CHAPTER **5**

The Multiple Regression Model

LEARNING OBJECTIVES

Based on the material in this chapter, you should be able to

1. Recognize a multiple regression model and be able to interpret the coefficients in that model.
2. Understand and explain the meanings of the assumptions for the multiple regression model.
3. Use your computer to find least squares estimates of the coefficients in a multiple regression model, and interpret those estimates.
4. Explain the meaning of the Gauss–Markov theorem.
5. Compute and explain the meaning of R^2 in a multiple regression model.
6. Explain the Frisch–Waugh–Lovell Theorem and estimate examples to show how it works.
7. Use your computer to obtain variance and covariance estimates, and standard errors, for the estimated coefficients in a multiple regression model.
8. Explain the circumstances under which coefficient variances (and standard errors) are likely to be relatively high, and those under which they are likely to be relatively low.
9. Find interval estimates for single coefficients and linear combinations of coefficients, and interpret the interval estimates.
10. Test hypotheses about single coefficients and about linear combinations of coefficients in a multiple regression model. In particular,
    a. What is the difference between a one-tail and a two-tail test?
    b. How do you compute the *p*-value for a one-tail test, and for a two-tail test?
    c. What is meant by ‘‘testing the significance of a coefficient’’?
    d. What is the meaning of the *t*-values and *p*-values that appear in your computer output?
    e. How do you compute the standard error of a linear combination of coefficient estimates?
11. Estimate and interpret multiple regression models with polynomial and interaction variables.
12. Find point and interval estimates and test hypotheses for marginal effects in polynomial regressions and models with interaction variables.
13. Explain the difference between finite and large sample properties of an estimator.
14. Explain what is meant by consistency and asymptotic normality.

15. Describe the circumstances under which we can use the finite sample properties of the least squares estimator, and the circumstances under which asymptotic properties are required.

16. Use your computer to compute the standard error of a nonlinear function of estimators. Use that standard error to find interval estimates and to test hypotheses about nonlinear functions of coefficients.

KEYWORDS

asymptotic normality
BLU estimator
consistency
covariance matrix of least squares estimators
critical value
delta method
error variance estimate
error variance estimator
explained sum of squares
FWL theorem
goodness-of-fit
interaction variable
interval estimate
least squares estimates
least squares estimation
least squares estimators
linear combinations
marginal effect
multiple regression model
nonlinear functions
one-tail test
p-value
polynomial
regression coefficients
standard errors
sum of squared errors
sum of squares due to regression
testing significance
total sum of squares
two-tail test

The model in Chapters 2–4 is called a simple regression model because the dependent variable y is related to only *one* explanatory variable x. Although this model is useful for a range of situations, in most economic models there are two or more explanatory variables that influence the dependent variable y. For example, in a demand equation the quantity demanded of a commodity depends on the price of that commodity, the prices of substitute and complementary goods, and income. Output in a production function will be a function of more than one input. Aggregate money demand will be a function of aggregate income and the interest rate. Investment will depend on the interest rate and on changes in income.

When we turn an economic model with more than one explanatory variable into its corresponding econometric model, we refer to it as a **multiple regression model**. Most of the results we developed for the simple regression model in Chapters 2–4 can be extended naturally to this general case. There are slight changes in the interpretation of the β parameters, the degrees of freedom for the t-distribution will change, and we will need to modify the assumption concerning the characteristics of the explanatory (x) variables. These and other consequences of extending the simple regression model to a multiple regression model are described in this chapter.

As an example for introducing and analyzing the multiple regression model, we begin with a model used to explain sales revenue for a fast-food hamburger chain with outlets in small U.S. cities.

