4

FINITE-SAMPLE PROPERTIES OF THE LEAST SQUARES ESTIMATOR

4.1 INTRODUCTION

Chapter 3 treated fitting the linear regression to the data as a purely algebraic exercise. We will now examine the least squares **estimator** from a statistical viewpoint. This chapter will consider exact, finite-sample results such as unbiased estimation and the precise distributions of certain test statistics. Some of these results require fairly strong assumptions, such as nonstochastic regressors or normally distributed disturbances. In the next chapter, we will turn to the properties of the least squares estimator in more general cases. In these settings, we rely on approximations that do not hold as exact results but which do improve as the sample size increases.

There are other candidates for estimating $\boldsymbol{\beta}$. In a two-variable case, for example, we might use the intercept, a, and slope, b, of the line between the points with the largest and smallest values of x. Alternatively, we might find the a and b that minimize the sum of absolute values of the residuals. The question of which estimator to choose is usually based on the **statistical properties** of the candidates, such as unbiasedness, efficiency, and precision. These, in turn, frequently depend on the particular distribution that we assume produced the data. However, a number of desirable properties can be obtained for the least squares estimator even without specifying a particular distribution for the disturbances in the regression.

In this chapter, we will examine in detail the least squares as an estimator of the model parameters of the classical model (defined in the following Table 4.1). We begin in Section 4.2 by returning to the question raised but not answered in Footnote 1, Chapter 3, that is, why least squares? We will then analyze the estimator in detail. We take Assumption A1, linearity of the model as given, though in Section 4.2, we will consider briefly the possibility of a different predictor for y. Assumption A2, the identification condition that the data matrix have full rank is considered in Section 4.9 where data complications that arise in practice are discussed. The near failure of this assumption is a recurrent problem in "real world" data. Section 4.3 is concerned with unbiased estimation. Assumption A3, that the disturbances and the independent variables are uncorrelated, is a pivotal result in this discussion. Assumption A4, homoscedasticity and nonautocorrelation of the disturbances, in contrast to A3, only has relevance to whether least squares is an optimal use of the data. As noted, there are alternative estimators available, but with Assumption A4, the least squares estimator is usually going to be preferable. Sections 4.4 and 4.5 present several statistical results for the least squares estimator that depend crucially on this assumption. The assumption that the data in $\mathbf{X}$ are nonstochastic, known constants, has some implications for how certain derivations

TABLE 4.1 Assumptions of the Classical Linear Regression Model

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

A2. Full rank: The $n \times K$ sample data matrix, **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.

A5. Exogenously generated data $(x_{i1}, x_{i2}, \ldots, x_{iK})\ i = 1, \ldots, n$.

A6. Normal distribution: The disturbances are normally distributed.

proceed, but in practical terms, is a minor consideration. Indeed, nearly all that we do with the regression model departs from this assumption fairly quickly. It serves only as a useful departure point. The issue is considered in Section 4.5. Finally, the normality of the disturbances assumed in A6 is crucial in obtaining the **sampling distributions** of several useful statistics that are used in the analysis of the linear model. We note that in the course of our analysis of the linear model as we proceed through Chapter 9, all six of these assumptions will be discarded.

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 **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

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

(The right-hand side is not a function of y so the expectation is taken only over **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

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

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 10 and 18 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

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

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

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

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

$$E_{\mathbf{x}} E_y[\mathbf{x}y] = E_{\mathbf{x}} E_y[\mathbf{x}\mathbf{x}']\boldsymbol{\gamma}. \qquad \textbf{(4-3)}$$

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

4.2.3 MINIMUM VARIANCE LINEAR UNBIASED ESTIMATION

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

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

4.3 UNBIASED ESTIMATION

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

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

Now, take expectations, iterating over **X**;

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

By Assumption A3, the second term is **0**, so

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

Therefore,

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

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

Example 4.1 The Sampling Distribution of a Least Squares Estimator
The following sampling experiment, which can be replicated in any computer program that provides a random number generator and a means of drawing a random sample of observations from a master data set, shows the nature of a sampling distribution and the implication of unbiasedness. We drew two samples of 10,000 random draws on w_i and x_i from the standard

