4
THE LEAST SQUARES ESTIMATOR

4.1 INTRODUCTION

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

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

The analysis assumes that the data in hand correspond to the assumptions of the model. In Section 4.7, we consider several practical problems that arise in analyzing nonexperimental data. Assumption A2, full rank of $\mathbf{X}$, 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 our results on asymptotic distributions. It may be helpful to review Appendix D before proceeding to this material.

TABLE 4.1 Assumptions of the Classical Linear Regression Model

A1. Linearity: $y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{iK}\beta_K$

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

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

A4. Homoscedasticity and nonautocorrelation: Each disturbance, ε_i has the same finite variance, σ^2, and is uncorrelated with every other disturbance, ε_j conditioned on xs.

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

A6. Normal distribution: The disturbances are normally distributed.

data, is examined in Sections 4.7.1 to 4.7.3. Missing data have the potential to derail the entire analysis. The benign case in which missing values are simply manageable random gaps in the data set is considered in Section 4.7.4. The more complicated case of nonrandomly missing data is discussed in Chapter 18. Finally, the problem of badly measured data is examined in Section 4.7.5.

4.2 MOTIVATING LEAST SQUARES

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

4.2.1 THE POPULATION ORTHOGONALITY CONDITIONS

Let $\mathbf{x}$ denote the vector of independent variables in the population regression model and for the moment, based on assumption A5, the data may be stochastic or nonstochastic. Assumption A3 states that the disturbances in the population are stochastically orthogonal to the independent variables in the model; that is, $E[\varepsilon \mid \mathbf{x}] = 0$. It follows that $\text{Cov}[\mathbf{x}, \varepsilon] = \mathbf{0}$. Since (by the law of iterated expectations—Theorem B.1) $E_{\mathbf{x}}\{E[\varepsilon \mid \mathbf{x}]\} = E[\varepsilon] = 0$, we may write this as

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

or

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

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

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

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 sums on the left-hand and right-hand sides of (4-2) are estimators of their counterparts in (4-1). Thus, by using least squares, we are mimicking in the sample the relationship in the population. We'll return to this approach to estimation in Chapters 12 and 13 under the subject of GMM estimation.

4.2.2 MINIMUM MEAN SQUARED ERROR PREDICTOR

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

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

This can be written as

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

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

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

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

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

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

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

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

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

4.2.3 MINIMUM VARIANCE LINEAR UNBIASED ESTIMATION

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

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

4.3 FINITE SAMPLE PROPERTIES OF LEAST SQUARES

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

Example 4.1 The Sampling Distribution of a Least Squares Estimator
The following sampling experiment shows the nature of a sampling distribution and the implication of unbiasedness. We drew two samples of 10,000 random draws on variables w_i and x_i from the standard normal population (mean zero, variance 1). We generated a set of 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, and with each one, computed the least squares slope, using at replication r, $b_r = \left[\Sigma_{j=1}^{100}(x_{ir} - \bar{x}_r)y_{ir}\right] / \left[\Sigma_{j=1}^{100}(x_{ir} - \bar{x}_r)^2\right]$. The histogram in Figure 4.1 shows the result of the experiment. Note that the distribution of slopes has a mean roughly equal to the "true value" of 0.5, and it has a substantial variance, reflecting the fact that the regression slope, like any other statistic computed from the sample, is a random variable. The concept of unbiasedness

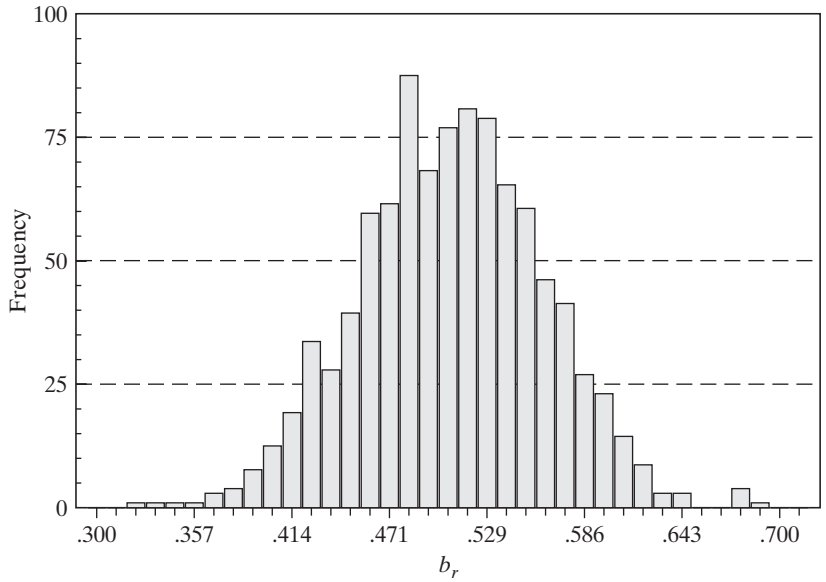


FIGURE 4.1 Histogram for Sampled Least Squares Regression Slopes.

relates to the central tendency of this distribution of values obtained in repeated sampling from the population. The shape of the histogram also suggests the normal distribution of the estimator that we will show theoretically in Section 4.3.8 (The experiment should be replicable with any regression program that provides a random number generator and a means of drawing a random sample of observations from a master data set.)

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 $\mathbf{X}$;

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

By Assumption A3, the second term is $\mathbf{0}$, so

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

Therefore,

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

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

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

4.3.2 BIAS CAUSED BY OMISSION OF RELEVANT VARIABLES

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

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

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

Suppose that a corrrectly specified regression model would be

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}, \tag{4-8}$$

where the two parts of $\mathbf{X}$ have K_1 and K_2 columns, respectively. If we regress $\mathbf{y}$ on $\mathbf{X}_1$ without including $\mathbf{X}_2$, then the estimator is

$$\mathbf{b}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y} = \boldsymbol{\beta}_1 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\boldsymbol{\beta}_2 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\boldsymbol{\varepsilon}. \tag{4-9}$$

Taking the expectation, we see that unless $\mathbf{X}_1'\mathbf{X}_2 = \mathbf{0}$ or $\boldsymbol{\beta}_2 = \mathbf{0}$, $\mathbf{b}_1$ is biased. The well-known result is the **omitted variable formula:**

$$E[\mathbf{b}_1 \mid \mathbf{X}] = \boldsymbol{\beta}_1 + \mathbf{P}_{1.2}\boldsymbol{\beta}_2, \tag{4-10}$$

where

$$\mathbf{P}_{1.2} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2. \tag{4-11}$$

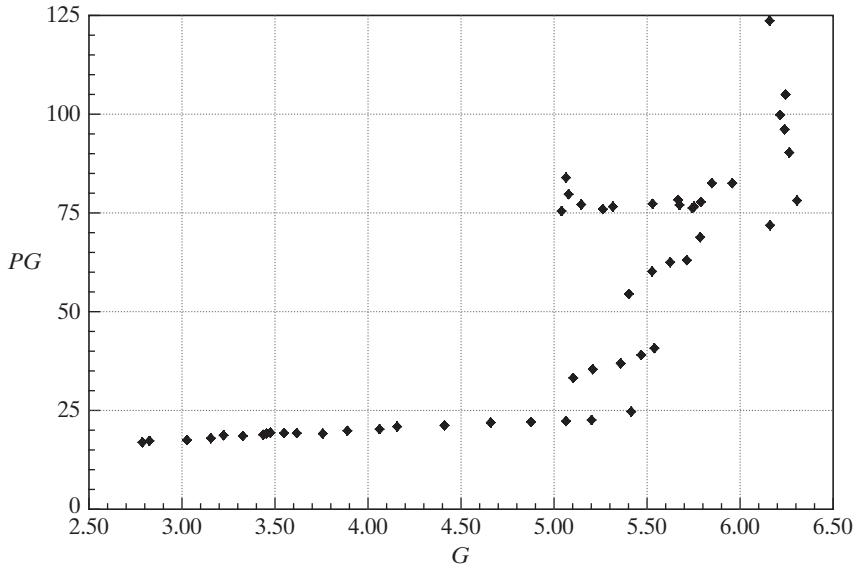


FIGURE 4.2 Per Capita Gasoline Consumption vs. Price, 1953–2004.

Each column of the $K_1 \times K_2$ matrix $\mathbf{P}_{1.2}$ is the column of slopes in the least squares regression of the corresponding column of $\mathbf{X}_2$ on the columns of $\mathbf{X}_1$.

Example 4.2 Omitted Variable

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

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

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

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

4.3.3 INCLUSION OF IRRELEVANT VARIABLES

If the regression model is correctly given by

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon} \tag{4-12}$$

and we estimate it as if (4-8) were correct (i.e., we include some extra variables), then it might seem that the same sorts of problems considered earlier would arise. In fact, this case is not true. We can view the omission of a set of relevant variables as equivalent to imposing an incorrect restriction on (4-8). In particular, omitting $\mathbf{X}_2$ is equivalent to *incorrectly* estimating (4-8) subject to the restriction $\boldsymbol{\beta}_2 = \mathbf{0}$. Incorrectly imposing a restriction produces a biased estimator. Another way to view this error is to note that it amounts to incorporating incorrect information in our estimation. Suppose, however, that our error is simply a failure to use some information that is *correct*.

The inclusion of the irrelevant variables $\mathbf{X}_2$ in the regression is equivalent to failing to impose $\boldsymbol{\beta}_2 = \mathbf{0}$ on (4-8) in estimation. But (4-8) is not incorrect; it simply fails to incorporate $\boldsymbol{\beta}_2 = \mathbf{0}$. Therefore, we do not need to prove formally that the least squares estimator of $\boldsymbol{\beta}$ in (4-8) is unbiased *even given* the restriction; we have already proved it. We can assert on the basis of all our earlier results that

$$E[\mathbf{b} \mid \mathbf{X}] = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \mathbf{0} \end{bmatrix}. \tag{4-13}$$

Then where is the problem? It would seem that one would generally want to "overfit" the model. From a theoretical standpoint, the difficulty with this view is that the failure to use correct information is always costly. In this instance, the cost will be reduced precision of the estimates. As we will show in Section 4.7.1, the covariance matrix in the short regression (omitting $\mathbf{X}_2$) is never larger than the covariance matrix for the estimator obtained in the presence of the superfluous variables.[2] Consider a single-variable comparison. If $\mathbf{x}_2$ is highly correlated with $\mathbf{x}_1$, then incorrectly including $\mathbf{x}_2$ in the regression will greatly inflate the variance of the estimator of $\boldsymbol{\beta}_1$.

