4

ESTIMATING THE REGRESSION MODEL BY LEAST SQUARES

4.1 INTRODUCTION

In this chapter, we will examine least squares in detail as an **estimator** of the parameters of the linear regression model (defined in Table 4.1). There are other candidates for estimating $\boldsymbol{\beta}$. For example, we might use the coefficients that minimize the sum of absolute values of the residuals. We begin in Section 4.2 by considering the question "Why should we use least squares?" We will then analyze the estimator in detail. 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 linear model is one of few settings in which the exact finite-sample properties of an estimator are known. In most cases, the only known properties are those that apply to large samples. We can approximate finite-sample behavior by using what we know about large-sample properties. In Section 4.4, we will examine the large-sample or **asymptotic properties** of the least squares estimator of the regression model.[1] Section 4.5 considers **robust inference**. The problem considered here is how to carry out inference when (real) data may not satisfy the assumptions of the basic linear model. Section 4.6 develops a method for inference based on functions of model parameters, rather than the estimates themselves.

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 4.7, 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 4.7. One of the most familiar applications of interval estimation is using the model to predict the dependent variable and to provide a plausible range of uncertainty for that prediction. Section 4.8 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 4.9, 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

[1] This discussion will use results on asymptotic distributions. It may be helpful to review Appendix D before proceeding to Section 4.4.

TABLE 4.1 Assumptions of the Classical Linear Regression Model

A1. Linearity: $y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{iK}\beta_K + \varepsilon_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$. For the sample, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.

A2. Full rank: The $n \times K$ sample data matrix, $\mathbf{X}$, has full column rank for every $n \geq$ K.

A3. Exogeneity of the independent variables: $E[\varepsilon_i | x_{j1}, x_{j2}, \ldots, x_{jK}] = 0, i, j = 1, \ldots, n$. There is no correlation between the disturbances and the independent variables. $E[\boldsymbol{\varepsilon}|\mathbf{X}] = \mathbf{0}$.

A4. Homoscedasticity and nonautocorrelation: Each disturbance, ε_i, has the same finite variance; $\text{E}[\varepsilon_i^2|\mathbf{X}] = \sigma^2$. Every disturbance ε_i is uncorrelated with every other disturbance, ε_j, conditioned on $\mathbf{X}$; $E[\varepsilon_i\varepsilon_j|\mathbf{X}] = 0, i \neq j$. $E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}] = \sigma^2\mathbf{I}$.

A5. Stochastic or nonstochastic data: $(x_{i1}, x_{i2}, \ldots, x_{iK}), i = 1, \ldots, n$.

A6. Normal distribution: The disturbances, ε_i, are normally distributed. $\boldsymbol{\varepsilon}|\mathbf{X} \sim N[\mathbf{0},\sigma^2\mathbf{I}]$.

data, is examined in Sections 4.9.1 and 4.9.2. Missing data have the potential to derail the entire analysis. The benign case in which missing values are simply unexplainable random gaps in the data set is considered in Section 4.9.3. The more complicated case of nonrandomly missing data is discussed in Chapter 19. Finally, the problems of badly measured and outlying observations are examined in Section 4.9.4 and 4.9.5.

This chapter describes the properties of estimators. The assumptions in Table 4.1 will provide the framework for the analysis. (The assumptions are discussed in greater detail in Chapter 3.) For the present, it is useful to assume that the data are a cross section of independent, identically distributed random draws from the joint distribution of $(y_i,\mathbf{x}_i)$ with A1–A3 which defines $E[y_i|\mathbf{x}_i]$. Later in the text (and in Section 4.5), we will consider more general cases. The leading exceptions, which all bear some similarity, are stratified samples, cluster samples, **panel data**, and spatially correlated data. In these cases, groups of related individual observations constitute the observational units. The time-series case in Chapters 20 and 21 will deal with data sets in which potentially all observations are correlated. These cases will be treated later when they are developed in more detail. Under random (cross-section) sampling, with little loss of generality, we can easily obtain very general statistical results such as consistency and asymptotic normality. Later, such as in Chapter 11, we will be able to accommodate the more general cases fairly easily.

4.2 MOTIVATING LEAST SQUARES

Ease of computation is one reason that is occasionally offered to motivate least squares. But, with modern software, ease of computation is a minor (usually trivial) virtue. There are several theoretical 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 equation fit by least squares is an optimal linear predictor for the explained variable. Thus, it enjoys a sort of robustness that other estimators do not. Finally, under the specific assumptions of the classical model, by one reasonable criterion, least squares will be the most efficient use of the data.

4.2.1 POPULATION ORTHOGONALITY CONDITIONS

Let $\mathbf{x}$ denote the vector of independent variables in the population regression model. Assumption A3 states that $E[\varepsilon|\mathbf{x}] = 0$. Three useful results follow from this. First, by iterated expectations (Theorem B.1), $E_{\mathbf{x}}E[\varepsilon|\mathbf{x}]] = E_{\mathbf{x}}[0] = E[\varepsilon] = 0$; ε has zero mean, conditionally and unconditionally. Second, by Theorem B.2, Cov$[\mathbf{x},\varepsilon] =$ Cov$[\mathbf{x},E[\varepsilon|\mathbf{x}]] =$ Cov$[\mathbf{x},0] = \mathbf{0}$ so $\mathbf{x}$ and ε are uncorrelated. Finally, combining the earlier results, $E[\mathbf{x}\varepsilon] =$ Cov$[\mathbf{x},\varepsilon] + E[\varepsilon]E[\mathbf{x}] = \mathbf{0}$. We write the third of these as $E[\mathbf{x}\varepsilon] = E[\mathbf{x}(y - \mathbf{x}'\beta)] = \mathbf{0}$ or

$$E[\mathbf{x}y] = E[\mathbf{x}\mathbf{x}']\boldsymbol{\beta}. \tag{4-1}$$

Now, recall the least squares normal equations (3-5) based on the sample of n observations, $\mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{X}\mathbf{b}$. Divide this by n and write it as a summation to obtain

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

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 means in (4-2) are estimators of their counterparts in (4-1). Thus, by using least squares, we are mimicking in the sample the relationship that holds in the population.

4.2.2 MINIMUM MEAN SQUARED ERROR PREDICTOR

Consider the problem of finding an **optimal linear predictor** for y. Once again, ignore Assumption A6 and, in addition, drop Assumption A1. The conditional mean function, $E[y|\mathbf{x}]$, might be nonlinear. 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 minimum mean squared error predictor would be the conditional mean function in all cases. Here, we consider only a linear predictor.) The expected squared error of the linear predictor is

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

This can be written as

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

We seek the γ that minimizes this expectation. The first term is not a function of γ, so only the second term needs to be minimized. The necessary condition is

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

We arrive at the equivalent condition

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

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

$$E[\mathbf{x}y] = E[\mathbf{x}\mathbf{x}']\boldsymbol{\gamma}. \tag{4-3}$$

This is the same as (4-1), which takes us back to the least squares condition. Assuming that these expectations exist, they would be estimated by the sums in (4-2), which means

THEOREM 4.1 Minimum Mean Squared Error Predictor
If the mechanism generating the data $(\mathbf{x}_i, y_i), i = 1, \ldots, n$, *is such that the law of large numbers applies to the estimators in (4-2) of the matrices in (4-1), then the slopes of the minimum expected squared error linear predictor of y are estimated by the least squares coefficient vector.*

that regardless of the form of the conditional mean, least squares is an estimator of the coefficients of the minimum expected squared error linear predictor of $y|\mathbf{x}$.

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.

4.3 STATISTICAL PROPERTIES OF THE LEAST SQUARES ESTIMATOR

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 it 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 is not obvious how one could know this if they were only going to analyze a single sample of data from the population. The argument adopted in 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.

The development of the properties of least squares as an estimator can be viewed in three stages. The **finite sample properties** based on Assumptions A1–A6 are precise, and are independent of the sample size. They establish the essential characteristics of the estimator, such as unbiasedness and the broad approach to be used to estimate the sampling variance. Finite sample results have two limiting aspects. First, they can only be obtained for a small number of statistics—essentially only for the basic least squares estimator. Second, the sharpness of the finite sample results is obtained by making assumptions about the data-generating process that we would prefer not to impose, such as normality of the disturbances (Assumption A6 in Table 4.1). Asymptotic properties of the estimator are obtained by deriving reliable results that will provide good approximations in moderate sized or large samples. For example, the large sample property of consistency of the least squares estimator is looser than unbiasedness in one respect, but at the same time, is more informative about how the estimator improves as more sample data are used. Finally, robust inference methods are a refinement of the asymptotic results. The essential asymptotic theory for least squares modifies the finite sample results after relaxing certain assumptions, mainly A5 (data-generating process for **X**) and A6 (normality). Assumption A4 (homoscedasticity and nonautocorrelation) remains a limitation on the generality of the model assumptions. Real-world data are likely to be heteroscedastic in ways that cannot be precisely quantified. They may also be autocorrelated as a consequence of the sample design, such as the within household correlation of panel data observations. These possibilities may taint the inferences that use standard errors that are based on A4. Robust methods are used to accommodate possible violations of Assumption A4 without redesigning the estimation strategy. That is, we continue to use least squares, but employ inference procedures that will be appropriate whether A4 is reasonable or not.

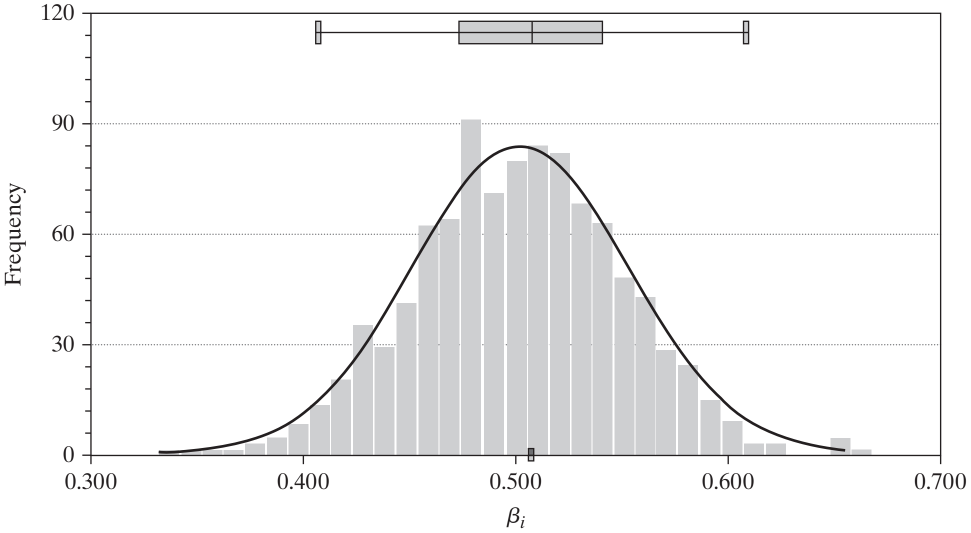
Example 4.1 The Sampling Distribution of a Least Squares Estimator

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 0, variance 1). We generated a set of ε_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 1,000 random samples of 100 observations on (y_i, x_i) from this population (without replacement), and with each one, computed the least squares slope, using at replication r,

$$b_r = \left[\Sigma_{i=1}^{100}(x_{ir} - \bar{x}_r)y_{ir} \right] / \left[\Sigma_{i=1}^{100}(x_{ir} - \bar{x}_r)^2 \right].$$

The histogram in Figure 4.1 shows the result of the experiment. Note that the distribution of slopes has mean and median 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 4.3.6.

FIGURE 4.1 Histogram for Sampled Least Squares Regression Slopes.



4.3.1 UNBIASED ESTIMATION

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

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

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

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

By Assumption A3, the expected value of the second term is $(\mathbf{X}'\mathbf{X}/n)^{-1}E[\Sigma_i\mathbf{x}_i\varepsilon_i/n|\mathbf{X}]$. Each term in the sum has expectation zero, which produces the result we need:

$$E[\mathbf{b}|\mathbf{X}] = \boldsymbol{\beta}. \tag{4-5}$$

Therefore,

$$E[\mathbf{b}] = E_{\mathbf{x}}\{E[\mathbf{b}|\mathbf{x}]\} = E_{\mathbf{x}}[\boldsymbol{\beta}] = \boldsymbol{\beta}. \tag{4-6}$$

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

4.3.2 OMITTED VARIABLE BIAS

Suppose that a correctly specified regression model would be

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{z}\gamma + \boldsymbol{\varepsilon}, \tag{4-7}$$

where the two parts have K and 1 columns, respectively. If we regress $\mathbf{y}$ on $\mathbf{X}$ without including the relevant variable, $\mathbf{z}$, then the estimator is

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{z}\gamma + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}. \tag{4-8}$$

(Note, "relevant" means $\gamma \neq 0$.) Taking the expectation, we see that unless $\mathbf{X}'\mathbf{z} = \mathbf{0}$, $\mathbf{b}$ is biased. The well-known result is the **omitted variable formula:**

$$E[\mathbf{b}|\mathbf{X},\mathbf{z}] = \boldsymbol{\beta} + \mathbf{p}_{\mathbf{X}.\mathbf{z}}\gamma, \tag{4-9}$$

where

$$\mathbf{p}_{\mathbf{X}.\mathbf{z}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{z}. \tag{4-10}$$

The vector $\mathbf{p}_{\mathbf{X}.\mathbf{z}}$ is the column of slopes in the least squares regression of $\mathbf{z}$ on $\mathbf{X}$. Theorem 3.2 (Frisch-Waugh) and Corollary 3.2.1 provide some insight for this result. For each coefficient in (4-9), we have

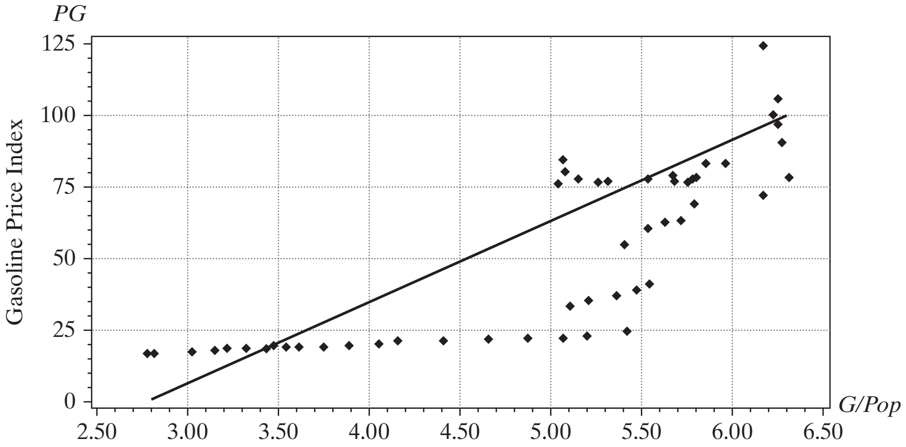
$$E[b_k|\mathbf{X},\mathbf{z}] = \beta_k + \gamma\left(\frac{\text{Cov}(z,x_k|\text{all other } x\text{'s}}{\text{Var}(x_k|\text{all other } x\text{'s})}\right) \tag{4-11}$$

Example 4.2 Omitted Variable in a Demand Equation

If a demand equation is estimated without the relevant income variable, then (4-11) shows how the estimated price elasticity will be biased. The gasoline market data we have examined in Example 2.3 provides a clear example. The base demand model is

$$\textit{Quantity} = \alpha + \beta\textit{Price} + \gamma\textit{Income} + \varepsilon.$$

FIGURE 4.2 Per Capita Gasoline Consumption Versus Price, 1953–2004.



Letting *b* be the slope coefficient in the regression of *Quantity* on *Price*, we obtain

$$E[b\,|\,Price, Income] = \beta + \gamma\frac{\text{Cov}[Price, Income]}{\text{Var}[Price]}.$$

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 γ would both 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* (in inverted Marshallian form). The plot disagrees with what one might expect. But a look at the data in Appendix Table F2.2 shows clearly what is at work. In these aggregate data, the simple correlations for (*G/Pop*, *Income/Pop*) and for (*PG*, *Income/Pop*) are 0.938 and 0.934, respectively. To see if the expected relationship between price and consumption shows up, we will have to purge our price and quantity data of the intervening effect of income. To do so, we rely on the Frisch–Waugh result in Theorem 3.2. In the simple regression of the 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, $\hat{\beta}$, is -0.16949 and the estimated income elasticity, $\hat{\gamma}$, is 0.96595. This agrees with expectations.

In this development, it is straightforward to deduce the directions of bias when there is a single included variable and one omitted variable, as in Example 4.2. It is important to note, however, that if more than one variable is included in **X**, then the terms in the omitted variable formula, (4-9) and (4-10), involve *multiple* regression coefficients, which have the signs of partial, not simple correlations. For example, in the demand model of the previous example, if the price of a closely related product, say new cars, had been included as well, then the simple correlation between gasoline 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:

$$\begin{aligned} E[b_{Gasoline\ Price}|\mathbf{X,z}] = \beta_{Gasoline\ Price} & \\ + \left(\frac{\text{Cov}(Income,\ Gasoline\ Price|New\ Cars\ Price}{\text{Var}(Gasoline\ Price|New\ Cars\ Price)}\right)\gamma. & \qquad \textbf{(4-12)} \end{aligned}$$

This sign might not be obvious, and it would become even less so as more regressors are added to the equation. However, (4-12) does suggest what would be needed for an argument that the least squares estimator remains unbiased, at least for coefficients that correspond to zero partial correlations.

4.3.3 INCLUSION OF IRRELEVANT VARIABLES

We can view the omission of a set of relevant variables as equivalent to imposing an incorrect restriction on (4-7). In particular, omitting **z** is equivalent to *incorrectly* estimating (4-7) subject to the restriction $\gamma = 0$. Incorrectly imposing a restriction produces a biased estimator. Suppose, however, that our error is a failure to use some information that is *correct*. If the regression model is correctly given by $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ and we estimate it as if (4-7) were correct [i.e., we include an (or some) extra variable(s)], then the inclusion of the irrelevant variable **z** in the regression is equivalent to failing to impose $\gamma = 0$ on (4-7) in estimation. But (4-7) is not incorrect; it simply fails to incorporate $\gamma = 0$. The least squares estimator of $(\boldsymbol{\beta}, \gamma)$ in (4-7) is still unbiased *even given* the restriction:

$$E\left[\begin{pmatrix}\mathbf{b}\\ c\end{pmatrix}|\mathbf{X,z}\right] = \begin{pmatrix}\boldsymbol{\beta}\\ \gamma\end{pmatrix} = \begin{pmatrix}\boldsymbol{\beta}\\ 0\end{pmatrix}. \qquad \textbf{(4-13)}$$

The broad result is that *including irrelevant variables in the estimation equation does not lead to bias in the estimation of the nonzero coefficients*. Then where is the problem? It would seem that to be conservative, one might generally want to overfit the model. As we will show in Section 4.9.1, the covariance matrix in the regression that properly omits the irrelevant **z** is generally smaller than the covariance matrix for the estimator obtained in the presence of the superfluous variables. *The cost of overspecifying the model is larger variances (less precision) of the estimators.*

4.3.4 VARIANCE OF THE LEAST SQUARES ESTIMATOR

The least squares coefficient vector is

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{\beta} + \mathbf{A}\boldsymbol{\varepsilon}, \qquad \textbf{(4-14)}$$

where $\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. By Assumption A4, $\text{E}[\varepsilon\varepsilon'|\mathbf{X}] = \text{Var}[\varepsilon|\mathbf{X}] = \sigma^2\mathbf{I}$. The conditional covariance matrix of the least squares slope estimator is

$$\begin{aligned} \text{Var}[\mathbf{b}|\mathbf{X}] &= E[(\mathbf{b} - \boldsymbol{\beta})(\mathbf{b} - \boldsymbol{\beta})'|\mathbf{X}] \\ &= E[\mathbf{A}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{A}'|\mathbf{X}] \\ &= \mathbf{A}E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}]\mathbf{A}' \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}. \end{aligned} \qquad \textbf{(4-15)}$$

If we wish to use **b** to test hypotheses about $\boldsymbol{\beta}$ or to form confidence intervals, then we will require a sample estimate of this matrix. The population parameter σ^2 remains to be estimated. Because σ^2 is the expected value of ε_i^2 and e_i is an estimate of ε_i, $\hat{\sigma}^2 = (1/\text{n})\sum_{i=1}^{n} e_i^2$ would seem to be the natural estimator. But the least squares residuals are imperfect estimates of their population counterparts; $e_i = y_i - \mathbf{x}_i'\mathbf{b} = \varepsilon_i - \mathbf{x}_i'(\mathbf{b} - \boldsymbol{\beta})$. The estimator $\hat{\sigma}^2$ is distorted because $\boldsymbol{\beta}$ must be estimated.

The least squares residuals are $\mathbf{e} = \mathbf{My} = \mathbf{M}[\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}] = \mathbf{M}\boldsymbol{\varepsilon}$, as $\mathbf{MX} = \mathbf{0}$. [See Definition 3.1 and (3-15).] An estimator of σ^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 1×1 matrix, so it is equal to its trace. By using (A-94), $E[\text{tr}(\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon})\,|\,\mathbf{X}] = E[\text{tr}(\mathbf{M}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')\,|\,\mathbf{X}]$. Because $\mathbf{M}$ 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,

$$E[\mathbf{e}'\mathbf{e}\,|\,\mathbf{X}] = (n - K)\sigma^2. \tag{4-16}$$

The natural estimator is biased toward zero, but the bias becomes smaller as the sample size increases. An unbiased estimator of σ^2 is

$$s^2 = \frac{\mathbf{e}'\mathbf{e}}{n - K}. \tag{4-17}$$

Like $\mathbf{b}$, s^2 is unbiased unconditionally, because $E[s^2] = E_{\mathbf{x}}\{E[s^2\,|\,\mathbf{X}]\} = E_{\mathbf{x}}[\sigma^2] = \sigma^2$. The **standard error of the regression** is s, the square root of s^2. We can then compute

$$\text{Est. Var}[\mathbf{b}\,|\,\mathbf{X}] = s^2(\mathbf{X}'\mathbf{X})^{-1}. \tag{4-18}$$

Henceforth, we shall use the notation Est.Var[Est.Var[·]] to indicate a sample estimate of the **sampling variance** of an estimator. The square root of the kth 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.

