from the standard normal distribution (mean zero, variance 1). We then generated a set of  $\varepsilon_i$ s equal to  $0.5w_i$  and  $y_i = 0.5 + 0.5x_i + \varepsilon_i$ . We take this to be our population. We then drew 500 random samples of 100 observations from this population, and with each one, computed the least squares slope (using at replication  $r$ ,  $b_r = [\sum_{j=1}^{100}(x_{jr} - \bar{x}_r)y_{jr}] / [\sum_{j=1}^{100}(x_{jr} - \bar{x}_r)^2]$ ). The histogram in Figure 4.1 shows the result of the experiment. Note that the distribution of slopes has a mean roughly equal to the "true value" of 0.5, and it has a substantial variance, reflecting the fact that the regression slope, like any other statistic computed from the sample, is a random variable. The concept of unbiasedness relates to the central tendency of this distribution of values obtained in repeated sampling from the population.

**FIGURE 4.1** Histogram for Sampled Least Squares Regression Slopes.



You might have noticed that in this section, we have done the analysis conditioning on  $\mathbf{X}$  – that is, conditioning on the entire sample, while in Section 4.2 we have conditioned  $y_i$  on  $x_i$ . (The sharp-eyed reader will have also noticed that in Table 4.1, in assumption A3, we have conditioned  $E[\varepsilon_i | \cdot]$  on  $x_i$ , that is, on all  $i$  and  $j$ , which is, once again, on  $\mathbf{X}$ , not just  $x_i$ . In Section 4.2, we have suggested a way to view the least squares estimator in the context of the joint distribution of a random variable,  $y$ , and a random vector,  $\mathbf{x}$ . For purpose of the discussion, this would be most appropriate if our data were going to be a cross section of independent observations. In this context, as shown in Section 4.2.2, the least squares estimator emerges as the sample counterpart to the slope vector of the minimum mean squared error predictor,  $\gamma$ , which is a feature of the population. In Section 4.3, we make a transition to an understanding of the process that is generating our observed sample of data. The statement that  $E[\mathbf{b} | \mathbf{X}] = \boldsymbol{\beta}$  is best understood from a Bayesian perspective; for the data that we have observed, we can expect certain behavior of the statistics that we compute, such as the least squares slope vector,  $\mathbf{b}$ . Much of the rest of this chapter, indeed much of the rest of this book, will examine the behavior of statistics as we consider whether what we learn from them in a particular sample can reasonably be extended to other samples if they were drawn under similar circumstances from the same population, or whether what we learn from a sample can be inferred to the full population. Thus, it is useful to think of the conditioning operation in  $E[\mathbf{b} | \mathbf{X}]$  in both of these ways at the same time, from the purely statistical viewpoint of deducing the properties of an estimator and from the methodological perspective of deciding how much can be learned about a broader population from a particular finite sample of data.



### 4.3.2 BIAS CAUSED BY OMISSION OF RELEVANT VARIABLES

The analysis has been based on the assumption that the correct specification of the regression model is known to be

$$y = X\beta + \varepsilon.$$

(4-7).

There are numerous types of **specification errors** that one might make in constructing the regression model. The most common ones are the **omission of relevant variables** and the **inclusion of superfluous (irrelevant) variables**.

Suppose that a correctly specified regression model would be

$$y = X_1\beta_1 + X_2\beta_2 + \varepsilon,$$

(4-8)

where the two parts of  $X$  have  $K_1$  and  $K_2$  columns, respectively. If we regress  $y$  on  $X_1$  without including  $X_2$ , then the estimator is

$$b_1 = (X_1'X_1)^{-1}X_1'y = \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 + (X_1'X_1)^{-1}X_1'\varepsilon.$$

(4-9)

Av: None of these three KT's is in the chapter list. Add them to the list or mark them for lightface here.

4-10

Aug. 11's  
term is  
not in  
the  
chapter  
list.

## 134 PART I ♦ The Linear Regression Model

Taking the expectation, we see that unless  $X_1'X_2 = 0$  or  $\beta_2 = 0$ ,  $b_1$  is biased. The well-known result is the **omitted variable formula**:

$$E[b_1 | X] = \beta_1 + P_{1.2}\beta_2,$$

where

$$P_{1.2} = (X_1'X_1)^{-1}X_1'X_2.$$

Each column of the  $K_1 \times K_2$  matrix  $P_{1.2}$  is the column of slopes in the least squares regression of the corresponding column of  $X_2$  on the columns of  $X_1$ .

**Example 7.1 Omitted Variables**

If a demand equation is estimated without the relevant income variable, then (7-4) shows how the estimated price elasticity will be biased. Letting  $b$  be the estimator, we obtain

$$E[b | \text{price, income}] = \beta + \frac{\text{Cov}[\text{price, income}]}{\text{Var}[\text{price}]} \gamma,$$

where  $\gamma$  is the income coefficient. In aggregate data, it is unclear whether the missing covariance would be positive or negative. The sign of the bias in  $b$  would be the same as this covariance, however, because  $\text{Var}[\text{price}]$  and  $\gamma$  would be positive for a normal good such as gasoline. (See Example 2.3.)

The gasoline market data we have examined in Examples 2.3 and 6.7 provide a striking example. Figure 6.5 showed a simple plot of per capita gasoline consumption,  $G/\text{Pop}$  against the price index  $P_G$ . The plot is considerably at odds with what one might expect. But a look at the data in Appendix Table F2.2 shows clearly what is at work. Holding per capita income,  $\text{Income}/\text{Pop}$ , and other prices constant, these data might well conform to expectations. In these data, however, income is persistently growing, and the simple correlations between  $G/\text{Pop}$  and  $\text{Income}/\text{Pop}$  and between  $P_G$  and  $\text{Income}/\text{Pop}$  are 0.938 and 0.934, respectively, which are quite large. To see if the expected relationship between price and consumption shows up, we will have to purge our data of the intervening effect of  $\text{Income}/\text{Pop}$ . To do so, we rely on the Frisch-Waugh result in Theorem 3.3. The regression results appear in Table 6.7. The first column shows the full regression model, with  $\ln P_G$ ,  $\ln \text{Income}$ , and several other variables. The estimated demand elasticity is  $-0.0539$ , which conforms with expectations. If income is omitted from this equation, the estimated price elasticity is  $+0.06788$  which has the wrong sign, but is what we would expect given the theoretical results above.

In this development, it is straightforward to deduce the directions of bias when there is a single included variable and one omitted variable. It is important to note, however, that if more than one variable is included, then the terms in the omitted variable formula involve multiple regression coefficients, which themselves have the signs of partial, not simple, correlations. For example, in the demand equation of the previous example, if the price of a closely related product had been included as well, then the simple correlation between price and income would be insufficient to determine the direction of the bias in the price elasticity. What would be required is the sign of the correlation between price and income net of the effect of the other price. This requirement might not be obvious, and it would become even less so as more regressors were added to the equation.

**7.2.2 PRETEST ESTIMATION**

The variance of  $b_1$  is that of the third term in (7-3), which is

$$\text{Var}[b_1 | X] = \sigma^2 (X_1'X_1)^{-1}. \quad (7-6)$$

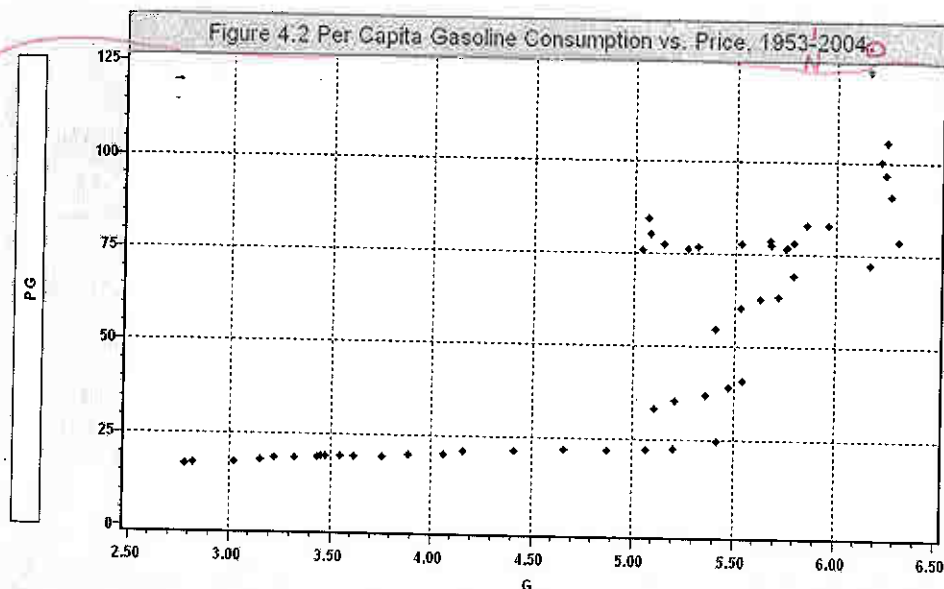
### Example 4.2 Omitted Variable

If a demand equation is estimated without the relevant income variable, then (4-10) shows how the estimated price elasticity will be biased. The gasoline market data we have examined in Example 2.3 provides a striking example. Letting  $b$  be the estimator, we obtain