4.3.4 THE VARIANCE OF THE LEAST SQUARES ESTIMATOR

If the regressors can be treated as nonstochastic, as they would be in an experimental situation in which the analyst chooses the values in $\mathbf{X}$, then the **sampling variance**

[2]There is no loss if $\mathbf{X}_1'\mathbf{X}_2 = \mathbf{0}$, which makes sense in terms of the information about $\mathbf{X}_1$ contained in $\mathbf{X}_2$ (here, none). This situation is not likely to occur in practice, however.

of the least squares estimator can be derived by treating $\mathbf{X}$ as a matrix of constants. Alternatively, we can allow $\mathbf{X}$ to be stochastic, do the analysis conditionally on the observed $\mathbf{X}$, then consider averaging over $\mathbf{X}$ as we did in obtaining (4-6) from (4-5). Using (4-4) again, we have

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}. \tag{4-14}$$

Since we can write $\mathbf{b} = \boldsymbol{\beta} + \mathbf{A}\boldsymbol{\varepsilon}$, where $\mathbf{A}$ is $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, $\mathbf{b}$ is a linear function of the disturbances, which by the definition we will use, makes it a **linear estimator.** As we have seen, the expected value of the second term in (4-14) is $\mathbf{0}$. Therefore, *regardless of the distribution of* $\boldsymbol{\varepsilon}$, *under our other assumptions,* $\mathbf{b}$ *is a linear, unbiased estimator of* $\boldsymbol{\beta}$. The conditional covariance matrix of the least squares slope estimator is

$$\begin{aligned}
\text{Var}[\mathbf{b} \mid \mathbf{X}] &= E\left[(\mathbf{b} - \boldsymbol{\beta})(\mathbf{b} - \boldsymbol{\beta})' \mid \mathbf{X}\right] \\
&= E\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mid \mathbf{X}\right] \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,E\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \mathbf{X}\right]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2\mathbf{I})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.
\end{aligned} \tag{4-15}$$

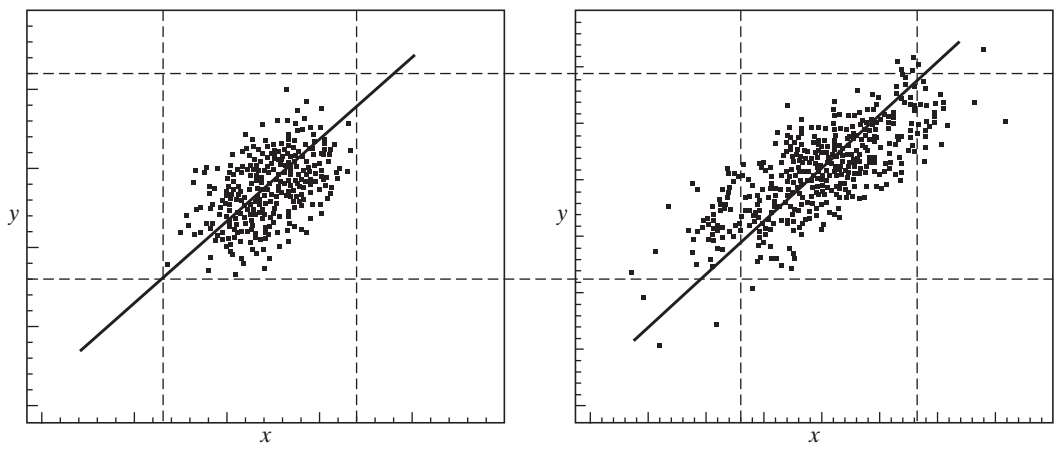
Example 4.3 Sampling Variance in the Two-Variable Regression Model

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

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

Note, in particular, the denominator of the variance of b. The greater the variation in x, the smaller this variance. For example, consider the problem of estimating the slopes of the two regressions in Figure 4.3. A more precise result will be obtained for the data in the right-hand panel of the figure.

FIGURE 4.3 Effect of Increased Variation in x Given the Same Conditional and Overall Variation in y.



4.3.5 THE GAUSS—MARKOV THEOREM

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

THEOREM 4.2 Gauss–Markov Theorem
In the linear regression model with regressor matrix $\mathbf{X}$, *the least squares estimator* $\mathbf{b}$ *is the minimum variance linear unbiased estimator of* $\boldsymbol{\beta}$. *For any vector of constants* $\mathbf{w}$, *the minimum variance linear unbiased estimator of* $\mathbf{w}'\boldsymbol{\beta}$ *in the regression model is* $\mathbf{w}'\mathbf{b}$, *where* $\mathbf{b}$ *is the least squares estimator.*

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

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

$$E[\mathbf{Cy} \mid \mathbf{X}] = E[(\mathbf{CX}\boldsymbol{\beta} + \mathbf{C}\boldsymbol{\varepsilon}) \mid \mathbf{X}] = \boldsymbol{\beta},$$

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

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

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

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

Since a quadratic form in $\mathbf{DD}'$ is $\mathbf{q}'\mathbf{DD}'\mathbf{q} = \mathbf{z}'\mathbf{z} \geq 0$, the conditional covariance matrix of $\mathbf{b}_0$ equals that of $\mathbf{b}$ plus a nonnegative definite matrix. Therefore, every quadratic form in $\text{Var}[\mathbf{b}_0 \mid \mathbf{X}]$ is larger than the corresponding quadratic form in $\text{Var}[\mathbf{b} \mid \mathbf{X}]$, which establishes the first result.

The proof of the second statement follows from the previous derivation, since the variance of $\mathbf{w}'\mathbf{b}$ is a quadratic form in $\text{Var}[\mathbf{b} \mid \mathbf{X}]$, and likewise for any $\mathbf{b}_0$ and proves 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 theorem is much broader than this, however, since the result also applies to every other linear combination of the elements of $\boldsymbol{\beta}$.

4.3.6 THE IMPLICATIONS OF STOCHASTIC REGRESSORS

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

integrating over) the conditional distributions. The crux of the argument is that if we can establish unbiasedness conditionally on an arbitrary $\mathbf{X}$, then we can average over $\mathbf{X}$'s to obtain an unconditional result. We have already used this approach to show the unconditional unbiasedness of $\mathbf{b}$ in Section 4.3.1, so we now turn to the conditional variance.

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

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

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

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

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

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

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

We showed in Section 4.3.5 that

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

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

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

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

The conclusion, therefore, is that the important results we have obtained thus far for the least squares estimator, unbiasedness, and the Gauss–Markov theorem hold whether or not we condition on the particular sample in hand or consider, instead, sampling broadly from the population.

THEOREM 4.3 Gauss–Markov Theorem (Concluded)
In the linear regression model, the least squares estimator $\mathbf{b}$ *is the minimum variance linear unbiased estimator of* $\boldsymbol{\beta}$ *whether* $\mathbf{X}$ *is stochastic or nonstochastic, so long as the other assumptions of the model continue to hold.*

4.3.7 ESTIMATING THE VARIANCE OF THE LEAST SQUARES ESTIMATOR

If we wish to test hypotheses about $\boldsymbol{\beta}$ or to form confidence intervals, then we will require a sample estimate of the covariance matrix $\text{Var}[\mathbf{b} \mid \mathbf{X}] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. The population

parameter σ^2 remains to be estimated. Since σ^2 is the expected value of ε_i^2 and e_i is an estimate of ε_i, by analogy,

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

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

The least squares residuals are

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

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

$$\mathbf{e}'\mathbf{e} = \boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}. \tag{4-16}$$

The expected value of this quadratic form is

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

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

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

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

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

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,

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

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

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

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

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

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

4.3.8 THE NORMALITY ASSUMPTION

To this point, our specification and analysis of the regression model are **semiparametric** (see Section 12.3). We have not used Assumption A6 (see Table 4.1), normality of $\boldsymbol{\varepsilon}$, in any of our results. The assumption is useful for constructing statistics for forming confidence intervals. In (4-4), $\mathbf{b}$ is a linear function of the disturbance vector $\boldsymbol{\varepsilon}$. If we assume that $\boldsymbol{\varepsilon}$ 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

$$\mathbf{b} \mid \mathbf{X} \sim N[\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}]. \tag{4-18}$$

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

$$b_k \mid \mathbf{X} \sim N\left[\beta_k, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{kk}\right]. \tag{4-19}$$

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

The distribution of $\mathbf{b}$ is conditioned on $\mathbf{X}$. The normal distribution of $\mathbf{b}$ in a finite sample is a consequence of our specific assumption of normally distributed disturbances. Without this assumption, and without some alternative specific assumption about the distribution of $\boldsymbol{\varepsilon}$, we will not be able to make any definite statement about the exact distribution of $\mathbf{b}$, conditional or otherwise. In an interesting result that we will explore at length in Section 4.4, we *will* be able to obtain an approximate normal distribution for $\mathbf{b}$, with or without assuming normally distributed disturbances and whether the regressors are stochastic or not.

4.4 LARGE SAMPLE PROPERTIES OF THE LEAST SQUARES ESTIMATOR

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

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

For this basic model, it is also straightforward to derive the large-sample, or asymptotic properties of the least squares estimator. The normality assumption, A6, becomes inessential at this point, and will be discarded save for discussions of maximum likelihood estimation in Section 4.4.6 and in Chapter 14.

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, it has two shortcomings. First, save for the least squares slope estimator we are discussing in this chapter, it is

relatively rare for an econometric estimator to be unbiased. In nearly all cases beyond the multiple regression model, the best one can hope for is that the estimator improves in the sense suggested by unbiasedness as more information (data) is brought to bear on the study. As such, we will need a broader set of tools to guide the econometric inquiry. Second, the property of unbiasedness does not, in fact, imply that more information is better than less in terms of estimation of parameters. The sample means of random samples of 2, 100, and 10,000 are all unbiased estimators of a population mean—by this criterion all are equally desirable. Logically, one would hope that a larger sample is better than a smaller one in some sense that we are about to define (and, by extension, an extremely large sample should be much better, or even perfect). The property of **consistency** improves on unbiasedness in both of these directions.

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

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

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

$$\operatorname*{plim}_{n\to\infty} \frac{\mathbf{X}'\mathbf{X}}{n} = \mathbf{Q}, \quad \text{a positive definite matrix.} \tag{4-20}$$

[We will return to (4-20) shortly.] The least squares estimator may be written

$$\mathbf{b} = \boldsymbol{\beta} + \left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\left(\frac{\mathbf{X}'\boldsymbol{\varepsilon}}{n}\right). \tag{4-21}$$

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

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

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

$$\frac{1}{n}\mathbf{X}'\boldsymbol{\varepsilon} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\varepsilon_i = \frac{1}{n}\sum_{i=1}^{n}\mathbf{w}_i = \overline{\mathbf{w}}. \tag{4-22}$$

Then

$$\operatorname{plim} \mathbf{b} = \boldsymbol{\beta} + \mathbf{Q}^{-1} \operatorname{plim} \overline{\mathbf{w}}.$$

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

$$\operatorname{Var}[\overline{\mathbf{w}} \mid \mathbf{X}] = E[\overline{\mathbf{w}}\,\overline{\mathbf{w}}' \mid \mathbf{X}] = \frac{1}{n}\mathbf{X}'E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \mathbf{X}]\mathbf{X}\frac{1}{n} = \left(\frac{\sigma^2}{n}\right)\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right).$$

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. No variable will degenerate to a sequence of zeros.