5.1 Introduction

5.1.1 The Economic Model

We will set up an economic model for a hamburger chain that we call Big Andy's Burger Barn.[1] Important decisions made by the management of Big Andy's include its pricing policy for different products and how much to spend on advertising. To assess the effect of different price

[1] The data we use reflect a real fast-food franchise whose identity we disguise under the name Big Andy's.

structures and different levels of advertising expenditure, Big Andy's Burger Barn sets different prices, and spends varying amounts on advertising, in different cities. Of particular interest to management is how sales revenue changes as the level of advertising expenditure changes. Does an increase in advertising expenditure lead to an increase in sales? If so, is the increase in sales sufficient to justify the increased advertising expenditure? Management is also interested in pricing strategy. Will reducing prices lead to an increase or decrease in sales revenue? If a reduction in price leads only to a small increase in the quantity sold, sales revenue will fall (demand is price-inelastic); a price reduction that leads to a large increase in quantity sold will produce an increase in revenue (demand is price-elastic). This economic information is essential for effective management.

The first step is to set up an economic model in which sales revenue depends on one or more explanatory variables. We initially hypothesize that sales revenue is linearly related to price and advertising expenditure. The economic model is

$$SALES = \beta_1 + \beta_2 PRICE + \beta_3 ADVERT \tag{5.1}$$

where *SALES* represents monthly sales (total) revenue in a given city, *PRICE* represents price in that city, and *ADVERT* is monthly advertising expenditure in that city. Both *SALES* and *ADVERT* are measured in terms of thousands of dollars. Because sales in bigger cities will tend to be greater than sales in smaller cities, we focus on smaller cities with comparable populations.

Since a hamburger outlet sells a number of products—burgers, fries, and shakes—and each product has its own price, it is not immediately clear what price should be used in (5.1). What we need is some kind of average price for all products and information on how this average price changes from city to city. For this purpose, management has constructed a single price index *PRICE*, measured in dollars and cents, that describes overall prices in each city.

The remaining symbols in (5.1) are the unknown parameters β_1, β_2, and β_3 that describe the dependence of sales (*SALES*) on price (*PRICE*) and advertising (*ADVERT*). To be more precise about the interpretation of these parameters, we move from the economic model in (5.1) to an econometric model that makes explicit assumptions about the way the data are generated.

5.1.2 The Econometric Model

When we collect data on *SALES*, *PRICE*, and *ADVERT* from the franchises in different cities, the observations will not exactly satisfy the linear relationship described in equation (5.1). The behavior of Andy's customers in different cities will not be such that the same prices and the same level of advertising expenditure will always lead to the same sales revenue. Other factors not in the equation likely to affect sales include the number and behavior of competing fast-food outlets, the nature of the population in each city—their age profile, income, and food preferences—and the location of Andy's burger barns—near a busy highway, downtown, and so on. To accommodate these factors, we include an error term e in the equation so that the model becomes

$$SALES = \beta_1 + \beta_2 PRICE + \beta_3 ADVERT + e \tag{5.2}$$

As discussed in Chapter 2, the way in which data are collected has a bearing on what assumptions are relevant and realistic for the error term e, the explanatory variables *PRICE* and *ADVERT*, and the dependent variable *SALES*. These assumptions in turn affect how we make inferences about the parameters β_1, β_2, and β_3.

Assume we take a random sample of 75 franchises in similar-sized cities in which Big Andy operates, and we observe their monthly sales, prices, and advertising expenditure. Thus, we have observations $\left(SALES_i, PRICE_i, ADVERT_i\right)$ for $i = 1, 2, \ldots, 75$. Because we do not know which cities will be chosen before we randomly sample, the triplet $\left(SALES_i, PRICE_i, ADVERT_i\right)$ is a three-dimensional random variable with a joint probability distribution. Also, the fact that we have a **random** sample implies that the observations from different cities are independent. That is,