$$E[b|price, income] = \beta + \frac{Cov[price, income]}{Var[price]} \gamma$$

where  $\gamma$  is the income coefficient. In aggregate data, it is unclear whether the missing covariance would be positive or negative. The sign of the bias in  $b$  would be the same as this covariance, however, because  $Var[price]$  and  $\gamma$  would be positive for a normal good such as gasoline. Figure 4.2 shows a simple plot of per capita gasoline consumption,  $G/Pop$ , against the price index  $PG$ . The plot is considerably at odds with what one might expect. But a look at the data in Appendix Table F2.2 shows clearly what is at work. Holding per capita income,  $Income/Pop$ , and other prices constant, these data might well conform to expectations. In these data, however, income is persistently growing, and the simple correlations between  $G/Pop$  and  $Income/Pop$  and between  $PG$  and  $Income/Pop$  are 0.938 and 0.934, respectively, which are quite large. To see if the expected relationship between price and consumption shows up, we will have to purge our data of the intervening effect of  $Income/Pop$ . To do so, we rely on the Frisch-Waugh result in Theorem 3.2. In the simple regression of log of per capita gasoline consumption on a constant and the log of the price index, the coefficient is 0.29904, which, as expected, has the "wrong" sign. In the multiple regression of the log of per capita gasoline consumption on a constant, the log of the price index and the log of per capita income, the estimated price elasticity,  $\beta$ , is -0.16949 and the estimated income elasticity,  $\gamma$ , is 0.96595. This conforms to expectations. The results are also broadly consistent with the widely observed result that in the U.S. market at least in this period (1953-2004), the main driver of changes in gasoline consumption was not changes in price, but the growth in income (output).

Please clarify your insert-2



In this development, it is straightforward to deduce the directions of bias when there is a single included variable and one omitted variable. It is important to note, however, that if more than one variable is included, then the terms in the omitted variable formula involve multiple regression coefficients, which themselves have the signs of partial, not simple, correlations. For example, in the demand equation of the previous example, if the price of a closely related product had been included as well, then the simple correlation between price and income would be insufficient to determine the direction of the bias in the price elasticity. What would be required is the sign of the correlation between price and income net of the effect of the other price. This requirement might not be obvious, and it would become even less so as more regressors were added to the equation.



## 136 PART I The Linear Regression Model

## 4.3.3

## INCLUSION OF IRRELEVANT VARIABLES

If the regression model is correctly given by

$$y = X_1\beta_1 + e$$

(4-12)

and we estimate it as if (7-2) were correct (i.e., we include some extra variables), then it might seem that the same sorts of problems considered earlier would arise. In fact, this case is not true. We can view the omission of a set of relevant variables as equivalent to imposing an incorrect restriction on (7-2). In particular, omitting  $X_2$  is equivalent to incorrectly estimating (7-2) subject to the restriction  $\beta_2 = 0$ . ~~As we discovered,~~ Incorrectly imposing a restriction produces a biased estimator. Another way to view this error is to note that it amounts to incorporating incorrect information in our estimation. Suppose, however, that our error is simply a failure to use some information that is correct.

The inclusion of the irrelevant variables  $X_2$  in the regression is equivalent to failing to impose  $\beta_2 = 0$  on (7-2) in estimation. But (7-2) is not incorrect; it simply fails to incorporate  $\beta_2 = 0$ . Therefore, we do not need to prove formally that the least squares estimator of  $\beta$  in (7-2) is unbiased even given the restriction; we have already proved it. We can assert on the basis of all our earlier results that

$$E[b|X] = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} \beta_1 \\ 0 \end{bmatrix}$$

(4-13)

By the same reasoning,  $s^2$  is also unbiased:

$$E\left[\frac{e'e}{n - K_1 - K_2} \middle| X\right] = \sigma^2 \quad (7-11)$$

Then where is the problem? It would seem that one would generally want to "overfit" the model. From a theoretical standpoint, the difficulty with this view is that the failure to use correct information is always costly. In this instance, the cost is ~~the~~ reduced precision of the estimates. As we have shown, the covariance matrix in the short regression (omitting  $X_2$ ) is never larger than the covariance matrix for the estimator obtained in the presence of the superfluous variables.<sup>2</sup> Consider again the single-variable comparison given earlier. If  $x_2$  is highly correlated with  $x_1$ , then incorrectly including  $x_2$  in the regression will greatly inflate the variance of the estimator. of  $\beta_1$

will be

a

## 7.2.4 MODEL BUILDING—A GENERAL TO SIMPLE STRATEGY

There has been a shift in the general approach to model building in the past 20 years or so, partly based on the results in the previous two sections. With an eye toward maintaining simplicity, model builders would generally begin with a small specification and gradually build up the model ultimately of interest by adding variables. But, based on the preceding results, we can surmise that just about any criterion that would be used to decide whether to add a variable to a current specification would be tainted by the biases caused by the incomplete specification at the early steps. Omitting variables from the equation seems generally to be the worse of the two errors. Thus, the simple-to-general approach to model building has little to recommend it. Building on the work

<sup>2</sup> There is no loss if  $X_1'X_2 = 0$ , which makes sense in terms of the information about  $X_1$  contained in  $X_2$  (here, none). This situation is not likely to occur in practice, however.

change 7-2 to  
4-8 →  
all 5 times →

will show in  
Section 4.7.1,

FN  
2



4-14

## 48 PART I ♦ The Linear Regression Model

4.3.4 ~~THE VARIANCE OF THE LEAST SQUARES ESTIMATOR AND THE GAUSS-MARKOV THEOREM~~

If the regressors can be treated as nonstochastic, as they would be in an experimental situation in which the analyst chooses the values in  $\mathbf{X}$ , then the sampling variance of the least squares estimator can be derived by treating  $\mathbf{X}$  as a matrix of constants. Alternatively, we can allow  $\mathbf{X}$  to be stochastic, do the analysis conditionally on the observed  $\mathbf{X}$ , then consider averaging over  $\mathbf{X}$  as we did in the preceding section. Using (4-4) again, we have

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{e}) = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}.$$

Since we can write  $\mathbf{b} = \boldsymbol{\beta} + \mathbf{A}\mathbf{e}$ , where  $\mathbf{A}$  is  $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ ,  $\mathbf{b}$  is a linear function of the disturbances, which by the definition we will use makes it a linear estimator. As we have seen, the expected value of the second term in (4-5) is 0. Therefore, regardless of the distribution of  $\mathbf{e}$ , under our other assumptions,  $\mathbf{b}$  is a linear, unbiased estimator of  $\boldsymbol{\beta}$ .

The covariance matrix of the least squares slope estimator is

$$\begin{aligned}\text{Var}[\mathbf{b} | \mathbf{X}] &= E[(\mathbf{b} - \boldsymbol{\beta})(\mathbf{b} - \boldsymbol{\beta})' | \mathbf{X}] \\ &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}\mathbf{e}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} | \mathbf{X}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\mathbf{e}\mathbf{e}' | \mathbf{X}]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2\mathbf{I})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.\end{aligned}$$

**Example 4.2 Sampling Variance in the Two-Variable Regression Model**  
Suppose that  $\mathbf{X}$  contains only a constant term (column of 1s) and a single regressor  $x$ . The lower right element of  $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$  is

$$\text{Var}[b | x] = \text{Var}[b - \beta | x] = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Note, in particular, the denominator of the variance of  $b$ . The greater the variation in  $x$ , the smaller this variance. For example, consider the problem of estimating the slopes of the two regressions in Figure 4.8. A more precise result will be obtained for the data in the right-hand panel of the figure.