4.3.5 THE GAUSS–MARKOV THEOREM

We will now obtain a general result for the class of linear unbiased estimators of $\boldsymbol{\beta}$. Because $\mathbf{b}\,|\,\mathbf{X} = \mathbf{Ay}$, where $\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, is a linear function of ε, by the definition we will use here, it is a **linear estimator** of $\boldsymbol{\beta}$. Because $E[\mathbf{A}\boldsymbol{\varepsilon}\,|\,\mathbf{X}] = \mathbf{0}$, *regardless of the distribution of* $\boldsymbol{\varepsilon}$, *under our other assumptions,* $\mathbf{b}$ *is a linear, unbiased estimator of* $\boldsymbol{\beta}$.

THEOREM 4.2 Gauss–Markov Theorem
In the linear regression model with given regressor matrix $\mathbf{X}$, (1) *the least squares estimator,* $\mathbf{b}$, *is the minimum variance linear unbiased estimator of* $\boldsymbol{\beta}$ and (2) *for any vector of constants* $\mathbf{w}$, *the minimum variance linear unbiased estimator of* $\mathbf{w}'\boldsymbol{\beta}$ *is* $\mathbf{w}'\mathbf{b}$.

Note that the theorem makes no use of Assumption A6, normality of the distribution of the disturbances. Only A1 to A4 are necessary. Let $\mathbf{b}_0 = \mathbf{Cy}$ be a different linear unbiased estimator of $\boldsymbol{\beta}$, where $\mathbf{C}$ is a $K \times n$ matrix. If $\mathbf{b}_0$ is unbiased, then $E[\mathbf{Cy}\,|\,\mathbf{X}] = E[(\mathbf{CX}\boldsymbol{\beta} + \mathbf{C}\boldsymbol{\varepsilon})\,|\,\mathbf{X}] = \boldsymbol{\beta}$, which implies that $\mathbf{CX} = \mathbf{I}$ and $\mathbf{b}_0 = \beta + \mathbf{C}\varepsilon$, so $\text{Var}[\mathbf{b}_0\,|\,\mathbf{X}] = \sigma^2\mathbf{CC}'$. Now, let $\mathbf{D} = \mathbf{C} - \mathbf{A}$ so $\mathbf{Dy} = \mathbf{b}_0 - \mathbf{b}$. Because $\mathbf{CX} = \mathbf{I}$ and

$\mathbf{AX} = \mathbf{I}, \mathbf{DX} = \mathbf{0}$ and $\mathbf{DA}' = \mathbf{0}$. Then, $\text{Var}[\mathbf{b}_0 | \mathbf{X}] = \sigma^2[(\mathbf{D} + \mathbf{A})(\mathbf{D} + \mathbf{A})']$. By multiplying the terms, we find

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

The quadratic form in $\mathbf{DD}'$ is $\mathbf{q}'\mathbf{DD}'\mathbf{q} = \mathbf{v}'\mathbf{v} \geq 0$. The conditional covariance matrix of $\mathbf{b}_0$ equals that of $\mathbf{b}$ plus a nonnegative definite matrix. Every quadratic form in $\text{Var}[\mathbf{b}_0 | \mathbf{X}]$ is larger than the corresponding quadratic form in $\text{Var}[\mathbf{b} | \mathbf{X}]$, which establishes result (1).

The proof of result (2) of the theorem follows from the previous derivation, because the variance of $\mathbf{w}'\mathbf{b}$ is a quadratic form in $\text{Var}[\mathbf{b} | \mathbf{X}]$, and likewise for any $\mathbf{b}_0$, and implies that each individual slope estimator b_k is the best linear unbiased estimator of β_k. (Let $\mathbf{w}$ be all zeros except for a one in the kth position.) The result applies to every linear combination of the elements of $\boldsymbol{\beta}$. *The implication is that under Assumptions A1–A5,* $\mathbf{b}$ *is the most efficient (linear unbiased) estimator of* $\boldsymbol{\beta}$.

4.3.6 THE NORMALITY ASSUMPTION

To this point, the specification and analysis of the regression model are **semiparametric** (see Section 12.3). We have not used Assumption A6, normality of $\boldsymbol{\varepsilon}$, in any of the results. In (4-4), $\mathbf{b}$ is a linear function of the disturbance vector, $\boldsymbol{\varepsilon}$. If $\boldsymbol{\varepsilon}$ 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

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

Each element of $\mathbf{b} | \mathbf{X}$ is normally distributed:

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

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

The exact distribution of $\mathbf{b}$ is conditioned on $\mathbf{X}$. The normal distribution of $\mathbf{b}$ in a finite sample is a consequence of the specific assumption of normally distributed disturbances. The normality assumption is useful for constructing test statistics and for forming **confidence intervals**. But we will ultimately find that we will be able to establish the results we need for inference about $\boldsymbol{\beta}$ based only on the sampling behavior of the statistics without tying the analysis to a narrow assumption of normality of $\boldsymbol{\varepsilon}$.

4.4 ASYMPTOTIC PROPERTIES OF THE LEAST SQUARES ESTIMATOR

The finite sample properties of the least squares estimator are helpful in suggesting the range of results that can be obtained from a sample of data. But the list of settings in which exact finite sample results can be obtained is extremely small. The assumption of normality likewise narrows the range of the applications. Estimation and inference can be based on approximate results that will be reliable guides in even moderately sized data sets, and require fewer assumptions.

4.4.1 CONSISTENCY OF THE LEAST SQUARES ESTIMATOR OF $\boldsymbol{\beta}$

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, unbiasedness has two shortcomings. First, save for the least squares slope estimator we are discussing in this chapter, it is rare for an econometric estimator to be unbiased. In nearly all cases beyond the multiple linear 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 two, 20 and 20,000 are all unbiased estimators of a population mean—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. The property of **consistency** improves on unbiasedness in both of these directions.

To begin, we leave the data-generating mechanism for **X** unspecified—**X** may be any mixture of constants and random variables generated independently of the process that generates $\boldsymbol{\varepsilon}$. We do make two crucial assumptions. The first is a modification of Assumption A5; **A5a**. $(\mathbf{x}_i, \varepsilon_i),\ i = 1, \ldots, n$ is a sequence of independent, identically distributed *observations*.

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

$$\underset{n \to \infty}{\text{plim}} \frac{\mathbf{X}'\mathbf{X}}{n} = \mathbf{Q}, \text{ a positive definite matrix.} \qquad \textbf{(4-19)}$$

Note how this extends A2. If every **X** has full column rank, then $\mathbf{X}'\mathbf{X}/n$ is a positive definite matrix in a specific sample of $n \geq K$ observations. Assumption (4-19) extends that to all samples with at least K observations. A straightforward way to reach (4-19) based on A5a is to assume

$$E[\mathbf{x}_i\mathbf{x}_i{}'] = \mathbf{Q},$$

so that by the law of large numbers, $(1/n)\Sigma_i\mathbf{x}_i\mathbf{x}_i'$ converges in probability to its expectation, $\mathbf{Q}$, and via Theorem D.14, $(\mathbf{X}'\mathbf{X}/n)^{-1}$ converges in probability to $\mathbf{Q}^{-1}$.

Time-series settings that involve 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 **X** that is broad enough to include most of these is the **Grenander Conditions** listed in Table 4.2.[2] 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.

At many points from here forward, we will make an assumption that the data are well behaved so that an estimator or statistic will converge to a result. Without repeating them in each instance, we will broadly rely on conditions such as those in Table 4.2.

The least squares estimator may be written

$$\mathbf{b} = \boldsymbol{\beta} + \left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\left(\frac{\mathbf{X}'\boldsymbol{\varepsilon}}{n}\right). \qquad \textbf{(4-20)}$$

Then,

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

[2] See Grenander (1956), Palma (2016, p. 373) and Judge et al. (1985, p. 162).

TABLE 4.2 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 $\lim_{n\to\infty} d_{nk}^2 = +\infty$. Hence, $\mathbf{x}_k$ does not degenerate to a sequence of zeros. *Sums of squares will continue to grow as the sample size increases.*

G2. $\text{Lim}_{n\to\infty}\, x_{ik}^2/d_{nk}^2 = 0$ for all $i = 1, \ldots, n$. No single observation will ever dominate $\mathbf{x}_k'\mathbf{x}_k$. *As $n \to \infty$, individual observations will become less important.*

G3. Let $\mathbf{C}_n$ be the sample correlation matrix of the columns of $\mathbf{X}$, excluding the constant term if there is one. Then $\lim_{n\to\infty}\mathbf{C}_n = \mathbf{C}$, 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. *This rank condition will not be violated as the sample size increases.*

We require the probability limit of the last term. In Section 4.2.1, we found that $E[\varepsilon|\mathbf{x}] = 0$ implies $E[\mathbf{x}\varepsilon] = \mathbf{0}$. Based on this result, again invoking D.4., we find $\mathbf{X}'\boldsymbol{\varepsilon}/n = (1/n)\Sigma_i\mathbf{x}_i\varepsilon_i$ converges in probability to its expectation of zero, so

$$\text{plim}\left(\frac{\mathbf{X}'\boldsymbol{\varepsilon}}{n}\right) = \mathbf{0}. \qquad \textbf{(4-21)}$$

It follows that

$$\text{plim}\,\mathbf{b} = \boldsymbol{\beta} + \mathbf{Q}^{-1}\cdot\mathbf{0} = \boldsymbol{\beta}. \qquad \textbf{(4-22)}$$

This result establishes that under Assumptions A1–A4 and the additional assumption (4-19), $\mathbf{b}$ is a **consistent estimator** of $\boldsymbol{\beta}$ in the linear regression model. Note how consistency improves on unbiasedness. The asymptotic result does not insist that $\mathbf{b}$ be unbiased. But, by the definition of consistency (see Definition D.6), it will follow that $\lim_{n\to\infty}\text{Prob}[|b_k - \beta_k| > \delta] = 0$ for any positive δ. This means that with increasing sample size, the estimator will be ever closer to the target. This is sometimes (loosely) labeled "asymptotic unbiasedness."

4.4.2 THE ESTIMATOR OF Asy. Var[b]

To complete the derivation of the asymptotic properties of $\mathbf{b}$, we will require an estimator of Asy. Var$[\mathbf{b}] = (\sigma^2/n)\mathbf{Q}^{-1}$. With (4-19), 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 σ^2. Expanding $s^2 = \boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}/(n - K)$ produces

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

The leading constant clearly converges to 1. We can apply (4-19), (4-21) (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 $\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2$. This is a narrow case in which the random variables ε_i^2 are independent with the same finite mean σ^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 ε_i have finite moments up to slightly greater than 2. Indeed, if we further assume that every ε_i has the same distribution, then by the Khinchine theorem (D.5) or the corollary to D8, finite moments (of ε_i) up to 2 is sufficient. So, under fairly weak conditions, the first term in brackets converges in probability to σ^2, which gives our result,

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

and, by the product rule,

$$\operatorname{plim} s^2(\mathbf{X}'\mathbf{X}/n)^{-1} = \sigma^2\mathbf{Q}^{-1}. \tag{4-23}$$

The appropriate *estimator* of the **asymptotic covariance matrix** of **b** is the familiar one,

$$\text{Est.Asy.Var}[\mathbf{b}] = s^2\,(\mathbf{X}'\mathbf{X})^{-1}. \tag{4-24}$$

4.4.3 ASYMPTOTIC NORMALITY OF THE LEAST SQUARES ESTIMATOR

By relaxing assumption A6, we will lose the exact normal distribution of the estimator that will enable us to form confidence intervals in Section 4.7. However, 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 (4-20) that

$$\sqrt{n}(\mathbf{b} - \boldsymbol{\beta}) = \left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\left(\frac{1}{\sqrt{n}}\right)\mathbf{X}'\boldsymbol{\varepsilon}. \tag{4-25}$$

If the limiting distribution of the random vector in (4-25) exists, then that limiting distribution is the same as that of

$$\left[\operatorname{plim}\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\right]\left(\frac{1}{\sqrt{n}}\right)\mathbf{X}'\boldsymbol{\varepsilon} = \mathbf{Q}^{-1}\left(\frac{1}{\sqrt{n}}\right)\mathbf{X}'\boldsymbol{\varepsilon}. \tag{4-26}$$

Thus, we must establish the limiting distribution of

$$\left(\frac{1}{\sqrt{n}}\right)\mathbf{X}'\varepsilon = \sqrt{n}(\overline{\mathbf{w}} - E[\overline{\mathbf{w}}]), \tag{4-27}$$

where $\mathbf{w}_i = \mathbf{x}_i\varepsilon_i$ and $E[\mathbf{w}_i] = E[\overline{\mathbf{w}}] = \mathbf{0}$. The mean vector $\overline{\mathbf{w}}$ is the average of n independent identically distributed random vectors with means $\mathbf{0}$ and variances

$$\text{Var}[\mathbf{x}_i\varepsilon_i] = \sigma^2 E[\mathbf{x}_i\mathbf{x}_i'] = \sigma^2\mathbf{Q}. \tag{4-28}$$

The variance of $\sqrt{n}\,\overline{\mathbf{w}} = \dfrac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i\varepsilon_i$ is

$$\sigma^2\left(\frac{1}{n}\right)[\mathbf{Q} + \mathbf{Q} + \cdots + \mathbf{Q}] = \sigma^2\mathbf{Q}. \tag{4-29}$$

We may apply the Lindeberg–Levy central limit theorem (D.18) to the vector $\sqrt{n}\,\overline{\mathbf{w}}$, as we did in Section D.3 for the univariate case $\sqrt{n}\,\overline{x}$. If $[\mathbf{x}_i\varepsilon_i]$, $i = 1, \ldots, n$ are independent vectors, each distributed with mean $\mathbf{0}$ and variance $\sigma^2\mathbf{Q} < \infty$, and if (4-19) holds, then

$$\left(\frac{1}{\sqrt{n}}\right)\mathbf{X}'\boldsymbol{\varepsilon} \xrightarrow{d} N[\mathbf{0}, \sigma^2\mathbf{Q}]. \tag{4-30}$$

It then follows that

$$\mathbf{Q}^{-1}\left(\frac{1}{\sqrt{n}}\right)\mathbf{X}'\boldsymbol{\varepsilon} \xrightarrow{d} N[\mathbf{Q}^{-1}\mathbf{0}, \mathbf{Q}^{-1}(\sigma^2\mathbf{Q})\mathbf{Q}^{-1}]. \tag{4-31}$$

Combining terms,

$$\sqrt{n}(\mathbf{b} - \boldsymbol{\beta}) \xrightarrow{d} N[\mathbf{0}, \sigma^2\mathbf{Q}^{-1}]. \tag{4-32}$$

Using the technique of Section D.3, we then obtain the **asymptotic distribution** of **b**:

THEOREM 4.3 Asymptotic Distribution of b with IID Observations
If $\{\varepsilon_i\}$ are independently distributed with mean zero and finite variance σ^2 and x_{ik} is such that the Grenander conditions are met, then

$$\mathbf{b} \overset{a}{\sim} N\left[\boldsymbol{\beta}, \frac{\sigma^2}{n}\mathbf{Q}^{-1}\right]. \tag{4-33}$$

The development here has relied on random sampling from $(\mathbf{x}_i, \varepsilon_i)$. If observations are not identically distributed, for example, if $E[\mathbf{x}_i\mathbf{x}_i'] = \boldsymbol{Q}_i$, then under suitable, more general assumptions, an argument could be built around the ***Lindeberg–Feller Central Limit Theorem*** *(D.19A). The essential results would be the same.*

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

If $\boldsymbol{\varepsilon}$ is normally distributed, then normality of $\mathbf{b}|\mathbf{X}$ 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.*

4.4.4 ASYMPTOTIC EFFICIENCY

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 linear regression model to those which might be biased, but which are consistent. Ultimately, we will be interested in nonlinear estimators as well. 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.

We can compare estimators based on their asymptotic variances. The complication in comparing two consistent estimators is that both converge to the true parameter as the sample size increases. Moreover, it usually happens (as in our Example 4.3), that they converge at the same rate—that is, in both cases, the asymptotic variances of the two estimators are of the same order, such as $O(1/n)$. In such a situation, we can sometimes compare the asymptotic variances for the same n to resolve the ranking. The least absolute deviations estimator as an alternative to least squares provides a leading example.

Example 4.3 Least Squares Vs. Least Absolute Deviations—A Monte Carlo Study

Least absolute deviations (LAD) is an alternative to least squares. (The LAD estimator is considered in more detail in Section 7.3.1.) The LAD estimator is obtained as

$$\mathbf{b}_{\text{LAD}} = \text{the minimizer of} \textstyle\sum_{i=1}^{n} |y_i - \mathbf{x}_i'\mathbf{b}_0|,$$

in contrast to the linear least squares estimator, which is

$$\mathbf{b}_{\text{LS}} = \text{the minimizer of} \textstyle\sum_{i=1}^{n} (y_i - \mathbf{x}_i'\mathbf{b_0})^2.$$

Suppose the regression model is defined by

$$y_i = \mathbf{x}_i'\beta + \varepsilon_i,$$

where the distribution of ε_i has conditional mean zero, constant variance σ^2, and conditional *median* zero as well—the distribution is symmetric—and $\text{plim}(1/n)\mathbf{X}'\varepsilon = \mathbf{0}.$ That is, all the usual regression assumptions, but with the normality assumption replaced by symmetry of the distribution. Then, under our assumptions, $\mathbf{b}_{\text{LS}}$ is a consistent and asymptotically normally distributed estimator with asymptotic covariance matrix given in Theorem 4.3, which we will call $\sigma^2\mathbf{A}.$ As Koenker and Bassett (1978, 1982), Huber (1987), Rogers (1993), and Koenker (2005) have discussed, under these assumptions, $\mathbf{b}_{\text{LAD}}$ is also consistent. A good estimator of the asymptotic variance of $\mathbf{b}_{\text{LAD}}$ would be $(1/2)^2[1/f(0)]^2\ \mathbf{A}$ where $f(0)$ is the density of ε at its median, zero. This means that we can compare these two estimators based on their asymptotic variances. The ratio of the asymptotic variance of the kth element of $\mathbf{b}_{\text{LAD}}$ to the corresponding element of $\mathbf{b}_{\text{LS}}$ would be

$$q_k = \text{Var}(b_{k,\text{LAD}})/\text{Var}(b_{k,\text{LS}}) = (1/2)^2(1/\sigma^2)[1/f(0)]^2.$$

If ε did actually have a normal distribution with mean (and median) zero, then $f(\varepsilon) = (2\pi\sigma^2)^{-1/2}\exp(-\varepsilon^2/(2\sigma^2))$ so $f(0) = (2\pi\sigma^2)^{-1/2}$ and for this special case $q_k = \pi/2$. If the disturbances are normally distributed, then LAD will be asymptotically less efficient by a factor of $\pi/2 = 1.573$.

The usefulness of the LAD estimator arises precisely in cases in which we cannot assume normally distributed disturbances. Then it becomes unclear which is the better estimator. It has been found in a long body of research that the advantage of the LAD estimator is most likely to appear in small samples and when the distribution of ε has thicker tails than the normal—that is, when outlying values of y_i are more likely. As the sample size grows larger, one can expect the LS estimator to regain its superiority. We will explore this aspect of the estimator in a small **Monte Carlo study**.

Examples 2.6 and 3.4 note an intriguing feature of the fine art market. At least in some settings, large paintings sell for more at auction than small ones. Appendix Table F4.1 contains the sale prices, widths, and heights of 430 Monet paintings. These paintings sold at auction for prices ranging from $10,000 to $33 million. A linear regression of the log of the price on a constant term, the log of the surface area, and the aspect ratio produces the results in the top line of Table 4.3. This is the focal point of our analysis. In order to study the different behaviors of the LS and LAD estimators, we will do the following Monte Carlo study: We will draw without replacement 100 samples of R observations from the 430. For each of the 100 samples, we will compute $\mathbf{b}_{\text{LS},r}$ and $\mathbf{b}_{\text{LAD},r}$. We then compute the average of the 100 vectors and the sample variance of the 100 observations.[3] The sampling variability of the 100 sets of results corresponds to the notion of "variation in repeated samples." For this experiment, we will do this for $R = 10$, 50, and 100. The overall sample size is fairly large, so it is reasonable to take the full sample results as at least approximately the "true parameters." The standard errors reported for the full sample LAD estimator are computed using **bootstrapping**. Briefly, the procedure is carried out by drawing B—we used $B = 100$—samples of n (430) observations *with replacement*, from the full sample of n observations. The estimated variance of the LAD estimator is then obtained by computing the mean squared deviation of these B estimates around the mean of the B estimates. This procedure is discussed in detail in Section 15.4.

TABLE 4.3 Estimated Equations for Art Prices

	Constant		*Log Area*		*Aspect Ratio*	
Full Sample	***Mean***	***Standard Error****	***Mean***	***Standard Error***	***Mean***	***Standard Error***
LS	−8.34327	0.67820	1.31638	0.09205	−0.09623	0.15784
LAD	−8.22726	0.82480	1.25904	0.13718	0.04195	0.22762
R = 10						
LS	−10.6218	8.39355	1.65525	1.21002	−0.07655	1.55330
LAD	−12.0635	11.1734	1.81531	1.53662	0.18269	2.11369
R = 50						
LS	−8.57755	1.94898	1.35026	0.27509	−0.08521	0.46600
LAD	−8.33638	2.18488	1.31408	0.36047	−0.06011	0.60910
R = 100						
LS	−8.38235	1.38332	1.32946	0.19682	−0.09378	0.33765
LAD	−8.37291	1.52613	1.31028	0.24277	−0.07908	0.47906

* For the full sample, standard errors for LS use (4-18). Standard errors for LAD are based on 100 bootstrap replications. For the $R = 10$, 50, and 100 experiments, standard errors are the sample standard deviations of the 100 sets of results from the runs of the experiments.

[3] The sample size R is not a negligible fraction of the population size, 430 for each replication. However, this does not call for a finite population correction of the variances in Table 4.3. We are not computing the variance of a sample of R observations drawn from a population of 430 paintings. We are computing the variance of a sample of R statistics, each computed from a different subsample of the full population. There about 10^{20} different samples of 10 observations we can draw. The number of different samples of 50 or 100 is essentially infinite.