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

G3. Let $\mathbf{R}_n$ be the sample correlation matrix of the columns of $\mathbf{X}$, excluding the constant term if there is one. Then $\lim_{n\to\infty} \mathbf{R}_n = \mathbf{C}$, a positive definite matrix. This condition implies that the full rank condition will always be met. We have already assumed that $\mathbf{X}$ has full rank in a finite sample, so this assumption ensures that the condition will never be violated.

Therefore,

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

The variance will collapse to zero if the expectation in parentheses is (or converges to) a constant matrix, so that the leading scalar will dominate the product as n increases. Assumption (4-20) should be sufficient. (Theoretically, the expectation could diverge while the probability limit does not, but this case would not be relevant for practical purposes.) It then follows that

$$\lim_{n\to\infty} \text{Var}[\overline{\mathbf{w}}] = 0 \cdot \mathbf{Q} = \mathbf{0}. \tag{4-23}$$

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

$$\text{plim}\frac{\mathbf{X}'\boldsymbol{\varepsilon}}{n} = \mathbf{0}, \tag{4-24}$$

so

$$\text{plim}\ \mathbf{b} = \boldsymbol{\beta} + \mathbf{Q}^{-1} \cdot \mathbf{0} = \boldsymbol{\beta}. \tag{4-25}$$

This result establishes that under Assumptions A1–A4 and the additional assumption (4-20), $\mathbf{b}$ is a **consistent estimator** of $\boldsymbol{\beta}$ in the linear regression model.

Time-series settings that involve time trends, polynomial time series, and trending variables often pose cases in which the preceding assumptions are too restrictive. A somewhat weaker set of assumptions about $\mathbf{X}$ that is broad enough to include most of these is the **Grenander conditions** listed in Table 4.2.[3] 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.[4]

4.4.2 ASYMPTOTIC NORMALITY OF THE LEAST SQUARES ESTIMATOR

As a guide to estimation, consistency is an improvement over unbiasedness. Since we are in the process of relaxing the more restrictive assumptions of the model, including A6, normality of the disturbances, we will also lose the normal distribution of the

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

[4] White (2001) continues this line of analysis.

estimator that will enable us to form confidence intervals in Section 4.5. It seems that the more general model we have built here has come at a cost. In this section, we will find that normality of the disturbances is not necessary for establishing the distributional results we need to allow statistical inference including confidence intervals and testing hypotheses. Under generally reasonable assumptions about the process that generates the sample data, large sample distributions will provide a reliable foundation for statistical inference in the regression model (and more generally, as we develop more elaborate estimators later in the book).

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

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

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

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

Thus, we must establish the limiting distribution of

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

where $E[\overline{\mathbf{w}}] = \mathbf{0}$. [See (4-22).] We can use the multivariate Lindeberg–Feller version of the central limit theorem (D.19.A) to obtain the limiting distribution of $\sqrt{n}\overline{\mathbf{w}}$.[5] Using that formulation, $\overline{\mathbf{w}}$ is the average of n independent random vectors $\mathbf{w}_i = \mathbf{x}_i\varepsilon_i$, with means $\mathbf{0}$ and variances

$$\operatorname{Var}[\mathbf{x}_i\varepsilon_i] = \sigma^2 E[\mathbf{x}_i\mathbf{x}_i'] = \sigma^2\mathbf{Q}_i. \tag{4-29}$$

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

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

As long as the sum is not dominated by any particular term and the regressors are well behaved, which in this case means that (4-20) holds,

$$\lim_{n\to\infty}\sigma^2\overline{\mathbf{Q}}_n = \sigma^2\mathbf{Q}. \tag{4-31}$$

Therefore, we may apply the Lindeberg–Feller central limit theorem to the vector $\sqrt{n}\,\overline{\mathbf{w}}$, as we did in Section D.3 for the univariate case $\sqrt{n}\overline{x}$. We now have the elements we need for a formal result. If $[\mathbf{x}_i\varepsilon_i]$, $i = 1, \ldots, n$ are independent vectors distributed with

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

mean $\mathbf{0}$ and variance $\sigma^2\mathbf{Q}_i < \infty$, and if (4-20) holds, then

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

It then follows that

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

Combining terms,

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

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

THEOREM 4.4 Asymptotic Distribution of b with Independent 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

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

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 result (4-18), 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.* We will consider other, more general cases in the sections to follow.

4.4.3 CONSISTENCY OF s^2 AND 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$[\mathbf{b}] = (\sigma^2/n)\mathbf{Q}^{-1}$.[6] With (4-20), 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

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

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 (4-20), (4-24) (twice), and the product rule for **probability limits** (Theorem D.14) to assert that the second term

[6] See McCallum (1973) for some useful commentary on deriving the asymptotic covariance matrix of the least squares estimator.

in the brackets converges to 0. That leaves

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

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

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

and, by the product rule,

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

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

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

4.4.4 ASYMPTOTIC DISTRIBUTION OF A FUNCTION OF 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

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

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. Thus, 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 4.5 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* $\boldsymbol{\Gamma} = \partial \mathbf{f}(\boldsymbol{\beta})/\partial \boldsymbol{\beta}'$ *and if Theorem 4.4 holds, then*

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

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

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

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

Example 4.4 Nonlinear Functions of Parameters: The Delta Method
A dynamic version of the demand for gasoline model in Example 2.3 would be used to separate the short- and long-term impacts of changes in income and prices. The model would be

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

where P_{nc} and P_{uc} are price indexes for new and used cars. In this model, the short-run price and income elasticities are β_2 and β_3. The long-run elasticities are $\phi_2 = \beta_2/(1-\gamma)$ and $\phi_3 = \beta_3/(1-\gamma)$, respectively. (See Section 21.3 for development of this model.) To estimate the long-run elasticities, we will estimate the parameters by least squares and then compute these two nonlinear functions of the estimates. We can use the delta method to estimate the standard errors.

Least squares estimates of the model parameters with standard errors and t ratios are given in Table 4.3. 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 partial derivatives of these functions with respect to the six parameters in the model:

$$\mathbf{g}_2' = \partial\phi_2/\partial\boldsymbol{\beta}' = [0, 1/(1-\gamma), 0, 0, 0, \beta_2/(1-\gamma)^2] = [0, 5.91613, 0, 0, 0, -2.43365],$$

$$\mathbf{g}_3' = \partial\phi_3/\partial\boldsymbol{\beta}' = [0, 0, 1/(1-\gamma), 0, 0, \beta_3/(1-\gamma)^2] = [0, 0, 5.91613, 0, 0, 5.74174].$$

Using (4-36), 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.

4.4.5 ASYMPTOTIC EFFICIENCY

We have not established any large-sample counterpart to the Gauss–Markov theorem. That is, it remains to establish whether the large-sample properties of the least squares estimator are optimal by any measure. The Gauss–Markov theorem establishes finite

TABLE 4.3 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	0.01973	−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*** $\mathbf{10^{-n}}$***)***

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

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 shall also be interested in nonlinear estimators. These cases extend beyond the reach of the Gauss–Markov theorem. To make any progress in this direction, we will require an alternative estimation criterion.

DEFINITION 4.1 Asymptotic Efficiency
An estimator is asymptotically efficient if it is consistent, asymptotically normally distributed, and has an asymptotic covariance matrix that is not larger than the asymptotic covariance matrix of any other consistent, asymptotically normally distributed estimator.

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 4.5), that they converge at the same rate—that is, in both cases, the asymptotic variance 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 an example.

Example 4.5 Least Squares vs. Least Absolute Deviations—A Monte Carlo Study

We noted earlier (Section 4.2) that while it enjoys several virtues, least squares is not the only available estimator for the parameters of the linear regresson model. Least absolute deviations (LAD) is an alternative. (The LAD estimator is considered in more detail in Section 7.3.1) The LAD estimator is obtained as

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

in contrast to the linear least squares estimator,

$$\mathbf{b}_{\text{LS}} = \text{the minimizer of } \sum_{i=1}^{n} (y_i - \mathbf{x}_i'\mathbf{b}_0)^2.$$

Suppose the regression model is defined by

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

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 plim$(1/n)\mathbf{X}'\boldsymbol{\varepsilon} = \mathbf{0}$. 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 4.4, 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 $q_k = \pi/2$. Thus, 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 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 up to as much as $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 4.4. 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:[7] 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

[7]Being a Monte Carlo study that uses a random number generator, there is a question of replicability. The study was done with NLOGIT and is replicable. The program can be found on the web site for the text. The qualitative results, if not the precise numerical values, can be reproduced with other programs that allow random sampling from a data set.

TABLE 4.4 Estimated Equations for Art Prices

	Constant		*Log Area*		*Aspect Ratio*	
Full Sample	*Mean*	*Standard Deviation*	*Mean*	*Standard Deviation*	*Mean*	*Standard Deviation*
LS	−8.42653	0.61184	1.33372	0.09072	−0.16537	0.12753
LAD	−7.62436	0.89055	1.20404	0.13626	−0.21260	0.13628
R = 10						
LS	−9.39384	6.82900	1.40481	1.00545	0.39446	2.14847
LAD	−8.97714	10.24781	1.34197	1.48038	0.35842	3.04773
R = 50						
LS	−8.73099	2.12135	1.36735	0.30025	−0.06594	0.52222
LAD	−8.91671	2.51491	1.38489	0.36299	−0.06129	0.63205
R = 100						
LS	−8.36163	1.32083	1.32758	0.17836	−0.17357	0.28977
LAD	−8.05195	1.54190	1.27340	0.21808	−0.20700	0.29465

the 100 vectors and the sample variance of the 100 observations.[8] 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 full sample LAD estimates (not the mean of the B estimates). This procedure is discussed in detail in Section 15.4.

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

1. Since both estimators are consistent, the averages should resemble the preceding main 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, a disadvantage of the LS estimator should become apparent.

A kernel density estimate for the distribution of the least squares residuals appears in Figure 4.4. 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 4.4. 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 $\sqrt{1.573} = 1.254$.

[8]Note that 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 4.4. 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 are a bit less than 10^{20} different samples of 10 observations we can draw. The number of different samples of 50 or 100 is essentially infinite.