## 4.3.5 The GAUSS-MARKOV THEOREM

We will now obtain a general result for the class of linear unbiased estimators of  $\boldsymbol{\beta}$ .

Let  $\mathbf{b}_0 = \mathbf{C}\mathbf{y}$  be another linear unbiased estimator of  $\boldsymbol{\beta}$ , where  $\mathbf{C}$  is a  $K \times n$  matrix. If  $\mathbf{b}_0$  is unbiased, then

$$E[\mathbf{C}\mathbf{y} | \mathbf{X}] = E[(\mathbf{C}\mathbf{X}\boldsymbol{\beta} + \mathbf{C}\mathbf{e}) | \mathbf{X}] = \boldsymbol{\beta},$$

which implies that  $\mathbf{C}\mathbf{X} = \mathbf{I}$ . There are many candidates. For example, consider using just the first  $K$  (or, any  $K$ ) linearly independent rows of  $\mathbf{X}$ . Then  $\mathbf{C} = [\mathbf{X}_0^{-1} : \mathbf{0}]$ , where  $\mathbf{X}_0^{-1}$  is the inverse of the matrix formed from the  $K$  rows of  $\mathbf{X}$ . The covariance matrix of  $\mathbf{b}_0$  can be found by replacing  $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$  with  $\mathbf{C}$  in (4-5); the result is  $\text{Var}[\mathbf{b}_0 | \mathbf{X}] = \sigma^2\mathbf{C}\mathbf{C}'$ . Now let  $\mathbf{D} = \mathbf{C} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$  so  $\mathbf{D}\mathbf{y} = \mathbf{b}_0 - \mathbf{b}$ . Then,

$$\text{Var}[\mathbf{b}_0 | \mathbf{X}] = \sigma^2[\mathbf{D} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'](\mathbf{D} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')'.$$

Insert next page (C)

~~We will now obtain a general result for the class of linear unbiased estimators of  $\beta$ .~~

### **THEOREM 4.2 Gauss-Markov Theorem**

*In the ~~classical~~ linear regression model with regressor matrix  $X$ , the least squares estimator  $b$  is the minimum variance linear unbiased estimator of  $\beta$ . For any vector of constants  $w$ , the minimum variance linear unbiased estimator of  $w'\beta$  in the ~~classical~~ regression model is  $w'b$ , where  $b$  is the least squares estimator.*

Note that the theorem makes no use of Assumption A6, normality of the distribution of the disturbances. Only A1 to A4 are necessary. A direct approach to proving this important theorem would be to define the class of linear and unbiased estimators ( $b_L = Cy$  such that  $E[b_L | X] = \beta$ ) and then find the member of that class that has the smallest variance. We will use an indirect method instead. We have already established that  $b$  is a linear unbiased estimator. We will now consider other linear unbiased estimators of  $\beta$  and show that any other such estimator has a larger variance.

## 48 PART I ♦ The Linear Regression Model

## 4.4 THE VARIANCE OF THE LEAST SQUARES ESTIMATOR AND THE GAUSS-MARKOV THEOREM

If the regressors can be treated as nonstochastic, as they would be in an experimental situation in which the analyst chooses the values in  $\mathbf{X}$ , then the **sampling variance** of the least squares estimator can be derived by treating  $\mathbf{X}$  as a matrix of constants. Alternatively, we can allow  $\mathbf{X}$  to be stochastic, do the analysis conditionally on the observed  $\mathbf{X}$ , then consider averaging over  $\mathbf{X}$  as we did in the preceding section. Using (4-4) again, we have

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{e}) = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}. \quad (4-5)$$

Since we can write  $\mathbf{b} = \boldsymbol{\beta} + \mathbf{A}\mathbf{e}$ , where  $\mathbf{A}$  is  $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ ,  $\mathbf{b}$  is a linear function of the disturbances, which by the definition we will use makes it a **linear estimator**. As we have seen, the expected value of the second term in (4-5) is 0. Therefore, *regardless of the distribution of  $\mathbf{e}$ , under our other assumptions,  $\mathbf{b}$  is a linear, unbiased estimator of  $\boldsymbol{\beta}$ .* The covariance matrix of the least squares slope estimator is

$$\begin{aligned} \text{Var}[\mathbf{b} | \mathbf{X}] &= E[(\mathbf{b} - \boldsymbol{\beta})(\mathbf{b} - \boldsymbol{\beta})' | \mathbf{X}] \\ &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}\mathbf{e}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} | \mathbf{X}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\mathbf{e}\mathbf{e}' | \mathbf{X}]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2\mathbf{I})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}. \end{aligned}$$

**Example 4.2 Sampling Variance in the Two-Variable Regression Model**

Suppose that  $\mathbf{X}$  contains only a constant term (column of 1s) and a single regressor  $x$ . The lower right element of  $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$  is

$$\text{Var}[b | x] = \text{Var}[b - \beta | x] = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Note, in particular, the denominator of the variance of  $b$ . The greater the variation in  $x$ , the smaller this variance. For example, consider the problem of estimating the slopes of the two regressions in Figure 4.2. A more precise result will be obtained for the data in the right-hand panel of the figure.

We will now obtain a general result for the class of linear unbiased estimators of  $\boldsymbol{\beta}$ .

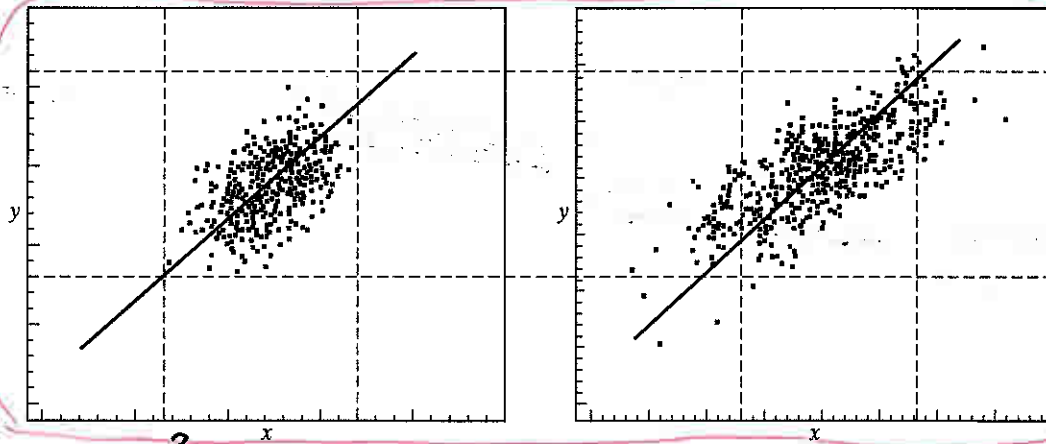
Let  $\mathbf{b}_0 = \mathbf{C}\mathbf{y}$  be another linear unbiased estimator of  $\boldsymbol{\beta}$ , where  $\mathbf{C}$  is a  $K \times n$  matrix. If  $\mathbf{b}_0$  is unbiased, then

$$E[\mathbf{C}\mathbf{y} | \mathbf{X}] = E[(\mathbf{C}\mathbf{X}\boldsymbol{\beta} + \mathbf{C}\mathbf{e}) | \mathbf{X}] = \boldsymbol{\beta},$$

which implies that  $\mathbf{C}\mathbf{X} = \mathbf{I}$ . There are many candidates. For example, consider using just the first  $K$  (or, any  $K$ ) linearly independent rows of  $\mathbf{X}$ . Then  $\mathbf{C} = [\mathbf{X}_0^{-1} : \mathbf{0}]$ , where  $\mathbf{X}_0^{-1}$  is the inverse of the matrix formed from the  $K$  rows of  $\mathbf{X}$ . The covariance matrix of  $\mathbf{b}_0$  can be found by replacing  $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$  with  $\mathbf{C}$  in (4-5); the result is  $\text{Var}[\mathbf{b}_0 | \mathbf{X}] = \sigma^2\mathbf{C}\mathbf{C}'$ . Now let  $\mathbf{D} = \mathbf{C} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$  so  $\mathbf{D}\mathbf{y} = \mathbf{b}_0 - \mathbf{b}$ . Then,  $\mathbf{14}$

$$\text{Var}[\mathbf{b}_0 | \mathbf{X}] = \sigma^2[(\mathbf{D} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')(\mathbf{D} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')'].$$

## CHAPTER 4 ♦ Statistical Properties of the Least Squares Estimator 49



**FIGURE 4.2** Effect of Increased Variation in  $x$  Given the Same Conditional and Overall Variation in  $y$ .

We know that  $CX = I = DX + (X'X)^{-1}(X'X)$ , so  $DX$  must equal  $0$ . Therefore,

$$\text{Var}[b_0 | X] = \sigma^2(X'X)^{-1} + \sigma^2DD' = \text{Var}[b | X] + \sigma^2DD'.$$

Since a quadratic form in  $DD'$  is  $q'DD'q = z'z \geq 0$ , the conditional covariance matrix of  $b_0$  equals that of  $b$  plus a nonnegative definite matrix. Therefore, every quadratic form in  $\text{Var}[b_0 | X]$  is larger than the corresponding quadratic form in  $\text{Var}[b | X]$ , which ~~implies a very important property of the least squares coefficient vector~~

establishes the first result.