If the assumptions underlying the regression model are correct, we should observe the following:

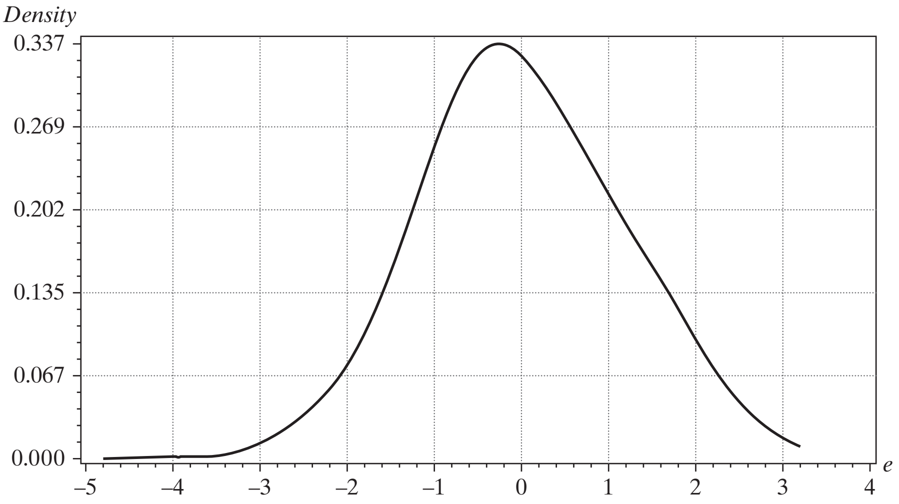
1. Because both estimators are consistent, the averages should resemble the full sample results, the more so as *R* increases.
2. As *R* increases, the sampling variance of the estimators should decline.
3. We should observe generally that the standard deviations of the LAD estimates are larger than the corresponding values for the LS estimator.
4. When *R* is small, the LAD estimator should compare more favorably to the LS estimator, but as *R* gets larger, the advantage of the LS estimator should become apparent.

A kernel density estimate for the distribution of the least squares residuals appears in Figure 4.3. There is a bit of skewness in the distribution, so a main assumption underlying our experiment may be violated to some degree. Results of the experiments are shown in Table 4.3. The force of the asymptotic results can be seen most clearly in the column for the coefficient on log Area. The decline of the standard deviation as *R* increases is evidence of the consistency of both estimators. In each pair of results (LS, LAD), we can also see that the estimated standard deviation of the LAD estimator is greater by a factor of about 1.2 to 1.4, which is also to be expected. Based on the normal distribution, we would have expected this ratio to be $\sqrt{\pi/2} = 1.253$.

4.4.5 LINEAR PROJECTIONS

Assumptions A1–A6 define the conditional mean function (CMF) in the joint distribution of $(y_i, \mathbf{x}_i)$, $E[y|\mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}$, and the conditional distribution of $y|\mathbf{x}$ (normal). Based on Assumptions A1–A6, we find that least squares is a consistent estimator of the slopes of the linear conditional mean under quite general conditions. A useful question for modeling is "What is estimated by linear least squares if the conditional mean function is not linear?" To consider this, we begin with a more

FIGURE 4.3 Kernel Density Estimator for Least Squares Residuals.



general statement of the structural model—this is sometimes labeled the "error form" of the model—in which

$$y = E[y|\mathbf{x}] + \varepsilon = \mu(\mathbf{x}) + \varepsilon.$$

We have shown earlier using the law of iterated expectations that $E[\varepsilon|\mathbf{x}] = E[\varepsilon] = 0$ regardless of whether $\mu(\mathbf{x})$ is linear or not. As a side result to modeling a conditional mean function without the linearity assumption, the modeler might use the results of linear least squares as an easily estimable, interesting feature of the population.

To examine the idea, we retain only the assumption of well-behaved data on **x**, A2, and A5, and assume, as well, that $(y_i,\mathbf{x}_i)$, $i = 1, \ldots, n$ are a random sample from the joint population of $(y,\mathbf{x})$. We leave the marginal distribution of **x** and the conditional distribution of $y|\mathbf{x}$ both unspecified, but assume that all variables in $(y_i, \mathbf{x}_i)$ have finite means, variances, and covariances. The **linear projection** of y on **x,** $Proj[y|\mathbf{x}]$, is defined by

$$y = \gamma_0 + \mathbf{x}'\boldsymbol{\gamma} + w = Proj[y|\mathbf{x}] + \text{w},$$

where $$\gamma_0 = E[y] - E[\mathbf{x}]'\boldsymbol{\gamma}$$

and $$\boldsymbol{\gamma} = (\text{Var}[\mathbf{x}])^{-1}\text{Cov}[\mathbf{x},y]. \quad \textbf{(4-34)}$$

As noted earlier, if $E[w|\mathbf{x}] = 0$, then this would define the CMF, but we have not assumed that. It does follow by inserting the expression for γ_0 in $E[y] = \gamma_0 + E[\mathbf{x}]'\boldsymbol{\gamma} + E[w]$ that $E[w] = 0$, and by expanding Cov[**x**,y] that $\text{Cov}[\mathbf{x},w] = \mathbf{0}$. The linear projection is a characteristic of the joint distribution of $(y_i,\mathbf{x}_i)$. As we have seen, if the CMF in the joint distribution is linear, then the projection will be the conditional mean. But, in the more general case, the linear projection will simply be a feature of the joint distribution. Some aspects of the linear projection function follow from the specification of the model:

1. Because the linear projection is generally not a structural model—that would usually be the CMF—the coefficients in the linear projection will generally not have a *causal* interpretation; indeed, the elements of $\boldsymbol{\gamma}$ will usually not have any direct economic interpretation other than as approximations (of uncertain quality) to the slopes of the CMF.
2. As we saw in Section 4.2.1, linear least squares regression of **y** on **X** (under the assumed sampling conditions) always estimates the γ_0 and $\boldsymbol{\gamma}$ of the projection regardless of the form of the conditional mean.
3. The CMF is the **minimum mean squared error** predictor of y in the joint distribution of $(y,\mathbf{x})$. We showed in Section 4.2.2 that the linear projection would be the minimum mean squared error *linear* predictor of y. Because both functions are predicting the same thing, it is tempting to infer that the linear projection is a linear approximation to the conditional mean function—and the approximation is exact if the conditional mean is linear. This approximation aspect of the projection function is a common motivation for its use. How effective it is likely to be is obviously dependent on the CMF—a linear function is only going to be able to approximate a nonlinear function locally, and how accurate that is will depend generally on how much curvature there is in the CMF. No generality seems possible; this would be application specific.
4. The interesting features in a structural model are often the partial effects or derivatives of the CMF—in the context of a structural model these are generally the objects of a search for causal effects. A widely observed empirical regularity that

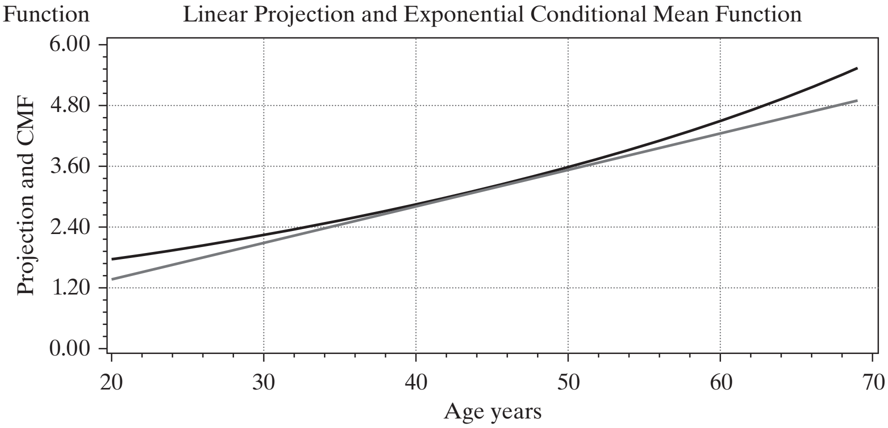
remains to be established with a firm theory is that $\boldsymbol{\gamma}$ in the linear projection often produces a good approximation to the average partial effects based on the CMF.

Example 4.4 Linear Projection: A Sampling Experiment

Table F7.1 describes panel data on 7,293 German households observed from 1 to 7 times for a total of 27,326 household-year observations. Looking ahead to Section 18.4, we examine a model with a nonlinear conditional mean function, a Poisson regression for the number of doctor visits by the household head, conditioned on the age of the survey respondent. We carried out the following experiment: Using all 27,326 observations, we fit a pooled Poisson regression by maximum likelihood in which the conditional mean function is $\lambda_i = \exp(\beta_0 + \beta_1 Age_i)$. The estimated values of (β_0,β_1) are [0.11384,0.02332]. We take this to be the population; $f(y_i | x_i) = \text{Poisson}(\lambda_i)$. We then used the observed data on age to (1) compute this true λ_i for each of the 27,326 observations. (2) We used a random number generator to draw 27,326 observations on y_i from the Poisson population with mean equal to this constructed λ_i. Note that the generated data conform exactly to the model with nonlinear conditional mean. The true value of the average partial effect is computed from $\partial E[y_i | \mathbf{x}_i]/\partial x_i = \beta_1\lambda_i$. We computed this for the full sample. The true APE is $(1/27{,}326)\Sigma_i\beta_1\lambda_i = 0.07384$. For the last step, we randomly sampled 1,000 observations from the population and fit the Poisson regression. The estimated coefficient was $b_1 = 0.02334$. The estimated average partial effect based on the MLEs is 0.07141. Finally, we linearly regressed the random draws y_i on Age_i using the 1,000 values. The estimated slope is 0.07163—nearly identical to the estimated average partial effect from the CMF. The estimated CMF and the linear projection are shown in Figure 4.4. The closest correspondence of the two functions occurs in the center of the data—the average age is 43 years. Several runs of the experiment (samples of 1,000 observations) produced the same result (not surprisingly).

As noted earlier, no firm theoretical result links the CMF to the linear projection save for the case when they are equal. As suggested by Figure 4.4, how good an approximation it provides will depend on the curvature of the CMF, and is an empirical question. For the present example, the fit is excellent in the middle of the data. Likewise, it is not possible to tie the slopes of the CMF at any particular point to the coefficients of the linear projection. The widely observed empirical regularity is that the linear projection can deliver good approximations to average partial effects in models with nonlinear CMFs. This is the underlying motivation

FIGURE 4.4 Nonlinear Conditional Mean Function and Linear Projection.



for recent applications of "linear probability models"—that is, for using linear least squares to fit a familiar nonlinear model. See Angrist and Pischke (2010) and Section 17.3 for further examination.

4.5 ROBUST ESTIMATION AND INFERENCE

Table 4.1 lists six assumptions that define the "Classical Linear Regression Model." A1–A3 define the linear regression framework. A5 suggests a degree of flexibility—the model is broad enough to encompass a wide variety of data generating processes. Assumptions A4 and A6, however, specifically narrow the situations in which the model applies. In particular, A4 seems to preclude the approach developed so far if the disturbances are heteroscedastic or autocorrelated, while A6 limits the stochastic specification to normally distributed disturbances. In fact, we have established all of the finite sample properties save for normality of $\mathbf{b}|\mathbf{X}$, and all of the asymptotic properties without actually using Assumption A6. As such, by these results, the least squares estimator is "robust" to violations of the normality assumption. In particular, it appears to be possible to establish the properties we need for least squares without any specific assumption about the distribution of ε (again, so long as the other assumptions are met).

An estimator of a model is said to be "robust" if it is insensitive to departures from the base assumptions of the model. In practical econometric terms, **robust estimators** retain their desirable properties in spite of violations of some of the assumptions of the model that motivate the estimator. We have seen, for example, that the unbiased least squares estimator is robust to a departure from the normality assumption, A6. In fact, the unbiasedness of least squares is also robust to violations of assumption A4. But, as regards unbiasedness, it is certainly not robust to violations of A3. Also, whether consistency for least squares can be established without A4 remains to be seen. Robustness is usually defined with respect to specific violations of the model assumptions. Estimators are not globally "robust." Robustness is not necessarily a precisely defined feature of an estimator, however. For example, the LAD estimator examined in Example 4.4 is often viewed as a more robust estimator than least squares, at least in small samples, because of its numerical insensitivity to the presence of outlying observations in the data.

For our practical purposes, we will take robustness to be a broad characterization of the asymptotic properties of certain estimators and procedures. We will specifically focus on and distinguish between **robust estimation** and **robust inference**. A robust estimator, in most settings, will be a consistent estimator that remains consistent in spite of violations of assumptions used to motivate it. To continue the example, with some fairly inocuous assumptions about the alternative specification, the least squares estimator will be robust to violations of the homoscedasticity assumption $\text{Var}[\varepsilon_i|\mathbf{x}_i] = \sigma^2$. In most applications, inference procedures are robust when they are based on estimators of asymptotic variances that are appropriate even when assumptions are violated.

Applications of econometrics rely heavily on robust estimation and inference. The development of robust methods has greatly simplified the development of models, as we shall see, by obviating assumptions that would otherwise limit their generality. We will develop a variety of robust estimators and procedures as we proceed.

4.5.1 CONSISTENCY OF THE LEAST SQUARES ESTIMATOR

In the context of A1–A6, we established consistency of $\mathbf{b}$ by invoking two results. Assumption A2 is an assumption about existence. Without A2, discussion of consistency is moot, because if $\mathbf{X}'\mathbf{X}/n$ does not have full rank, $\mathbf{b}$ does not exist. We also relied on A4. The central result is plim $\mathbf{X}'\boldsymbol{\varepsilon}/n = \mathbf{0}$, which we could establish if $E[\mathbf{x}_i\varepsilon_i] = \mathbf{0}$. The remaining element would be a law of large numbers by which the sample mean would converge to its population counterpart. Collecting terms, it turns out that normality, homoscedasticity and nonautocorrelation are not needed for consistency of $\mathbf{b}$, so, in turn, consistency of the least squares estimator is robust to violations of these three assumptions. Broadly, random sampling is sufficient.

4.5.2 A HETEROSCEDASTICITY ROBUST COVARIANCE MATRIX FOR LEAST SQUARES

The derivations in Sections 4.4.2 of Asy.Var$[\mathbf{b}] = (\sigma^2/\text{n})\mathbf{Q}^{-1}$ relied specifically on Assumption A4. In the analysis of a cross section, in which observations are uncorrelated, the issue will be the implications of violations of the homoscedasticity assumption. (We will consider the heteroscedasticity case here. Autocorrelation in time-series data is examined in Section 20.5.2.) For the most general case, suppose Var$[\varepsilon_i | \mathbf{x}_i] = \sigma_i^2$, with variation assumed to be over $\mathbf{x}_i$. In this case,

$$\mathbf{b} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\sum_i \mathbf{x}_i\varepsilon_i.$$

Then,

$$\text{Var}[\mathbf{b}|\mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1}\left[\sum_i \sigma_i^2\mathbf{x}_i\mathbf{x}_i'\right](\mathbf{X}'\mathbf{X})^{-1}. \tag{4-35}$$

Based on this finite sample result, the asymptotic variance will be

$$\text{Asy.Var}[\mathbf{b}] = \frac{1}{n}\mathbf{Q}^{-1}\left[\text{plim}\frac{1}{n}\sum_i \sigma_i^2\mathbf{x}_i\mathbf{x}_i'\right]\mathbf{Q}^{-1} = \frac{1}{n}\mathbf{Q}^{-1}\mathbf{Q}^*\mathbf{Q}^{-1}. \tag{4-36}$$

Two points to consider are (1) is $s^2(\mathbf{X}'\mathbf{X})^{-1}$ likely to be a valid estimator of Asy.Var[$\mathbf{b}$] in this case? and, if not, (2) is there a strategy available that is "robust" to unspecified heteroscedasticity? The first point is pursued in detail in Section 9.3. The answer to the second is yes. What is required is a feasible estimator of $\mathbf{Q}^*$. White's (1980) heteroscedasticity robust estimator of $\mathbf{Q}^*$ is

$$\mathbf{W}_{het} = \frac{1}{n}\sum_i e_i^2\mathbf{x}_i\mathbf{x}_i',$$

where e_i is the least squares residual, $y_i - \mathbf{x}_i{}'\mathbf{b}$. With $\mathbf{W}_{het}$ in hand, an estimator of Asy.Var[$\mathbf{b}$] that is robust to unspecified heteroscedasticity is

$$\text{Est.Asy.Var}[\mathbf{b}] = n(\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{W}_{het}(\mathbf{X}'\mathbf{X})^{-1}. \tag{4-37}$$

The implication to this point will be that we can discard the homoscedasticity assumption in A4 and recover appropriate standard errors by using (4-37) to estimate the asymptotic standard errors for the coefficients.

4.5.3 ROBUSTNESS TO CLUSTERING

Settings in which the sample data consist of groups of related observations are increasingly common. Panel data applications such as that in Example 4.5 and in Chapter 11 are an obvious case. Samples of firms grouped by industries, students in schools, home prices in neighborhoods, and so on are other examples. In this application, we suppose that the sample consists of C groups, or "clusters" of observations, labeled $c = 1,\ldots,C$. There are N_c observations in cluster c where N_c is one or more. The n observations in the entire sample therefore comprise $n = \sum_c N_c$ observations. The regression model is

$$y_{i,c} = \mathbf{x}_{i,c}'\boldsymbol{\beta} + \varepsilon_{i,c}.$$

The observations within a cluster are grouped by the correlation across observations within the group. Consider, for example, student test scores where students are grouped by their class. The common teacher will induce a cross-student correlation of $\varepsilon_{i,c}$. An intuitively appealing formulation of such teacher effects would be the "random effects" formulation,

$$y_{i,c} = \mathbf{x}_{i,c}'\boldsymbol{\beta} + w_c + u_{i,c}. \qquad \textbf{(4-38)}$$

By this formulation, the common within cluster effect (e.g., the common teacher effect) would induce the same correlation across all members of the group. This random effects specification is considered in detail in Chapter 11. For present purposes, the assumption is stronger than necessary—note that in (4-38), assuming $u_{i,c}$ is independent across observations, $\text{Cov}(\varepsilon_{i,c},\varepsilon_{j,c}) = \sigma_w^2$. At this point, we prefer to allow the correlation to be unspecified, and possibly vary for different pairs of observations.

The least squares estimator is

$$\mathbf{b} = \boldsymbol{\beta} + \left(\sum_{c=1}^{C}\mathbf{X}_c'\mathbf{X}_c\right)^{-1}\left[\sum_{c=1}^{C}\left(\sum_{i=1}^{N}\mathbf{x}_{i,c}\varepsilon_{i,c}\right)\right] = \boldsymbol{\beta} + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\left[\sum_{c=1}^{C}\left(\mathbf{X}_c'\boldsymbol{\varepsilon}_c\right)\right],$$

where $\mathbf{X}_c$ is the $N_c \times K$ matrix of exogenous variables for cluster c and $\boldsymbol{\varepsilon}_c$ is the N_c disturbances for the group. Assuming that the clusters are independent,

$$\text{Var}[\mathbf{b}\,|\,\mathbf{X}] = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\left[\sum_{c=1}^{C}\mathbf{X}_c\boldsymbol{\Omega}_c\mathbf{X}_c'\right]\left(\mathbf{X}'\mathbf{X}\right)^{-1}. \qquad \textbf{(4-39)}$$

Like σ_i^2 before, $\boldsymbol{\Omega}_c$ is not meant to suggest a particular set of population parameters. Rather, $\boldsymbol{\Omega}_c$ represents the possibly unstructured correlations allowed among the N_c disturbances in cluster c. The construction is essentially the same as the White estimator, though $\boldsymbol{\Omega}_c$ is the matrix of variances and covariances for the full vector ε_c. (It would be identical to the White estimator if each cluster contained one observation.) Taking the same approach as before, we obtain the asymptotic variance

$$\text{Asy.Var}[\mathbf{b}] = \frac{1}{C}\mathbf{Q}^{-1}\left[\text{plim}\frac{1}{C}\sum_{c=1}^{C}\mathbf{X}_c\boldsymbol{\Omega}_c\mathbf{X}_c'\right]\mathbf{Q}^{-1}.^{4} \qquad \textbf{(4-40)}$$

[4] Since the observations in a cluster are not assumed to be independent, the number of observations in the sample is no longer n. Logically, the sample would now consist of C multivariate observations. In order to employ the asymptotic theory used to obtain Asy.Var[**b**], we are implicitly assuming that C is large while N_c is relatively small, and asymptotic results would relate to increasing C, not n. In practical applications, the number of clusters is often rather small, and the group sizes relatively large. We will revisit these complications in Section 11.3.3.

A feasible estimator of the bracketed matrix based on the least squares residuals is

$$\mathbf{W}_{cluster} = \frac{1}{C}\sum\nolimits_{c=1}^{C}\left(\mathbf{X}_c'\mathbf{e}_c\right)\left(\mathbf{e}_c'\mathbf{X}_c\right) = \frac{1}{C}\sum\nolimits_{c=1}^{C}\left(\sum\nolimits_{i=1}^{N_c}\mathbf{x}_{ic}e_{ic}\right)\left(\sum\nolimits_{i=1}^{N_c}\mathbf{x}_{ic}e_{ic}\right)'. \quad \textbf{(4-41)}$$

Then,

$$\text{Est.Asy.Var}[\mathbf{b}] = C(\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{W}_{cluster}(\mathbf{X}'\mathbf{X})^{-1}. \quad \textbf{(4-42)}$$

[A refinement intended to accommodate a possible downward bias induced by a small number of clusters is to multiply $\mathbf{W}_{cluster}$ by $C/(C - 1)$ (*SAS*) or by $[C/(C - 1)] \times [(n - 1)/(n - K)]$ (*Stata, NLOGIT*).]

Example 4.5 Robust Inference About the Art Market

The Monet paintings examined in Example 4.3 were sold at auction over 1989–2006. Our model thus far is

$$\ln Price_{it} = \beta_1 + \beta_2 \ln Area_{it} + \beta_3 AspectRatio_{it} + \varepsilon_{it}$$

The subscript "*it*" uniquely identifies the painting and when it was sold. Prices in open outcry auctions reflect (at least) three elements, the common (public), observable features of the item, the public unobserved (by the econometrician) elements of the asset, and the private unobservable preferences of the winning bidder. For example, it will turn out (in a later example) that whether the painting is signed or not has a large and significant influence on the price. For now, we assume (for sake of the example), that we do not observe whether the painting is signed or not, though, of course, the winning bidders do observe this. It does seem reasonable to suggest that the presence of a signature is uncorrelated with the two attributes we do observe, area and aspect ratio. We respecify the regression as

$$\ln Price_{it} = \beta_1 + \beta_2 \ln Area_{it} + \beta_3 AspectRatio_{it} + w_{it} + u_{it},$$

where w_{it} represents the intrinsic, unobserved features of the painting and u_{it} represents the unobserved preferences of the buyer. In fact, the sample of 430 sales involves 376 unique paintings. Several of the sales are repeat sales of the same painting. The numbers of sales per painting were one, 333; two, 34; three, 7; and four, 2. Figure 4.5 shows the configuration of the sample. For those paintings that sold more than once, the terms w_{it} do relate to the same *i,* and, moreover, would naturally be correlated. [They needn't be identical as in (4-38), however. The valuation of attributes of paintings or other assets sold at auction could vary over time.]