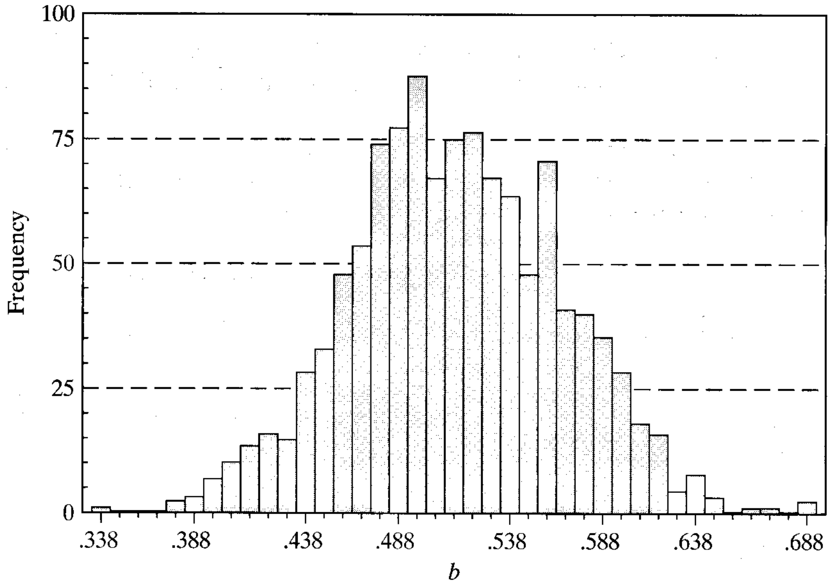


FIGURE 4.1 Histogram for Sampled Least Squares Regression Slopes.

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

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

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

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

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 (4-5) 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 covariance matrix of the least squares slope estimator is

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

Example 4.2 Sampling Variance in the Two-Variable Regression Model

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

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

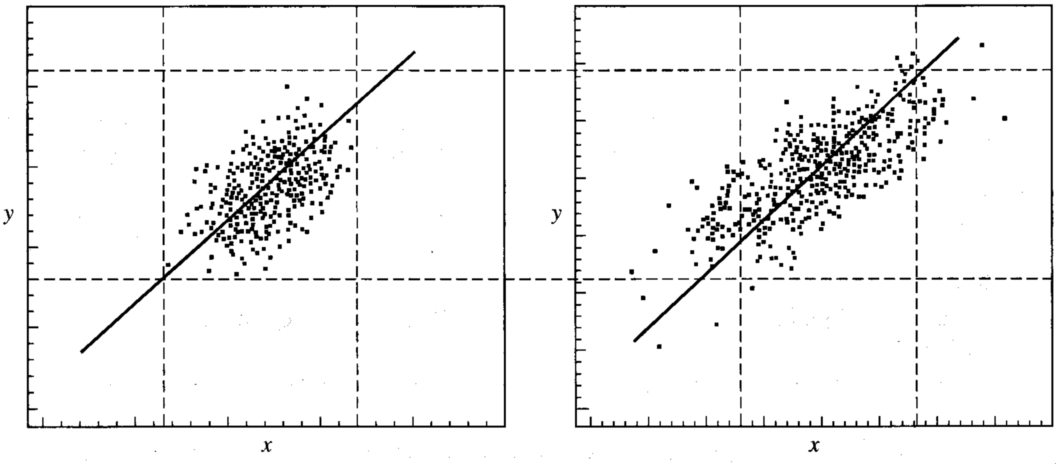
Note, in particular, the denominator of the variance of b. The greater the variation in x, the smaller this variance. For example, consider the problem of estimating the slopes of the two regressions in Figure 4.2. A more precise result will be obtained for the data in the right-hand panel of the figure.

We will now obtain a general result for the class of linear unbiased estimators of $\boldsymbol{\beta}$. Let $\mathbf{b}_0 = \mathbf{C}\mathbf{y}$ be another linear unbiased estimator of $\boldsymbol{\beta}$, where $\mathbf{C}$ is a $K\times n$ matrix. If $\mathbf{b}_0$ is unbiased, then

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

which implies that $\mathbf{C}\mathbf{X} = \mathbf{I}$. There are many candidates. For example, consider using just the first K (or, any K) linearly independent rows of $\mathbf{X}$. Then $\mathbf{C} = [\mathbf{X}_0^{-1} : \mathbf{0}]$, where $\mathbf{X}_0^{-1}$

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



is the transpose of the matrix formed from the K rows of $\mathbf{X}$. The covariance matrix of $\mathbf{b}_0$ can be found by replacing $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ with $\mathbf{C}$ in (4-5); the result is $\text{Var}[\mathbf{b}_0 \mid \mathbf{X}] = \sigma^2\mathbf{CC}'$. 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 implies a very important property of the least squares coefficient vector.

THEOREM 4.2 Gauss–Markov Theorem
In the classical 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 classical regression model is* $\mathbf{w}'\mathbf{b}$, *where* $\mathbf{b}$ *is the least squares estimator.*

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 $\boldsymbol{\beta}$.

4.5 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, 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, 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-70):

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

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 Chapter 5.

We showed in Section 4.4 that

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

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

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

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 regard $\mathbf{X}$ as stochastic.