#### **THEOREM 4.2 Gauss-Markov Theorem**

*In the classical linear regression model with regressor matrix  $X$ , the least squares estimator  $b$  is the minimum variance linear unbiased estimator of  $\beta$ . For any vector of constants  $w$ , the minimum variance linear unbiased estimator of  $w'\beta$  in the classical regression model is  $w'b$ , where  $b$  is the least squares estimator.*

The proof of the second statement follows from the previous derivation, since the variance of  $w'b$  is a quadratic form in  $\text{Var}[b | X]$ , and likewise for any  $b_0$ , and proves that each individual slope estimator  $b_k$  is the best linear unbiased estimator of  $\beta_k$ . (Let  $w$  be all zeros except for a one in the  $k$ th position.) The theorem is much broader than this, however, since the result also applies to every other linear combination of the elements of  $\beta$ .

4.3.6  
4.8

#### **THE IMPLICATIONS OF STOCHASTIC REGRESSORS**

The preceding analysis is done conditionally on the observed data. A convenient method of obtaining the unconditional statistical properties of  $b$  is to obtain the desired results conditioned on  $X$  first, then find the unconditional result by “averaging” (e.g., by

## 50 PART I ♦ The Linear Regression Model

integrating over) the conditional distributions. The crux of the argument is that if we can establish unbiasedness conditionally on an arbitrary  $\mathbf{X}$ , then we can average over  $\mathbf{X}$ 's to obtain an unconditional result. We have already used this approach to show the unconditional unbiasedness of  $\mathbf{b}$  in Section 4.3, so we now turn to the conditional variance.

The conditional variance of  $\mathbf{b}$  is

$$\text{Var}[\mathbf{b} | \mathbf{X}] = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}.$$

For the exact variance, we use the decomposition of variance of (B-69):

$$\text{Var}[\mathbf{b}] = E_{\mathbf{X}}[\text{Var}[\mathbf{b} | \mathbf{X}]] + \text{Var}_{\mathbf{X}}[E[\mathbf{b} | \mathbf{X}]].$$

The second term is zero since  $E[\mathbf{b} | \mathbf{X}] = \boldsymbol{\beta}$  for all  $\mathbf{X}$ , so

$$\text{Var}[\mathbf{b}] = E_{\mathbf{X}}[\sigma^2 (\mathbf{X}'\mathbf{X})^{-1}] = \sigma^2 E_{\mathbf{X}}[(\mathbf{X}'\mathbf{X})^{-1}].$$

Our earlier conclusion is altered slightly. We must replace  $(\mathbf{X}'\mathbf{X})^{-1}$  with its expected value to get the appropriate covariance matrix, which brings a subtle change in the interpretation of these results. The unconditional variance of  $\mathbf{b}$  can only be described in terms of the average behavior of  $\mathbf{X}$ , so to proceed further, it would be necessary to make some assumptions about the variances and covariances of the regressors. We will return to this subject in Section 4.4.

We showed in Section 4.3 that

$$\text{Var}[\mathbf{b} | \mathbf{X}] \leq \text{Var}[\mathbf{b}_0 | \mathbf{X}]$$

for any  $\mathbf{b}_0 \neq \mathbf{b}$  and for the specific  $\mathbf{X}$  in our sample. But if this inequality holds for every particular  $\mathbf{X}$ , then it must hold for

$$\text{Var}[\mathbf{b}] = E_{\mathbf{X}}[\text{Var}[\mathbf{b} | \mathbf{X}]].$$

That is, if it holds for every particular  $\mathbf{X}$ , then it must hold over the average value(s) of  $\mathbf{X}$ .

The conclusion, therefore, is that the important results we have obtained thus far for the least squares estimator, unbiasedness, and the Gauss-Markov theorem hold whether or not we regard  $\mathbf{X}$  as stochastic. Condition on the particular sample in hand or consider, instead, sampling broadly from the population.

**THEOREM 4.3 Gauss-Markov Theorem (Concluded)**

In the ~~classical~~ linear regression model, the least squares estimator  $\mathbf{b}$  is the minimum variance linear unbiased estimator of  $\boldsymbol{\beta}$  whether  $\mathbf{X}$  is stochastic or nonstochastic, so long as the other assumptions of the model continue to hold.

linear and unbiased



## CHAPTER 4 ♦ Statistical Properties of the Least Squares Estimator 51

4.3.7 ~~4.3~~ ESTIMATING THE VARIANCE OF THE LEAST SQUARES ESTIMATOR

If we wish to test hypotheses about  $\beta$  or to form confidence intervals, then we will require a sample estimate of the covariance matrix  $\text{Var}[\mathbf{b} | \mathbf{X}] = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$ . The population parameter  $\sigma^2$  remains to be estimated. Since  $\sigma^2$  is the expected value of  $\varepsilon_i^2$  and  $e_i$  is an estimate of  $\varepsilon_i$ , by analogy,

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n e_i^2$$

would seem to be a natural estimator. But the least squares residuals are imperfect estimates of their population counterparts;  $e_i = y_i - \mathbf{x}_i' \mathbf{b} = \varepsilon_i - \mathbf{x}_i' (\mathbf{b} - \beta)$ . The estimator is distorted (as might be expected) because  $\beta$  is not observed directly. The expected square on the right-hand side involves a second term that might not have expected value zero.

The least squares residuals are

$$\mathbf{e} = \mathbf{My} = \mathbf{M}[\mathbf{X}\beta + \boldsymbol{\varepsilon}] = \mathbf{M}\boldsymbol{\varepsilon},$$

as  $\mathbf{MX} = \mathbf{0}$ . [See (3-15).] An estimator of  $\sigma^2$  will be based on the sum of squared residuals:

$$\mathbf{e}'\mathbf{e} = \boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}.$$

The expected value of this quadratic form is

$$E[\mathbf{e}'\mathbf{e} | \mathbf{X}] = E[\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon} | \mathbf{X}].$$

The scalar  $\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}$  is a  $1 \times 1$  matrix, so it is equal to its trace. By using the result on cyclic permutations (A-94),

$$E[\text{tr}(\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}) | \mathbf{X}] = E[\text{tr}(\mathbf{M}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') | \mathbf{X}].$$

Since  $\mathbf{M}$  is a function of  $\mathbf{X}$ , the result is

$$\text{tr}(\mathbf{M}E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' | \mathbf{X}]) = \text{tr}(\mathbf{M}\sigma^2\mathbf{I}) = \sigma^2 \text{tr}(\mathbf{M}).$$

The trace of  $\mathbf{M}$  is

$$\text{tr}[\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] = \text{tr}(\mathbf{I}_n) - \text{tr}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}] = \text{tr}(\mathbf{I}_n) - \text{tr}(\mathbf{I}_K) = n - K.$$

Therefore,

$$E[\mathbf{e}'\mathbf{e} | \mathbf{X}] = (n - K)\sigma^2,$$

so the natural estimator is biased toward zero, although the bias becomes smaller as the sample size increases. An unbiased estimator of  $\sigma^2$  is

$$s^2 = \frac{\mathbf{e}'\mathbf{e}}{n - K}.$$

The estimator is unbiased unconditionally as well, since  $E[s^2] = E_{\mathbf{X}}\{E[s^2 | \mathbf{X}]\} = E_{\mathbf{X}}[\sigma^2] = \sigma^2$ . The **standard error of the regression** is  $s$ , the square root of  $s^2$ . With  $s^2$ , we can then compute

$$\text{Est. Var}[\mathbf{b} | \mathbf{X}] = s^2 (\mathbf{X}'\mathbf{X})^{-1}.$$

## 52 PART I ♦ The Linear Regression Model

Henceforth, we shall use the notation  $\text{Est. Var}[\cdot]$  to indicate a sample estimate of the sampling variance of an estimator. The square root of the  $k$ th diagonal element of this matrix,  $\{[s^2(\mathbf{X}'\mathbf{X})^{-1}]_{kk}\}^{1/2}$ , is the **standard error** of the estimator  $b_k$ , which is often denoted simply "the standard error of  $b_k$ ."