$(SALES_i, PRICE_i, ADVERT_i)$ is independent of $(SALES_j, PRICE_j, ADVERT_j)$ for $i \neq j$. Associated with each observation is another random variable, the unobservable error term e_i that reflects the effect of factors other than *PRICE* and *ADVERT* on *SALES*. The model for the *i*th observation is written as

$$SALES_i = \beta_1 + \beta_2 PRICE_i + \beta_3 ADVERT_i + e_i \tag{5.3}$$

We assume that the effect of e_i on sales, averaged over all cities in the population, is zero, and that knowing *PRICE* and *ADVERT* for a given city does not help us predict the value of *e* for that city. At each $(PRICE_i, ADVERT_i)$ pair of values the average of the random errors is zero, that is,

$$E(e_i | PRICE_i, ADVERT_i) = 0 \tag{5.4}$$

This assumption, when combined with the assumption of independent observations generated from a random sample, implies that e_i is **strictly exogenous**. How do we check whether this is a reasonable assumption? We need to ask whether e_i includes any variables that have an effect on *SALES* (are correlated with *SALES*), *and* are also correlated with *PRICE* or *ADVERT*. If the answer is yes, strict exogeneity is violated. This might happen, for example, if the pricing and advertising behavior of Andy's competitors affects his sales, and is correlated with his own pricing and advertising policies. At the moment, it is convenient if we abstract from such a situation and continue with the strict exogeneity assumption.[2]

Using equations (5.3) and (5.4), we can write

$$E(SALES | PRICE, ADVERT) = \beta_1 + \beta_2 PRICE + \beta_3 ADVERT \tag{5.5}$$

Equation (5.5) is the *conditional mean* or *conditional expectation* of *SALES* given *PRICE* and *ADVERT* and is known as the **multiple regression function** or simply the **regression function**. It shows how the population average or population mean value for *SALES* changes depending on the settings for price and advertising expenditure. For given values of *PRICE* and *ADVERT*, some *SALES* values will fall above the mean and some below. We have dropped the subscript *i* for convenience and to emphasize that we assume this relationship holds for all cities in the population.

With this background, how do we interpret each of the parameters β_1, β_2, and β_3? Mathematically, the intercept parameter β_1 is the expected value of the dependent variable when each of the independent, explanatory variables takes the value zero. However, in many cases this parameter has no clear economic interpretation. In this particular case, it is not realistic to have a situation in which $PRICE = ADVERT = 0$. Except in very special circumstances, we always include an intercept in the model, even if it has no direct economic interpretation. Omitting it can lead to a model that fits the data poorly and that does not predict well.

The other parameters in the model measure the change in the expected value of the dependent variable given a unit change in an explanatory variable, *all other variables held constant*.

β_2 = the change in expected monthly *SALES* (\$1000) when the price index *PRICE* is increased by one unit (\$1), and advertising expenditure *ADVERT* is held constant

$$= \left.\frac{\Delta E(SALES|PRICE, ADVERT)}{\Delta PRICE}\right|_{(ADVERT \text{ held constant})} = \frac{\partial E(SALES|PRICE, ADVERT)}{\partial PRICE}$$

The symbol "∂" stands for "partial differentiation." Those of you familiar with calculus may have seen this operation. In the context above, the partial derivative of average *SALES* with respect to

[2] How to cope with violations of this assumption is considered in Chapter 10.

PRICE is the rate of change of average *SALES* as *PRICE* changes, with other factors, in this case *ADVERT*, held constant. Further details can be found in Appendix A.3.5. We will occasionally use partial derivatives, but not to an extent that will disadvantage you if you have not had a course in calculus. Rules for differentiation are provided in Appendix A.3.1.

The sign of β_2 could be positive or negative. If an increase in price leads to an increase in sales revenue, then $\beta_2 > 0$, and the demand for the chain's products is price-inelastic. Conversely, a price-elastic demand exists if an increase in price leads to a decline in revenue, in which case $\beta_2 < 0$. Thus, knowledge of the *sign* of β_2 provides information on the price-elasticity of demand. The *magnitude* of β_2 measures the amount of change in revenue for a given price change.