FIGURE 4.5 Repeat Sales of Monet Paintings.

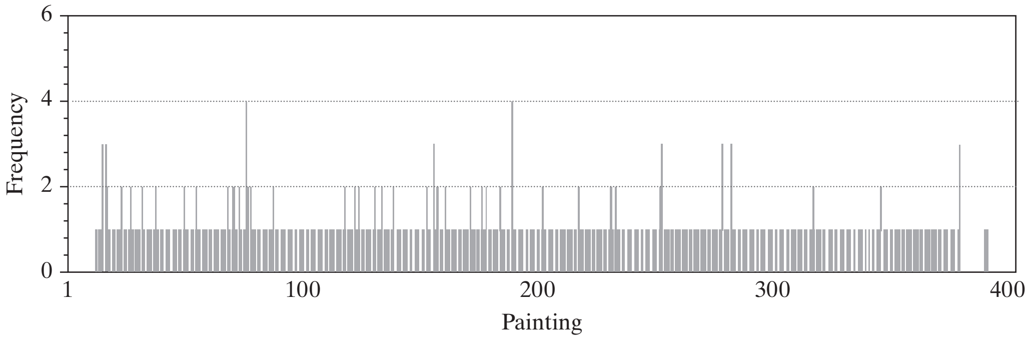


TABLE 4.4 Robust Standard Errors

	Estimated	*LS Standard*	*Heteroscedasticity*	*Cluster Robust*
Variable	*Coefficient*	*Error*	*Robust Std.Error*	*Std.Error*
Constant	−8.34237	0.67820	0.73342	0.75873
ln *Area*	1.31638	0.09205	0.10598	0.10932
Aspect Ratio	−0.09623	0.15784	0.16706	0.17776

The least squares estimates and three sets of estimated standard errors are shown in Table 4.4. Even with only a small amount of clustering, the correction produces a tangible adjustment of the standard errors. Perhaps surprisingly, accommodating possible heteroscedasticity produces a more pronounced effect than the cluster correction. Note, finally, in contrast to common expectations, the robust covariance matrix does not always have larger standard errors. The standard errors do increase slightly in this example, however.

4.5.4 BOOTSTRAPPED STANDARD ERRORS WITH CLUSTERED DATA

The sampling framework that underlies the treatment of clustering in the preceding section assumes that the sample consists of a reasonably large number of clusters, drawn randomly from a very large population of clusters. Within each cluster reside a number of observations generated by the linear regression model. Thus,

$$y_{i,c} = \mathbf{x}_{i,c}'\boldsymbol{\beta} + \varepsilon_{i,c},$$

where within each cluster, $E[\varepsilon_{i,c},\varepsilon_{j,c}]$ may be nonzero—observations may be freely correlated. Clusters are assumed to be independent. Each cluster consists of N_c observations, $(\mathbf{y}_c,\mathbf{X}_c,\boldsymbol{\varepsilon}_c)$ and the cluster is the unit of observation. For example, we might be examining student test scores in a state where students are grouped by classroom, and there are potentially thousands of classrooms in the state. The sample consists of a sample of classrooms. (Higher levels of grouping, such as classrooms in a school, and schools in districts, would require some extensions. We will consider this possibility later in Chapter 11.) The essential feature of the data is the likely correlation across observations in the group. Another natural candidate for this type of process would be a panel data set such as the labor market data examined in Example 4.6, where a sample of 595 individuals is each observed in 7 consecutive years. The common feature is the large number of relatively small or moderately sized clusters in the sample.

The method of estimating a robust asymptotic covariance matrix for the least squares estimator that was introduced in the preceding section involves a method of using the data and the least squares residuals to build a covariance matrix. **Bootstrapping** is another method that is likely to be effective under these assumed sampling conditions. (We emphasize, if the number of clusters is quite small and/or group sizes are very large relative to the number of clusters, then bootstrapping, like the previous method, is likely not to be effective.[5] Bootstrapping was introduced in Example 4.3 where we used the

[5] See, for example, Wooldridge (2010, Chapter 20).

method to estimate an asymptotic covariance matrix for the LAD estimator. The basic steps in the methodology are:

1. For R repetitions, draw a random sample of N_c observations from the full sample of N_c observations *with replacement*. Estimate the parameters of the regression model with each of the R constructed samples.
2. The estimator of the asymptotic covariance matrix is the sample variance of the R sets of estimated coefficients.

Keeping in mind that in the current case, the cluster is the unit of observation, we use a **block bootstrap**. In the example below, the block is the 7 observations for individual i, so each observation in the bootstrap replication is a block of 7 observations. Example 4.6 below illustrates the use of block bootstrap.

Example 4.6 Clustering and Block Bootstrapping

Cornwell and Rupert (1988) examined the returns to schooling in a panel data set of 595 heads of households observed in seven years, 1976–1982. The sample data (Appendix Table F8.1) are drawn from years 1976 to 1982 from the *Non-Survey of Economic Opportunit*y from the Panel Study of Income Dynamics. A slightly modified version of their regression model is

$$\begin{aligned}\ln Wage_{it} = {} & \beta_1 + \beta_2 Exp_{it} + \beta_3 Exp_{it}^2 + \beta_4 Wks_{it} + \beta_5 Occ_{it} + \beta_6 Ind_{it} + \beta_7 South_{it} \\ & + \beta_8 SMSA_{it} + \beta_9 MS_{it} + \beta_{10} Union_{it} + \beta_{11} Ed_i + \beta_{12} Fem_i + \beta_{13} Blk_i + \varepsilon_{it}.\end{aligned}$$

The variables in the model are as follows:

Exp	=	years of full time work experience,
Wks	=	weeks worked,
Occ	=	1 if blue-collar occupation, 0 if not,
Ind	=	1 if the individual works in a manufacturing industry, 0 if not,
South	=	1 if the individual resides in the south, 0 if not,
SMSA	=	1 if the individual resides in an SMSA, 0 if not,
MS	=	1 if the individual is married, 0 if not,
Union	=	1 if the individual wage is set by a union contract, 0 if not,
Ed	=	years of education as of 1976,
Fem	=	1 if the individual is female, 0 if not,
Blk	=	1 if the individual is black.

See Appendix Table F8.1 for the data source.

Table 4.5 presents the least squares and three sets of asymptotic standard errors. The first is the conventional results based on $s^2(\mathbf{X}'\mathbf{X})^{-1}$. Compared to the other estimates, it appears that the uncorrected standard errors substantially understate the variability of the least squares estimator. The clustered standard errors are computed using (4-42). The values are 50%–100% larger. The bootstrapped standard errors are quite similar to the robust estimates, as would be expected.

4.6 ASYMPTOTIC DISTRIBUTION OF A FUNCTION OF b: THE DELTA METHOD

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

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

TABLE 4.5 Clustered, Robust, and Bootstrapped Standard Errors

Variable	*Least Squares Estimate*	*Standard Error*	*Clustered Std.Error*	*Bootstrapped Std.Error*	*White Hetero. Robust Std.Error*
Constant	5.25112	0.07129	0.12355	0.11171	0.07435
Exp	0.00401	0.00216	0.00408	0.00434	0.00216
ExpSq	−0.00067	0.00005	0.00009	0.00010	0.00005
Wks	0.00422	0.00108	0.00154	0.00164	0.00114
Occ	−0.14001	0.01466	0.02724	0.02555	0.01494
Ind	0.04679	0.01179	0.02366	0.02153	0.01199
South	−0.05564	0.01253	0.02616	0.02414	0.01274
SMSA	0.15167	0.01207	0.02410	0.02323	0.01208
MS	0.04845	0.02057	0.04094	0.03749	0.02049
Union	0.09263	0.01280	0.02367	0.02553	0.01233
Ed	0.05670	0.00261	0.00556	0.00483	0.00273
Fem	−0.36779	0.02510	0.04557	0.04460	0.02310
Blk	−0.16694	0.02204	0.04433	0.05221	0.02075

where $\mathbf{C}$ is the $J \times K$ matrix whose jth row is the vector of derivatives of the jth 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 a linear Taylor series approach, 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 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. The mean of the asymptotic distribution is 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 Asymptotic Distribution of a Function of b

If $\mathbf{f}(\mathbf{b})$ *is a set of continuous and continuously differentiable functions of* $\mathbf{b}$ *such that* $\mathbf{f}(\text{plim } \mathbf{b})$ exists and $\boldsymbol{\Gamma} = \partial \mathbf{f}(\boldsymbol{\beta})/\partial \boldsymbol{\beta}'$ *and if Theorem 4.4 holds, then*

$$\mathbf{f}(\mathbf{b}) \overset{a}{\sim} N\left[\mathbf{f}(\boldsymbol{\beta}), \boldsymbol{\Gamma}\left(\text{Asy.Var}[\mathbf{b}]\right)\boldsymbol{\Gamma}'\right]. \qquad \textbf{(4-43)}$$

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

$$\text{Est.Asy.Var}[\mathbf{f}(\mathbf{b})] = \mathbf{C}\{\text{Est. Asy.Var}[\mathbf{b}]\}\mathbf{C}'.$$

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, $\mathbf{f}(\mathbf{b})$ is a consistent estimator of $\mathbf{f}(\boldsymbol{\beta})$, and the asymptotic covariance matrix is readily available.

Example 4.7 Nonlinear Functions of Parameters: The Delta Method

A dynamic version of the demand for gasoline model in Example 2.3 would be used to separate the short- and long-term impacts of changes in income and prices. The model would be

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

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 β_2 and β_3. The long-run elasticities are $\phi_2 = \beta_2/(1 - \gamma)$ and $\phi_3 = \beta_3/(1 - \gamma)$, respectively. 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 4.6. (Because these are aggregate time-series data, we have not computed a robust covariance matrix.) The estimated short-run elasticities are the estimates given in the table. The two estimated long-run elasticities are $f_2 = b_2/(1 - c) = -0.069532/(1 - 0.830971) = -0.411358$ and $f_3 = 0.164047/(1 - 0.830971) = 0.970522$. To compute the estimates of the standard errors, we need the estimated partial derivatives of these functions with respect to the six parameters in the model:

$$\begin{aligned}\hat{\boldsymbol{\Gamma}}_2' &= \partial\phi_2(\hat{\boldsymbol{\beta}})/\partial\hat{\boldsymbol{\beta}}' = [0, 1/(1 - \hat{\gamma}), 0, 0, 0, \hat{\boldsymbol{\beta}}_2/(1 - \hat{\gamma})^2] = [0, 5.91613, 0, 0, 0, -2.43365], \\ \hat{\boldsymbol{\Gamma}}_3' &= \partial\phi_3(\hat{\boldsymbol{\beta}})/\partial\hat{\boldsymbol{\beta}}' = [0, 0, 1/(1 - \hat{\gamma}), 0, 0, \hat{\boldsymbol{\beta}}_3/(1 - \hat{\gamma})^2] = [0, 0, 5.91613, 0, 0, 5.74174].\end{aligned}$$

Using (4-43), we can now compute the estimates of the asymptotic variances for the two estimated long-run elasticities by computing $\mathbf{g}_2'[s^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{g}_2$ and $\mathbf{g}_3'[s^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{g}_3$. The results are 0.023194 and 0.0263692, respectively. The two asymptotic standard errors are the square roots, 0.152296 and 0.162386.

TABLE 4.6 Regression Results for a Demand Equation

Sum of squared residuals:	0.0127352
Standard error of the regression:	0.0168227
R^2 based on 51 observations	0.9951081

Variable	*Coefficient*	*Standard Error*	*t Ratio*
Constant	−3.123195	0.99583	−3.136
ln P_G	−0.069532	0.01473	−4.720
ln *Income / Pop*	0.164047	0.05503	2.981
ln P_{nc}	−0.178395	0.05517	−3.233
ln P_{uc}	0.127009	0.03577	3.551
last period ln *G / Pop*	0.830971	0.04576	18.158

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

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

4.7 INTERVAL ESTIMATION

The objective of interval estimation is to present the best estimate of a parameter with an explicit expression of the uncertainty attached to that estimate. A general approach for estimation of a parameter θ would be

$$\hat{\theta} \pm \text{ sampling variability.} \tag{4-44}$$

(We are assuming that the interval of interest would be symmetric around $\hat{\theta}$.) Following the logic that the range of the sampling variability should convey the degree of (un) certainty, we consider the logical extremes. We can be absolutely (100%) certain that the true value of the parameter we are estimating lies in the range $\hat{\theta} \pm \infty$. Of course, this is not particularly informative. At the other extreme, we should place no certainty (0.0%) on the range $\hat{\theta} \pm 0$. The probability that our estimate precisely hits the true parameter value should be considered zero. The point is to choose a value of α—0.05 or 0.01 is conventional—such that we can attach the desired confidence (probability), $100(1 - \alpha)\%$, to the interval in (4-44). We consider how to find that range and then apply the procedure to three familiar problems, calculating an interval for one of the regression parameters, estimating a function of the parameters, and predicting the value of the dependent variable in the regression using a specific setting of the independent variables. For this latter purpose, we will rely on the asymptotic normality of the estimator.

4.7.1 FORMING A CONFIDENCE INTERVAL FOR A COEFFICIENT

If the disturbances are normally distributed, then for any particular element of **b**,

$$b_k \sim N[\beta_k, \sigma^2 S^{kk}],$$

where S^{kk} denotes the kth diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$. By standardizing the variable, we find

$$z_k = \frac{b_k - \beta_k}{\sqrt{\sigma^2 S^{kk}}} \tag{4-45}$$

has a standard normal distribution. Note that z_k, which is a function of b_k, β_k, σ^2, and S^{kk}, nonetheless has a distribution that involves none of the model parameters or the data. Using the conventional 95% confidence level, we know that $\text{Prob}[-1.96 \le z_k \le 1.96] = 0.95$. By a simple manipulation, we find that

$$\text{Prob}[b_k - 1.96\sqrt{\sigma^2 S^{kk}} \le \beta_k \le b_k + 1.96\sqrt{\sigma^2 S^{kk}}] = 0.95. \qquad \textbf{(4-46)}$$

This states the probability that the random interval, $[b_k \pm$ the sampling variability], contains β_k, not the probability that β_k lies in the specified interval. If we wish to use some other level of confidence, not 95%, then the 1.96 in (4-46) is replaced by the appropriate $z_{(1-\alpha/2)}$. (We are using the notation $z_{(1-\alpha/2)}$ to denote the value of z such that for the standard normal variable z, $\text{Prob}[z \le z_{(1-\alpha/2)}] = 1 - \alpha/2$. Thus, $z_{0.975} = 1.96$, which corresponds to $\alpha = 0.05$.)

We would have the desired confidence interval in (4-46), save for the complication that σ^2 is not known, so the interval is not operational. Using s^2 from the regression instead, the ratio

$$t_k = \frac{b_k - \beta_k}{\sqrt{s^2 S^{kk}}} \qquad \textbf{(4-47)}$$

has a t distribution with $(n - K)$ degrees of freedom.[6] We can use t_k to test hypotheses or form confidence intervals about the individual elements of β. A confidence interval for β_k would be formed using

$$\text{Prob}\left[b_k - t_{(1-\alpha/2),[n-K]}\sqrt{s^2 S^{kk}} \le \beta_k \le b_k + t_{(1-\alpha/2),[n-K]}\sqrt{s^2 S^{kk}}\right] = 1 - \alpha, \qquad \textbf{(4-48)}$$

where $t_{(1-\alpha/2),[n-K]}$ is the appropriate critical value from the t distribution. The distribution of the pivotal statistic depends on the sample size through $(n - K)$, but, once again, not on the parameters or the data.

If the disturbances are not normally distributed, then the theory for the t distribution in (4-48) does not apply. But, the large sample results in Section 4.4 provide an alternative approach. Based on the development that we used to obtain Theorem 4.3 and (4-33), the limiting distribution of the statistic

$$z_k = \frac{\sqrt{n}(b_k - \beta_k)}{\sqrt{\sigma^2 Q^{kk}}}$$

is standard normal, where $\mathbf{Q} = [\text{plim}(\mathbf{X}'\mathbf{X}/n)]^{-1}$ and Q^{kk} is the kth diagonal element of $\mathbf{Q}$. Based on the Slutsky theorem (D.16), we may replace σ^2 with a consistent estimator, s^2, and obtain a statistic with the same limiting distribution. We estimate $\mathbf{Q}$ with $(\mathbf{X}'\mathbf{X}/n)^{-1}$. This gives us precisely (4-47), which states that under the assumptions in Section 4.4, the "t" statistic in (4-47) converges to standard normal even if the disturbances are not normally distributed. The implication would be that to employ the asymptotic distribution of $\mathbf{b}$, we should use (4-48) to compute the confidence interval but use the critical values from the standard normal table (e.g., 1.96) rather than from the t distribution. In practical terms, if the degrees of freedom in (4-48) are moderately large, say greater than 100, then the t distribution will be indistinguishable from the standard normal, and this large sample result would apply in any event. For smaller sample sizes, however, in the interest of conservatism, one might be advised to use the critical

[6] See (B-36) in Section B.4.2. It is the ratio of a standard normal variable to the square root of a chi-squared variable divided by its degrees of freedom.

values from the *t* table rather the standard normal, even in the absence of the normality assumption. In the application in Example 4.8, based on a sample of 52 observations, we form a confidence interval for the income elasticity of demand using the critical value of 2.012 from the *t* table with 47 degrees of freedom. If we chose to base the interval on the asymptotic normal distribution, rather than the standard normal, we would use the 95% critical value of 1.96. One might think this is a bit optimistic, however, and retain the value 2.012, again, in the interest of conservatism.

The preceding analysis starts from Assumption A6, normally distributed disturbance, then shows how the procedure is adjusted to rely on the asymptotic properties of the estimator rather than the narrow possibly unwarranted assumption of normally distributed disturbances. It continues to rely on the homoscedasticity assumption in A4. (For the present, we are assuming away possible autocorrelation.) Section 4.5 showed how the estimator of the asymptotic covariance matrix can be refined to allow for unspecified heteroscedasticity or cluster effects. The final adjustment of the confidence intervals would be to replace (4-48) with

$$\begin{aligned}\text{Prob}[b_k - z_{(1-\alpha/2)}\sqrt{Est.Asy.Var[b_k]} \leq \beta_k \leq b_k \\ + z_{(1-\alpha/2)}\sqrt{Est.Asy.Var[b_k]}] = 1-\alpha,\end{aligned} \qquad \textbf{(4-49)}$$

Example 4.8 Confidence Interval for the Income Elasticity of Demand for Gasoline

Using the gasoline market data discussed in Examples 4.2 and 4.4, we estimated the following demand equation using the 52 observations:

$$\ln(G/Pop) = \beta_1 + \beta_2 \ln P_G + \beta_3 \ln(Income/Pop) + \beta_4 \ln P_{nc} + \beta_5 \ln P_{uc} + \varepsilon.$$

Least squares estimates of the model parameters with standard errors and *t* ratios are given in Table 4.7. To form a confidence interval for the income elasticity, we need the critical value from the *t* distribution with $n - K = 52 - 5 = 47$ degrees of freedom. The 95% critical value is 2.012. Therefore a 95% confidence interval for β_3 is 1.095874 ± 2.012 (0.07771) = [0.9395,1.2522].

4.7.2 CONFIDENCE INTERVAL FOR A LINEAR COMBINATION OF COEFFICIENTS: THE OAXACA DECOMPOSITION

In Example 4.8, we showed how to form a confidence interval for one of the elements of $\boldsymbol{\beta}$. By extending those results, we can show how to form a confidence interval for a

TABLE 4.7 Regression Results for a Demand Equation

Sum of squared residuals:	0.120871		
Standard error of the regression:	0.050712		
R^2 based on 52 observations	0.958443		

Variable	*Coefficient*	*Standard Error*	*t Ratio*
Constant	−21.21109	0.75322	−28.160
ln P_G	−0.02121	0.04377	−0.485
ln *Income/Pop*	1.09587	0.07771	14.102
ln P_{nc}	−0.37361	0.15707	−2.379
ln P_{uc}	0.02003	0.10330	0.194

linear function of the parameters. **Oaxaca's (1973) and Blinder's (1973) decomposition** provides a frequently used application.[7]

Let $\mathbf{w}$ denote a $K \times 1$ vector of known constants. Then, the linear combination $c = \mathbf{w}'\mathbf{b}$ is asymptotically normally distributed with mean $\gamma = \mathbf{w}'\boldsymbol{\beta}$ and variance $\sigma_c^2 = \mathbf{w}'[Asy.Var[\mathbf{b}]]\mathbf{w}$, which we estimate with $s_c^2 = \mathbf{w}'[Est.Asy.Var[\mathbf{b}]]\mathbf{w}$. With these in hand, we can use the earlier results to form a confidence interval for γ :

$$\text{Prob}[c - z_{(1-\alpha/2)}s_c \leq \gamma \leq c + z_{(1-\alpha/2)}s_c] = 1 - \alpha. \qquad \textbf{(4-50)}$$

This general result can be used, for example, for the sum of the coefficients or for a difference.

Consider, then, Oaxaca's (1973) application. In a study of labor supply, separate wage regressions are fit for samples of n_m men and n_f women. The underlying regression models are

$$\ln wage_{m,i} = \mathbf{x}'_{m,i}\boldsymbol{\beta}_m + \varepsilon_{m,i}, \quad i = 1, \ldots, n_m$$

and

$$\ln wage_{f,j} = \mathbf{x}'_{f,j}\,\boldsymbol{\beta}_f + \varepsilon_{f,j}, \quad j = 1, \ldots, n_f.$$

The regressor vectors include sociodemographic variables, such as age, and human capital variables, such as education and experience. We are interested in comparing these two regressions, particularly to see if they suggest wage discrimination. Oaxaca suggested a comparison of the regression functions. For any two vectors of characteristics,

$$\begin{aligned} E[\ln \text{wage}_{m,i}|\mathbf{x}_{m,i}] - E[\ln \text{wage}_{f,j}|\mathbf{x}_{f,i}] &= \mathbf{x}'_{m,i}\boldsymbol{\beta}_m - \mathbf{x}'_{f,j}\boldsymbol{\beta}_f \\ &= \mathbf{x}'_{m,i}\boldsymbol{\beta}_m - \mathbf{x}'_{m,i}\boldsymbol{\beta}_f + \mathbf{x}'_{m,i}\boldsymbol{\beta}_f - \mathbf{x}'_{f,j}\boldsymbol{\beta}_f \\ &= \mathbf{x}'_{m,i}(\boldsymbol{\beta}_m - \boldsymbol{\beta}_f) + (\mathbf{x}_{m,i} - \mathbf{x}_{f,i})'\boldsymbol{\beta}_f. \end{aligned}$$

The second term in this decomposition is identified with differences in human capital that would explain wage differences naturally, assuming that labor markets respond to these differences in ways that we would expect. The first term shows the differential in log wages that is attributable to differences unexplainable by human capital; holding these factors constant at $\mathbf{x}_m$ makes the first term attributable to other factors. Oaxaca suggested that this decomposition be computed at the means of the two regressor vectors, $\overline{\mathbf{x}}_m$ and $\overline{\mathbf{x}}_f$, and the least squares coefficient vectors, $\mathbf{b}_m$ and $\mathbf{b}_f$. If the regressions contain constant terms, then this process will be equivalent to analyzing $\overline{\ln y_m} - \overline{\ln y_f}$.