4.3.8 ~~THE NORMALITY ASSUMPTION AND BASIC STATISTICAL INFERENCE~~12.3 ~~12.3~~Confidence  
Intervals.

To this point, our specification and analysis of the regression model is **semiparametric** (see Section 14.3). We have not used Assumption A6 (see Table 4.1), normality of  $\mathbf{e}$ , in any of our results. The assumption is useful for constructing statistics for **testing** hypotheses. In (4.5),  $\mathbf{b}$  is a linear function of the disturbance vector  $\mathbf{e}$ . If we assume that  $\mathbf{e}$  has a multivariate normal distribution, then we may use the results of Section B.10.2 and the mean vector and covariance matrix derived earlier to state that

forming

4-4

$$\mathbf{b}|\mathbf{X} \sim N[\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}].$$

4-18

(4-8)

This specifies a multivariate normal distribution, so each element of  $\mathbf{b}|\mathbf{X}$  is normally distributed:

4-19

(4-9)

$$b_k|\mathbf{X} \sim N[\beta_k, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{kk}].$$

We found evidence of this result in Figure 4.1 example 4.1.

new  
paragraph

4.4

The distribution of  $\mathbf{b}$  is conditioned on  $\mathbf{X}$ . The normal distribution of  $\mathbf{b}$  in a finite sample is a consequence of our specific assumption of normally distributed disturbances. Without this assumption, and without some alternative specific assumption about the distribution of  $\mathbf{e}$ , we will not be able to make any definite statement about the exact distribution of  $\mathbf{b}$ , conditional or otherwise. In an interesting result that we will explore at length in Section 4.8, we will be able to obtain an approximate normal distribution for  $\mathbf{b}$ , with or without assuming normally distributed disturbances and whether the regressors are stochastic or not.

4.7.1 ~~FORMING A CONFIDENCE INTERVAL FOR A TESTING A HYPOTHESIS ABOUT A COEFFICIENT~~

Let  $S^{kk}$  be the  $k$ th diagonal element of  $(\mathbf{X}'\mathbf{X})^{-1}$ . Then, assuming normality and conditioned on  $\mathbf{X}$ ,

$$z_k = \frac{b_k - \beta_k}{\sqrt{\sigma^2 S^{kk}}} \quad (4-10)$$

has a standard normal distribution. If  $\sigma^2$  were known, then statistical inference about  $\beta_k$  could be based on  $z_k$ . By using  $s^2$  instead of  $\sigma^2$ , we can derive a statistic to use in place of  $z_k$  in (4-10). The quantity

$$\frac{(n-K)s^2}{\sigma^2} = \frac{\mathbf{e}'\mathbf{e}}{\sigma^2} = \left(\frac{\mathbf{e}}{\sigma}\right)' \mathbf{M} \left(\frac{\mathbf{e}}{\sigma}\right) \quad (4-11)$$

is an idempotent quadratic form in a standard normal vector  $(\mathbf{e}/\sigma)$ . Therefore, it has a chi-squared distribution with rank  $(\mathbf{M}) = \text{trace}(\mathbf{M}) = n - K$  degrees of freedom.<sup>2</sup> The

<sup>2</sup>This result is proved in Section B.11.4.

## 4.4 LARGE SAMPLE PROPERTIES OF THE LEAST SQUARES ESTIMATOR

Using only assumptions A1 through A4 of the classical model listed in Table 4.1, we have established the following exact, finite sample properties for the least squares estimators  $\mathbf{b}$  and  $s^2$  of the unknown parameters  $\beta$  and  $\sigma^2$ :

- $E[\mathbf{b}|\mathbf{X}] = E[\mathbf{b}] = \beta$  — the least squares coefficient estimator is unbiased;
- $E[s^2|\mathbf{X}] = E[s^2] = \sigma^2$  — the disturbance variance estimator is unbiased;
- $\text{Var}[\mathbf{b}|\mathbf{X}] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$  and  $\text{Var}[\mathbf{b}] = \sigma^2 E[(\mathbf{X}'\mathbf{X})^{-1}]$ ;
- Gauss – Markov theorem: The MVLUE of  $\mathbf{w}'\beta$  is  $\mathbf{w}'\mathbf{b}$  for any vector of constants,  $\mathbf{w}$ .

For this basic model, it is <sup>also</sup> straightforward to derive the large-sample, or asymptotic properties of the least squares estimator. The normality assumption, A6, becomes inessential at this point, and will be discarded save for discussions of maximum likelihood estimation in Section 4.4.6 and in Chapter 13.

### 4.4.1 CONSISTENCY OF THE LEAST SQUARES ESTIMATOR OF $\beta$

Unbiasedness is a useful starting point for assessing the virtues of an estimator. It assures the analyst that their estimator will not persistently miss its target, either systematically too high or too low. However, as a guide to estimation strategy, it has two shortcomings. First, save for the least squares slope estimator we are discussing in this chapter, it is relatively rare for an econometric estimator to be unbiased. In nearly all cases beyond the multiple regression model, the best one can hope for is that the estimator improves in the sense suggested by unbiasedness as more information (data) is brought to bear on the study. As such, we will need a broader set of tools to guide the econometric inquiry. Second, the property of unbiasedness does not, in fact, imply that more information is better than less, in terms of estimation of parameters. The sample means of random samples of 2, 100, and 10,000 are all unbiased estimators of a population mean  $\mu$ ; by this criterion all are equally desirable. Logically, one would hope that a larger sample is better than a smaller one in some sense that we are about to define (and, by extension, an extremely large sample should be much better, or even perfect). The property of consistency improves on unbiasedness in both of these directions.

A0: "asymptotic properties" was a hold KT on msp 4-1. Here also?

## 64 PART I ♦ The Linear Regression Model

it is straightforward to derive the large-sample, or **asymptotic properties** of the least squares estimator. The normality assumption, A6, becomes inessential at this point, and will be discarded save for discussions of maximum likelihood estimation in Chapter 16. This section will also consider various forms of Assumption A6, the data generating mechanism.

4.9.1 CONSISTENCY OF THE LEAST SQUARES ESTIMATOR OF  $\beta$ 

To begin, we leave the data generating mechanism for  $\mathbf{X}$  unspecified— $\mathbf{X}$  may be any mixture of constants and random variables generated independently of the process that generates  $\varepsilon$ . We do make two crucial assumptions. The first is a modification of Assumption A5 in Table 4.1;

**A5a.**  $(\mathbf{x}_i, \varepsilon_i)$   $i = 1, \dots, n$  is a sequence of independent observations.

The second concerns the behavior of the data in large samples;

$$\text{plim}_{n \rightarrow \infty} \frac{\mathbf{X}'\mathbf{X}}{n} = \mathbf{Q}, \quad \text{a positive definite matrix.} \quad (4-21)$$

[We will return to (4-21) shortly.] The least squares estimator may be written

$$\mathbf{b} = \beta + \left( \frac{\mathbf{X}'\mathbf{X}}{n} \right)^{-1} \left( \frac{\mathbf{X}'\varepsilon}{n} \right). \quad (4-22)$$

If  $\mathbf{Q}^{-1}$  exists, then

$$\text{plim } \mathbf{b} = \beta + \mathbf{Q}^{-1} \text{plim} \left( \frac{\mathbf{X}'\varepsilon}{n} \right)$$

because the inverse is a continuous function of the original matrix. (We have invoked Theorem D.14.) We require the probability limit of the last term. Let

$$\frac{1}{n} \mathbf{X}'\varepsilon = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \varepsilon_i = \frac{1}{n} \sum_{i=1}^n \mathbf{w}_i = \bar{\mathbf{w}}. \quad (4-23)$$

Then

$$\text{plim } \mathbf{b} = \beta + \mathbf{Q}^{-1} \text{plim } \bar{\mathbf{w}}.$$