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

4.6 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:

$$\mathbf{e}'\mathbf{e} = \boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}. \qquad \textbf{(4-6)}$$

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),

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

Since **M** is a function of **X**, the result is

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

The trace of **M** is

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

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

$$s^2 = \frac{\mathbf{e}'\mathbf{e}}{n - K}. \qquad \textbf{(4-7)}$$

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

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

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, $\{[s^2(\mathbf{X}'\mathbf{X})^{-1}]_{kk}\}^{1/2}$, is the **standard error** of the estimator b_k, which is often denoted simply "the standard error of b_k."

4.7 THE NORMALITY ASSUMPTION AND BASIC STATISTICAL INFERENCE

To this point, our specification and analysis of the regression model is **semiparametric** (see Section 16.3). 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 testing hypotheses. In (4-5), $\mathbf{b}$ 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

$$\mathbf{b} \mid \mathbf{X} \sim N[\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}]. \tag{4-8}$$

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

$$b_k \mid \mathbf{X} \sim N\big[\beta_k, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{kk}\big]. \tag{4-9}$$

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 Chapter 5, we *will* be able to obtain an approximate normal distribution for $\mathbf{b}$, with or without assuming normally distributed disturbances and whether the regressors are stochastic or not.

4.7.1 TESTING A HYPOTHESIS ABOUT A COEFFICIENT

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

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

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

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

is an idempotent quadratic form in a standard normal vector $(\boldsymbol{\varepsilon}/\sigma)$. Therefore, it has a chi-squared distribution with rank $(\mathbf{M}) = \text{trace}(\mathbf{M}) = n - K$ degrees of freedom.[1] The chi-squared variable in (4-11) is independent of the standard normal variable in (4-10). To prove this, it suffices to show that

$$\frac{\mathbf{b}-\boldsymbol{\beta}}{\sigma} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right) \tag{4-12}$$

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 an idempotent quadratic

[1] This fact is proved in Section B.10.3.

form $\mathbf{x}'\mathbf{A}\mathbf{x}$ 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 4.4 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

$$t_k = \frac{(b_k - \beta_k)/\sqrt{\sigma^2 S^{kk}}}{\sqrt{[(n-K)s^2/\sigma^2]/(n-K)}} = \frac{b_k - \beta_k}{\sqrt{s^2 S^{kk}}} \qquad \textbf{(4-13)}$$

has a t distribution with $(n - K)$ degrees of freedom.[2] We can use t_k to test hypotheses or form confidence intervals about the individual elements of $\boldsymbol{\beta}$.

A common test is whether a parameter β_k is significantly different from zero. The appropriate test statistic

$$t = \frac{b_k}{s_{b_k}} \qquad \textbf{(4-14)}$$

is presented as standard output with the other results by most computer programs. The test is done in the usual way. This statistic is usually labeled the ***t* ratio** for the estimator b_k. If $|b_k|/s_{b_k} > t_{\alpha/2}$, where $t_{\alpha/2}$ is the $100(1 - \alpha/2)$ percent critical value from the t distribution with $(n - K)$ degrees of freedom, then the hypothesis is rejected and the coefficient is said to be "statistically significant." The value of 1.96, which would apply for the 5 percent significance level in a large sample, is often used as a benchmark value when a table of critical values is not immediately available. The t ratio for the test of the hypothesis that a coefficient equals zero is a standard part of the regression output of most computer programs.

Example 4.3 Earnings Equation

Appendix Table F4.1 contains 753 observations used in Mroz's (1987) study of labor supply behavior of married women. We will use these data at several points below. Of the 753 individuals in the sample, 428 were participants in the formal labor market. For these individuals, we will fit a semilog earnings equation of the form suggested in Example 2.2;

$$\ln earnings = \beta_1 + \beta_2\, age + \beta_3\, age^2 + \beta_4\, education + \beta_5\, kids + \varepsilon,$$

where *earnings* is *hourly wage* times *hours worked*, *education* is measured in years of schooling and *kids* is a binary variable which equals one if there are children under 18 in the household. (See the data description in Appendix F for details.) Regression results are shown in Table 4.2. There are 428 observations and 5 parameters, so the t statistics have 423 degrees

[2] 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.