We are interested in forming a confidence interval for the first term, which will require two applications of our result. We will treat the two vectors of sample means as known vectors. Assuming that we have two independent sets of observations, our two estimators, $\mathbf{b}_m$ and $\mathbf{b}_f$, are independent with means $\boldsymbol{\beta}_m$ and $\boldsymbol{\beta}_f$ and estimated asymptotic covariance matrices $Est.Asy.Var[\mathbf{b}_m]$ and $Est.Asy.Var[\mathbf{b}_f]$. The covariance matrix of the difference is the sum of these two matrices. We are forming a confidence interval for $\overline{\mathbf{x}}'_m\,\mathbf{d}$ where $\mathbf{d} = \mathbf{b}_m - \mathbf{b}_f$. The estimated covariance matrix is

$$Est.Asy.Var[\mathbf{d}] = Est.Asy.Var[\mathbf{b}_m] + Est.Asy.Var[\mathbf{b}_f]. \qquad \textbf{(4-51)}$$

Now we can apply the result above. We can also form a confidence interval for the second term; just define $\mathbf{w} = \overline{\mathbf{x}}_m - \overline{\mathbf{x}}_f$ and apply the earlier result to $\mathbf{w}'\mathbf{b}_f$.

[7] See Bourgignon et al. (2002) for an extensive application.

Example 4.9 Oaxaca Decomposition of Home Sale Prices

The town of Shaker Heights, Ohio, a suburb of Cleveland, developed in the twentieth century as a patchwork of neighborhoods associated with neighborhood-based school districts. Responding to changes in the demographic composition of the city, in 1987, Shaker Heights redistricted the neighborhoods. Some houses in some neighborhoods remained in the same school district while others in the same neighborhood were removed to other school districts. Bogart and Cromwell (2000) examined how this abrupt policy change affected home values in Shaker Heights by studying sale prices of houses before and after the change. Several econometric approaches were used.

- **Difference in Differences Regression:** Houses that did not change districts constituted a control group while those that did change constitute a treatment group. Sales take place both before and after the treatment date, 1987. A hedonic regression of home sale prices on attributes and the treatment and policy dummy variables reveals the causal effect of the policy change. (We will examine this method in Chapter 6.)
- **Repeat Sales:** Some homes were sold more than once. For those that sold both before and after the redistricting, a regression of the form

$$\text{lnPrice}_{i1} - \text{lnPrice}_{i0} = \text{time effects} + \text{school effects} + \Delta\text{redistricted}.$$

  The advantage of the first difference regression is that it effectively controls for and eliminates the characteristics of the house, and leaves only the persistent school effects and the effect of the policy change.
- **Oaxaca Decomposition:** Two hedonic regressions based on house characteristics are fit for different parts of neighborhoods where there are both houses that are in the neighborhood school areas and houses that are districted to other schools. The decomposition approach described above is applied to the two groups. The differences in the means of the sale prices are decomposed into a component that can be explained by differences in the house attributes and a residual effect that is suggested to be related to the benefit of having a neighborhood school. Figure 4.6 below shows the authors' main results for this part of the analysis.[8]

FIGURE 4.6 Results of Oaxaca Decomposition.

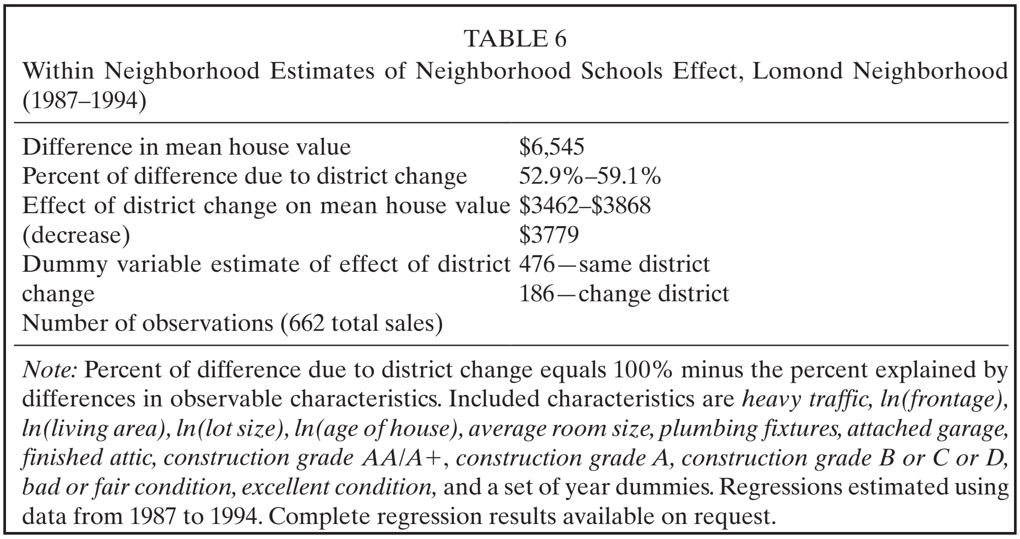


TABLE 6

Within Neighborhood Estimates of Neighborhood Schools Effect, Lomond Neighborhood (1987–1994)

Difference in mean house value	$6,545
Percent of difference due to district change	52.9%–59.1%
Effect of district change on mean house value	$3462–$3868
(decrease)	$3779
Dummy variable estimate of effect of district	476—same district
change	186—change district
Number of observations (662 total sales)	

Note: Percent of difference due to district change equals 100% minus the percent explained by differences in observable characteristics. Included characteristics are *heavy traffic, ln(frontage), ln(living area), ln(lot size), ln(age of house), average room size, plumbing fixtures, attached garage, finished attic, construction grade AA/A+, construction grade A, construction grade B or C or D, bad or fair condition, excellent condition,* and a set of year dummies. Regressions estimated using data from 1987 to 1994. Complete regression results available on request.

[8] Bogart and Cromwell (2000, p. 298).

4.8 PREDICTION AND FORECASTING

After the estimation of the model parameters, a common use of regression modeling is for prediction of the dependent variable. We make a distinction between *prediction* and *forecasting* most easily based on the difference between cross section and time-series modeling. **Prediction** (which would apply to either case) involves using the regression model to compute fitted (predicted) values of the dependent variable, either within the sample or for observations outside the sample. The same set of results will apply to cross sections, panels, and time series. We consider these methods first. **Forecasting**, while largely the same exercise, explicitly gives a role to time and often involves lagged dependent variables and disturbances that are correlated with their past values. This exercise usually involves predicting future outcomes. An important difference between predicting and forecasting (as defined here) is that for predicting, we are usually examining a scenario of our own design. Thus, in the example below in which we are predicting the prices of Monet paintings, we might be interested in predicting the price of a hypothetical painting of a certain size and aspect ratio, or one that actually exists in the sample. In the time-series context, we will often try to forecast an event such as real investment next year, not based on a hypothetical economy but based on our best estimate of what economic conditions will be next year. We will use the term **ex post prediction** (or **ex post forecast**) for the cases in which the data used in the regression equation to make the prediction are either observed or constructed experimentally by the analyst. This would be the first case considered here. An **ex ante forecast** (in the time-series context) will be one that requires the analyst to forecast the independent variables first before it is possible to forecast the dependent variable. In an exercise for this chapter, real investment is forecasted using a regression model that contains real GDP and the consumer price index. In order to forecast real investment, we must first forecast real GDP and the price index. Ex ante forecasting is considered briefly here and again in Chapter 20.

4.8.1 PREDICTION INTERVALS

Suppose that we wish to predict the value of y^0 associated with a regressor vector $\mathbf{x}^0$. The actual value would be

$$y^0 = \mathbf{x}^{0\prime}\boldsymbol{\beta} + \varepsilon^0.$$

It follows from the Gauss–Markov theorem that

$$\hat{y}^0 = \mathbf{x}^{0\prime}\mathbf{b} \tag{4-52}$$

is the minimum variance linear unbiased estimator of $E[y^0|\mathbf{x}^0] = \mathbf{x}^{0\prime}\boldsymbol{\beta}$. The **prediction error** is

$$e^0 = \hat{y}^0 - y^0 = (\mathbf{b} - \boldsymbol{\beta})'\mathbf{x}^0 - \varepsilon^0.$$

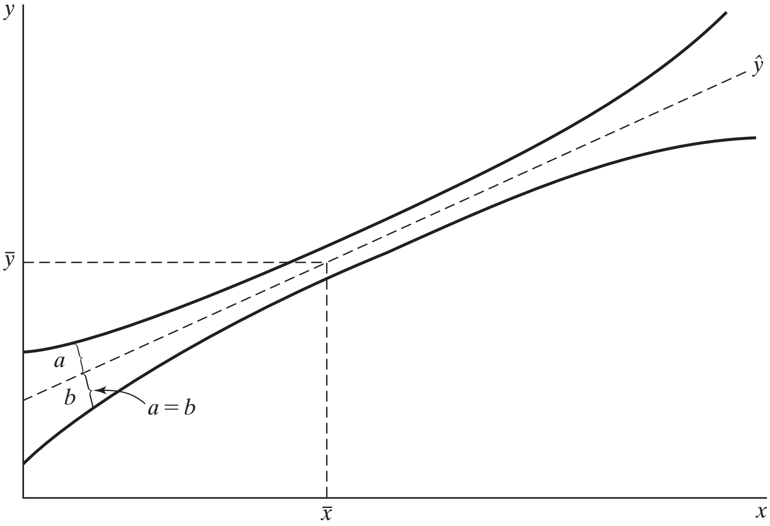
The **prediction variance** of this estimator based on (4-15) is

$$\text{Var}[e^0|\mathbf{X},\mathbf{x}^0] = \sigma^2 + \text{Var}[(\mathbf{b} - \boldsymbol{\beta})'\mathbf{x}^0|\mathbf{X},\mathbf{x}^0] = \sigma^2 + \mathbf{x}^{0\prime}[\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{x}^0. \tag{4-53}$$

If the regression contains a constant term, then an equivalent expression is

$$\text{Var}[e^0|\mathbf{X},\mathbf{x}^0] = \sigma^2\left[1 + \frac{1}{n} + \sum_{j=1}^{K-1},\sum_{k=1}^{K-1}(x_j^0 - \overline{x}_j)(x_k^0 - \overline{x}_k)(\mathbf{Z}'\mathbf{M}^0\mathbf{Z})^{jk}\right], \tag{4-54}$$

FIGURE 4.7 Prediction Intervals.



where $\mathbf{Z}$ is the $K - 1$ columns of $\mathbf{X}$ not including the constant, $\mathbf{Z}'\mathbf{M}^0\mathbf{Z}$ is the matrix of sums of squares and products for the columns of $\mathbf{X}$ in deviations from their means [see (3-21)], and the "*jk*" superscript indicates the *jk* element of the inverse of the matrix. This result suggests that the width of a confidence interval (i.e., a **prediction interval**) depends on the distance of the elements of $\mathbf{x}^0$ from the center of the data. Intuitively, this idea makes sense; the farther the forecasted point is from the center of our experience, the greater is the degree of uncertainty. Figure 4.7 shows the effect for the bivariate case. Note that the prediction variance is composed of three parts. The second and third become progressively smaller as we accumulate more data (i.e., as n increases). But, the first term, σ^2 is constant, which implies that no matter how much data we have, we can never predict perfectly.

The prediction variance can be estimated by using s^2 in place of σ^2. A confidence (prediction) interval for y^0 would then be formed using

$$\text{prediction interval} = \hat{y}^0 \pm t_{(1-\alpha/2),[n-K]} se(e^0), \tag{4-55}$$

where $t_{(1-\alpha/2),[n-K]}$ is the appropriate critical value for $100(1 - \alpha)$ % significance from the t table for $n - K$ degrees of freedom and $se(e^0)$ is the square root of the estimated prediction variance.

4.8.2 PREDICTING *Y* WHEN THE REGRESSION MODEL DESCRIBES LOG *y*

It is common to use the regression model to describe a function of the dependent variable, rather than the variable, itself. In Example 4.5 we model the sale prices of Monet paintings using

$$\ln Price = \beta_1 + \beta_2 \ln Area + \beta_3\, Aspect\ Ratio + \varepsilon.$$

The log form is convenient in that the coefficient provides the elasticity of the dependent variable with respect to the independent variable, that is, in this model,

$\beta_2 = \partial E\,[\ln Price \mid \ln Area, AspectRatio]/\partial \ln Area$. However, the equation in this form is less interesting for prediction purposes than one that predicts the price itself. The natural approach for a predictor of the form

$$\ln y^0 = \mathbf{x}^{0\prime}\mathbf{b}$$

would be to use

$$\hat{y}^0 = \exp(\mathbf{x}^{0\prime}\mathbf{b}).$$

The problem is that $E[y \mid \mathbf{x}^0]$ is not equal to $\exp(E[\ln y \mid \mathbf{X}^0])$. The appropriate conditional mean function would be

$$E[y \mid \mathbf{x}^0] = E[\exp(\mathbf{x}^{0\prime}\beta + \varepsilon^0) \mid \mathbf{x}^0] \; = \exp(\mathbf{x}^{0\prime}\beta)\, E[\exp(\varepsilon^0) \mid \mathbf{x}^0].$$

The second term is not $\exp(E[\varepsilon^0 \mid \mathbf{x}^0]) = 1$ in general. The precise result if $\varepsilon^0 \mid \mathbf{x}^0$ is normally distributed with mean zero and variance σ^2 is $E[\exp(\varepsilon^0) \mid \mathbf{x}^0] = \exp(\sigma^2/2)$. (See Section B.4.4.) The implication for normally distributed disturbances would be that an appropriate predictor for the conditional mean would be

$$\hat{y}^0 = \exp(\mathbf{x}^{0\prime}\mathbf{b} + s^2/2) > \exp(\mathbf{x}^{0\prime}\mathbf{b}), \qquad \textbf{(4-56)}$$

which would seem to imply that the naïve predictor would systematically underpredict y. However, this is not necessarily the appropriate interpretation of this result. The inequality implies that the naïve predictor will systematically underestimate the conditional mean function, not necessarily the realizations of the variable itself. The pertinent question is whether the conditional mean function is the desired predictor for the exponent of the dependent variable in the log regression. The conditional median might be more interesting, particularly for a financial variable such as income, expenditure, or the price of a painting. If the distribution of the variable in the log regression is symmetrically distributed (as they are when the disturbances are normally distributed), then the exponent will be asymmetrically distributed with a long tail in the positive direction, and the mean will exceed the median, possibly vastly so. In such cases, the median is often a preferred estimator of the center of a distribution. For estimating the median, rather then the mean, we would revert to the original naïve predictor, $\hat{y}^0 = \exp(\mathbf{x}^{0\prime}\mathbf{b})$.

Given the preceding, we consider estimating $E[\exp(y) \mid \mathbf{x}^0]$. If we wish to avoid the normality assumption, then it remains to determine what one should use for $E[\exp(\varepsilon^0) \mid \mathbf{x}^0]$. Duan (1983) suggested the consistent estimator (assuming that the expectation is a constant, that is, that the regression is homoscedastic),

$$\hat{E}[\exp(\varepsilon^0) \mid \mathbf{x}^0] = h^0 = \frac{1}{n}\sum\nolimits_{i=1}^{n} \exp(e_i), \qquad \textbf{(4-57)}$$

where e_i is a least squares residual in the original log form regression. Then, Duan's **smearing estimator** for prediction of y^0 is

$$\hat{y}^0 = h^0 \exp(\mathbf{x}^{0\prime}\mathbf{b}).$$

4.8.3 PREDICTION INTERVAL FOR *Y* WHEN THE REGRESSION MODEL DESCRIBES LOG *y*

We obtained a prediction interval in (4-55) for $\ln y \mid \mathbf{x}^0$ in the loglinear model $\ln y = \mathbf{x}'\beta + \varepsilon$,

$$[\ln \hat{y}^0_{LOWER}, \ln \hat{y}^0_{UPPER}] = \left[\mathbf{x}^{0\prime}\mathbf{b} - t_{(1-\alpha/2),[n-K]}se(e^0), \mathbf{x}^{0\prime}\mathbf{b} + t_{(1-\alpha/2),\,[n-K]}se(e^0)\right].$$

For a given choice of α, say, 0.05, these values give the 0.025 and 0.975 quantiles of the distribution of ln $y|\mathbf{x}^0$. If we wish specifically to estimate these quantiles of the distribution of $y|\mathbf{x}^0$, not ln$y|\mathbf{x}^0$, then we would use:

$$\left[\hat{y}^0_{LOWER}, \hat{y}^0_{UPPER}\right] = \left\{\exp\left[\mathbf{x}^{0\prime}\mathbf{b} - t_{(1-\alpha/2),[n-K]}se(e^0)\right], \exp\left[\mathbf{x}^{0\prime}\mathbf{b} + t_{(1-\alpha/2),\,[n-K]}se(e^0)\right]\right\}. \quad \textbf{(4-58)}$$

This follows from the result that if Prob[ln $y \leq$ ln L] $= 1 - \alpha/2$, then Prob[$y \leq L$] $= 1 - \alpha/2$. The result is that the natural estimator is the right one for estimating the specific quantiles of the distribution of the original variable. However, if the objective is to find an interval estimator for $y|\mathbf{x}^0$ that is as narrow as possible, then this approach is not optimal. If the distribution of y is asymmetric, as it would be for a loglinear model with normally distributed disturbances, then the naïve interval estimator is longer than necessary. Figure 4.8 shows why. We suppose that (L, U) in the figure is the prediction interval formed by (4-58). Then the probabilities to the left of L and to the right of U each equal $\alpha/2$. Consider alternatives $L_0 = 0$ and U_0 instead. As we have constructed the figure, the area (probability) between L_0 and L equals the area between U_0 and U. But, because the density is so much higher at L, the distance (0, U_0), the dashed interval, is visibly shorter than that between (L, U). The sum of the two tail probabilities is still equal to α, so this provides a shorter prediction interval. We could improve on (4-58) by using, instead, (0, U_0), where U_0 is simply exp[$\mathbf{x}^{0\prime}\mathbf{b} + t_{(1-\alpha),\,[n-K]}se(e^0)$] (i.e., we put the entire tail area to the right of the upper value). However, while this is an improvement, it goes too far, as we now demonstrate.

Consider finding directly the shortest prediction interval. We treat this as an optimization problem,

$$\text{Minimize}(L, U){:}I = U - L \text{ subject to } F(L) + [1 - F(U)] = \alpha,$$

where F is the cdf of the random variable y (not ln y). That is, we seek the shortest interval for which the two tail probabilities sum to our desired α (usually 0.05). Formulate this as a Lagrangean problem,

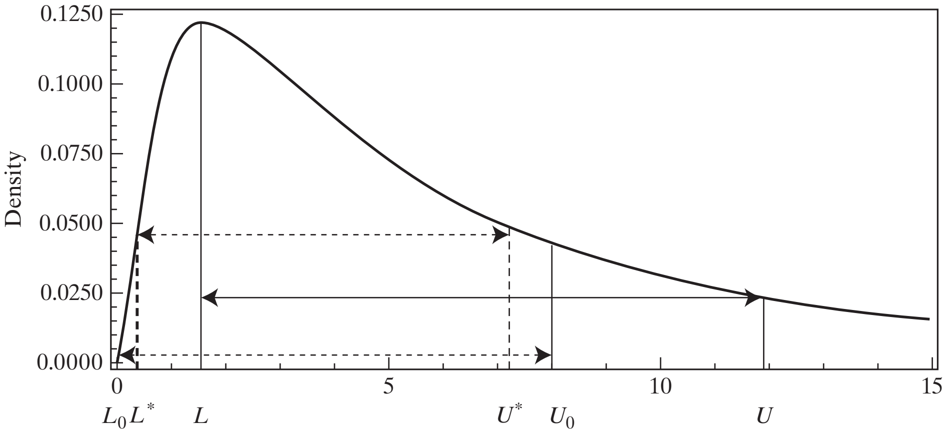
$$\text{Minimize}(L, U, \lambda){:}I^* = U - L + \lambda[F(L) + (1 - F(U)) - \alpha].$$

The solutions are found by equating the three partial derivatives to zero:

$$\partial I^*/\partial L = -1 + \lambda f(L) = 0,$$
$$\partial I^*/\partial U = 1 - \lambda f(U) = 0,$$
$$\partial I^*/\partial \lambda = F(L) + [1 - F(U)] - \alpha = 0,$$

where $f(L) = F'(L)$ and $f(U) = F'(U)$ are the derivatives of the cdf, which are the densities of the random variable at L and U, respectively. The third equation enforces the restriction that the two tail areas sum to α but does not force them to be equal. By adding the first two equations, we find that $\lambda[f(L) - f(U)] = 0$, which, if λ is not zero, means that the solution is obtained by locating (L^*, U^*) such that the tail areas sum to α

FIGURE 4.8 Lognormal Distribution for Prices of Monet Paintings.



and the densities are equal. Looking again at Figure 4.8, we can see that the solution we would seek is (L^*, U^*) where $0 < L^* < L$ and $U^* < U_0$. This is the shortest interval, and it is shorter than both $[0, U_0]$ and $[L, U]$.

This derivation would apply for any distribution, symmetric or otherwise. For a symmetric distribution, however, we would obviously return to the symmetric interval in (4-58). It provides the correct solution for when the distribution is asymmetric. In Bayesian analysis, the counterpart when we examine the distribution of a parameter conditioned on the data, is the **highest posterior density interval**. (See Section 16.4.2.) For practical application, this computation requires a specific assumption for the distribution of $y|\mathbf{x}^0$, such as lognormal. Typically, we would use the smearing estimator specifically to avoid the distributional assumption. There also is no simple formula to use to locate this interval, even for the lognormal distribution. A crude grid search would probably be best, though each computation is very simple. What this derivation does establish is that one can do substantially better than the naïve interval estimator, for example, using $[0, U_0]$.