From the exogeneity Assumption A3, we have  $E[\mathbf{w}_i] = E_{\mathbf{x}}[E[\mathbf{w}_i | \mathbf{x}_i]] = E_{\mathbf{x}}[\mathbf{x}_i E[\varepsilon_i | \mathbf{x}_i]] = \mathbf{0}$ , so the exact expectation is  $E[\bar{\mathbf{w}}] = \mathbf{0}$ . For any element in  $\mathbf{x}_i$  that is nonstochastic, the zero expectations follow from the marginal distribution of  $\varepsilon_i$ . We now consider the variance. By (B-70),  $\text{Var}[\bar{\mathbf{w}}] = E[\text{Var}[\bar{\mathbf{w}} | \mathbf{X}]] + \text{Var}[E[\bar{\mathbf{w}} | \mathbf{X}]]$ . The second term is zero because  $E[\varepsilon_i | \mathbf{x}_i] = 0$ . To obtain the first, we use  $E[\varepsilon \varepsilon' | \mathbf{X}] = \sigma^2 \mathbf{I}$ , so

$$\text{Var}[\bar{\mathbf{w}} | \mathbf{X}] = E[\bar{\mathbf{w}} \bar{\mathbf{w}}' | \mathbf{X}] = \frac{1}{n} \mathbf{X}' E[\varepsilon \varepsilon' | \mathbf{X}] \mathbf{X} \frac{1}{n} = \left( \frac{\sigma^2}{n} \right) \left( \frac{\mathbf{X}'\mathbf{X}}{n} \right).$$

Therefore,

$$\text{Var}[\bar{\mathbf{w}}] = \left( \frac{\sigma^2}{n} \right) E \left( \frac{\mathbf{X}'\mathbf{X}}{n} \right).$$

## CHAPTER 4 ♦ Statistical Properties of the Least Squares Estimator 65

TABLE 4.6 Grenander Conditions for Well-Behaved Data

- G1.** For each column of  $\mathbf{X}$ ,  $\mathbf{x}_k$ , if  $d_{nk}^2 = \mathbf{x}_k' \mathbf{x}_k$ , then  $\lim_{n \rightarrow \infty} d_{nk}^2 = +\infty$ . Hence,  $\mathbf{x}_k$  does not degenerate to a sequence of zeros. Sums of squares will continue to grow as the sample size increases. No variable will degenerate to a sequence of zeros.
- G2.**  $\lim_{n \rightarrow \infty} x_{ik}^2 / d_{nk}^2 = 0$  for all  $i = 1, \dots, n$ . This condition implies that no single observation will ever dominate  $\mathbf{x}_k \mathbf{x}_k'$ , and as  $n \rightarrow \infty$ , individual observations will become less important.
- G3.** Let  $\mathbf{R}_n$  be the sample correlation matrix of the columns of  $\mathbf{X}$ , excluding the constant term if there is one. Then  $\lim_{n \rightarrow \infty} \mathbf{R}_n = \mathbf{C}$ , a positive definite matrix. This condition implies that the full rank condition will always be met. We have already assumed that  $\mathbf{X}$  has full rank in a finite sample, so this assumption ensures that the condition will never be violated.

The variance will collapse to zero if the expectation in parentheses is (or converges to) a constant matrix, so that the leading scalar will dominate the product as  $n$  increases. Assumption (4.21) should be sufficient. (Theoretically, the expectation could diverge while the probability limit does not, but this case would not be relevant for practical purposes.) It then follows that

$$\lim_{n \rightarrow \infty} \text{Var}[\bar{\mathbf{w}}] = \mathbf{0} \cdot \mathbf{Q} = \mathbf{0}.$$

(4-23)

Since the mean of  $\bar{\mathbf{w}}$  is identically zero and its variance converges to zero,  $\bar{\mathbf{w}}$  converges in mean square to zero, so  $\text{plim } \bar{\mathbf{w}} = \mathbf{0}$ . Therefore,

$$\text{plim } \frac{\mathbf{X}'\mathbf{e}}{n} = \mathbf{0}.$$

(4-24)

so

$$\text{plim } \mathbf{b} = \boldsymbol{\beta} + \mathbf{Q}^{-1} \cdot \mathbf{0} = \boldsymbol{\beta}.$$

(4-25)

This result establishes that under Assumptions A1-A4 and the additional assumption (4.21),  $\mathbf{b}$  is a **consistent estimator** of  $\boldsymbol{\beta}$  in the classical regression model.

Time-series settings that involve time trends, polynomial time series, and trending variables often pose cases in which the preceding assumptions are too restrictive. A somewhat weaker set of assumptions about  $\mathbf{X}$  that is broad enough to include most of these is the **Grenander conditions** listed in Table 4.6. The conditions ensure that the data matrix is "well behaved" in large samples. The assumptions are very weak and likely to be satisfied by almost any data set encountered in practice.

Av: This bold face term is not in chap. list. Add or mark lightface here.

## 4.2 ASYMPTOTIC NORMALITY OF THE LEAST SQUARES ESTIMATOR

To derive the asymptotic distribution of the least squares estimator, we shall use the results of Section D.3. We will make use of some basic central limit theorems, so in addition to Assumption A3 (uncorrelatedness), we will assume that the observations

3 Judge et al. (1985, p. 162).

4 White (2001) continues this line of analysis.

insert next page

TB 4.2

FN 3  
FN 4



#### 4.4.2 Asymptotic Normality of the Least Squares Estimator

4.5 As a guide to estimation, consistency is an improvement over unbiasedness. Since we are in the process of relaxing the more restrictive assumptions of the model, including A.6, normality of the disturbances, we will also lose the normal distribution of the estimator that enables us to form confidence intervals in Section 4.7. It seems that the more general model we have built here has come at a cost. In this section, we will find that normality of the disturbances is not necessary for establishing the distributional results we need to allow statistical inference including confidence intervals and testing hypotheses. Under generally reasonable assumptions about the process that generates the sample data, large sample distributions will provide a reliable foundation for statistical inference in the regression model (and more generally, as we develop more elaborate estimators later in the book).

To derive the asymptotic distribution of the least squares estimator, we shall use the results of Section D.3. We will make use of some basic central limit theorems, so in addition to Assumption A3 (uncorrelatedness), we will assume that observations are independent. It follows from (4-21) that

## 66 PART I ♦ The Linear Regression Model

~~are independent. It follows from (4-22) that~~

$$\sqrt{n}(\mathbf{b} - \boldsymbol{\beta}) = \left( \frac{\mathbf{X}'\mathbf{X}}{n} \right)^{-1} \left( \frac{1}{\sqrt{n}} \right) \mathbf{X}'\boldsymbol{\varepsilon}. \quad (4-26)$$

Since the inverse matrix is a continuous function of the original matrix,  $\text{plim}(\mathbf{X}'\mathbf{X}/n)^{-1} = \mathbf{Q}^{-1}$ . Therefore, if the limiting distribution of the random vector in (4-26) exists, then that limiting distribution is the same as that of

$$\left[ \text{plim} \left( \frac{\mathbf{X}'\mathbf{X}}{n} \right)^{-1} \right] \left( \frac{1}{\sqrt{n}} \right) \mathbf{X}'\boldsymbol{\varepsilon} = \mathbf{Q}^{-1} \left( \frac{1}{\sqrt{n}} \right) \mathbf{X}'\boldsymbol{\varepsilon}. \quad (4-27)$$

Thus, we must establish the limiting distribution of

$$\left( \frac{1}{\sqrt{n}} \right) \mathbf{X}'\boldsymbol{\varepsilon} = \sqrt{n}(\bar{\mathbf{w}} - E[\bar{\mathbf{w}}]), \quad (4-28)$$

where  $E[\bar{\mathbf{w}}] = \mathbf{0}$ . [See (4-23).] We can use the multivariate Lindeberg-Feller version of the central limit theorem (D.19.A) to obtain the limiting distribution of  $\sqrt{n}\bar{\mathbf{w}}$ . Using that formulation,  $\bar{\mathbf{w}}$  is the average of  $n$  independent random vectors  $\mathbf{w}_i = \mathbf{x}_i \varepsilon_i$ , with means  $\mathbf{0}$  and variances

$$\text{Var}[\mathbf{x}_i \varepsilon_i] = \sigma^2 E[\mathbf{x}_i \mathbf{x}_i'] = \sigma^2 \mathbf{Q}_i. \quad (4-29)$$