The parameter β_3 describes the response of expected sales revenue to a change in the level of advertising expenditure. That is,

β_3 = the change in expected monthly *SALES*($1000) when advertising expenditure *ADVERT* is increased by one unit ($1000), and the price index *PRICE* is held constant

$$= \left.\frac{\Delta E(SALES|PRICE, ADVERT)}{\Delta ADVERT}\right|_{(PRICE\text{ held constant})} = \frac{\partial E(SALES|PRICE, ADVERT)}{\partial ADVERT}$$

We expect the sign of β_3 to be positive. That is, we expect that an increase in advertising expenditure, unless the advertising is offensive, will lead to an increase in sales revenue. Whether or not the increase in revenue is sufficient to justify the added advertising expenditure, as well as the added cost of producing more hamburgers, is another question. With $\beta_3 < 1$, an increase of $1000 in advertising expenditure will yield an increase in revenue that is less than $1000. For $\beta_3 > 1$, it will be greater. Thus, in terms of the chain's advertising policy, knowledge of β_3 is very important.

Critical to the above interpretations for β_2 and β_3 is the strict exogeneity assumption $E\left(e_i|PRICE_i, ADVERT_i\right) = 0$. It implies that β_2, for example, can be interpreted as the effect of *PRICE* on *SALES*, holding all other factors constant, including the unobservable factors that form part of the error term e. We can say that a one-unit change in *PRICE causes* mean *SALES* to change by β_2 units. If the exogeneity assumption does not hold, the parameters cannot be given this causal interpretation. When $E\left(e_i|PRICE_i\right) \neq 0$, a change in price is correlated with the error term and hence the effect of a change in price cannot be captured by β_2 alone. For example, suppose that Big Andy's main competitor is Little Jim's Chicken House. And suppose that every time Andy changes his burger price, Jim responds by changing his chicken price. Because Jim's chicken price is not explicitly included in the equation, but is likely to impact on Andy's sales, its effect will be included in the error term. Also, because Jim's price is correlated with Andy's price, $E\left(e_i|PRICE_i\right) \neq 0$. Thus, a change in Andy's price (*PRICE*) will impact on *SALES* through both β_2 and the error term. Note, however, if Jim's price is added to the equation as another variable, instead of forming part of the error term, and the new error term satisfies the exogeneity assumption, then the causal interpretation of the parameter is retained.

Similar remarks can be made about the parameter for *ADVERT*, β_3.

EXAMPLE 5.1 | Data for Hamburger Chain

In the simple regression model in Chapters 2–4, the regression function was represented graphically by a line describing the relationship between $E(y|x)$ and x. With the multiple regression model with two explanatory variables, equation (5.5) describes not a line but a *plane*. As illustrated in Figure 5.1, the plane intersects the vertical axis at β_1. The parameters β_2 and β_3 measure the slope of the plane in the directions of the "price axis" and the "advertising axis," respectively. Representative observations for sales revenue, price, and advertising for some cities are displayed in Table 5.1. The complete set of observations can be found in the data file *andy* and is represented by the dots in Figure 5.1. These data do not fall exactly on a plane but instead resemble a "cloud."

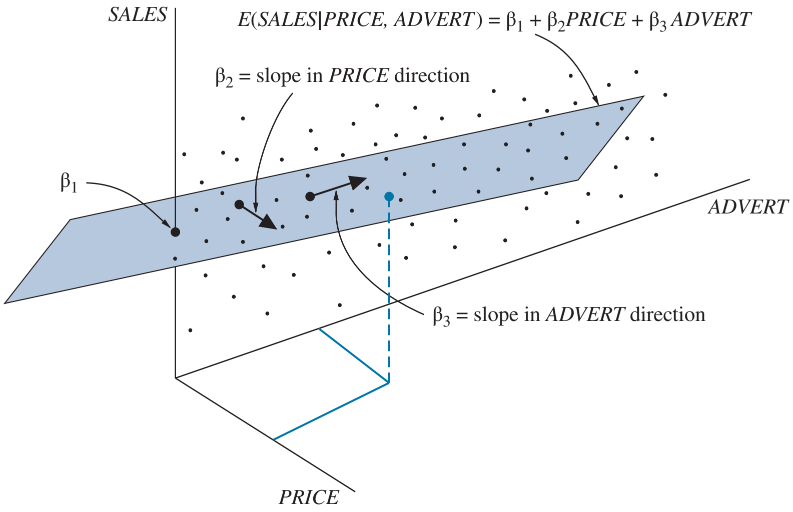
SALES
$E(SALES|PRICE, ADVERT) = \beta_1 + \beta_2 PRICE + \beta_3 ADVERT$
β_2 = slope in *PRICE* direction
β_1
ADVERT
β_3 = slope in *ADVERT* direction
PRICE