Example 4.10 Pricing Art

In Examples 4.3 and 4.5, we examined an intriguing feature of the market for Monet paintings, that larger paintings sold at auction for more than smaller ones. Figure 4.9 shows a histogram for the sample of sale prices (in $million). Figure 4.10 shows a histogram for the logs of the prices. Results of the linear regression of lnPrice on lnArea (height times width) and Aspect Ratio (height divided by width) are given in Table 4.8.

We consider using the regression model to predict the price of one of the paintings, a 1903 painting of Charing Cross Bridge that sold for $3,522,500. The painting is 25.6″ high and 31.9″ wide. (This is observation 58 in the sample.) The log area equals ln(25.6 × 31.9) = 6.705198 and the aspect ratio equals 31.9/25.6 = 1.246094. The prediction for the log of the price would be

$$\ln P|\mathbf{x}^0 = -8.34327 + 1.31638(6.705198) - 0.09623(1.246094) = 0.3643351$$

Note that the mean log price is 0.33274, so this painting is expected to sell for roughly 9.5% more than the average painting, based on its dimensions. The estimate of the prediction variance is computed using (4-53); $s_p = 1.105640$ The sample is large enough to use the

FIGURE 4.9 Histogram for Sale Prices of 430 Monet Paintings ($million).

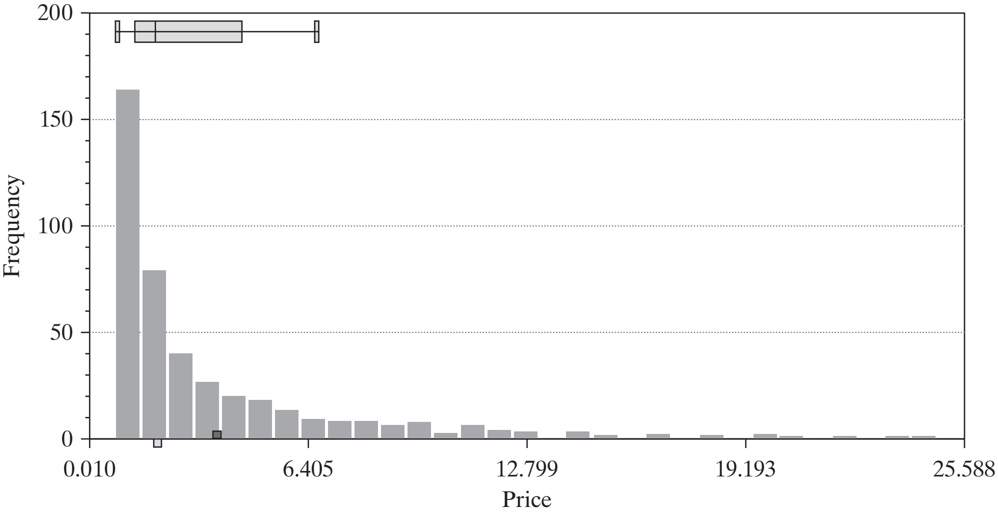
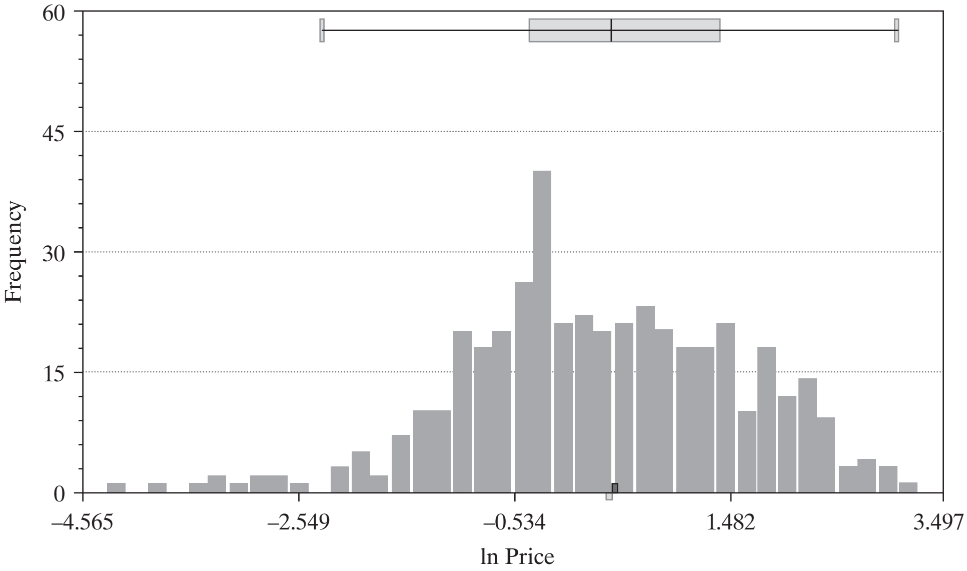


FIGURE 4.10 Histogram of Logs of Auction Prices for Monet Paintings.



critical value from the standard normal table, 1.96, for a 95% confidence interval. A prediction interval for the log of the price is therefore

$$0.364331 \pm 1.96(1.10564) = [-1.80272, 2.53140].$$

For predicting the price, the naïve predictor would be exp(0.3643351) = $1.43956M, which is far under the actual sale price of $3,522,500. To compute the smearing estimator, we require

TABLE 4.8 Estimated Equation for ln Price

Mean of ln Price	0.33274
Sum of squared residuals	520.765
Standard error of regression	1.10435
R-squared	0.33417
Adjusted *R*-squared	0.33105
Number of observations	430

Variable	*Coefficient*	*Standard Error*	*t Ratio*	*Mean of X*
Constant	−8.34327	0.67820	−12.30	1.00000
ln *Area*	1.31638	0.09205	14.30	6.68007
Aspect Ratio	−0.09623	0.15784	−0.61	1.23066

Estimated Asymptotic Covariance Matrix

	Constant	*ln Area*	*Aspect Ratio*
Constant	0.45996		
ln *Area*	−0.05969	0.00847	
Aspect Ratio	−0.04744	0.00251	0.02491

the mean of the exponents of the residuals, which is 1.81661. The revised point estimate for the price would thus be 1.81661 × 1.43956 = \$2.61511M—this is better, but still fairly far off. This particular painting seems to have sold for relatively more than history (the data) would have predicted.

4.8.4 FORECASTING

The preceding discussion assumes that $\mathbf{x}^0$ is known with certainty, ex post, or has been forecast perfectly, ex ante. If $\mathbf{x}^0$ must, itself, be forecast (an ex ante forecast), then the formula for the forecast variance in (4-46) would have to be modified to incorporate the uncertainty in forecasting $\mathbf{x}^0$. This would be analogous to the term σ^2 in the prediction variance that accounts for the implicit prediction of ε^0. This will vastly complicate the computation. Many authors view it as simply intractable. Beginning with Feldstein (1971), derivation of firm analytical results for the correct forecast variance for this case remain to be derived except for simple special cases. The one qualitative result that seems certain is that (4-53) will understate the true variance. McCullough (1996) presents an alternative approach to computing appropriate forecast standard errors based on the method of bootstrapping. (See Chapter 15.)

Various measures have been proposed for assessing the predictive accuracy of forecasting models.[9] Most of these measures are designed to evaluate ex post forecasts; that is, forecasts for which the independent variables do not themselves have to be forecast. Two measures that are based on the residuals from the forecasts are the **root mean squared error**,

$$\text{RMSE} = \sqrt{\frac{1}{n^0}\sum_i (y_i - \hat{y}_i)^2},$$

[9] See Theil (1961) and Fair (1984).

and the **mean absolute error**,

$$\text{MAE} = \frac{1}{n^0}\sum_i |y_i - \hat{y}_i|,$$

where n^0 is the number of periods being forecasted. (Note that both of these, as well as the following measure below, are backward looking in that they are computed using the observed data on the independent variable.) These statistics have an obvious scaling problem—multiplying values of the dependent variable by any scalar multiplies the measure by that scalar as well. Several measures that are scale free are based on the **Theil *U* statistic**:[10]

$$U = \sqrt{\frac{(1/n^0)\sum_i (y_i - \hat{y}_i)^2}{(1/n^0)\sum_i y_i^2}}.$$

This measure is related to R^2 but is not bounded by zero and one. Large values indicate a poor forecasting performance.

4.9 DATA PROBLEMS

The analysis to this point has assumed that the data in hand, **X** and **y**, are well measured and correspond to the assumptions of the model and to the variables described by the underlying theory. At this point, we consider several ways that real-world observed nonexperimental data fail to meet the assumptions. Failure of the assumptions generally has implications for the performance of the estimators of the model parameters—unfortunately, none of them good. The cases we will examine are:

- **Multicollinearity:** Although the full rank assumption, A2, is met, it almost fails. (*Almost* is a matter of degree, and sometimes a matter of interpretation.) Multicollinearity leads to imprecision in the estimator, though not to any systematic biases in estimation.
- **Missing values:** Gaps in **X** and/or **y** can be harmless. In many cases, the analyst can (and should) simply ignore them, and just use the complete data in the sample. In other cases, when the data are missing for reasons that are related to the outcome being studied, ignoring the problem can lead to inconsistency of the estimators.
- **Measurement error:** Data often correspond only imperfectly to the theoretical construct that appears in the model—individual data on income and education are familiar examples. Measurement error is never benign. The least harmful case is measurement error in the dependent variable. In this case, at least under probably reasonable assumptions, the implication is to degrade the fit of the model to the data compared to the (unfortunately hypothetical) case in which the data are accurately measured. Measurement error in the regressors is malignant—it produces systematic biases in estimation that are difficult to remedy.

[10] Theil (1961).

4.9.1 MULTICOLLINEARITY

The Gauss–Markov theorem states that among all linear unbiased estimators, the least squares estimator has the smallest variance. Although this result is useful, it does not assure us that the least squares estimator has a small variance in any absolute sense. Consider, for example, a model that contains two explanatory variables and a constant. For either slope coefficient,

$$\text{Var}[b_k|\ \mathbf{X}] = \frac{\sigma^2}{(1 - r_{12}^2)\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2} = \frac{\sigma^2}{(1 - r_{12}^2)S_{kk}}, k = 1, 2. \qquad \textbf{(4-59)}$$

If the two variables are perfectly correlated, then the variance is infinite. The case of an exact linear relationship among the regressors is a serious failure of the assumptions of the model, not of the data. The more common case is one in which the variables are highly, but not perfectly, correlated. In this instance, the regression model retains all its assumed properties, although potentially severe statistical problems arise. The problem faced by applied researchers when regressors are highly, although not perfectly, correlated include the following symptoms:

- Small changes in the data produce wide swings in the parameter estimates.
- Coefficients may have very high standard errors and low significance levels even though they are jointly significant and the R^2 for the regression is quite high.
- Coefficients may have the "wrong" sign or implausible magnitudes.

For convenience, define the data matrix, $\mathbf{X}$, to contain a constant and $K - 1$ other variables measured in deviations from their means. Let $\mathbf{x}_k$ denote the kth variable, and let $\mathbf{X}_{(k)}$ denote all the other variables (including the constant term). Then, in the inverse matrix, $(\mathbf{X}'\mathbf{X})^{-1}$, the kth diagonal element is

$$\begin{aligned}(\mathbf{x}_k'\mathbf{M}_{(k)}\mathbf{x}_k)^{-1} &= [\mathbf{x}_k'\mathbf{x}_k - \mathbf{x}_k'\mathbf{X}_{(k)}(\mathbf{X}_{(k)}'\mathbf{X}_{(k)})^{-1}\mathbf{X}_{(k)}'\mathbf{x}_k]^{-1} \\ &= \left[\mathbf{x}_k'\mathbf{x}_k\left(1 - \frac{\mathbf{x}_k'\mathbf{X}_{(k)}(\mathbf{X}_{(k)}'\mathbf{X}_{(k)})^{-1}\mathbf{X}_{(k)}'\mathbf{x}_k}{\mathbf{x}_k'\mathbf{x}_k}\right)\right]^{-1} \\ &= \frac{1}{(1 - R_{k.}^2)\, S_{kk}}, \qquad \textbf{(4-60)}\end{aligned}$$

where $R_{k.}^2$ is the R^2 in the regression of x_k on all the other variables. In the multiple regression model, the variance of the kth least squares coefficient estimator is σ^2 times this ratio. It then follows that the more highly correlated a variable is with the other variables in the model (collectively), the greater its variance will be. In the most extreme case, in which $\mathbf{x}_k$ can be written as a linear combination of the other variables, so that $R_{k.}^2 = 1$, the variance becomes infinite. The result,

$$\text{Var}[b_k|\ \mathbf{X}] = \frac{\sigma^2}{(1 - R_{k.}^2)\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2}, \qquad \textbf{(4-61)}$$

shows the three ingredients of the precision of the kth least squares coefficient estimator:

- Other things being equal, the greater the correlation of x_k with the other variables, the higher the variance will be, due to multicollinearity.
- Other things being equal, the greater the variation in x_k, the lower the variance will be.
- Other things being equal, the better the overall fit of the regression, the lower the variance will be. This result would follow from a lower value of σ^2.

Because nonexperimental data will never be orthogonal ($R^2_{k.} = 0$), to some extent multicollinearity will always be present. When is multicollinearity a problem? That is, when are the variances of our estimates so adversely affected by this intercorrelation that we should be "concerned"? Some computer packages report a **variance inflation factor (VIF)**, $1/(1 - R^2_{k.})$, for each coefficient in a regression as a diagnostic statistic. As can be seen, the VIF for a variable shows the increase in $\text{Var}[b_k]$ that can be attributable to the fact that this variable is not orthogonal to the other variables in the model. Another measure that is specifically directed at $\mathbf{X}$ is the **condition number** of $\mathbf{X}'\mathbf{X}$, which is the square root of the ratio of the largest characteristic root of $\mathbf{X}'\mathbf{X}$ to the smallest after scaling each column so that it has unit length. Values in excess of 20 are suggested as indicative of a problem [Belsley, Kuh, and Welsh (1980)]. (The condition number for the Longley data of Example 4.11 is over 15,000!)

Example 4.11 Multicollinearity in the Longley Data

The data in Appendix Table F4.2 were assembled by J. Longley (1967) for the purpose of assessing the accuracy of least squares computations by computer programs. (These data are still widely used for that purpose.[11]) The Longley data are notorious for severe multicollinearity. Note, for example, the last year of the data set. The last observation does not appear to be unusual. But the results in Table 4.9 show the dramatic effect of dropping this single observation from a regression of employment on a constant and the other variables. The last coefficient rises by 600%, and the third rises by 800%.

Several strategies have been proposed for finding and coping with multicollinearity.[12] Under the view that a multicollinearity problem arises because of a shortage of information, one suggestion is to obtain more data. One might argue that if analysts had such additional information available at the outset, they ought to have used it before reaching this juncture. More information need not mean more observations,

TABLE 4.9 Longley Results: Dependent Variable Is Employment

	1947–1961	*Variance Inflation*	*1947–1962*
Constant	1,459,415		1,169,087
Year	−721.756	143.4638	−576.464
GNP Deflator	−181.123	75.6716	−19.7681
GNP	0.0910678	132.467	0.0643940
Armed Forces	−0.0749370	1.55319	−0.0101453

[11] Computing the correct least squares coefficients with the Longley data is not a particularly difficult task by modern standards. The current standard benchmark is set by the NIST's "Filipelli Data." See www.itl.nist.gov/div898/strd/data/Filip.shtml. This application is considered in the Exercises.

[12] See Hill and Adkins (2001) for a description of the standard set of tools for diagnosing collinearity.

however. The obvious practical remedy (and surely the most frequently used) is to drop variables suspected of causing the problem from the regression—that is, to impose on the regression an assumption, possibly erroneous, that the *problem* variable does not appear in the model. If the variable that is dropped actually belongs in the model (in the sense that its coefficient, β_k, is not zero), then estimates of the remaining coefficients will be biased, possibly severely so. On the other hand, overfitting—that is, trying to estimate a model that is too large—is a common error, and dropping variables from an excessively specified model might have some virtue.

Using diagnostic tools to detect multicollinearity could be viewed as an attempt to distinguish a bad model from bad data. But, in fact, the problem only stems from a prior opinion with which the data seem to be in conflict. A finding that suggests multicollinearity is adversely affecting the estimates seems to suggest that, but for this effect, all the coefficients would be statistically significant and of the right sign. Of course, this situation need not be the case. If the data suggest that a variable is unimportant in a model, then, the theory notwithstanding, the researcher ultimately has to decide how strong the commitment is to that theory. Suggested remedies for multicollinearity might well amount to attempts to force the theory on the data.

As a response to what appears to be a multicollinearity problem, it is often difficult to resist the temptation to drop what appears to be an offending variable from the regression. This strategy creates a subtle dilemma for the analyst. Consider the partitioned multiple regression

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{z}\gamma + \varepsilon.$$

If we regress **y** only on **X**, the estimator is biased:

$$E[\mathbf{b}|\;\;\mathbf{X}] = \boldsymbol{\beta} + \mathbf{p}_{\mathrm{X.z}}\gamma.$$

The covariance matrix of this estimator is

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

(Keep in mind, this variance is around $E[\mathbf{b}|\mathbf{X}]$, not around $\boldsymbol{\beta}$.) If γ is not actually zero, then in the multiple regression of **y** on (**X**, **z**), the variance of $\mathbf{b}_{\mathrm{X.z}}$ around its mean, β would be

$$\begin{aligned}\mathrm{Var}[\mathbf{b}_{\mathrm{X.z}}|\mathbf{X},\mathbf{z}] &= \sigma^2(\mathbf{X}'\mathbf{M}_z\mathbf{X})^{-1}\\ &= \sigma^2[\mathbf{X}'\mathbf{X} - \mathbf{X}'\mathbf{z}(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{X}]^{-1}.\end{aligned}$$

To compare the two covariance matrices, it is simpler to compare their inverses. [See result (A-120).] Thus,

$$\{\mathrm{Var}[\mathbf{b}|\mathbf{X}]\}^{-1} - \{\mathrm{Var}[\mathbf{b}_{\mathrm{X.z}}|\mathbf{X},\mathbf{z}]\}^{-1} = (1/\sigma^2)\mathbf{X}'\mathbf{z}(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{X},$$

which is a nonnegative definite matrix. The implication is that the variance of **b** is not larger than the variance of $\mathbf{b}_{\mathbf{X.z}}$ (because its inverse is at least as large). It follows that although **b** is biased, its variance is never larger than the variance of the unbiased estimator. In any realistic case (i.e., if $\mathbf{X}'\mathbf{z}$ is not zero), in fact, it will be smaller. We get a useful comparison from a simple regression with two variables, x and z, measured as deviations from their means. Then, $\mathrm{Var}[b|\mathbf{x}] = \sigma^2/S_{xx}$ where $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$ while

$\text{Var}[b_{\mathbf{x.z}}|\mathbf{x,z}] = \sigma^2/[S_{xx}(1 - r_{xz}^2)]$ where r_{xz}^2 is the squared correlation between x and z. Clearly, $\text{Var}[b_{\text{x.z}}|\mathbf{x,z}]$ is larger.

The result in the preceding paragraph poses a bit of a dilemma for applied researchers. The situation arises frequently in the search for a model specification. Faced with a variable that a researcher suspects should be in the model, but that is causing a problem of multicollinearity, the analyst faces a choice of omitting the relevant variable or including it and estimating its (and all the other variables') coefficient imprecisely. This presents a choice between two estimators, the biased but precise b_1 and the unbiased but imprecise $b_{1.2}$. There is no accepted right answer to this dilemma, but as a general rule, the methodology leans away from estimation strategies that include ad hoc remedies for multicollinearity. For this particular case, there would be a general preference to retain z in the estimated model.

4.9.2 PRINCIPAL COMPONENTS

A device that has been suggested for reducing multicollinearity is to use a small number, say L, of **principal components** constructed as linear combinations of the K original variables.[13] (The mechanics are illustrated in Example 4.11.) The argument against using this approach is that if the original specification in the form $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ were correct, then it is unclear what one is estimating when one regresses $\mathbf{y}$ on some small set of linear combinations of the columns of $\mathbf{X}$. For a set of $L < K$ principal components, if we regress $\mathbf{y}$ on $\mathbf{Z} = \mathbf{XC}_L$ to obtain $\mathbf{d}$, it follows that $E[\mathbf{d}] = \boldsymbol{\delta} = \mathbf{C}_L'\boldsymbol{\beta}$. (The proof is considered in the exercises.) In an economic context, if $\boldsymbol{\beta}$ has an interpretation, then it is unlikely that $\boldsymbol{\delta}$ will. For example, how do we interpret the price elasticity minus twice the income elasticity?

This orthodox interpretation cautions the analyst about mechanical devices for coping with multicollinearity that produce uninterpretable mixtures of the coefficients. But there are also situations in which the model is built on a platform that might well involve a mixture of some measured variables. For example, one might be interested in a regression model that contains *ability*, ambiguously defined. As a measured counterpart, the analyst might have in hand standardized scores on a set of tests, none of which individually has any particular meaning in the context of the model. In this case, a mixture of the measured test scores might serve as one's preferred proxy for the underlying variable. The study in Example 4.11 describes another natural example.

Example 4.12 Predicting Movie Success

Predicting the box office success of movies is a favorite exercise for econometricians.[14] The traditional predicting equation takes the form

$$\textit{Box Office Receipts} = f(\textit{Budget}, \textit{Genre}, \textit{MPAA Rating}, \textit{Star Power}, \textit{Sequel}, \textit{etc.}) + \varepsilon.$$

Coefficients of determination on the order of 0.4 are fairly common. Notwithstanding the relative power of such models, the common wisdom in Hollywood is "nobody knows." There is tremendous randomness in movie success, and few really believe they can forecast it with any reliability. Versaci (2009) added a new element to the model, "Internet buzz."

[13] See, for example, Gurmu, Rilstone, and Stern (1999).

[14] See, for example, Litman (1983), Ravid (1999), De Vany (2003), De Vany and Walls (1999, 2002, 2003), and Simonoff and Sparrow (2000).