The variance of  $\sqrt{n}\bar{\mathbf{w}}$  is

$$\sigma^2 \bar{\mathbf{Q}}_n = \sigma^2 \left( \frac{1}{n} \right) [\mathbf{Q}_1 + \mathbf{Q}_2 + \cdots + \mathbf{Q}_n]. \quad (4-30)$$

As long as the sum is not dominated by any particular term and the regressors are well behaved, which in this case means that (4-21) holds,

$$\lim_{n \rightarrow \infty} \sigma^2 \bar{\mathbf{Q}}_n = \sigma^2 \mathbf{Q}. \quad (4-31)$$

Therefore, we may apply the Lindeberg-Feller central limit theorem to the vector  $\sqrt{n}\bar{\mathbf{w}}$ , as we did in Section D.3 for the univariate case  $\sqrt{n}\bar{x}$ . We now have the elements we need for a formal result. If  $[\mathbf{x}_i \varepsilon_i]$ ,  $i = 1, \dots, n$  are independent vectors distributed with mean  $\mathbf{0}$  and variance  $\sigma^2 \mathbf{Q}_i < \infty$ , and if (4-21) holds, then

$$\left( \frac{1}{\sqrt{n}} \right) \mathbf{X}'\boldsymbol{\varepsilon} \xrightarrow{d} N[\mathbf{0}, \sigma^2 \mathbf{Q}]. \quad (4-32)$$

It then follows that

$$\mathbf{Q}^{-1} \left( \frac{1}{\sqrt{n}} \right) \mathbf{X}'\boldsymbol{\varepsilon} \xrightarrow{d} N[\mathbf{Q}^{-1}\mathbf{0}, \mathbf{Q}^{-1}(\sigma^2 \mathbf{Q})\mathbf{Q}^{-1}]. \quad (4-33)$$

Combining terms,

$$\sqrt{n}(\mathbf{b} - \boldsymbol{\beta}) \xrightarrow{d} N[\mathbf{0}, \sigma^2 \mathbf{Q}^{-1}]. \quad (4-34)$$

Note that the Lindeberg-Levy version does not apply because  $\text{Var}[\mathbf{w}_i]$  is not necessarily constant.

## CHAPTER 4 ♦ Statistical Properties of the Least Squares Estimator 67

Using the technique of Section D.3, we obtain the asymptotic distribution of  $\mathbf{b}$ :

**THEOREM 4.3** <sup>4</sup> **Asymptotic Distribution of  $\mathbf{b}$  with Independent Observations**

If  $\{\varepsilon_i\}$  are independently distributed with mean zero and finite variance  $\sigma^2$  and  $x_{ik}$  is such that the Grenander conditions are met, then

$$\mathbf{b} \stackrel{a}{\sim} N \left[ \boldsymbol{\beta}, \frac{\sigma^2}{n} \mathbf{Q}^{-1} \right]. \quad (4-35)$$

In practice, it is necessary to estimate  $(1/n)\mathbf{Q}^{-1}$  with  $(\mathbf{X}'\mathbf{X})^{-1}$  and  $\sigma^2$  with  $\mathbf{e}'\mathbf{e}/(n-K)$ .

If  $\mathbf{e}$  is normally distributed, then Result ES.7 in Table 4.4, Section 4.8 holds in every sample, so it holds asymptotically as well. The important implication of this derivation is that if the regressors are well behaved and observations are independent, then the asymptotic normality of the least squares estimator does not depend on normality of the disturbances; it is a consequence of the central limit theorem. We will consider other, more general cases in the sections to follow.

result (4-18)  
normality of  $\mathbf{b}|\mathbf{X}$

**4.3.3 CONSISTENCY OF  $s^2$  AND THE ESTIMATOR OF Asy. Var[ $\mathbf{b}$ ]**

To complete the derivation of the asymptotic properties of  $\mathbf{b}$ , we will require an estimator of  $\text{Asy. Var}[\mathbf{b}] = (\sigma^2/n)\mathbf{Q}^{-1}$ . With (4-21), it is sufficient to restrict attention to  $s^2$ , so the purpose here is to assess the consistency of  $s^2$  as an estimator of  $\sigma^2$ . Expanding

$$s^2 = \frac{1}{n-K} \mathbf{e}'\mathbf{M}\mathbf{e}$$

produces

$$s^2 = \frac{1}{n-K} [\mathbf{e}'\mathbf{e} - \mathbf{e}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}] = \frac{n}{n-k} \left[ \frac{\mathbf{e}'\mathbf{e}}{n} - \left( \frac{\mathbf{e}'\mathbf{X}}{n} \right) \left( \frac{\mathbf{X}'\mathbf{X}}{n} \right)^{-1} \left( \frac{\mathbf{X}'\mathbf{e}}{n} \right) \right].$$

The leading constant clearly converges to 1. We can apply (4-21), (4-24) (twice), and the product rule for probability limits (Theorem D.14) to assert that the second term in the brackets converges to 0. That leaves

$$\overline{\varepsilon^2} = \frac{1}{n} \sum_{i=1}^n \varepsilon_i^2.$$

This is a narrow case in which the random variables  $\varepsilon_i^2$  are independent with the same finite mean  $\sigma^2$ , so not much is required to get the mean to converge almost surely to  $\sigma^2 = E[\varepsilon_i^2]$ . By the Markov theorem (D.8), what is needed is for  $E[|\varepsilon_i^2|^{1+\delta}]$  to be finite, so the minimal assumption thus far is that  $\varepsilon_i$  have finite moments up to slightly greater than 2. Indeed, if we further assume that every  $\varepsilon_i$  has the same distribution, then by the Khinchine theorem (D.5) or the corollary to D.8, finite moments (of  $\varepsilon_i$ ) up to 2 is

<sup>6</sup> See McCallum (1973) for some useful commentary on deriving the asymptotic covariance matrix of the least squares estimator.

## 68 PART I ♦ The Linear Regression Model

sufficient. **Mean square convergence** would require  $E[\varepsilon_i^4] = \phi_\varepsilon < \infty$ . Then the terms in the sum are independent, with mean  $\sigma^2$  and variance  $\phi_\varepsilon - \sigma^4$ . So, under fairly weak conditions, the first term in brackets converges in probability to  $\sigma^2$ , which gives our result,

$$\text{plim } s^2 = \sigma^2,$$

and, by the product rule,

$$\text{plim } s^2 (\mathbf{X}'\mathbf{X}/n)^{-1} = \sigma^2 \mathbf{Q}^{-1}.$$

The appropriate *estimator* of the asymptotic covariance matrix of  $\mathbf{b}$  is

$$\text{Est. Asy. Var}[\mathbf{b}] = s^2 (\mathbf{X}'\mathbf{X})^{-1}.$$

#### 4.4.4 ASYMPTOTIC DISTRIBUTION OF A FUNCTION OF $\mathbf{b}$ : THE DELTA METHOD AND THE METHOD OF KRINSKY AND HODGE

We can extend Theorem D.22 to functions of the least squares estimator. Let  $\mathbf{f}(\mathbf{b})$  be a set of  $J$  continuous, linear, or nonlinear and continuously differentiable functions of the least squares estimator, and let

$$\mathbf{C}(\mathbf{b}) = \frac{\partial \mathbf{f}(\mathbf{b})}{\partial \mathbf{b}'},$$

where  $\mathbf{C}$  is the  $J \times K$  matrix whose  $j$ th row is the vector of derivatives of the  $j$ th function with respect to  $\mathbf{b}'$ . By the Slutsky theorem (D.12),

$$\text{plim } \mathbf{f}(\mathbf{b}) = \mathbf{f}(\boldsymbol{\beta})$$

and

$$\text{plim } \mathbf{C}(\mathbf{b}) = \frac{\partial \mathbf{f}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'} = \boldsymbol{\Gamma}.$$

Using our usual linear Taylor series approach (see Section 5.5), we expand this set of functions in the approximation

$$\mathbf{f}(\mathbf{b}) = \mathbf{f}(\boldsymbol{\beta}) + \boldsymbol{\Gamma} \times (\mathbf{b} - \boldsymbol{\beta}) + \text{higher-order terms}.$$

The higher-order terms become negligible in large samples if  $\text{plim } \mathbf{b} = \boldsymbol{\beta}$ . Then, the asymptotic distribution of the function on the left-hand side is the same as that on the right. Thus, the mean of the asymptotic distribution is  $\text{plim } \mathbf{f}(\mathbf{b}) = \mathbf{f}(\boldsymbol{\beta})$ , and the asymptotic covariance matrix is  $\{\boldsymbol{\Gamma}[\text{Asy. Var}(\mathbf{b} - \boldsymbol{\beta})]\boldsymbol{\Gamma}'\}$ , which gives us the following theorem:

#### THEOREM 4.4.5 Asymptotic Distribution of a Function of $\mathbf{b}$

If  $\mathbf{f}(\mathbf{b})$  is a set of continuous and continuously differentiable functions of  $\mathbf{b}$  such that  $\boldsymbol{\Gamma} = \partial \mathbf{f}(\boldsymbol{\beta}) / \partial \boldsymbol{\beta}'$  and if Theorem 4.5 holds, then

$$\mathbf{f}(\mathbf{b}) \stackrel{d}{\sim} N \left[ \mathbf{f}(\boldsymbol{\beta}), \boldsymbol{\Gamma} \left( \frac{\sigma^2}{n} \mathbf{Q}^{-1} \right) \boldsymbol{\Gamma}' \right]. \quad (4-36)$$

In practice, the estimator of the asymptotic covariance matrix would be

$$\text{Est. Asy. Var}[\mathbf{f}(\mathbf{b})] = \mathbf{C}[s^2 (\mathbf{X}'\mathbf{X})^{-1}] \mathbf{C}'.$$

## CHAPTER 4 ♦ Statistical Properties of the Least Squares Estimator 69

If any of the functions are nonlinear, then the property of unbiasedness that holds for  $\mathbf{b}$  may not carry over to  $\mathbf{f}(\mathbf{b})$ . Nonetheless, it follows from (4-25) that  $\mathbf{f}(\mathbf{b})$  is a consistent estimator of  $\mathbf{f}(\beta)$ , and the asymptotic covariance matrix is readily available.

**Example 4.4 Nonlinear Functions of Parameters: The Delta Method**

A dynamic version of the demand for gasoline model in Example 4.4 would be used to separate the short and long term impacts of changes in income and prices. The model would be

$$\ln(G/Pop)_t = \beta_1 + \beta_2 \ln P_{G,t} + \beta_3 \ln(Income/Pop)_t + \beta_4 \ln P_{nc,t} + \beta_5 \ln P_{uc,t} + \gamma \ln(G/Pop)_{t-1} + \varepsilon_t,$$

where  $P_{nc}$  and  $P_{uc}$  are price indexes for new and used cars.

In this model, the short run price and income elasticities are  $\beta_2$  and  $\beta_3$ . The long run elasticities are  $\phi_2 = \beta_2/(1 - \gamma)$  and  $\phi_3 = \beta_3/(1 - \gamma)$ , respectively. (See Section 20.3 for development of this model.) To estimate the long run elasticities, we will estimate the parameters by least squares and then compute these two nonlinear functions of the estimates. We can use the delta method to estimate the standard errors.

Least squares estimates of the model parameters with standard errors and  $t$  ratios are given in Table 4.7. Note the much improved fit compared to the model in Example 4.4—this is typical when lagged values of the dependent variable are added to the regression.

The estimated short run elasticities are the estimates given in the table. The two estimated long run elasticities are  $f_2 = \beta_2/(1 - \gamma) = -0.069532/(1 - 0.830971) = -0.411358$  and  $f_3 = 0.164047/(1 - 0.830971) = 0.970522$ . (Note how close this estimate is to the estimate from the static equation in Example 4.4.) To compute the estimates of the standard errors, we need the partial derivatives of these functions with respect to the six parameters in the model:

$$g_2' = \partial \phi_2 / \partial \beta' = [0, 1/(1 - \gamma), 0, 0, 0, \beta_2/(1 - \gamma)^2] = [0, 5.91613, 0, 0, 0, -2.43365],$$

$$g_3' = \partial \phi_3 / \partial \beta' = [0, 0, 1/(1 - \gamma), 0, 0, \beta_3/(1 - \gamma)^2] = [0, 0, 5.91613, 0, 0, 5.74174].$$

TABLE 4.7 Regression Results for a Demand Equation

Sum of squared residuals:	0.0127352		
Standard error of the regression:	0.0168227		
$R^2$ based on 52 observations	0.9951081		
Variable	Coefficient	Standard Error	t Ratio
Constant	-3.123195	0.99583	-3.136
$\ln P_G$	-0.069532	0.04377	-4.720
$\ln Income/Pop$	0.164047	0.07771	2.981
$\ln P_{nc}$	-0.1783975	0.15707	-3.377
$\ln P_{uc}$	0.127009	0.10338	3.551
last period $\ln G/Pop$	0.830971	0.04576	18.158

Estimated Covariance Matrix for  $\mathbf{b}$  ( $e - n = \text{times } 10^{-n}$ )

Constant	$\ln P_G$	$\ln(Income/Pop)$	$\ln P_{nc}$	$\ln P_{uc}$	$\ln(G/Pop)_{t-1}$
0.99168					
-0.0012088	0.00021705				
-0.052602	1.62165e-5	0.0030279			
0.0051016	-0.00021705	-0.00024708	0.0030440		
0.0091672	-4.0551e-5	-0.00060624	-0.0016782	0.0012795	
0.043915	-0.0001109	-0.0021881	0.00068116	8.57001e-5	0.0020943

Pos: check lettered subscripts, OK marked for italics?



Using (4-36), we can now compute the estimates of the asymptotic variances for the two estimated long run elasticities by computing  $g_2' [s^2(X'X)^{-1}] g_2$  and  $g_3' [s^2(X'X)^{-1}] g_3$ . The results are 0.023194 and 0.0263692, respectively. The two asymptotic standard errors are the square roots, 0.152296 and 0.162386. We can also form confidence intervals in

the same way that we did in Example 4.4. The 95 percent confidence interval for the long run price elasticity would be  $-0.411358 \pm 2.014 (0.152296) = [-0.718098, -0.104618]$ . The interval for the income elasticity is  $0.870523 \pm 2.014 (0.162386) = [0.643460, 1.297585]$ .

## CHAPTER 4 ♦ Statistical Properties of the Least Squares Estimator 71

TABLE 4.8 Simulation Results

Parameter	Estimate	Std. Error	Sample Mean	Sample Std. Dev.
$\beta_2$	-0.069332	0.04377	-0.069453	0.035074
$\beta_3$	0.164047	0.07771	0.16410	0.053602
$\gamma$	0.830971	0.04576	0.83083	0.044533
$\phi_2$	-0.41361	0.152296	-0.44913	0.19444
$\phi_3$	0.97527	0.162382	0.96313	0.18787

Krinsky and Robb (1986) report huge differences in the standard errors produced by the delta method compared to the simulation based estimator. In a subsequent paper (1990), they report that the entire difference can be attributed to a bug in the software they used—upon redoing the computations, their estimates are essentially the same with the two methods. It is difficult to draw a conclusion about the effectiveness of the delta method based on the received results—it does seem at this juncture that the delta method remains an effective device that can often be employed with a hand calculator as opposed to the much more computation intensive Krinsky and Robb (1986) technique. Unfortunately, the results of any comparison will depend on the data, the model, and the functions being computed. The amount of nonlinearity in the sense of the complexity of the functions seems not to be the answer. Krinsky and Robb's case was motivated by the extreme complexity of the translog elasticities. In another study, Hole (2006) examines a similarly complex problem, and finds that the delta method still appears to be the most accurate. For another (now classic) application, see Example 6.5.

## 4.4.5 ASYMPTOTIC EFFICIENCY

We have not established any large-sample counterpart to the Gauss–Markov theorem. That is, it remains to establish whether the large-sample properties of the least squares estimator are optimal by any measure. The Gauss–Markov theorem establishes finite sample conditions under which least squares is optimal. The requirements that the estimator be linear and unbiased limit the theorem's generality, however. One of the main purposes of the analysis in this chapter is to broaden the class of estimators in the classical model to those which might be biased, but which are consistent. Ultimately, we shall also be interested in nonlinear estimators. These cases extend beyond the reach of the Gauss–Markov theorem. To make any progress in this direction, we will require an alternative estimation criterion.

**DEFINITION 4.1 Asymptotic Efficiency**

*An estimator is asymptotically efficient if it is consistent, asymptotically normally distributed, and has an asymptotic covariance matrix that is not larger than the asymptotic covariance matrix of any other consistent, asymptotically normally distributed estimator.*

linear  
regression