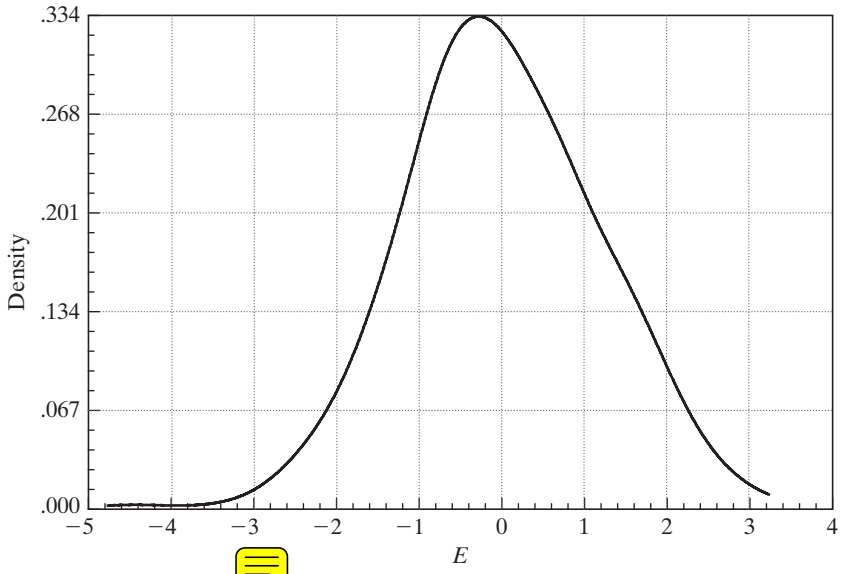


FIGURE 4.4 Kernal Dansity Estimator for Least Squares Residuals.

4.4.6 MAXIMUM LIKELIHOOD ESTIMATION

We have motivated the least squares estimator in two ways: First, we obtained Theorem 4.1 which states that the least squares estimator mimics the coefficients in the minimum mean squared error predictor of y in the joint distribution of y and $\mathbf{x}$. Second, Theorem 4.2, the Gauss–Markov theorem, states that the least squares estimator is the *minimum variance linear unbiased* estimator of $\boldsymbol{\beta}$ under the assumptions of the model. Neither of these results relies on Assumption A6, normality of the distribution of ε. A natural question at this point would be, what is the role of this assumption? There are two. First, the assumption of normality will produce the basis for determining the appropriate endpoints for confidence intervals in Sections 4.5 and 4.6. But, we found in Section 4.4.2 that based on the central limit theorem, we could base inference on the asymptotic normal distribution of $\mathbf{b}$, even if the disturbances were not normally distributed. That would seem to make the normality assumption no longer necessary, which is largely true but for a second result.

If the disturbances are normally distributed, then the least squares estimator is also the **maximum likelihood estimator (MLE)**. We will examine maximum likelihood estimation in detail in Chapter 14, so we will describe it only briefly at this point. The end result is that by virtue of being an MLE, least squares is *asymptotically efficient among consistent and asymptotically normally distributed estimators*. This is a large sample counterpart to the Gauss–Markov theorem (known formally as the Cramér–Rao bound). What the two theorems have in common is that they identify the least squares estimator as the most efficient estimator in the assumed class of estimators. They differ in the class of estimators assumed:

Gauss–Markov:	Linear and unbiased estimators
ML:	Based on normally distributed disturbances, consistent and asymptotically normally distributed estimators

These are not "nested." Notice, for example, that the MLE result does not require unbiasedness or linearity. Gauss–Markov does not require normality or consistency. The Gauss–Markov theorem is a finite sample result while the Cramér–Rao bound is an asymptotic (large-sample) property. The important aspect of the development concerns the efficiency property. Efficiency, in turn, relates to the question of how best to use the sample data for statistical inference. In general, it is difficult to establish that an estimator is efficient without being specific about the candidates. The Gauss–Markov theorem is a powerful result for the linear regression model. However, it has no counterpart in any other modeling context, so once we leave the linear model, we will require different tools for comparing estimators. The principle of maximum likelihood allows the analyst to assert asymptotic efficiency for the estimator, but only for the specific distribution assumed. Example 4.6 establishes that **b** is the MLE in the regression model with normally distributed disturbances. Example 4.7 then considers a case in which the regression disturbances are not normally distributed and, consequently, **b** is less efficient than the MLE.

Example 4.6 MLE with Normally Distributed Disturbances

With normally distributed disturbances, $y_i|\mathbf{x}_i$ is normally distributed with mean $\mathbf{x}_i'\boldsymbol{\beta}$ and variance σ^2, so the density of $y_i|\mathbf{x}_i$ is

$$f(y_i|\mathbf{x}_i) = \frac{\exp\left[-\frac{1}{2}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2\right]}{\sqrt{2\pi\sigma^2}}$$

The log likelihood for a sample of n independent observations is equal to the log of the joint density of the observed random variables. For a random sample, the joint density would be the product, so the log likelihood, given the data, which is written $\ln L(\boldsymbol{\beta}, \sigma^2|\mathbf{y},\mathbf{X})$ would be the sum of the logs of the densities. This would be (after a bit of manipulation)

$$\ln L(\boldsymbol{\beta}, \sigma^2|\mathbf{y},\mathbf{X}) = -(n/2)[\ln\sigma^2 + \ln 2\pi + (1/\sigma^2)\tfrac{1}{n}\Sigma_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2].$$

The values of $\boldsymbol{\beta}$ and σ^2 that maximize this function are the maximum likelihood estimators of $\boldsymbol{\beta}$ and σ^2. As we will explore further in Chapter 14, the functions of the data that maximize this function with respect to $\boldsymbol{\beta}$ and σ^2 are the least squares coefficient vector, **b**, and the mean squared residual, $\mathbf{e}'\mathbf{e}/n$. Once again, we leave for Chapter 14 a derivation of the following result,

$$\text{Asy.Var}\left[\hat{\boldsymbol{\beta}}_{ML}\right] = -E[\partial^2 \ln L/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}']^{-1} = \sigma^2 E[(\mathbf{X}'\mathbf{X})^{-1}],$$

which is exactly what appears in Section 4.3.6. This shows that the least squares estimator is the maximum likelihood estimator. It is consistent, asymptotically (and exactly) normally distributed, and, under the assumption of normality, by virtue of Theorem 14.4, asymptotically efficient.

It is important to note that the properties of an MLE depend on the specific distribution assumed for the observed random variable. If some nonnormal distribution is specified for ε and it emerges that **b** is not the MLE, then least squares may not be efficient. The following example illustrates.

Example 4.7 The Gamma Regression Model

Greene (1980a) considers estimation in a regression model with an asymmetrically distributed disturbance,

$$y = (\alpha + \sigma\sqrt{P}) + \mathbf{x}'\boldsymbol{\beta} + (\varepsilon - \sigma\sqrt{P}) = \alpha^* + \mathbf{x}'\boldsymbol{\beta} + \varepsilon^*,$$

where ε has the gamma distribution in Section B.4.5 [see (B-39)] and $\sigma = \sqrt{P}/\lambda$ is the standard deviation of the disturbance. In this model, the covariance matrix of the least squares estimator of the slope coefficients (not including the constant term) is

$$\text{Asy. Var}[\mathbf{b} \mid \mathbf{X}] = \sigma^2(\mathbf{X}'\mathbf{M}^0\mathbf{X})^{-1},$$

whereas for the maximum likelihood estimator (which is not the least squares estimator),[9]

$$\text{Asy. Var}[\hat{\boldsymbol{\beta}}_{ML}] \approx [1 - (2/P)]\sigma^2(\mathbf{X}'\mathbf{M}^0\mathbf{X})^{-1}.$$

But for the asymmetry parameter, this result would be the same as for the least squares estimator. We conclude that the estimator that accounts for the asymmetric disturbance distribution is more efficient asymptotically.

Another example that is somewhat similar to the model in Example 4.7 is the stochastic frontier model developed in Chapter 18. In these two cases in particular, the distribution of the disturbance is asymmetric. The maximum likelihood estimators are computed in a way that specifically accounts for this while the least squares estimator treats observations above and below the regression line symmetrically. That difference is the source of the asymptotic advantage of the MLE for these two models.

4.5 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

$$\hat{\theta} \pm \text{ sampling variability.} \qquad \textbf{(4-37)}$$

(We are assuming that the interval of interest would be symmetic 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 percent) 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 percent) 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), $100(1 - \alpha)$ percent, to the interval in (4-37). We consider how to find that range and then apply the procedure to three familiar problems, interval estimation 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 purpose, we depart from Assumption A6 that the disturbances are normally distributed. We will then relax that assumption and rely instead on the asymptotic normality of the estimator.

[9]The matrix $\mathbf{M}^0$ produces data in the form of deviations from sample means. (See Section A.2.8.) In Greene's model, P must be greater than 2.

4.5.1 FORMING A CONFIDENCE INTERVAL FOR A COEFFICIENT

From(4-18), we have that $\mathbf{b}|\mathbf{X} \sim N[\boldsymbol{\beta},\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]$. It follows that for any particular element of $\mathbf{b}$, say b_k,

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

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

$$z_k = \frac{b_k - \beta_k}{\sqrt{\sigma^2 S^{kk}}} \tag{4-38}$$

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; z_k is a **pivotal statistic**. Using our conventional 95 percent confidence level, we know that Prob$[-1.96 \leq z_k \leq 1.96]$. By a simple manipulation, we find that

$$\text{Prob}\left[b_k - 1.96\sqrt{\sigma^2 S^{kk}} \leq \beta_k \leq b_k + 1.96\sqrt{\sigma^2 S^{kk}}\right] = 0.95. \tag{4-39}$$

Note that this is a statement about the probability that the random interval $b_k\pm$ the sampling variability contains $\boldsymbol{\beta}_k$, not the probability that $\boldsymbol{\beta}_k$ lies in the specified interval. If we wish to use some other level of confidence, not 95 percent, then the 1.96 in (4-39) 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, Prob$[z \leq z_{(1-\alpha/2)}] = 1 - \alpha/2$. Thus, $z_{0.975} = 1.96$, which corresponds to $\alpha = 0.05$.)