Internet buzz is vaguely defined to be Internet traffic and interest on familiar Web sites such as RottenTomatoes.com, ImDB.com, Fandango.com, and traileraddict.com. None of these by itself defines Internet buzz. But, collectively, activity on these Web sites, say three weeks before a movie's opening, might be a useful predictor of upcoming success. Versaci's data set (Table F4.3) contains data for 62 movies released in 2009, including four Internet buzz variables, all measured three weeks prior to the release of the movie:

$buzz_1$ = number of Internet views of movie trailer at traileraddict.com

$buzz_2$ = number of message board comments about the movie at ComingSoon.net

$buzz_3$ = total number of "can't wait" (for release) plus "don't care" votes at Fandango.com

$buzz_4$ = percentage of Fandango votes that are "can't wait"

We have aggregated these into a single principal component as follows: We first computed the logs of $buzz_1 - buzz_3$ to remove the scale effects. We then standardized the four variables, so z_k contains the original variable minus its mean, $\bar{z}_k$, then divided by its standard deviation, s_k. Let **Z** denote the resulting 62 × 4 matrix (z_1, z_2, z_3, z_4). Then $\mathbf{V} = (1/61)\mathbf{Z}'\mathbf{Z}$ is the sample correlation matrix. Let c_1 be the characteristic vector of **V** associated with the largest characteristic root. The first principal component (the one that explains most of the variation of the four variables) is $\mathbf{Zc}_1$. (The roots are 2.4142, 0.7742, 0.4522, and 0.3585, so the first principal component explains 2.4142/4 or 60.3% of the variation. Table 4.10 shows the regression results for the sample of 62 2009 movies. It appears that Internet buzz adds substantially to the predictive power of the regression. The R^2 of the regression nearly doubles, from 0.34 to 0.59, when Internet buzz is added to the model. As we will discuss in Chapter 5, buzz is also a highly significant predictor of success.

4.9.3 MISSING VALUES AND DATA IMPUTATION

It is common for data sets to have gaps for a variety of reasons. Perhaps the most frequent occurrence of this problem is in survey data, in which respondents may simply

TABLE 4.10 Regression Results for Movie Success

	Internet Buzz Model			*Traditional Model*		
$e'e$	22.30215			35.66514		
R^2	0.58883			0.34247		
Variable	*Coefficient*	*Std.Error*	*t*	*Coefficient*	*Std.Error*	*t*
Constant	15.4002	0.64273	23.96	13.5768	0.68825	19.73
Action	−0.86932	0.29333	−2.96	−0.30682	0.34401	−0.89
Comedy	−0.01622	0.25608	−0.06	−0.03845	0.32061	−0.12
Animated	−0.83324	0.43022	−1.94	−0.82032	0.53869	−1.52
Horror	0.37460	0.37109	1.01	1.02644	0.44008	2.33
G	0.38440	0.55315	0.69	0.25242	0.69196	0.36
PG	0.53359	0.29976	1.78	0.32970	0.37243	0.89
PG13	0.21505	0.21885	0.98	0.07176	0.27206	0.26
ln *Budget*	0.26088	0.18529	1.41	0.70914	0.20812	3.41
Sequel	0.27505	0.27313	1.01	0.64368	0.33143	1.94
Star Power	0.00433	0.01285	0.34	0.00648	0.01608	0.40
Buzz	0.42906	0.07839	5.47	–	–	–

fail to respond to the questions. In a time series, the data may be missing because they do not exist at the frequency we wish to observe them; for example, the model may specify monthly relationships, but some variables are observed only quarterly. In panel data sets, the gaps in the data may arise because of **attrition** from the study. This is particularly common in health and medical research, when individuals choose to leave the study—possibly because of the success or failure of the treatment that is being studied.

There are several possible cases to consider, depending on why the data are missing. The data may be simply unavailable, for reasons unknown to the analyst and unrelated to the completeness or the values of the other observations in the sample. This is the most benign situation. If this is the case, then the complete observations in the sample constitute a usable data set, and the only issue is what possibly helpful information could be salvaged from the incomplete observations. Griliches (1986) calls this the **ignorable case** in that, for purposes of estimation, if we are not concerned with efficiency, then we may simply delete the incomplete observations and ignore the problem. Rubin (1976, 1987), Afifi and Elashoff (1966, 1967), and Little and Rubin (1987, 2002) label this case **missing completely at random (MCAR)**. A second case, which has attracted a great deal of attention in the econometrics literature, is that in which the gaps in the data set are not benign but are systematically related to the phenomenon being modeled. This case happens most often in surveys when the data are self-selected or self-reported. For example, if a survey were designed to study expenditure patterns and if high-income individuals tended to withhold information about their income, then the gaps in the data set would represent more than just missing information. The clinical trial case is another instance. In this (worst) case, the complete observations would be qualitatively different from a sample taken at random from the full population. The missing data in this situation are termed **not missing at random (NMAR)**. We treat this second case in Chapter 19 with the subject of **sample selection**, so we shall defer our discussion until later.

The intermediate case is that in which there is information about the missing data contained in the complete observations that can be used to improve inference about the model. The incomplete observations in this **missing at random (MAR)** case are also ignorable, in the sense that unlike the NMAR case, simply using the complete data does not induce any biases in the analysis, as long as the underlying process that produces the missingness in the data does not share parameters with the model that is being estimated, which seems likely.[15] This case is unlikely, of course, if "missingness" is based on the values of the dependent variable in a regression. Ignoring the incomplete observations when they are MAR but not MCAR does ignore information that is in the sample and therefore sacrifices some efficiency. Researchers have used a variety of **data imputation** methods to fill gaps in data sets. The (by far) simplest case occurs when the gaps occur in the data on the regressors. For the case of missing data on the regressors, it helps to consider the simple regression and multiple regression cases separately. In the first case, $\mathbf{X}$ has two columns: the column of 1s for the constant and a column with some blanks where the missing data would be if we had them. The **zero-order method** of replacing each missing x with $\bar{x}$ based on the observed data results in no changes and is equivalent to dropping the incomplete data. (See Exercise 7 in Chapter 3.) However, the R^2 will be lower. An alternative, **modified zero-order regression**, fills the second column of $\mathbf{X}$ with zeros and adds a variable that takes the value one for **missing observations**

[15] See Allison (2002).

and zero for complete ones. We leave it as an exercise to show that this is algebraically identical to simply filling the gaps with $\overline{x}$. These same methods can be used when there are multiple regressors. Once again, it is tempting to replace missing values of $\mathbf{x}_k$ with simple means of complete observations or with the predictions from linear regressions based on other variables in the model for which data are available when $\mathbf{x}_k$ is missing. In most cases in this setting, a general characterization can be based on the principle that for any missing observation, the *true* unobserved x_{ik} is being replaced by an erroneous proxy that we might view as $\hat{x}_{ik} = x_{ik} + u_{ik}$, that is, in the framework of **measurement error**. Generally, the least squares estimator is biased (and inconsistent) in the presence of measurement error such as this. (We will explore the issue in Chapter 8.) A question does remain: Is the bias likely to be reasonably small? As intuition should suggest, it depends on two features of the data: (1) how good the prediction of x_{ik} is in the sense of how large the variance of the measurement error, u_{ik}, is compared to that of the actual data, x_{ik}, and (2) how large a proportion of the sample the analyst is filling.

The regression method replaces each missing value on an $\mathbf{x}_k$ with a single prediction from a linear regression of $\mathbf{x}_k$ on other exogenous variables—in essence, replacing the missing x_{ik} with an estimate of it based on the regression model. In a Bayesian setting, some applications that involve unobservable variables (such as our example for a binary choice model in Chapter 17) use a technique called **data augmentation** to treat the unobserved data as unknown parameters to be estimated with the structural parameters, such as $\boldsymbol{\beta}$ in our regression model. Building on this logic researchers, for example, Rubin (1987) and Allison (2002), have suggested taking a similar approach in classical estimation settings. The technique involves a data imputation step that is similar to what was suggested earlier, but with an extension that recognizes the variability in the estimation of the regression model used to compute the predictions. To illustrate, we consider the case in which the independent variable, $\mathbf{x}_k$, is drawn in principle from a normal population, so it is a continuously distributed variable with a mean, a variance, and a joint distribution with other variables in the model. Formally, an imputation step would involve the following calculations:

1. Using as much information (complete data) as the sample will provide, linearly regress $\mathbf{x}_k$ on other variables in the model (and/or outside it, if other information is available), $\mathbf{Z}_k$, and obtain the coefficient vector $\mathbf{d}_k$ with associated asymptotic covariance matrix $\mathbf{A}_k$ and estimated disturbance variance s_k^2.
2. For purposes of the imputation, we draw an observation from the estimated asymptotic normal distribution of $\mathbf{d}_k$; that is, $\mathbf{d}_{k,m} = \mathbf{d}_k + \mathbf{v}_k$ where v_k is a vector of random draws from the normal distribution with mean zero and covariance matrix $\mathbf{A}_k$.
3. For each missing observation in $\mathbf{x}_k$ that we wish to impute, we compute $x_{i,k,m} = \mathbf{d}'_{k,m}\mathbf{z}_{i,k} + s_{k,m}u_{i,k}$, where $s_{k,m}$ is s_k divided by a random draw from the chi-squared distribution with degrees of freedom equal to the number of degrees of freedom in the imputation regression.

At this point, the iteration is the same as considered earlier, where the missing values are imputed using a regression, albeit a much more elaborate procedure. The regression is then computed, using the complete data and the imputed data for the missing observations, to produce coefficient vector, $\mathbf{b}_m$, and estimated covariance matrix, $\mathbf{V}_m$. This constitutes a single round. The technique of *multiple imputation* involves repeating this set of steps M times. The estimators of the parameter vector and the appropriate asymptotic covariance matrix are

$$\hat{\boldsymbol{\beta}} = \overline{\mathbf{b}} = \frac{1}{M}\sum\nolimits_{m=1}^{M}\mathbf{b}_{\mathbf{m}}, \tag{4-61}$$

$$\hat{\mathbf{V}} = \overline{\mathbf{V}} + \mathbf{B} = \frac{1}{M}\sum\nolimits_{m=1}^{M}\mathbf{V}_m + \left(1 + \frac{1}{M}\right)\left(\frac{1}{M-1}\right)\sum\nolimits_{m=1}^{M}(\mathbf{b}_m - \overline{\mathbf{b}})(\mathbf{b}_m - \overline{\mathbf{b}})'. \tag{4-62}$$

Researchers differ on the effectiveness or appropriateness of multiple imputation. When all is said and done, the measurement error in the imputed values remains. It takes very strong assumptions to establish that the multiplicity of iterations will suffice to average away the effect of this error. Very elaborate techniques have been developed for the special case of joint normally distributed cross sections of regressors such as those suggested above. However, the typical application to survey data involves gaps due to nonresponse to qualitative questions with binary answers. The efficacy of the theory is much less well developed for imputation of binary, ordered, count, or other qualitative variables.

Example 4.13 Imputation in the Survey of Consumer Finances[16]

The Survey of Consumer Finances (SCF) is a survey of U.S. households sponsored every three years by the Board of Governors of the Federal Reserve System with the cooperation of the U.S. Department of the Treasury. SCF interviews are conducted by NORC at the University of Chicago. Data from the SCF are used to inform monetary policy, tax policy, consumer protection, and a variety of other policy issues. The most recent release of the survey was in 2013. The 2016 survey is in process as of this writing. Missing data in the survey have been imputed five times using a multiple imputation technique. The information is stored in five separate imputation replicates (implicates). Thus, for the 6,026 families interviewed for the current survey, there are 30,130 records in the data set.[17] Rhine et al. (2016) used the Survey of Consumer Finances to examine savings behavior in the United States during the Great Recession of 2007–2009.

The more manageable case is missing values of the dependent variable, y_i. Once again, it must be the case that y_i is at least MAR and that the mechanism that is determining presence in the sample does not share parameters with the model itself. Assuming the data on $\mathbf{x}_i$ are complete for all observations, one might consider filling the gaps in the data on y_i by a two-step procedure: (1) estimate $\boldsymbol{\beta}$ with $\mathbf{b}_c$ using the complete observations, $\mathbf{X}_c$ and $\mathbf{y}_c$, then (2) fill the missing values, $\mathbf{y}_m$, with predictions, $\hat{\mathbf{y}}_m = \mathbf{X}_m\mathbf{b}_c$, and recompute the coefficients. We leave as an exercise (Exercise 17) to show that the second step estimator is exactly equal to the first. However, the variance estimator at the second step, s^2, must underestimate σ^2, intuitively because we are adding to the sample a set of observations that are fit perfectly.[18] So, this is not a beneficial way to proceed.

[16] See http://www.federalreserve.gov/econresdata/scf/scfindex.htm

[17] The Federal Reserve's download site for the SCF provides the following caution: *WARNING: Please review the following PDF for instructions on how to calculate correct standard errors. As a result of multiple imputation, the dataset you are downloading contains five times the number of actual observations. Failure to account for the imputations and the complex sample design will result in incorrect estimation of standard errors.* (Ibid.)

[18] See Cameron and Trivedi (2005, Chapter 27).

The flaw in the method comes back to the device used to impute the missing values for y_i. Recent suggestions that appear to provide some improvement involve using a randomized version, $\hat{\mathbf{y}}_m = \mathbf{X}_m\mathbf{b}_c + \hat{\boldsymbol{\varepsilon}}_m$, where $\hat{\boldsymbol{\varepsilon}}_m$ are random draws from the (normal) population with zero mean and estimated variance $s^2[\mathbf{I} + \mathbf{X}_m(\mathbf{X}_c'\mathbf{X}_c)^{-1}\mathbf{X}_m']$. (The estimated variance matrix corresponds to $\mathbf{X}_m\mathbf{b}_c + \boldsymbol{\varepsilon}_m$.) This defines an iteration. After reestimating $\boldsymbol{\beta}$ with the augmented data, one can return to re-impute the augmented data with the new $\hat{\boldsymbol{\beta}}$, then recompute $\mathbf{b}$, and so on. The process would continue until the estimated parameter vector stops changing. (A subtle point to be noted here: The same random draws should be used in each iteration. If not, there is no assurance that the iterations would ever converge.)

In general, not much is known about the properties of estimators based on using predicted values to fill missing values of y. Those results we do have are largely from simulation studies based on a particular data set or pattern of missing data. The results of these Monte Carlo studies are usually difficult to generalize. The overall conclusion seems to be that in a single-equation regression context, filling in missing values of y leads to biases in the estimator which are difficult to quantify. The only reasonably clear result is that imputations are more likely to be beneficial if the proportion of observations that are being filled is small—the smaller the better.

4.9.4 MEASUREMENT ERROR

There are any number of cases in which observed data are imperfect measures of their theoretical counterparts in the regression model. Examples include income, education, ability, health, the interest rate, output, capital, and so on. Mismeasurement of the variables in a model will generally produce adverse consequences for least squares estimation. Remedies are complicated and sometimes require heroic assumptions. In this section, we will provide a brief sketch of the issues. We defer to Section 8.8 for a more detailed discussion of the problem of measurement error, the most common solution (instrumental variables estimation), and some applications.

It is convenient to distinguish between measurement error in the dependent variable and measurement error in the regressor(s). For the second case, it is also useful to consider the simple regression case and then extend it to the multiple regression model. Consider a model to describe expected income in a population,

$$I^* = \mathbf{x}'\boldsymbol{\beta} + \varepsilon, \tag{4-63}$$

where I^* is the intended total income variable. Suppose the observed counterpart is I, earnings. How I relates to I^* is unclear; it is common to assume that the measurement error is additive, so $I = I^* + w$. Inserting this expression for I into (4-63) gives

$$\begin{aligned} I &= \mathbf{x}'\boldsymbol{\beta} + \varepsilon + w \\ &= \mathbf{x}'\boldsymbol{\beta} + v, \end{aligned} \tag{4-64}$$

which appears to be a slightly more complicated regression, but otherwise similar to what we started with. As long as w and $\mathbf{x}$ are uncorrelated, that is the case. If w is a homoscedastic zero mean error that is uncorrelated with $\mathbf{x}$, then the only difference between the models in (4-63) and (4-64) is that the disturbance variance in (4-64) is $\sigma_w^2 + \sigma_\varepsilon^2 > \sigma_\varepsilon^2$. Otherwise both are regressions and evidently $\boldsymbol{\beta}$ can be estimated consistently by least squares in either case. The cost of the measurement error is in the precision of the estimator because the asymptotic variance of the estimator in (4-64) is $(\sigma_v^2/n)[\text{plim}(\mathbf{X}'\mathbf{X}/n)]^{-1}$, while it is $(\sigma_\varepsilon^2/n)[\text{plim}(\mathbf{X}'\mathbf{X}/n)]^{-1}$ if $\boldsymbol{\beta}$ is estimated using (4-63). The measurement error also costs some fit. To see this, note that the R^2 in the sample regression in (4-63) is

$$R_*^2 = 1 - (\mathbf{e}'\mathbf{e}/n)/(\mathbf{I}^{*'}\mathbf{M}^0\mathbf{I}^*/n).$$

The numerator converges to σ_ε^2 while the denominator converges to the total variance of I^*, which would approach $\sigma_\varepsilon^2 + \boldsymbol{\beta}'\mathbf{Q}\boldsymbol{\beta}$ where $\mathbf{Q} = \text{plim}(\mathbf{X}'\mathbf{X}/n)$. Therefore,

$$\text{plim}R_*^2 = \boldsymbol{\beta}'Q\boldsymbol{\beta}/[\sigma_\varepsilon^2 + \boldsymbol{\beta}'\mathbf{Q}\boldsymbol{\beta}.$$

The counterpart for (4-64), R^2, differs only in that σ_ε^2 is replaced by $\sigma_v^2 > \sigma_\varepsilon^2$ in the denominator. It follows that

$$\text{plim } R_*^2 - \text{plim } R^2 > 0.$$

This implies that the fit of the regression in (4-64) will, at least broadly in expectation, be inferior to that in (4-63). (The preceding is an asymptotic approximation that might not hold in every finite sample.)

These results demonstrate the implications of measurement error in the dependent variable. We note, in passing, that if the measurement error is not additive, if it is correlated with **x**, or if it has any other features such as heteroscedasticity, then the preceding results are lost, and nothing in general can be said about the consequence of the measurement error. Whether there is a *solution* is likewise an ambiguous question. The preceding explanation shows that it would be better to have the underlying variable if possible. In the absence, would it be preferable to use a proxy? Unfortunately, I is already a proxy, so unless there exists an available I' which has smaller measurement error variance, we have reached an impasse. On the other hand, it does seem that the outcome is fairly benign. The sample does not contain as much information as we might hope, but it does contain sufficient information consistently to estimate β and to do appropriate statistical inference based on the information we do have.

The more difficult case occurs when the measurement error appears in the independent variable(s). For simplicity, we retain the symbols I and I^* for our observed and theoretical variables. Consider a simple regression,

$$y = \beta_1 + \beta_2 I^* + \varepsilon,$$

where y is the perfectly measured dependent variable and the same measurement equation, $I = I* + w$, applies now to the independent variable. Inserting I into the equation and rearranging a bit, we obtain

$$\begin{aligned} y &= \beta_1 + \beta_2 I + (\varepsilon - \beta_2 w) \\ &= \beta_1 + \beta_2 I + v. \end{aligned} \tag{4-65}$$

It appears that we have obtained (4-64) once again. Unfortunately, this is not the case, because $\text{Cov}[I, v] = \text{Cov}[I^* + w, \varepsilon - \beta_2 w] = -\beta_2\sigma_w^2$. Because the regressor in (4-65) is correlated with the disturbance, least squares regression in this case is inconsistent. There is a bit more that can be derived—this is pursued in Section 8.5, so we state it here without proof. In this case,

$$\text{plim } b_2 = \beta_2[\sigma_*^2/(\sigma_*^2 + \sigma_w^2)],$$

where σ_*^2 is the marginal variance of I^*. The scale factor is less than one, so the least squares estimator is biased toward zero. The larger the measurement error variance, the worse is the bias. (This is called **least squares attenuation**.) Now, suppose there are additional variables in the model:

$$y = \mathbf{x}'\boldsymbol{\beta}_1 + \beta_2 I^* + \varepsilon.$$

In this instance, almost no useful theoretical results are forthcoming. The following fairly general conclusions can be drawn—once again, proofs are deferred to Section 8.5:

1. The least squares estimator of β_2 is still biased toward zero.
2. All the elements of the estimator of $\boldsymbol{\beta}_1$ are biased, in unknown directions, even though the variables in **x** are not measured with error.

Solutions to the "measurement error problem" come in two forms. If there is outside information on certain model parameters, then it is possible to deduce the scale factors (using the **method of moments**) and undo the bias. For the obvious example, in (4-65), if σ_w^2 were known, then it would be possible to deduce σ_*^2 from $\text{Var}[I] = \sigma_*^2 + \sigma_w^2$ and thereby compute the necessary scale factor to undo the bias. This sort of information is generally not available. A second approach that has been used in many applications is the technique of instrumental variables. This is developed in detail for this application in Section 8.5.

4.9.5 OUTLIERS AND INFLUENTIAL OBSERVATIONS

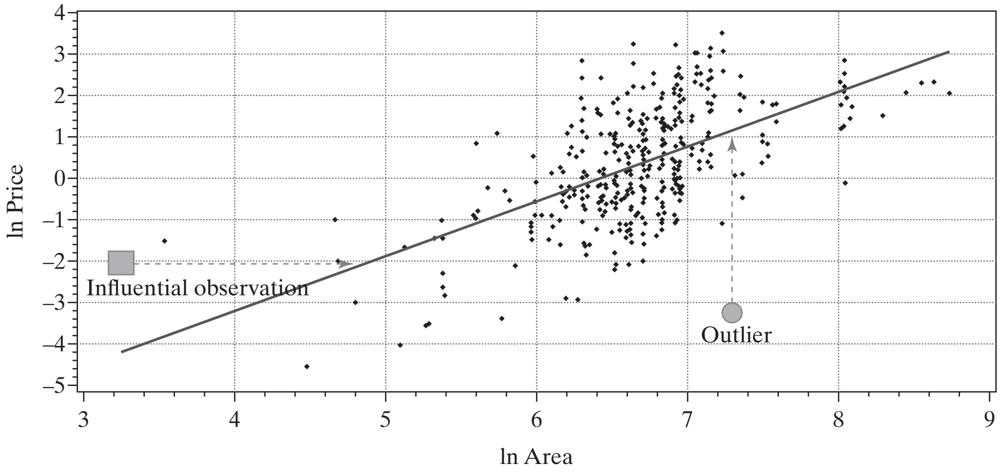
Figure 4.10 shows a scatter plot of the data on sale prices of Monet paintings that were used in Example 4.5. Two points have been highlighted. The one noted with the square overlay shows the smallest painting in the data set. The circle highlights a painting that fetched an unusually low price, at least in comparison to what the regression would have predicted. (It was not the least costly painting in the sample, but it was the one most poorly predicted by the regression.) Because least squares is based on squared deviations, the estimator is likely to be strongly influenced by extreme observations such as these, particularly if the sample is not very large.