TABLE 4.2 Regression Results for an Earnings Equation

Sum of squared residuals:	599.4582
Standard error of the regression:	1.19044
R^2 based on 428 observations	0.040995

Variable	*Coefficient*	*Standard Error*	*t Ratio*
Constant	3.24009	1.7674	1.833
Age	0.20056	0.08386	2.392
Age^2	−0.0023147	0.00098688	−2.345
Education	0.067472	0.025248	2.672
Kids	−0.35119	0.14753	−2.380

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

Constant	*Age*	*Age^2*	*Education*	*Kids*
3.12381				
−0.14409	0.0070325			
0.0016617	−8.23237e−5	9.73928e−7		
−0.0092609	5.08549e−5	−4.96761e−7	0.00063729	
0.026749	−0.0026412	3.84102e−5	−5.46193e−5	0.021766

of freedom. For 95 percent significance levels, the standard normal value of 1.96 is appropriate when the degrees of freedom are this large. By this measure, all variables are statistically significant and signs are consistent with expectations. It will be interesting to investigate whether the effect of Kids is on the wage or hours, or both. We interpret the schooling variable to imply that an additional year of schooling is associated with a 6.7 percent increase in earnings. The quadratic age profile suggests that for a given education level and family size, earnings rise to the peak at $-b_2/(2b_3)$ which is about 43 years of age, at which they begin to decline. Some points to note: (1) Our selection of only those individuals who had positive hours worked is not an innocent sample selection mechanism. Since individuals chose whether or not to be in the labor force, it is likely (almost certain) that earnings potential was a significant factor, along with some other aspects we will consider in Chapter 22. (2) The earnings equation is a mixture of a labor supply equation—hours worked by the individual, and a labor demand outcome—the wage is, presumably, an accepted offer. As such, it is unclear what the precise nature of this equation is. Presumably, it is a hash of the equations of an elaborate structural equation system.

4.7.2 CONFIDENCE INTERVALS FOR PARAMETERS

A confidence interval for β_k would be based on (4-13). We could say that

$$\text{Prob}(b_k - t_{\alpha/2}s_{b_k} \leq \beta_k \leq b_k + t_{\alpha/2}s_{b_k}) = 1 - \alpha,$$

where $1 - \alpha$ is the desired level of confidence and $t_{\alpha/2}$ is the appropriate critical value from the t distribution with $(n - K)$ degrees of freedom.

Example 4.4 Confidence Interval for the Income Elasticity of Demand for Gasoline

Using the gasoline market data discussed in Example 2.3, we estimated following demand equation using the 36 observations. Estimated standard errors, computed as shown above, are given in parentheses below the least squares estimates.

$$\begin{aligned}\ln(G/\text{pop}) &= \underset{(0.6749)}{-7.737} - \underset{(0.03248)}{0.05910}\ln P_G + \underset{(0.075628)}{1.3733}\ln \text{income}\\ &\quad \underset{(0.12699)}{-0.12680}\ln P_{nc} - \underset{(0.081337)}{0.11871}\ln P_{uc} + e.\end{aligned}$$

To form a confidence interval for the income elasticity, we need the critical value from the t distribution with $n - K = 36 - 5$ degrees of freedom. The 95 percent critical value is 2.040. Therefore, a 95 percent confidence interval for β_I is $1.3733 \pm 2.040(0.075628)$, or $[1.2191, 1.5276]$.

We are interested in whether the demand for gasoline is income inelastic. The hypothesis to be tested is that β_I is less than 1. For a one-sided test, we adjust the critical region and use the t_α critical point from the distribution. Values of the sample estimate that are greatly inconsistent with the hypothesis cast doubt upon it. Consider testing the hypothesis

$$H_0 : \beta_I < 1 \quad \text{versus} \quad H_1 : \beta_I \geq 1.$$

The appropriate test statistic is

$$t = \frac{1.3733 - 1}{0.075628} = 4.936.$$

The critical value from the t distribution with 31 degrees of freedom is 2.04, which is far less than 4.936. We conclude that the data are not consistent with the hypothesis that the income elasticity is less than 1, so we reject the hypothesis.

4.7.3 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 4.4, 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) **decomposition** provides a frequently used application.

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:

$$\text{Prob}[c - t_{\alpha/2}s_c \leq \gamma \leq c + t_{\alpha/2}s_c] = 1 - \alpha.$$

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

Consider, then, Oaxaca's 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,

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

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 suggested that this decomposition be computed at the means of the two regressor vectors, $\bar{\mathbf{x}}_m$ and $\bar{\mathbf{x}}_f$, 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 $\bar{\mathbf{x}}'_m\mathbf{d}$ where $\mathbf{d} = \mathbf{b}_m - \mathbf{b}_f$. The estimated covariance matrix is

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

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

4.7.4 TESTING THE SIGNIFICANCE OF THE REGRESSION

A question that is usually of interest is whether the regression equation as a whole is significant. This test is a joint test of the hypotheses that *all* the coefficients except the constant term are zero. If all the slopes are zero, then the multiple correlation coefficient is zero as well, so we can base a test of this hypothesis on the value of R^2. The central result needed to carry out the test is the distribution of the statistic

$$F[K-1, n-K] = \frac{R^2/(K-1)}{(1-R^2)/(n-K)}. \qquad \textbf{(4-15)}$$

If the hypothesis that $\boldsymbol{\beta}_2 = \mathbf{0}$ (the part of $\boldsymbol{\beta}$ not including the constant) is true and the disturbances are normally distributed, then this statistic has an F distribution with $K-1$ and $n-K$ degrees of freedom.[3] Large values of F give evidence against the validity of the hypothesis. Note that a large F is induced by a large value of R^2.