We would have our desired confidence interval in (4-39), save for the complication that σ^2 is not known, so the interval is not operational. It would seem natural to use s^2 from the regression. This is, indeed, an appropriate approach. The quantity

$$\frac{(n-K)s^2}{\sigma^2} = \frac{\boldsymbol{e}'\boldsymbol{e}}{\sigma^2} = \left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right)' \mathbf{M} \left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right) \tag{4-40}$$

is an idempotent quadratic form in a standard normal vector, $(\boldsymbol{\varepsilon}/\sigma)$. Therefore, it has a chi-squared distribution with degrees of freedom equal to the rank$(\mathbf{M})$ = trace$(\mathbf{M})$ = $n-K$. (See Section B11.4 for the proof of this result.) The chi-squared variable in (4-40) is independent of the standard normal variable in (38). To prove this, it suffices to show that

$$\left(\frac{\mathbf{b} - \boldsymbol{\beta}}{\sigma}\right) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right)$$

is independent of $(n-K)s^2/\sigma^2$. In Section B.11.7 (Theorem B.12), we found that a sufficient condition for the independence of a linear form $\mathbf{Lx}$ and a[illegible]empotent quadratic form $\mathbf{x}'\mathbf{Ax}$ in a standard normal vector $\mathbf{x}$ is that $\mathbf{LA} = \mathbf{0}$. Letting $\boldsymbol{\varepsilon}/\sigma$ be the $\mathbf{x}$, we find that the requirement here would be that $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{M} = \mathbf{0}$. It does, as seen in (3-15). The general result is central in the derivation of many test statistics in regression analysis.

THEOREM 4.6 Independence of b and s^2
If $\boldsymbol{\varepsilon}$ is normally distributed, then the least squares coefficient estimator **b** *is statistically independent of the residual vector* **e** *and therefore, all functions of* **e**, *including* s^2.

Therefore, the ratio

$$t_k = \frac{(b_k - \beta_k)/\sqrt{\sigma^2 S^{kk}}}{\sqrt{[(n-K)s^2/\sigma^2]/(n-K)}} = \frac{b_k - \beta_k}{\sqrt{s^2 S^{kk}}} \tag{4-41}$$

has a t distribution with $(n - K)$ degrees of freedom.[10] We can use t_k to test hypotheses or form confidence intervals about the individual elements of $\boldsymbol{\beta}$.

The result in (4-41) differs from (38) in the use of s^2 instead of σ^2, and in the pivotal distribution, t with $(n - K)$ degrees of freedom, rather than standard normal. It follows that a confidence interval for β_k can be formed using

$$\text{Prob}\left[b_k - t_{(1-\alpha/2),[n-K]}\sqrt{s^2 S^{kk}} \le \beta_k \le b_k + t_{(1-\alpha/2),[n-K]}\sqrt{s^2 S^{kk}}\right] = 1 - \alpha, \tag{4-42}$$

where $t_{(1-\alpha/2),[n-K]}$ is the appropriate critical value from the t distribution. Here, the distribution of the pivotal statistic depends on the sample size through $(n - K)$, but, once again, not on the parameters or the data. The practical advantage of (4-42) is that it does not involve any unknown parameters. A confidence interval for β_k can be based on (4-42)

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 4.5.

TABLE 4.5 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	−0.021206	0.04377	−0.485
ln $Income/Pop$	1.095874	0.07771	14.102
ln P_{nc}	−0.373612	0.15707	−2.379
ln P_{uc}	0.02003	0.10330	0.194

[10] 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.

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 percent critical value is 2.012. Therefore a 95 percent confidence interval for β_3 is $1.095874 \pm 2.012\ (0.07771) = [0.9395, 1.2522]$.

4.5.2 CONFIDENCE INTERVALS BASED ON LARGE SAMPLES

If the disturbances are not normally distributed, then the development in the previous section, which departs from this assumption, is not usable. But, the large sample results in Section 4.4 provide an alternative approach. Based on the development that we used to obtain Theorem 4.4 and (4-35), we have that the limiting distribution of the statistic

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

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. And, of course, we estimate $\mathbf{Q}$ with $(\mathbf{X}'\mathbf{X}/n)^{-1}$. This gives us precisely (4-41), which states that under the assumptions in Section 4.4, the "t" statistic in (4-41) 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 (4-42) 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 (4-42) 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 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 formed 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 percent 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.

Example 4.9 Confidence Interval Based on the Asymptotic Distribution

In Example 4.4, we analyzed a dynamic form of the demand equation for gasoline,

$$\ln(G/Pop)_t = \beta_1 + \beta_2 \ln P_{G,t} + \beta_3 \ln(Income/Pop) + \cdots + \gamma \ln(G/POP)_{t-1} + \varepsilon_t.$$

In this model, the long-run price and income elasticities are $\theta_P = \beta_2/(1-\gamma)$ and $\theta_I = \beta_3/(1-\gamma)$. We computed estimates of these two nonlinear functions using the least squares and the delta method, Theorem 4.5. The point estimates were -0.411358 and 0.970522, respectively. The estimated asymptotic standard errors were 0.152296 and 0.162386. In order to form confidence intervals for θ_P and θ_I, we would generally use the asymptotic distribution, not the finite-sample distribution. Thus, the two confidence intervals are

$$\hat{\theta}_P = -0.411358 \pm 1.96(0.152296) = [-0.709858, -0.112858]$$

and

$$\hat{\theta}_I = 0.970523 \pm 1.96(0.162386) = [0.652246, 1.288800].$$

In a sample of 51 observations, one might argue that using the critical value for the limiting normal distribution might be a bit optimistic. If so, using the critical value for the t distribution with $51 - 6 = 45$ degrees of freedom would give a slightly wider interval. For example, for the the income elasticity the interval would be $0.970523 \pm 2.014(0.162386) = [0.643460, 1.297585]$. We do note this is a practical adjustment. The statistic based on the asymptotic standard error does not actually have a t distribution with 45 degrees of freedom.

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

With normally distributed disturbances, the least squares coefficient estimator, $\mathbf{b}$, is normally distributed with mean $\boldsymbol{\beta}$ and covariance matrix $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. In Example 4.8, we showed how to use this result 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 linear function of the parameters. **Oaxaca's** (1973) and **Blinder's** (1973) **decomposition** provides a frequently used application.[11]

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

$$\text{Prob}[c - t_{(1-\alpha/2),[n-k]}s_c \leq \gamma \leq c + t_{(1-\alpha/2),[n-k]}s_c] = 1 - \alpha. \qquad \textbf{(4-43)}$$

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

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

and

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

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,

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

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

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

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 covariance matrices $\sigma_m^2(\mathbf{X}_m'\mathbf{X}_m)^{-1}$ and $\sigma_f^2(\mathbf{X}_f'\mathbf{X}_f)^{-1}$. 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

$$\text{Est. Var}[\mathbf{d}] = s_m^2(\mathbf{X}_m'\mathbf{X}_m)^{-1} + s_f^2(\mathbf{X}_f'\mathbf{X}_f)^{-1}. \tag{4-44}$$

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 $\mathbf{w}'\mathbf{b}_f$.

4.6 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.

4.6.1 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

$$\hat{y}^0 = \mathbf{x}^{0\prime}\mathbf{b} \tag{4-45}$$

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 is

$$\text{Var}[e^0|\mathbf{X}, \mathbf{x}^0] = \sigma^2 + \text{Var}[(\mathbf{b} - \boldsymbol{\beta})'\mathbf{x}^0|\mathbf{X}, \mathbf{x}^0] = \sigma^2 + \mathbf{x}^{0\prime}\left[\sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right]\mathbf{x}^0. \tag{4-46}$$

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

$$\text{Var}[e^0|\mathbf{X}, \mathbf{x}^0] = \sigma^2\left[1 + \frac{1}{n} + \sum_{j=1}^{K-1}\sum_{k=1}^{K-1}\left(x_j^0 - \bar{x}_j\right)\left(x_k^0 - \bar{x}_k\right)\left(\mathbf{Z}'\mathbf{M}^0\mathbf{Z}\right)^{jk}\right], \tag{4-47}$$

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 4.5 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

$$\text{prediction interval} = \hat{y}^0 \pm t_{(1-\alpha/2),[n-K]}se\left(e^0\right) \tag{4-48}$$

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

4.6.2 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

$$\ln Price = \beta_1 + \beta_2 \ln Area + \beta_3 AspectRatio + \varepsilon$$

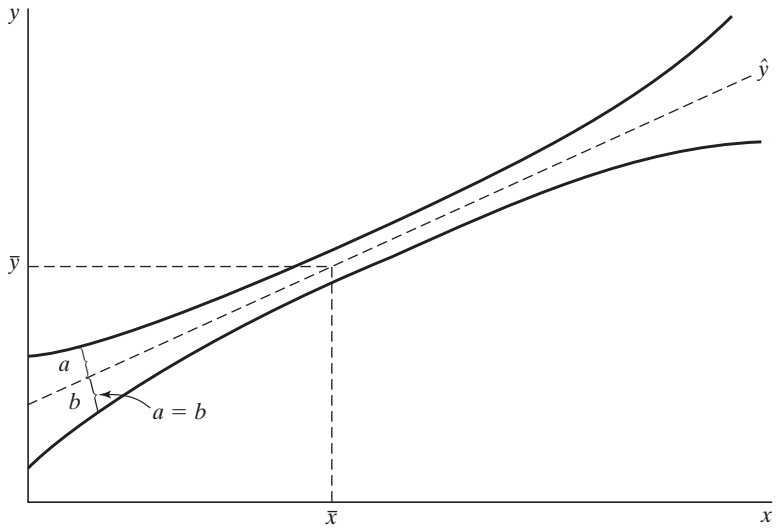


FIGURE 4.5 Prediction Intervals.