FIGURE 5.1 **The multiple regression plane.**

TABLE 5.1 **Observations on Monthly Sales, Price, and Advertising**

	SALES	*PRICE*	*ADVERT*
City	$1000 units	$1 units	$1000 units
1	73.2	5.69	1.3
2	71.8	6.49	2.9
3	62.4	5.63	0.8
4	67.4	6.22	0.7
5	89.3	5.02	1.5
.	.	.	.
.	.	.	.
.	.	.	.
73	75.4	5.71	0.7
74	81.3	5.45	2.0
75	75.0	6.05	2.2
Summary statistics			
Sample mean	77.37	5.69	1.84
Median	76.50	5.69	1.80
Maximum	91.20	6.49	3.10
Minimum	62.40	4.83	0.50
Std. Dev.	6.49	0.52	0.83

5.1.3 The General Model

It is useful to digress for a moment and summarize how the concepts developed so far relate to the general case. Working in this direction, let

$$y_i = SALES_i \quad x_{i2} = PRICE_i \quad x_{i3} = ADVERT_i$$

Then, equation (5.3) can be written as

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i \tag{5.6}$$

You might wonder why we have defined x_{i2} and x_{i3}, and not x_{i1}. We can think of the first term on the right-hand side of the equation as $\beta_1 x_{i1}$ where $x_{i1} = 1$, that is, x_{i1} is equal to 1 for all observations; it is called the **constant term**.

In Chapter 2, we used the notation $\mathbf{x}$ to denote all sample observations on a single variable x. Now that we have observations on two explanatory variables, we use the notation $\mathbf{X}$ to denote all observations on both variables as well as the constant term x_{i1}. That is, $\mathbf{X} = \left\{\left(1, x_{i2}, x_{i3}\right), i = 1, 2, \ldots, N\right\}$. In the Burger Barn example, $N = 75$. Also, it will sometimes be convenient to denote the ith observation as $\mathbf{x}_i = \left(1, x_{i2}, x_{i3}\right)$. Given this setup, the strict exogeneity assumption for the Burger Barn example, where we have a random sample with independent $\mathbf{x}_i$, is $E\left(e_i|\mathbf{x}_i\right) = 0$. For more general data generating processes where the different sample observations on $\mathbf{x}_i$ are correlated with each other, the strict exogeneity assumption is written as $E\left(e_i|\mathbf{X}\right) = 0$. If you need a refresher on the difference between $E\left(e_i|\mathbf{x}_i\right) = 0$ and $E\left(e_i|\mathbf{X}\right) = 0$, please go back and reread Section 2.2. Correlation between different observations $\left(\text{different } \mathbf{x}_i\right)$ typically exists when using time-series data. In the Burger Barn example, it could occur if our sample was not random, but taken as a collection of Barns from each of a number of states, and the pricing-advertising policies were similar for all Barns within a particular state.

We have noted the implications of the strict exogeneity assumption for the interpretation of the parameters β_2 and β_3. Later, we discuss the implications for estimator properties and inference.