The logic of the test is that the F statistic is a measure of the loss of fit (namely, all of R^2) that results when we impose the restriction that all the slopes are zero. If F is large, then the hypothesis is rejected.

[3] The proof of the distributional result appears in Section 6.3.1. The F statistic given above is the special case in which $\mathbf{R} = [\mathbf{0} \mid \mathbf{I}_{K-1}]$.

Example 4.5 F Test for the Earnings Equation

The F ratio for testing the hypothesis that the four slopes in the earnings equation are all zero is

$$F[4, 423] = \frac{0.040995/4}{(1 - 0.040995)/(428 - 5)} = 4.521,$$

which is far larger than the 95 percent critical value of 2.37. We conclude that the data are inconsistent with the hypothesis that all the slopes in the earnings equation are zero.

We might have expected the preceding result, given the substantial t ratios presented earlier. But this case need not always be true. Examples can be constructed in which the individual coefficients are statistically significant, while jointly they are not. This case can be regarded as pathological, but the opposite one, in which none of the coefficients is significantly different from zero while R^2 is highly significant, is relatively common. The problem is that the interaction among the variables may serve to obscure their individual contribution to the fit of the regression, whereas their joint effect may still be significant. We will return to this point in Section 4.9.1 in our discussion of multicollinearity.

4.7.5 MARGINAL DISTRIBUTIONS OF THE TEST STATISTICS

We now consider the relation between the sample test statistics and the data in **X**. First, consider the conventional t statistic in (4-14) for testing $H_0 : \beta_k = \beta_k^0$,

$$t \,|\, \mathbf{X} = \frac{\left(b_k - \beta_k^0\right)}{\left[s^2(\mathbf{X}'\mathbf{X})_{kk}^{-1}\right]^{1/2}}.$$

Conditional on **X**, if $\beta_k = \beta_k^0$ (i.e., under H_0), then $t \,|\, \mathbf{X}$ has a t distribution with $(n - K)$ degrees of freedom. What interests us, however, is the marginal, that is, the unconditional, distribution of t. As we saw, **b** is only normally distributed conditionally on **X**; the marginal distribution may not be normal because it depends on **X** (through the conditional variance). Similarly, because of the presence of **X**, the denominator of the t statistic is not the square root of a chi-squared variable divided by its degrees of freedom, again, except conditional on this **X**. But, because the distributions of $\left\{(b_k - \beta_k)/[\sigma^2(\mathbf{X}'\mathbf{X})_{kk}^{-1}]^{1/2}\right\} \,|\, \mathbf{X}$ and $[(n - K)s^2/\sigma^2] \,|\, \mathbf{X}$ are still independent $N[0, 1]$ and $\chi^2[n - K]$, respectively, which do not involve **X**, we have the surprising result that, regardless of the distribution of **X**, or even of whether **X** is stochastic or nonstochastic, the marginal distributions of t is still t, even though the marginal distribution of b_k may be nonnormal. This intriguing result follows because $f(t \,|\, \mathbf{X})$ is not a function of **X**. The same reasoning can be used to deduce that the usual F ratio used for testing linear restrictions is valid whether **X** is stochastic or not. This result is very powerful. The implication is that *if the disturbances are normally distributed, then we may carry out tests and construct confidence intervals for the parameters without making any changes in our procedures, regardless of whether the regressors are stochastic, nonstochastic, or some mix of the two.*

4.8 FINITE-SAMPLE PROPERTIES OF LEAST SQUARES

A summary of the results we have obtained for the least squares estimator appears in Table 4.3. For constructing confidence intervals and testing hypotheses, we derived some additional results that depended explicitly on the normality assumption. Only

TABLE 4.3 Finite Sample Properties of Least Squares

General results:

FS1. $E[\mathbf{b} \mid \mathbf{X}] = E[\mathbf{b}] = \boldsymbol{\beta}$. Least squares is unbiased.

FS2. $\text{Var}[\mathbf{b} \mid \mathbf{X}] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$; $\text{Var}[\mathbf{b}] = \sigma^2 E[(\mathbf{X}'\mathbf{X})^{-1}]$.