An *influential observation* is one that is likely to have a substantial impact on the least squares regression coefficient(s). For a simple regression such as the one shown in Figure 4.11, Belsley, Kuh, and Welsh (1980) defined an influence measure, for observation x_i,

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x}_{(i)})^2}{\Sigma_{j=1, j \neq i}^{n}(x_j - \bar{x}_{(i)})^2}, \qquad \textbf{(4-66)}$$

where $\bar{x}_{(i)}$ and the summation in the denominator of the fraction are computed without this observation. (The measure derives from the difference between **b** and $\mathbf{b}_{(i)}$ where the latter is computed without the particular observation. We will return to this shortly.) It is suggested that an observation should be noted as influential if $h_i > 2/n$. The decision is whether to drop the observation or not. We should note observations with high leverage are arguably not outliers (which remains to be defined) because the analysis is conditional on x_i. To underscore the point, referring to Figure 4.11, this observation would be marked even if it fell precisely on the regression line—the source of the influence is the numerator of the second term in h_i, which is unrelated to the distance of the point from the line. In our example, the influential observation happens to be the

FIGURE 4.11 Log Price Versus Log Area for Monet Paintings.



result of Monet's decision to paint a small painting. The point is that in the absence of an underlying theory that explains (and justifies) the extreme values of x_i, eliminating such observations is an algebraic exercise that has the effect of forcing the regression line to be fitted with the values of x_i closest to the means.

The change in the linear regression coefficient vector in a multiple regression when an observation is added to the sample is

$$\mathbf{b} - \mathbf{b}_{(i)} = \Delta\mathbf{b} = \frac{1}{1 + \mathbf{x}_i'(\mathbf{X}_{(i)}'\mathbf{X}_{(i)})^{-1}\mathbf{x}_i}(\mathbf{X}_{(i)}'\mathbf{X}_{(i)})^{-1}\mathbf{x}_i(\mathbf{y}_i - \mathbf{x}_i'\mathbf{b}_{(i)}), \qquad \textbf{(4-67)}$$

where $\mathbf{b}$ is computed with observation i in the sample, $\mathbf{b}_{(i)}$ is computed without observation i, and $\mathbf{X}_{(i)}$ does not include observation i. (See Exercise 5 in Chapter 3.) It is difficult to single out any particular feature of the observation that would drive this change. The influence measure,

$$\begin{aligned} h_{ii} &= \mathbf{x}_i'(\mathbf{X}_{(i)}'\mathbf{X}_{(i)})^{-1}\mathbf{x}_i \\ &= \frac{1}{n} + \sum_{j=1}^{K-1}\sum_{k=1}^{K-1}(x_{i,j} - \overline{x}_{n,j})(x_{i,k} - \overline{x}_k)(\mathbf{Z}_{(i)}'\mathbf{M}^0\mathbf{Z}_{(i)})^{jk}, \end{aligned} \qquad \textbf{(4-68)}$$

has been used to flag influential observations.[19] In this instance, the selection criterion would be $h_{ii} > 2(K - 1)/n$. Squared deviations of the elements of $\mathbf{x}_i$ from the means of the variables appear in h_{ii}, so it is also operating on the difference of $\mathbf{x}_i$ from the center of the data. (See expression (4-54) for the forecast variance in Section 4.8.1 for an application.)

In principle, an outlier is an observation that appears to be outside the reach of the model, perhaps because it arises from a different data-generating process. The outlier

[19] See, once again, Belsley, Kuh, and Welsh (1980) and Cook (1977).

in Figure 4.11 appears to be a candidate. Outliers could arise for several reasons. The simplest explanation would be actual data errors. Assuming the data are not erroneous, it then remains to define what constitutes an outlier. Unusual residuals are an obvious choice. But, because the distribution of the disturbances would anticipate a certain small percentage of extreme observations in any event, simply singling out observations with large residuals is actually a dubious exercise. On the other hand, one might suspect that the outlying observations are actually generated by a different population. *Studentized residuals* are constructed with this in mind by computing the regression coefficients and the residual variance without observation *i* for each observation in the sample and then standardizing the modified residuals. The *i*th studentized residual is

$$e(i) = \frac{e_i}{\sqrt{1 - h_{ii}}} \Big/ \sqrt{\frac{\mathbf{e}'\mathbf{e} - e_i^2/(1 - h_{ii})}{n - 1 - K}}, \tag{4-69}$$

where **e** is the residual vector for the full sample, based on **b**, including e_i the residual for observation *i*. In principle, this residual has a *t* distribution with $n - 1 - K$ degrees of freedom (or a standard normal distribution asymptotically). Observations with large studentized residuals, that is, greater than 2.0, would be singled out as outliers.

There are several complications that arise with isolating outlying observations in this fashion. First, there is no a priori assumption of which observations are from the alternative population, if this is the view. From a theoretical point of view, this would suggest a skepticism about the model specification. If the sample contains a substantial proportion of outliers, then the properties of the estimator based on the reduced sample are difficult to derive. In the next application, the suggested procedure deletes 4.2% of the sample (18 observations). Finally, it will usually occur that observations that were not outliers in the original sample will become outliers when the original set of outliers is removed. It is unclear how one should proceed at this point. (Using the Monet paintings data, the first round of studentizing the residuals removes 18 observations. After 11 iterations, the sample size stabilizes at 364 of the original 430 observations, a reduction of 15.3%.) Table 4.11 shows the original results (from Table 4.4) and the modified results with 18 outliers removed. Given that the 430 is a relatively large sample, the modest change in the results is to be expected.

TABLE 4.11 Estimated Equations for Log Price

Number of observations	430	412
Mean of log price	0.33274	0.36328
Sum of squared residuals	520.765	393.845
Standard error of regression	1.10435	0.98130
R-squared	0.33417	0.38371
Adjusted R-squared	0.33105	0.38070

	Coefficient		***Standard Error***		***t***	
Variable	$n = 430$	$n = 412$	$n = 430$	$n = 412$	$n = 430$	$n = 412$
Constant	−8.34237	−8.62152	0.67820	0.62524	−12.30	−13.79
ln *Area*	1.31638	1.35777	0.09205	0.08612	14.30	15.77
Aspect Ratio	−0.09623	−0.08346	0.15784	0.14569	−0.61	−0.57

It is difficult to draw firm general conclusions from this exercise. It remains likely that in very small samples, some caution and close scrutiny of the data are called for. If it is suspected at the outset that a process prone to large observations is at work, it may be useful to consider a different estimator altogether, such as least absolute deviations, or even a different model specification that accounts for this possibility. For example, the idea that the sample may contain some observations that are generated by a different process lies behind the latent class model that is discussed in Chapters 14 and 18.

4.10 SUMMARY AND CONCLUSIONS

This chapter has examined a set of properties of the least squares estimator that will apply in all samples, including unbiasedness and efficiency among unbiased estimators. The formal assumptions of the linear model are pivotal in the results of this chapter. All of them are likely to be violated in more general settings than the one considered here. For example, in most cases examined later in the book, the estimator has a possible bias, but that bias diminishes with increasing sample sizes. For purposes of forming confidence intervals and testing hypotheses, the assumption of normality is narrow, so it was necessary to extend the model to allow nonnormal disturbances. These and other "large-sample" extensions of the linear model were considered in Section 4.4. The crucial results developed here were the consistency of the estimator and a method of obtaining an appropriate covariance matrix and large-sample distribution that provides the basis for forming confidence intervals and testing hypotheses. Statistical inference in the form of interval estimation for the model parameters and for values of the dependent variable was considered in Sections 4.6 and 4.7. This development will continue in Chapter 5 where we will consider hypothesis testing and model selection.

Finally, we considered some practical problems that arise when data are less than perfect for the estimation and analysis of the regression model, including multicollinearity, missing observations, measurement error, and outliers.

Key Terms and Concepts

- Assumptions
- Asymptotic covariance matrix
- Asymptotic distribution
- Asymptotic efficiency
- Asymptotic normality
- Asymptotic properties
- Attrition
- Bootstrapping
- Condition number
- Confidence intervals
- Consistency
- Consistent estimator
- Data imputation
- Efficient scale
- Estimator
- Ex ante forecast
- Ex post forecast
- Ex post predication
- Finite sample properties
- Gauss–Markov theorem
- Grenander conditions
- Highest posterior density interval
- Ignorable case
- Interval estimation
- Least squares attenuation
- Lindeberg–Feller Central Limit Theorem
- Linear estimator
- Linear unbiased estimator
- Mean absolute error
- Mean squared error
- Measurement error
- Method of moments
- Minimum mean squared error
- Minimum variance linear unbiased estimator
- Missing at random (MAR)
- Missing completely at random (MCAR)
- Missing observations
- Modified zero-order regression

- Monte Carlo study
- Multicollinearity
- Not missing at random (NMAR)
- Oaxaca's and Blinder's decomposition
- Optimal linear predictor
- Panel data
- Point estimation
- Prediction error
- Prediction interval
- Prediction variance
- Principal components
- Probability limit
- Root mean squared error
- Sample selection
- Sampling distribution
- Sampling variance
- Semiparametric
- Smearing estimator
- Standard error
- Standard error of the regression
- Statistical properties
- Theil *U* statistic
- Variance inflation factor (VIF)
- Zero-order method

Exercises

1. Suppose that you have two independent unbiased estimators of the same parameter θ, say $\hat{\theta}_1$ and $\hat{\theta}_2$, with different variances v_1 and v_2. What linear combination $\hat{\theta} = c_1\hat{\theta}_1 + c_2\hat{\theta}_2$ is the minimum variance unbiased estimator of θ?
2. Consider the simple regression $y_i = \beta x_i + \varepsilon_i$ where $E[\varepsilon|x] = 0$ and $E[\varepsilon^2|x] = \sigma^2$
   a. What is the minimum mean squared error linear estimator of β? [*Hint:* Let the estimator be ($\hat{\beta} = \mathbf{c}'\mathbf{y}$). Choose $\mathbf{c}$ to minimize $\text{Var}(\hat{\beta}) + (E(\hat{\beta} - \beta))^2$. The answer is a function of the unknown parameters.]
   b. For the estimator in part a, show that ratio of the mean squared error of $\hat{\beta}$ to that of the ordinary least squares estimator b is
   $$\frac{\text{MSE}[\hat{\beta}]}{\text{M}SE[b]} = \frac{\tau^2}{(1 + \tau^2)}, \text{ where } \tau^2 = \frac{\beta^2}{[\sigma^2/\mathbf{X}'\mathbf{X}]}.$$
   Note that τ is the population analog to the "t ratio" for testing the hypothesis that $\beta = 0$, which is given in (5-11). How do you interpret the behavior of this ratio as $\tau \rightarrow \infty$?
3. Suppose that the classical regression model applies but that the true value of the constant is zero. Compare the variance of the least squares slope estimator computed without a constant term with that of the estimator computed with an unnecessary constant term.
4. Suppose that the regression model is $y_i = \alpha + \beta x_i + \varepsilon_i$, where the disturbances ε_i have $f(\varepsilon_i) = (1/\lambda)\exp(-\varepsilon_i/\lambda)$, $\varepsilon_i \geq 0$. This model is rather peculiar in that all the disturbances are assumed to be nonnegative. Note that the disturbances have $E[\varepsilon_i|x_i] = \lambda$ and $\text{Var}[\varepsilon_i|x_i] = \lambda^2$. Show that the least squares slope estimator is unbiased but that the intercept estimator is biased.
5. Prove that the least squares intercept estimator in the classical regression model is the minimum variance linear unbiased estimator.
6. As a profit-maximizing monopolist, you face the demand curve $Q = \alpha + \beta P + \varepsilon$. In the past, you have set the following prices and sold the accompanying quantities:

Q	3	3	7	6	10	15	16	13	9	15	9	15	12	18	21
P	18	16	17	12	15	15	4	13	11	6	8	10	7	7	7

Suppose that your marginal cost is 10. Based on the least squares regression, compute a 95% confidence interval for the expected value of the profit-maximizing output.

7. The following sample moments for $x = [1, x_1, x_2, x_3]$ were computed from 100 observations produced using a random number generator:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 100 & 123 & 96 & 109 \\ 123 & 252 & 125 & 189 \\ 96 & 125 & 167 & 146 \\ 109 & 189 & 146 & 168 \end{bmatrix}, \qquad \mathbf{X}'\mathbf{y} = \begin{bmatrix} 460 \\ 810 \\ 615 \\ 712 \end{bmatrix}, \qquad \mathbf{y}'\mathbf{y} = 3924.$$

The true model underlying these data is $y = x_1 + x_2 + x_3 + \varepsilon$.

a. Compute the simple correlations among the regressors.

b. Compute the ordinary least squares coefficients in the regression of y on a constant x_1, x_2, and x_3.

c. Compute the ordinary least squares coefficients in the regression of y on a constant, x_1 and x_2, on a constant, x_1 and x_3, and on a constant, x_2 and x_3.

d. Compute the variance inflation factor associated with each variable.

e. The regressors are obviously badly collinear. Which is the problem variable? Explain.

8. Consider the multiple regression of $\mathbf{y}$ on K variables $\mathbf{X}$ and an additional variable $\mathbf{z}$. Prove that under the assumptions A1 through A6 of the classical regression model, the true variance of the least squares estimator of the slopes on $\mathbf{X}$ is larger when $\mathbf{z}$ is included in the regression than when it is not. Does the same hold for the sample estimate of this covariance matrix? Why or why not? Assume that $\mathbf{X}$ and $\mathbf{z}$ are nonstochastic and that the coefficient on $\mathbf{z}$ is nonzero.

9. For the classical normal regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with no constant term and K regressors, assuming that the true value of β is zero, what is the exact expected value of $F[K, n - K] = (R^2/K)/[(1 - R^2)/(n - K)]$?

10. Prove that $E[\mathbf{b}'\mathbf{b}] = \boldsymbol{\beta}'\boldsymbol{\beta} + \sigma^2 \sum_{k=1}^{K}(1/\lambda_k)$, where $\mathbf{b}$ is the ordinary least squares estimator and λ_k is a characteristic root of $\mathbf{X}'\mathbf{X}$.

11. For the classical normal regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with no constant term and K regressors, what is $\operatorname{plim} F[K, n - K] = \operatorname{plim}\frac{R^2/K}{(1 - R^2)/(n - K)}$, assuming that the true value of $\boldsymbol{\beta}$ is zero?

12. Let e_i be the ith residual in the ordinary least squares regression of $\mathbf{y}$ on $\mathbf{X}$ in the classical regression model, and let ε_i be the corresponding true disturbance. Prove that $\operatorname{plim}(e_i - \varepsilon_i) = 0$.

13. For the simple regression model $y_i = \mu + \varepsilon_i$, $\varepsilon_i \sim N[0, \sigma^2]$, prove that the sample mean is consistent and asymptotically normally distributed. Now consider the alternative estimator $\hat{\mu} = \sum_i w_i y_i$, $w_i = \frac{i}{(n(n+1)/2)} = \frac{i}{\sum_i i}$. Note that $\sum_i w_i = 1$.

Prove that this is a consistent estimator of μ and obtain its asymptotic variance. [*Hint:* $\sum_i i^2 = n(n + 1)(2n + 1)/6$.]

14. Consider a data set consisting of n observations, n_c complete and n_m incomplete, for which the dependent variable, y_i, is missing. Data on the independent variables, $\mathbf{x}_i$, are complete for all n observations, $\mathbf{X}_c$ and $\mathbf{X}_m$. We wish to use the data to estimate the parameters of the linear regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Consider the following the imputation strategy: Step 1: Linearly regress $\mathbf{y}_c$ on $\mathbf{X}_c$ and compute $\mathbf{b}_c$.

Step 2: Use $\mathbf{X}_m$ to predict the missing $\mathbf{y}_m$ with $\mathbf{X}_m\mathbf{b}_c$. Then regress the full sample of observations, $(\mathbf{y}_c, \mathbf{X}_m\mathbf{b}_c)$, on the full sample of regressors, $(\mathbf{X}_c, \mathbf{X}_m)$.

a. Show that the first and second step least squares coefficient vectors are identical.
b. Is the second step coefficient estimator unbiased?
c. Show that the sum of squared residuals is the same at both steps.
d. Show that the second step estimator of σ^2 is biased downward.

15. In (4-13), we find that when superfluous variables $\mathbf{X}_2$ are added to the regression of $\mathbf{y}$ on $\mathbf{X}_1$ the least squares coefficient estimator is an unbiased estimator of the true parameter vector, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \mathbf{0}')'$. Show that, in this long regression, $\mathbf{e}'\mathbf{e}/(n - K_1 - K_2)$ is also unbiased as estimator of σ^2.
16. In Section 4.9.2, we consider regressing $\mathbf{y}$ on a set of principal components, rather than the original data. For simplicity, assume that $\mathbf{X}$ does not contain a constant term, and that the K variables are measured in deviations from the means and are standardized by dividing by the respective standard deviations. We consider regression of $\mathbf{y}$ on L principal components, $\mathbf{Z} = \mathbf{X}\mathbf{C}_L$, where $L < K$. Let $\mathbf{d}$ denote the coefficient vector. The regression model is $\mathbf{y} = \mathbf{X}\beta + \varepsilon$. In the discussion, it is claimed that $E[\mathbf{d}] = \mathbf{C}_L'\beta$. Prove the claim.
17. Example 4.10 presents a regression model that is used to predict the auction prices of Monet paintings. The most expensive painting in the sample sold for \$33.0135M (ln = 17.3124). The height and width of this painting were 35″ and 39.4″, respectively. Use these data and the model to form prediction intervals for the log of the price and then the price for this painting.

Applications

1. Data on U.S. gasoline consumption for the years 1953 to 2004 are given in Table F2.2. Note the consumption data appear as total expenditure. To obtain the per capita quantity variable, divide GASEXP by GASP times Pop. The other variables do not need transformation.
   a. Compute the multiple regression of per capita consumption of gasoline on per capita income, the price of gasoline, the other prices, and a time trend. Report all results. Do the signs of the estimates agree with your expectations?
   b. Test the hypothesis that at least in regard to demand for gasoline, consumers do not differentiate between changes in the prices of new and used cars.
   c. Estimate the own price elasticity of demand, the income elasticity, and the cross-price elasticity with respect to changes in the price of public transportation. Do the computations at the 2004 point in the data.
   d. Reestimate the regression in logarithms so that the coefficients are direct estimates of the elasticities. (Do not use the log of the time trend.) How do your estimates compare with the results in the previous question? Which specification do you prefer?
   e. Compute the simple correlations of the price variables. Would you conclude that multicollinearity is a problem for the regression in part a or part d?
   f. Notice that the price index for gasoline is normalized to 100 in 2000, whereas the other price indices are anchored at 1983 (roughly). If you were to renormalize the indices so that they were all 100.00 in 2004, then how would the results of

the regression in part a change? How would the results of the regression in part d change?

g. This exercise is based on the model that you estimated in part d. We are interested in investigating the change in the gasoline market that occurred in 1973. First, compute the average values of log of per capita gasoline consumption in the years 1953–1973 and 1974–2004 and report the values and the difference. If we divide the sample into these two groups of observations, then we can decompose the change in the expected value of the log of consumption into a change attributable to change in the regressors and a change attributable to a change in the model coefficients, as shown in Section 4.7.2. Using the Oaxaca–Blinder approach described there, compute the decomposition by partitioning the sample and computing separate regressions. Using your results, compute a confidence interval for the part of the change that can be attributed to structural change in the market, that is, change in the regression coefficients.

2. Christensen and Greene (1976) estimated a "generalized Cobb–Douglas" cost function for electricity generation of the form

$$\ln C = \alpha + \beta \ln Q + \gamma[\tfrac{1}{2}(\ln Q)^2] + \delta_k \ln P_k + \delta_l \ln P_l + \delta_f \ln P_f + \varepsilon.$$

P_k, P_l, and P_f indicate unit prices of capital, labor, and fuel, respectively, Q is output, and C is total cost. To conform to the underlying theory of production, it is necessary to impose the restriction that the cost function be homogeneous of degree one in the three prices. This is done with the restriction $\delta_k + \delta_l + \delta_f = 1$, or $\delta_f = 1 - \delta_k - \delta_l$. Inserting this result in the cost function and rearranging terms produces the estimating equation,

$$\ln(C/P_f) = \alpha + \beta \ln Q + \gamma[\tfrac{1}{2}(\ln Q)^2] + \delta_k \ln(P_k/P_f) + \delta_l \ln(P_l/P_f) + \varepsilon.$$

The purpose of the generalization was to produce a U-shaped average total cost curve. We are interested in the **efficient scale**, which is the output at which the cost curve reaches its minimum. That is the point at which $(\partial \ln C/\partial \ln Q)_{|Q=Q^*} = 1$ or $Q^* = \exp[(1 - \beta)/\gamma]$.

a. Data on 158 firms extracted from Christensen and Greene's study are given in Table F4.4. Using all 158 observations, compute the estimates of the parameters in the cost function and the estimate of the asymptotic covariance matrix.

b. Note that the cost function does not provide a direct estimate of δ_f. Compute this estimate from your regression results, and estimate the asymptotic standard error.

c. Compute an estimate of Q^* using your regression results and then form a confidence interval for the estimated efficient scale.

d. Examine the raw data and determine where in the sample the efficient scale lies. That is, determine how many firms in the sample have reached this scale, and whether, in your opinion, this scale is large in relation to the sizes of firms in the sample. Christensen and Greene approached this question by computing the proportion of total output in the sample that was produced by firms that had not yet reached efficient scale. (*Note:* There is some double counting in the data set—more than 20 of the largest "firms" in the sample we are using for this exercise are holding companies and power pools that are aggregates of other firms in the sample. We will ignore that complication for the purpose of our numerical exercise.)

3. The Filipelli data mentioned in Footnote 11 are used to test the accuracy of computer programs in computing least squares coefficients. The 82 observations on (x,y) are given in Appendix Table F4.5. The regression computation involves regression of y on a constant and the first 10 powers of x. (The condition number for this 11-column data matrix is 0.3×10^{10}.) The correct least squares solutions are given on the NIST Website. Using the software you are familiar with, compute the regression using these data.