There are many multiple regression models where we have more than two explanatory variables. For example, the Burger Barn model could include the price of Little Jim's Chicken, and an indicator variable equal to 1 if a Barn is near a major highway interchange, and zero otherwise. The ith observation for the general model with $K - 1$ explanatory variables and a constant term can be written as

$$y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + e_i$$

The definitions of $\mathbf{X}$ and $\mathbf{x}_i$ extend readily to this general case with $\mathbf{X} = \left\{\left(1, x_{i2}, \ldots, x_{iK}\right), i = 1, 2, \ldots, N\right\}$ and $\mathbf{x}_i = \left(1, x_{i2}, \ldots, x_{iK}\right)$. If strict exogeneity $E\left(e_i|\mathbf{X}\right) = 0$ holds, the multiple regression function is

$$E\left(y_i|\mathbf{X}\right) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \cdots + \beta_K x_{iK} \tag{5.7}$$

The unknown parameters $\beta_2, \beta_3, \ldots, \beta_K$ correspond to the explanatory variables $x_2, x_3, \ldots, x_K$. Because of this correspondence, we will also refer to $\beta_2, \beta_3, \ldots, \beta_K$ as the **coefficients** of $x_2, x_3, \ldots, x_K$. A single coefficient, call it β_k, measures the effect of a change in the variable x_k upon the expected value of y, all other variables held constant. In terms of partial derivatives,

$$\beta_k = \left.\frac{\Delta E\left(y|x_2, x_3, \ldots, x_K\right)}{\Delta x_k}\right|_{\text{other } x\text{'s held constant}} = \frac{\partial E\left(y|x_2, x_3, \ldots, x_K\right)}{\partial x_k}$$

The parameter β_1 is the intercept term. We use K to denote the total number of unknown parameters in (5.7). For a large part of this chapter, we will introduce point and interval estimation in terms of the model with $K = 3$. The results generally hold for models with more explanatory variables ($K > 3$).

5.1.4 Assumptions of the Multiple Regression Model

To complete our specification of the multiple regression model, we make further assumptions about the error term and the explanatory variables. These assumptions align with those made for the simple regression model in Section 2.2. Their purpose is to establish a framework for estimating the unknown parameters β_k, deriving the properties of the estimator for the β_k, and testing hypotheses of interest about those unknown coefficients. As we travel through the book, we discover that some of the assumptions are too restrictive for some samples of data, requiring us to weaken many of the assumptions. We will examine the implications of changes to the assumptions for estimation and hypothesis testing.

MR1: Econometric Model Observations on $\left(y_i, \mathbf{x}_i\right) = \left(y_i, x_{i2}, x_{i3}, \ldots x_{iK}\right)$ satisfy the population relationship

$$y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + e_i$$

MR2: Strict Exogeneity The conditional expectation of the random error e_i, given all explanatory variable observations $\mathbf{X} = \left\{\mathbf{x}_i, i = 1, 2, \ldots, N\right\}$, is zero.

$$E\left(e_i|\mathbf{X}\right) = 0$$

This assumption implies $E\left(e_i\right) = 0$ and $\text{cov}\left(e_i, x_{jk}\right) = 0$ for $k = 1, 2, \ldots, K$ and $(i, j) = 1, 2, \ldots, N$. Each random error has a probability distribution with zero mean. Some errors will be positive, some will be negative; over a large number of observations they will average out to zero. Also, all the explanatory variables are uncorrelated with the error; knowing values of the explanatory variables does not help predict the value of e_i. Thus, the observations will be scattered evenly above and below a plane like the one depicted in Figure 5.1. Fitting a plane through the data will make sense. Another implication of the strict exogeneity assumption is that the multiple regression function is given by

$$E\left(y_i|\mathbf{X}\right) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \cdots + \beta_K x_{iK}$$

The mean of the conditional distribution of the dependent variable y_i is a linear function of the explanatory variables $\mathbf{x}_i = \left(x_{i2}, x_{i3}, \ldots, x_{iK}\right)$.