FS3. Gauss–Markov theorem: The MVLUE of $\mathbf{w}'\boldsymbol{\beta}$ is $\mathbf{w}'\mathbf{b}$.

FS4. $E[s^2 \mid \mathbf{X}] = E[s^2] = \sigma^2$.

FS5. $\text{Cov}[\mathbf{b}, \mathbf{e} \mid \mathbf{X}] = E[(\mathbf{b} - \boldsymbol{\beta})\mathbf{e}' \mid \mathbf{X}] = E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{M} \mid \mathbf{X}] = \mathbf{0}$ as $\mathbf{X}'(\sigma^2\mathbf{I})\mathbf{M} = \mathbf{0}$.

Results that follow from Assumption A6, normally distributed disturbances:

FS6. **b** and **e** are statistically independent. It follows that **b** and s^2 are uncorrelated and statistically independent.

FS7. The exact distribution of $\mathbf{b} \mid \mathbf{X}$, is $N[\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}]$.

FS8. $(n - K)s^2/\sigma^2 \sim \chi^2[n - K]$. s^2 has mean σ^2 and variance $2\sigma^4/(n - K)$.

Test Statistics based on results FS6 through FS8:

FS9. $t[n - K] = (b_k - \beta_k)/[s^2(\mathbf{X}'\mathbf{X})^{-1}_{kk}]^{1/2} \sim t[n - K]$ independently of **X**.

FS10. The test statistic for testing the null hypothesis that all slopes in the model are zero, $F[K-1, n-K] = [R^2/(K-1)]/[(1-R^2)/(n-K)]$ has an F distribution with $K-1$ and $n-K$ degrees of freedom when the null hypothesis is true.

FS7 depends on whether **X** is stochastic or not. If so, then the *marginal* distribution of **b** depends on that of **X**. Note the distinction between the properties of **b** established using A1 through A4 and the additional inference results obtained with the further assumption of normality of the disturbances. The primary result in the first set is the Gauss–Markov theorem, which holds regardless of the distribution of the disturbances. The important additional results brought by the normality assumption are FS9 and FS10.

4.9 DATA PROBLEMS

In this section, we consider three practical problems that arise in the setting of regression analysis, multicollinearity, missing observations and outliers.

4.9.1 MULTICOLLINEARITY

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

$$\text{Var}[b_k] = \frac{\sigma^2}{\left(1 - r_{12}^2\right)\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2} = \frac{\sigma^2}{\left(1 - r_{12}^2\right)S_{kk}}, \quad k = 1, 2. \qquad \textbf{(4-16)}$$

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

$$\begin{aligned}\left(\mathbf{x}_k'\mathbf{M}_{(k)}\mathbf{x}_k\right)^{-1} &= \left[\mathbf{x}_k'\mathbf{x}_k - \mathbf{x}_k'\mathbf{X}_{(k)}\left(\mathbf{X}_{(k)}'\mathbf{X}_{(k)}\right)^{-1}\mathbf{X}_{(k)}'\mathbf{x}_k\right]^{-1} \\ &= \left[\mathbf{x}_k'\mathbf{x}_k\left(1 - \frac{\mathbf{x}_k'\mathbf{X}_{(k)}\left(\mathbf{X}_{(k)}'\mathbf{X}_{(k)}\right)^{-1}\mathbf{X}_{(k)}'\mathbf{x}_k}{\mathbf{x}_k'\mathbf{x}_k}\right)\right]^{-1} \\ &= \frac{1}{\left(1 - R_{k.}^2\right)S_{kk}},\end{aligned} \tag{4-17}$$

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

$$\text{Var}[b_k] = \frac{\sigma^2}{\left(1 - R_{k.}^2\right)\sum_{i=1}^n (x_{ik} - \bar{x}_k)^2}, \tag{4-18}$$

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 4.2.
- 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 4.2 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

TABLE 4.4 Longley Results: Dependent Variable is Employment

	1947–1961	*Variance Inflation*	*1947–1962*
Constant	1,459,415		1,169,087
Year	−721.756	251.839	−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

square root ratio of the largest characteristic root of **X**′**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 4.6 is over 15,000!)