(*Area* is width times height of the painting and aspect ratio is the height divided by the width.) 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, $\beta_2 = \partial E[\ln Price|\ln Area, AspectRatio]/\partial \ln Area$. 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 $E[y|\mathbf{x}^0]$ is not equal to $\exp(E[\ln y|\mathbf{x}^0])$. The appropriate conditional mean function would be

$$\begin{aligned} E[y|\mathbf{x}^0] &= E[\exp(\mathbf{x}^{0\prime}\boldsymbol{\beta} + \varepsilon^0)|\mathbf{x}^0] \\ &= \exp(\mathbf{x}^{0\prime}\boldsymbol{\beta})\, E[\exp(\varepsilon^0)|\mathbf{x}^0]. \end{aligned}$$

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

$$\hat{y}^0 = \exp(\mathbf{x}^{0\prime}\mathbf{b} + s^2/2) > \exp(\mathbf{x}^{0\prime}\mathbf{b}), \qquad \textbf{(4-49)}$$

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

$$\hat{E}[\exp(\varepsilon^0)|\mathbf{x}^0] = h^0 = \frac{1}{n}\sum_{i=1}^{n}\exp(e_i), \tag{4-50}$$

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

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

4.6.3 PREDICTION INTERVAL FOR *y* WHEN THE REGRESSION MODEL DESCRIBES LOG *y*

We obtained a prediction interval in (4-48) for $\ln y|\mathbf{x}^0$ in the loglinear model $\ln y = \mathbf{x}'\boldsymbol{\beta} + \varepsilon$,

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

For a given choice of α, say, 0.05, these values give the .025 and .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;

$$\left[\hat{y}^0_{LOWER}, \hat{y}^0_{UPPER}\right] = \left\{\exp\left[\mathbf{x}^{0\prime}\mathbf{b} - t_{(1-\alpha/2),[n-K]}se\left(e^0\right)\right], \exp\left[\mathbf{x}^{0\prime}\mathbf{b} + t_{(1-\alpha/2),[n-K]}se\left(e^0\right)\right]\right\}. \tag{4-51}$$

This follows from the result that if $\text{Prob}[\ln y \le \ln L] = 1 - \alpha/2$, then $\text{Prob}[y \le L] = 1 - \alpha/2$. 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 4.6 shows why. We suppose that (L, U) in the figure is the prediction interval formed by (4-51). 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, 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 (4-51) 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,

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

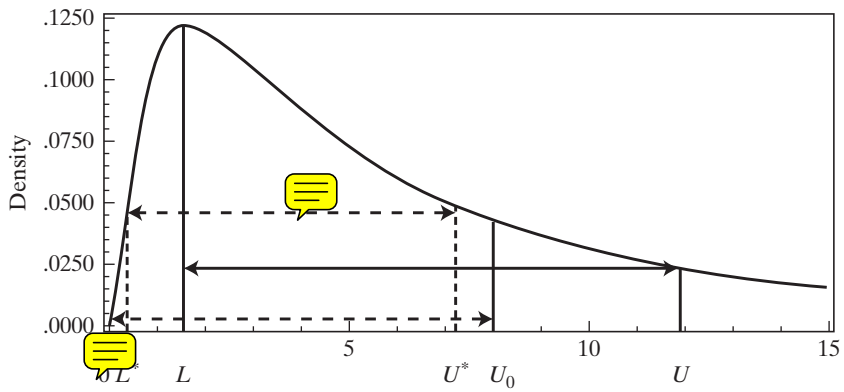
The solutions are found by equating the three partial derivatives to zero:

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

where $f(L) = F'(L)$ and $f(U) = F'(U)$ are the derivtives 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 and the densities are equal. Looking again at Figure 4.6, 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 (4-51). It provides the correct solution for when the distribution is asymmetric. In Bayesian analysis, the counterpart when we examine the distribution of a parameter

FIGURE 4.6 Lognorms Distrution for Prices of Monet Paintings.



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

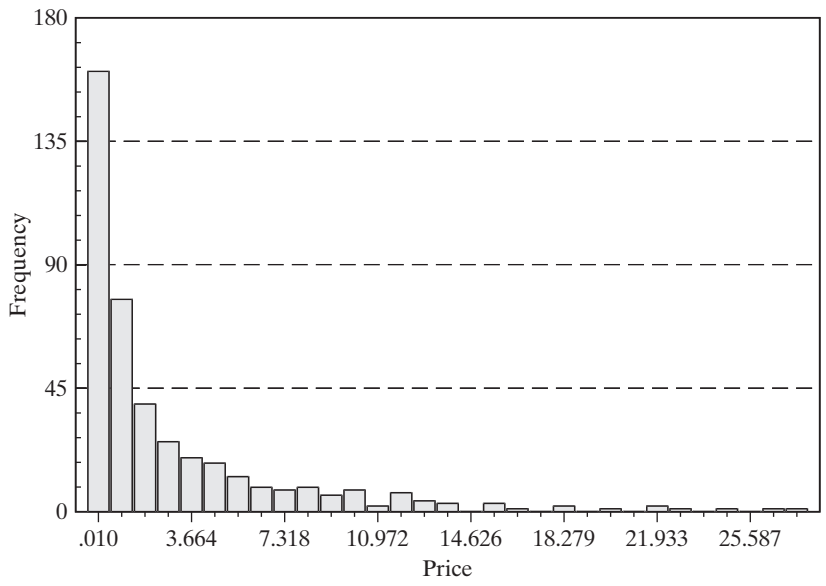
In Example 4.5, we suggested an intriguing feature of the market for Monet paintings, that larger paintings sold at auction for more than than smaller ones. In this example, we will examine that proposition empirically. Table F4.1 contains data on 430 auction prices for Monet paintings, with data on the dimensions of the paintings and several other variables that we will examine in later examples. Figure 4.7 shows a histogram for the sample of sale prices (in $million). Figure 4.8 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 4.6.

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 60 in the sample.) The log area equals $\ln(25.6 \times 31.9) = 6.705198$ and the aspect ratio equals $25.6/31.9 = 0.802508$. The prediction for the log of the price would be

$$\ln P|\mathbf{x}^0 = -8.42653 + 1.33372(6.705198) - 0.16537(0.802508) = 0.383636.$$

FIGURE 4.7 Histogram for Sale Prices of 430 Monet Paintings ($million).



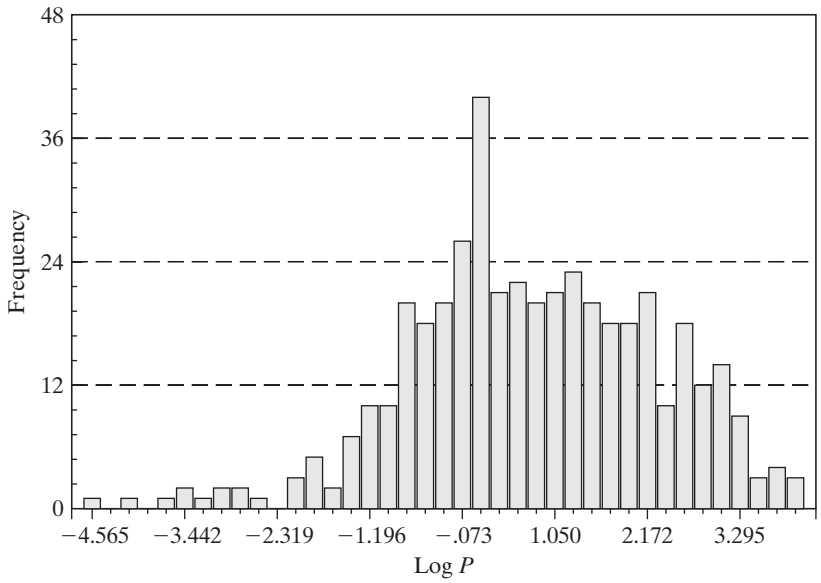


FIGURE 4.8 Histogram of Logs of Auction Prices for Monet Paintings.

TABLE 4.6 Estimated Equation for Log Price

Mean of log Price	.33274
Sum of squared residuals	519.17235
Standard error of regression	1.10266
R-squared	.33620
Adjusted *R*-squared	.33309
Number of observations	430

Variable	*Coefficient*	*Standard Error*	*t*	*Mean of X*
Constant	−8.42653	.61183	−13.77	1.00000
LOGAREA	1.33372	.09072	14.70	6.68007
ASPECT	−.16537	.12753	−1.30	0.90759

Estimated	*Asymptotic Constant*	*Covariance LogArea*	*Matrix AspectRatio*
Constant	.37434	−.05429	−.00974
LogArea	−.05429	.00823	−.00075
AspectRatio	−.00974	−.00075	.01626

Note that the mean log price is 0.33274, so this painting is expected to be sell for roughly 5 percent more than the average painting, based on its dimensions. The estimate of the prediction variance is computed using (4-47); $s_p = 1.104027$. The sample is large enough to use the critical value from the standard normal table, 1.96, for a 95 percent confidence

interval. A prediction interval for the log of the price is therefore

$$0.383636 \pm 1.96(1.104027) = [-1.780258, 2.547529].$$

For predicting the price, the naïve predictor would be exp(0.383636) = $1.476411M, which is far under the actual sale price of $3.5225M. To compute the smearing estimator, we require the mean of the exponents of the residuals, which is 1.813045. The revised point estimate for the price would thus be 1.813045 × 1.47641 = $2.660844M—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.

To compute an interval estimate for the price, we begin with the naïve prediction by simply exponentiating the lower and upper values for the log price, which gives a prediction interval for 95 percent confidence of [$0.168595M, $12.77503M]. Using the method suggested in Section 4.6.3, however, we are able to narrow this interval to [0.021261, 9.027], a range of $9M compared to the range based on the simple calculation of $12.2M. The interval divides the .05 tail probability into 0.00063 on the left and .04937 on the right. The search algorithm is outlined next.

Grid Search Algorithm for Optimal Prediction Interval [LO, UO]
$\mathbf{x}^0 = (1, log(25.6 \times 31.9), 25.6/31.9)'$;
$\hat{\mu}^0 = exp(\mathbf{x}^{0\prime}\mathbf{b})$, $\hat{\sigma}_p^0 = \sqrt{s^2 + \mathbf{x}^{0\prime}[s^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{x}^0}$;
Confidence interval for $\log P|\mathbf{x}^0$: [Lower, Upper] = $[\hat{\mu}^0 - 1.96\hat{\sigma}_p^0, \hat{\mu}^0 + 1.96\hat{\sigma}_p^0]$;
Naïve confidence interval for Price$|\mathbf{x}^0$: L1 = exp(Lower) ; U1 = exp(Upper);
Initial value of L was .168595, LO = this value;
Grid search for optimal interval, decrement by $\Delta = .005$ (chosen ad hoc);
Decrement LO and compute companion UO until densities match;
(*) LO = LO − Δ = new value of LO;
$f(\text{LO}) = \left[LO\hat{\sigma}_p^0\sqrt{2\pi}\right]^{-1} \exp\left[-\frac{1}{2}\left((\ln LO - \hat{\mu}^0)/\hat{\sigma}_p^0\right)^2\right]$;
$F(\text{LO}) = \Phi((\ln(\text{LO}) - \hat{\mu}^0)/\hat{\sigma}_p^0)$ = left tail probability;
$\text{UO} = \exp(\hat{\sigma}_p^0\Phi^{-1}[F(\text{LO}) + .95] + \hat{\mu}^0)$ = next value of UO;
$f(\text{UO}) = \left[UO\hat{\sigma}_p^0\sqrt{2\pi}\right]^{-1} \exp\left[-\frac{1}{2}\left((\ln UO - \hat{\mu}^0)/\hat{\sigma}_p^0\right)^2\right]$;
$1 - F(\text{UO}) = 1 - \Phi((\ln(\text{UO}) - \hat{\mu}^0)/\hat{\sigma}_p^0)$ = right tail probability;
Compare f(LO) to f(UO). If not equal, return to (*). If equal, exit.