Example 4.6 Multicollinearity in the Longley Data

The data in 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 be unusual. But, the results in Table 4.4 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.[4] 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 8.2. 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. Several other practical approaches have also been suggested. The **ridge regression estimator** is $\mathbf{b}_r = [\mathbf{X}'\mathbf{X} + r\mathbf{D}]^{-1}\mathbf{X}'\mathbf{y}$, where **D** is a diagonal matrix. This biased estimator has a covariance matrix unambiguously smaller than that of **b**. The tradeoff of some bias for smaller variance may be worth making (see Judge et al., 1985), but, nonetheless, economists are generally averse to biased estimators, so this approach has seen little practical use. Another approach sometimes used [see, e.g., Gurmu, Rilstone, and Stern (1999)] is to use a small number, say L, of **principal components** constructed from the K original variables. [See Johnson and Wichern (1999).] The problem here is that if the original model in the form $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ were correct, then it is unclear what one is estimating when one regresses **y** on some

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

small set of linear combinations of the columns of $\mathbf{X}$. Algebraically, it is simple; at least for the principal components case, in which we regress $\mathbf{y}$ on $\mathbf{Z} = \mathbf{X}\mathbf{C}_L$ to obtain $\mathbf{d}$, it follows that $E[\mathbf{d}] = \boldsymbol{\delta} = \mathbf{C}_L\mathbf{C}'_L\boldsymbol{\beta}$. In an economic context, if $\boldsymbol{\beta}$ has an interpretation, then it is unlikely that $\boldsymbol{\delta}$ will. (How do we interpret the price elasticity plus minus twice the income elasticity?)

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.

4.9.2 MISSING OBSERVATIONS

It is fairly common for a data set to have gaps, for a variety of reasons. Perhaps the most common occurrence of this problem is in survey data, in which it often happens that respondents simply fail to answer 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.

There are two possible cases to consider, depending on why the data are missing. One is that the data are simply unavailable, for reasons unknown to the analyst and unrelated to the completeness of the other observations in the sample. 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 ignore the problem. 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."[5] 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. In this case, the complete observations would be qualitatively different. We treat this second case in Chapter 22, so we shall defer our discussion until later.

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

[5] 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 first was presented with much greater modesty.

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.

For the case of missing data in the regressors, it helps to consider the simple regression and multiple regression cases separately. In the first case, **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. Several schemes have been suggested for filling the blanks. The zero-order method of replacing each missing x with $\bar{x}$ 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* is to fill the second column of **X** with zeros and add a variable that takes the value one for missing observations and zero for complete ones.[6] We leave it as an exercise to show that this is algebraically identical to simply filling the gaps with $\bar{x}$ Last, there is 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.[7]

4.9.3 REGRESSION DIAGNOSTICS AND INFLUENTIAL DATA POINTS

Even in the absence of multicollinearity or other data problems, it is worthwhile to examine one's data closely for two reasons. First, the identification of **outliers** in the data is useful, particularly in relatively small cross sections in which the identity and perhaps even the ultimate source of the data point may be known. Second, it may be possible to ascertain which, if any, particular observations are especially influential in the results obtained. As such, the identification of these data points may call for further study. It is worth emphasizing, though, that there is a certain danger in singling out particular observations for scrutiny or even elimination from the sample on the basis of statistical results that are based on those data. At the extreme, this step may invalidate the usual inference procedures.

Of particular importance in this analysis is the **projection matrix** or **hat matrix:**

$$\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'. \tag{4-19}$$

This matrix appeared earlier as the matrix that projects any $n \times 1$ vector into the column space of **X**. For any vector **y**, **Py** is the set of fitted values in the least squares regression of **y** on **X**. The least squares residuals are

$$\mathbf{e} = \mathbf{My} = \mathbf{M}\boldsymbol{\varepsilon} = (\mathbf{I} - \mathbf{P})\boldsymbol{\varepsilon},$$

so the covariance matrix for the least squares residual vector is

$$E[\mathbf{ee}'] = \sigma^2\mathbf{M} = \sigma^2(\mathbf{I} - \mathbf{P}).$$

To identify which residuals are significantly large, we first standardize them by dividing

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

[7] Afifi and Elashoff (1966, 1967) and Haitovsky (1968). Griliches (1986) considers a number of other possibilities.