4.6.4 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. Most 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 (4-46) 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.[12] 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,**

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

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 measures, 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:**[13]

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

This measure is related to R^2 but is not bounded by zero and one. Large values indicate a poor forecasting performance. An alternative is to compute the measure in terms of the changes in y:

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

where $\Delta y_i = y_i - y_{i-1}$ and $\Delta \hat{y}_i = \hat{y}_i - y_{i-1}$, or, in percentage changes, $\Delta y_i = (y_i - y_{i-1})/y_{i-1}$ and $\Delta \hat{y}_i = (\hat{y}_i - y_{i-1})/y_{i-1}$. These measures will reflect the model's ability to track turning points in the data.

4.7 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 in Table 2.1 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

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

[13] Theil (1961).

model parameters—unfortunately, none of them good. The cases we will examne 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 $\mathbf{X}$ and/or $\mathbf{y}$ 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.

4.7.1 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,

$$\text{Var}[b_k \mid \mathbf{X}] = \frac{\sigma^2}{\left(1 - r_{12}^2\right) \sum_{i=1}^{n} (x_{ik} - \overline{x}_k)^2} = \frac{\sigma^2}{\left(1 - r_{12}^2\right) S_{kk}}, \quad k = 1, 2. \tag{4-52}$$

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

$$\begin{aligned}\left(\mathbf{x}_k'\mathbf{M}_{(k)}\mathbf{x}_k\right)^{-1} &= \left[\mathbf{x}_k'\mathbf{x}_k - \mathbf{x}_k'\mathbf{X}_{(k)}\left(\mathbf{X}_{(k)}'\mathbf{X}_{(k)}\right)^{-1}\mathbf{X}_{(k)}'\mathbf{x}_k\right]^{-1} \\ &= \left[\mathbf{x}_k'\mathbf{x}_k\left(1 - \frac{\mathbf{x}_k'\mathbf{X}_{(k)}\left(\mathbf{X}_{(k)}'\mathbf{X}_{(k)}\right)^{-1}\mathbf{X}_{(k)}'\mathbf{x}_k}{\mathbf{x}_k'\mathbf{x}_k}\right)\right]^{-1} \\ &= \frac{1}{\left(1 - R_{k.}^2\right)S_{kk}},\end{aligned} \tag{4-53}$$

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

$$\text{Var}[b_k \mid \mathbf{X}] = \frac{\sigma^2}{\left(1 - R_{k.}^2\right)\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2}, \tag{4-54}$$

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. This result is shown in Figure 4.3.
- 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. We have yet to develop this implication, but it can be suggested by Figure 4.3 by imagining the identical figure in the right panel but with all the points moved closer to the regression line.

Since nonexperimental data will never be orthogonal ($R_{k.}^2 = 0$), 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), $1/(1 - R_{k.}^2)$, 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 $\mathbf{X}'\mathbf{X}$ (after scaling each column so that it has unit length) to the smallest. Values in excess of 20 are suggested as indicative of a problem [Belsley, Kuh, and Welsch (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.) 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

TABLE 4.7 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

be unusual. But, the results in Table 4.7 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 percent, and the third rises by 800 percent.

Several strategies have been proposed for finding and coping with multicollinearity.[14] 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, 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. In doing so, one encounters the problems of specification that we will discuss in Section 4.7.2. 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.

4.7.2 PRETEST ESTIMATION

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, if it seems to be the one causing the problem. This "strategy" creates a subtle dilemma for the analyst. Consider the partitioned multiple regression

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}.$$

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

If we regress $\mathbf{y}$ only on $\mathbf{X}_1$, the estimator is biased;

$$E[\mathbf{b}_1|\mathbf{X}] = \boldsymbol{\beta}_1 + \mathbf{P}_{1.2}\boldsymbol{\beta}_2.$$

The covariance matrix of this estimator is

$$\text{Var}[\mathbf{b}_1|\mathbf{X}] = \sigma^2(\mathbf{X}_1'\mathbf{X}_1)^{-1}.$$

(Keep in mind, this variance is around the $E[\mathbf{b}_1|\mathbf{X}]$, not around $\boldsymbol{\beta}_1$.) If $\boldsymbol{\beta}_2$ is not actually zero, then in the multiple regression of $\mathbf{y}$ on $(\mathbf{X}_1, \mathbf{X}_2)$, the variance of $\mathbf{b}_{1.2}$ around its mean, $\boldsymbol{\beta}_1$ would be

$$\text{Var}[\mathbf{b}_{1.2}|\mathbf{X}] = \sigma^2(\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1}$$

where

$$\mathbf{M}_2 = \mathbf{I} - \mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2',$$

or

$$\text{Var}[\mathbf{b}_{1.2}|\mathbf{X}] = \sigma^2[\mathbf{X}_1'\mathbf{X}_1 - \mathbf{X}_1'\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{X}_1]^{-1}.$$

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

$$\{\text{Var}[\mathbf{b}_1|\mathbf{X}]\}^{-1} - \{\text{Var}[\mathbf{b}_{1.2}|\mathbf{X}]\}^{-1} = (1/\sigma^2)\mathbf{X}_1'\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{X}_1,$$

which is a nonnegative definite matrix. The implication is that the variance of $\mathbf{b}_1$ is not larger than the variance of $\mathbf{b}_{1.2}$ (since its inverse is at least as large). It follows that although $\mathbf{b}_1$ is biased, its variance is never larger than the variance of the unbiased estimator. In any realistic case (i.e., if $\mathbf{X}_1'\mathbf{X}_2$ is not zero), in fact it will be smaller. We get a useful comparison from a simple regression with two variables measured as deviations from their means. Then, $\text{Var}[\mathbf{b}_1|\mathbf{X}] = \sigma^2/S_{11}$ where $S_{11} = \sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2$ and $\text{Var}[b_{1.2}|\mathbf{X}] = \sigma^2/[S_{11}(1 - r_{12}^2)]$ where r_{12}^2 is the squared correlation between x_1 and x_2.

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, b_1 and $b_{1.2}$. In fact, what researchers usually do actually creates a third estimator. It is common to include the problem variable provisionally. If its t ratio is sufficiently large, it is retained; otherwise it is discarded. This third estimator is called a **pretest estimator.** What is known about pretest estimators is not encouraging. Certainly they are biased. How badly depends on the unknown parameters. Analytical results suggest that the pretest estimator is the least precise of the three when the researcher is most likely to use it. [See Judge et al. (1985).] The conclusion to be drawn is that as a general rule, the methodology leans away from estimation strategies that include ad hoc remedies for multicollinearity.

4.7.3 PRINCIPAL COMPONENTS

A device that has been suggested for "reducing" multicollinearity [see, e.g., Gurmu, Rilstone, and Stern (1999)] is to use a small number, say L, of **principal components**

constructed as linear combinations of the K original variables. [See Johnson and Wichern (2005, Chapter 8).] (The mechanics are illustrated in Example 4.12.) 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 $\mathbf{Z} = \mathbf{X}\mathbf{C}_L$ 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. (E.g., 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 4.12 describes another natural example.

Example 4.12 Predicting Movie Success

Predicting the box office success of movies is a favorite exercise for econometricians. [See, e.g., Litman (1983), Ravid (1999), De Vany (2003), De Vany and Walls (1999, 2002, 2003), and Simonoff and Sparrow (2000).] The traditional predicting equation takes the form

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

Coefficients of determination on the order of .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.[15] Versaci (2009) added a new element to the model, "Internet buzz." 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 $\mathbf{Z}$ 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 $\mathbf{c}_1$ be the characteristic vector of $\mathbf{V}$

[15] The assertion that "nobody knows" will be tested on a newly formed (April 2010) futures exchange where investors can place early bets on movie success (and producers can hedge their own bets). See http://www.cantorexchange.com/ for discussion. The real money exchange was created by Cantor Fitzgerald, Inc. after they purchased the popular culture web site *Hollywood Stock Exchange*.

TABLE 4.8 Regression Results for Movie Success

	Internet Buzz Model			*Traditional Model*		
	22.30215			*35.66514*		
$e'e$	*.58883*			*.34247*		
R^2						
Variable	*Coefficient*	*Std.Error*	*t*	*Coefficient*	*Std.Error*	*t*
Constant	15.4002	.64273	23.96	13.5768	.68825	19.73
ACTION	−.86932	.29333	−2.96	−.30682	.34401	−.89
COMEDY	−.01622	.25608	−.06	−.03845	.32061	−.12
ANIMATED	−.83324	.43022	−1.94	−.82032	.53869	−1.52
HORROR	.37460	.37109	1.01	1.02644	.44008	2.33
G	.38440	.55315	.69	.25242	.69196	.36
PG	.53359	.29976	1.78	.32970	.37243	.89
PG13	.21505	.21885	.98	.07176	.27206	.26
LOGBUDGT	.26088	.18529	1.41	.70914	.20812	3.41
SEQUEL	.27505	.27313	1.01	.64368	.33143	1.94
STARPOWR	.00433	.01285	.34	.00648	.01608	.40
BUZZ	.42906	.07839	5.47			

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, 0.3585 so the first principal component explains 2.4142/4 or 60.3 percent of the variation. Table 4.8 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 .34 to .58 when Internet buzz is added to the model. As we will discuss in Chapter 5, buzz is also a highly "significant" predictor of success.

4.7.4 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 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) and Little and Rubin (1987, 2002) label this case **missing completely at random,** or 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."[16] 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,** or NMAR. We treat this second case in Chapter 18 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. [See Allison (2002).] 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 and zero for complete ones.[17] We leave this as an exercise to show that this is algebraically identical to simply filling the gaps with $\bar{x}$. There is also the possibility of computing fitted values for the missing x's by a regression of x on y in the complete data. The sampling properties of the resulting estimator are largely unknown, but what evidence there is suggests that this is not a beneficial way to proceed.[18]

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

[16]The vast surveys of Americans' opinions about sex by Ann Landers (1984, passim) and Shere Hite (1987) constitute two celebrated studies that were surely tainted by a heavy dose of self-selection bias. The latter was pilloried in numerous publications for purporting to represent the population at large instead of the opinions of those strongly enough inclined to respond to the survey. The former was presented with much greater modesty.

[17]See Maddala (1977a, p. 202).

[18]Afifi and Elashoff (1966, 1967) and Haitovsky (l968). Griliches (1986) considers a number of other possibilities.

"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: (a) 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 (b) 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

$$\hat{\boldsymbol{\beta}} = \bar{\mathbf{b}} = \frac{1}{M}\sum_{m=1}^{M} \mathbf{b}_m,$$

$$\hat{\mathbf{V}} = \bar{\mathbf{V}} + \mathbf{B} = \frac{1}{M}\sum_{m=1}^{M} \mathbf{V}_m + \left(1 + \frac{1}{M}\right)\left(\frac{1}{M-1}\right)\sum_{m=1}^{M} \left(\mathbf{b}_m - \bar{\mathbf{b}}\right)\left(\mathbf{b}_m - \bar{\mathbf{b}}\right)'.$$

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.

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. [See Cameron and Trivedi (2005, Chapter 27).] So, this is not a beneficial way to proceed. 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.

4.7.5 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 8.5 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,

$$I^* = \mathbf{x}'\boldsymbol{\beta} + \varepsilon \tag{4-55}$$

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 the expression for I into (4-55) gives

$$\begin{aligned} I &= \mathbf{x}'\boldsymbol{\beta} + \varepsilon + w \\ &= \mathbf{x}'\boldsymbol{\beta} + v, \end{aligned} \tag{4-56}$$

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 (4-55) and (4-56) is that the disturbance variance in (4-56) 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, since the asymptotic variance of the estimator in (4-56) 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 (4-55). The measurement error also costs some fit. To see this, note that the R^2 in the sample regression in (4-55) is

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

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,

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

The counterpart for (4-56), 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 (4-56) will, at least broadly in expectation, be inferior to that in (4-55). (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 $\mathbf{x}$, 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 the 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 $\boldsymbol{\beta}$ 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

$$\begin{aligned} y &= \beta_1 + \beta_2 I + (\varepsilon - \beta_2 w) \\ &= \beta_1 + \beta_2 I + v. \end{aligned} \tag{4-57}$$

It appears that we have obtained (4-56) once again. Unfortunately, this is not the case, because $\text{Cov}[I, v] = \text{Cov}[I^* + w, \varepsilon - \beta_2 w] = -\beta_2 \sigma_w^2$. Since the regressor in (4-57) 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 is the measurement error variance, the worse is the bias. (This is called l**east squares attenuation**.) Now, suppose there are additional variables in the model;

$$y = \mathbf{x}'\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2 I^* + \varepsilon.$$

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 $\boldsymbol{\beta}_2$ 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 $\mathbf{x}$ 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 (4-57), 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 setting in Section 8.5.

4.7.6 OUTLIERS AND INFLUENTIAL OBSERVATIONS

Figure 4.9 shows a scatter plot of the data on sale prices of Monet paintings that were used in Example 4.10. Two points have been highlighted. The one marked "I" and noted with the square overlay shows the smallest painting in the data set. The circle marked

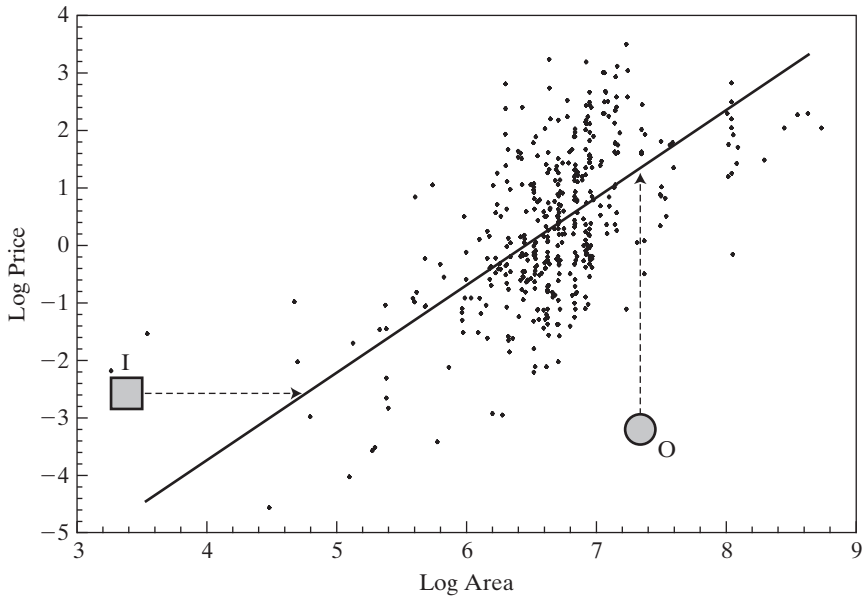


FIGURE 4.9 Log Price vs. Log Area for Monet Paintings.

"O" 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.) Since 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 4.9, Belsley, Kuh and Welsh (1980) defined an influence measure, for observation i,

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x}_n)^2}{\Sigma_{j=1}^{n}(x_j - \bar{x}_n)^2} \qquad \textbf{(4-58)}$$

where $\bar{x}_n$ and the summation in the denominator of the fraction are computed without this observation. (The measure derives from the difference between $\mathbf{b}$ 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 4.9, 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 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

$$\mathbf{b} - \mathbf{b}_{(i)} = \Delta\mathbf{b} = \frac{1}{1 + \mathbf{x}_i' \left(\mathbf{X}_{(i)}'\mathbf{X}_{(i)}\right)^{-1}\mathbf{x}_i} \left(\mathbf{X}_{(i)}'\mathbf{X}_{(i)}\right)^{-1}\mathbf{x}_i \left(\mathbf{y}_i - \mathbf{x}_i'\mathbf{b}_{(i)}\right) \tag{4-59}$$

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 6 in Chapter 3.) It is difficult to single out any particular feature of the observation that would drive this change. The influence measure,

$$\begin{aligned} h_{ii} &= \mathbf{x}_i' \left(\mathbf{X}_{(i)}'\mathbf{X}_{(i)}\right)^{-1}\mathbf{x}_i \\ &= \frac{1}{n} + \sum_{j=1}^{K-1}\sum_{k=1}^{K-1} \left(x_{i,j} - \bar{x}_{n,j}\right)\left(x_{i,k} - \bar{x}_k\right)\left(\mathbf{Z}_{(i)}'\mathbf{M}^0\mathbf{Z}_{(i)}\right)^{jk}, \end{aligned} \tag{4-60}$$

has been used to flag influential observations. [See, once again, Belsley, Kuh and Welsh (1980) and Cook (1977).] 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 the expression for the forecast variance in Section 4.6.1 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. Point "O" in Figure 4.9 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, since 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 ith studentized residual is

$$e(i) = \frac{e_i}{(1 - h_{ii})} \Bigg/ \sqrt{\frac{\mathbf{e}'\mathbf{e} - e_i^2/(1 - h_{ii})}{n - 1 - K}} \tag{4-61}$$

where $\mathbf{e}$ is the residual vector for the full sample, based on $\mathbf{b}$, 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 following, application, following, the procedure

TABLE 4.9 Estimated Equations for Log Price

Number of observations	430	410
Mean of log Price	0.33274	.36043
Sum of squared residuals	519.17235	383.17982
Standard error of regression	1.10266	0.97030
R-squared	0.33620	0.39170
Adjusted R-squared	0.33309	0.38871

	Coefficient		*Standard Error*		*t*	
Variable	$n = 430$	$n = 410$	$n = 430$	$n = 410$	$n = 430$	$n = 410$
Constant	−8.42653	−8.67356	.61183	.57529	−13.77	−15.08
LOGAREA	1.33372	1.36982	.09072	.08472	14.70	16.17
ASPECT	−.16537	−.14383	.12753	.11412	−1.30	−1.26

deletes 4.7 percent of the sample (20 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 20 observations. After 16 iterations, the sample size stabilizes at 316 of the original 430 observations, a reduction of 26.5 percent.) Table 4.9 shows the original results (from Table 4.6) and the modified results with 20 outliers removed. Since 430 is a relatively large sample, the modest change in the results is to be expected.

It is difficult to draw a 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.

4.8 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 4.5 and 4.6. 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
- Bootstrap
- condition number
- Confidence interval
- Consistency
- Consistent estimator
- Data imputation
- Efficient scale
- Ergodic
- Estimator
- Ex ante forecast
- Ex post forecast
- Finite sample properties
- Gauss–Markov theorem
- Grenander conditions
- Highest posterior density interval
- Identification
- Ignorable case
- Inclusion of superfluous (irrelevant) variables
- Indicator
- Interval estimation
- Least squares attenuation
- Lindeberg–Feller central limit theorem
- Linear estimator
- Linear unbiased estimator
- Maximum likelihood estimator
- Mean absolute error
- Mean square convergence
- Mean squared error
- Measurement error
- Method of moments
- Minimum mean squared error
- Minimum variance linear unbiased estimator
- Missing at random
- Missing completely at random
- Missing observations
- Modified zero-order regression
- Monte Carlo study
- Multicollinearity
- Not missing at random
- Oaxaca's and Blinder's decomposition
- Omission of relevant variables
- Optimal linear predictor
- Orthogonal random variables
- Panel data
- Pivotal statistic
- Point estimation
- Prediction error
- Prediction interval
- Prediction variance
- Pretest estimator
- Principal components
- Probability limit
- Root mean squared error
- Sample selection
- Sampling distribution
- Sampling variance
- Semiparametric
- Smearing estimator
- Specification errors
- Standard error
- Standard error of the regression
- Stationary process
- Statistical properties
- Stochastic regressors
- Theil U statistic
- t ratio
- Variance inflation factor
- 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 $E[\varepsilon \mid x] = 0$ and $E[\varepsilon^2 \mid x] = \sigma^2$
   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

$$\frac{\text{MSE}[\hat{\beta}]}{\text{MSE}[b]} = \frac{\tau^2}{(1+\tau^2)}, \quad \text{where } \tau^2 = \frac{\beta^2}{[\sigma^2/\mathbf{x}'\mathbf{x}]}.$$

Note that τ is the square of the population analogue to the "t ratio" for testing the hypothesis that $\beta = 0$, which is given in (4-14). How do you interpret the behavior of this ratio as $\tau \to \infty$?

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 $E[\varepsilon_i \mid x_i] = \lambda$ and $\text{Var}[\varepsilon_i \mid x_i] = \lambda^2$. Show that the least squares slope is unbiased but that the intercept 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 percent 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:

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

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 collinear. Which is the problem variable?

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 $\boldsymbol{\beta}$ 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 plim $F[K, n-K] = \text{plim } \frac{R^2/K}{(1-R^2)/(n-K)}$, 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 $\text{plim}(e_i - \varepsilon_i) = 0$.
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 4.7.3, 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 $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. In the discussion, it is claimed that $E[\mathbf{d}] = \mathbf{C}_L'\boldsymbol{\beta}$. Prove the claim.
17. Example 4.9 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 (log = 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, all 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 4.5.3. 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

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

   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 produces the estimating equation,

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

The purpose of the generalization was to produce a U-shaped average total cost curve. [See Example 6.6 for discussion of Nerlove's (1963) predecessor to this study.] 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.)