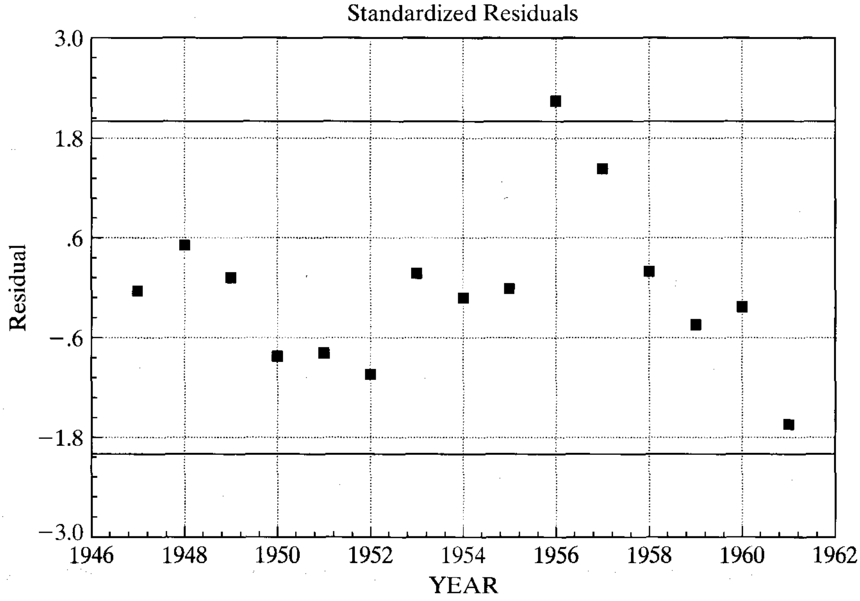


FIGURE 4.3 Standardized Residuals for the Longley Data.

by the appropriate standard deviations. Thus, we would use

$$\hat{e}_i = \frac{e_i}{[s^2(1 - p_{ii})]^{1/2}} = \frac{e_i}{(s^2 m_{ii})^{1/2}}, \tag{4-20}$$

where e_i is the ith least squares residual, $s^2 = \mathbf{e}'\mathbf{e}/(n-K)$, p_{ii} is the ith diagonal element of $\mathbf{P}$ and m_{ii} is the ith diagonal element of $\mathbf{M}$. It is easy to show (we leave it as an exercise) that $e_i/m_{ii} = y_i - \mathbf{x}_i'\mathbf{b}(i)$ where $\mathbf{b}(i)$ is the least squares slope vector computed without this observation, so the standardization is a natural way to investigate whether the particular observation differs substantially from what should be expected given the model specification. Dividing by s^2, or better, $s(i)^2$ scales the observations so that the value 2.0 [suggested by Belsley, et al. (1980)] provides an appropriate benchmark. Figure 4.3 illustrates for the Longley data of the previous example. Apparently, 1956 was an unusual year according to this "model." (What to do with outliers is a question. Discarding an observation in the middle of a time series is probably a bad idea, though we may hope to learn something about the data in this way. For a cross section, one may be able to single out observations that do not conform to the model with this technique.)

4.10 SUMMARY AND CONCLUSIONS

This chapter has examined a set of properties of the least squares estimator that will apply in all samples, including unbiasedness and efficiency among unbiased estimators. The assumption of normality of the disturbances produces the distributions of some useful test statistics which are useful for a statistical assessment of the validity of the regression model. The finite sample results obtained in this chapter are listed in Table 4.3.

We also considered some practical problems that arise when data are less than perfect for the estimation and analysis of the regression model, including multicollinearity and missing observations.

The formal assumptions of the classical 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. Also, we are going to be interested in hypothesis tests of the type considered here, but at the same time, the assumption of normality is narrow, so it will be necessary to extend the model to allow nonnormal disturbances. These and other 'large sample' extensions of the linear model will be considered in Chapter 5.

Key Terms and Concepts

- Assumptions
- Condition number
- Confidence interval
- Estimator
- Gauss-Markov Theorem
- Hat matrix
- Ignorable case
- Linear estimator
- Linear unbiased estimator
- Mean squared error
- Minimum mean squared error
- Minimum variance linear unbiased estimator
- Missing observations
- Multicollinearity
- Oaxaca's decomposition
- Optimal linear predictor
- Orthogonal random variables
- Principal components
- Projection matrix
- Sampling distribution
- Sampling variance
- Semiparametric
- Standard Error
- Standard error of the regression
- Statistical properties
- Stochastic regressors
- *t* ratio

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 analog 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 $f(\varepsilon_i) = (1/\lambda)\exp(-\lambda\varepsilon_i)$, $\varepsilon_i \geq 0$. This model is rather peculiar in that all the disturbances are assumed to be positive. 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:

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

   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 **y** on K variables **X** and an additional variable **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 **X** is larger when **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 **X** and **z** are nonstochastic and that the coefficient on **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 $E[\mathbf{b'b}] = \boldsymbol{\beta}'\boldsymbol{\beta} + \sigma^2 \sum_{k=1}^{K}(1/\lambda_k)$ where **b** is the ordinary least squares estimator and λ_k is a characteristic root of **X′X**.
11. Data on U.S. gasoline consumption for the years 1960 to 1995 are given in Table F2.2.
    a. Compute the multiple regression of per capita consumption of gasoline, G/pop, on all the other explanatory variables, including the time trend, and 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.

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. Notice that the price indices for the automobile market are normalized to 1967, whereas the aggregate price indices are anchored at 1982. Does this discrepancy affect the results? How? If you were to renormalize the indices so that they were all 1.000 in 1982, then how would your results change?