MR3: Conditional Homoskedasticity The variance of the error term, conditional on $\mathbf{X}$, is a constant.

$$\text{var}\left(e_i|\mathbf{X}\right) = \sigma^2$$

This assumption implies $\text{var}\left(y_i|\mathbf{X}\right) = \sigma^2$ is a constant. The variability of y_i around its conditional mean function $E\left(y_i|\mathbf{X}\right) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_K x_{iK}$ does not depend on $\mathbf{X}$. The errors are not more or less likely to be larger for some values of the explanatory variables than for others. Errors with this property are said to be *homoskedastic*.[3]

MR4: Conditionally Uncorrelated Errors The covariance between different error terms e_i and e_j, conditional on $\mathbf{X}$, is zero.

$$\text{cov}\left(e_i, e_j|\mathbf{X}\right) = 0 \quad \text{for} \quad i \neq j$$

[3] Because $E\left(e_i|\mathbf{X}\right) = 0$, the unconditional variance of e_i is also constant. That is, $\text{var}\left(e_i\right) = \sigma^2$. We cannot make the same statement about the unconditional variance for y_i, however. See Appendix B, equation (B.27) for the relationship between conditional and unconditional variances.

All pairs of errors are uncorrelated. The covariance between two random errors corresponding to any two different observations is zero for all values of $\mathbf{X}$. There is no covariation or co-movement in the errors in the sense that the size of an error for one observation has no bearing on the likely size of an error for another observation. With cross-sectional data, this assumption implies that there is no spatial correlation between the errors. With time-series data, it implies there is no correlation in the errors over time. When it exists, correlation over time is referred to as serial or autocorrelation. We typically use subscripts t and s with time-series data and hence the assumption of no serial correlation can be written alternatively as $\text{cov}\left(e_t, e_s|\mathbf{X}\right) = 0$ for $t \neq s$.[4]

MR5: No Exact Linear Relationship Exists Between the Explanatory Variables

It is not possible to express one of the explanatory variables as an exact linear function of the others. Mathematically, we write this assumption as saying: The only values of $c_1, c_2, \ldots, c_K$ for which

$$c_1x_{i1} + c_2x_{i2} + \cdots + c_Kx_{iK} = 0 \quad \text{for all observations} \quad i = 1, 2, \ldots, N \tag{5.8}$$

are the values $c_1 = c_2 = \cdots = c_K = 0$. If (5.8) holds and one or more of the c_k's can be nonzero, the assumption is violated. To appreciate why this assumption is necessary, it is useful to consider some special case violations. First, suppose $c_2 \neq 0$ and the other c_k are zero. Then, (5.8) implies $x_{i2} = 0$ for all observations. If $x_{i2} = 0$, then we cannot hope to estimate β_2, which measures the effect of a change in x_{i2} on y_i, with all other factors held constant. As a second special case, suppose c_2, c_3, and c_4 are nonzero and the other c_k are zero. Then, from (5.8) we can write $x_{i2} = -\left(c_3/c_2\right)x_{i3} - \left(c_4/c_2\right)x_{i4}$. In this case, x_{i2} is an exact linear function of x_{i3} and x_{i4}. This relationship presents problems because changes in x_{i2} are completely determined by changes in x_{i3} and x_{i4}. It is not possible to separately estimate the effects of changes in each of these three variables. Put another way, there is no independent variation in x_{i2} that will enable us to estimate β_2. Our third special case relates to assumption SR5 of the simple regression model, which stated that the explanatory variable must vary. Condition (5.8) includes this case. Suppose that there is no variation in x_{i3} such that we can write $x_{i3} = 6$ for all i. Then, recalling that $x_{i1} = 1$, we can write $6x_{i1} = x_{i3}$. This outcome violates (5.8), with $c_1 = 6, c_3 = -1$ and the other c_k equal to zero.

MR6: Error Normality (*optional*)

Conditional on $\mathbf{X}$, the errors are normally distributed

$$e_i|\mathbf{X} \sim N\left(0, \sigma^2\right)$$

This assumption implies that the conditional distribution of y is also normally distributed, $y_i|\mathbf{X} \sim N\left(E\left(y_i|\mathbf{X}\right), \sigma^2\right)$. It is useful for hypothesis testing and interval estimation when samples are relatively small. However, we call it optional for two reasons. First, it is not necessary for many of the good properties of the least squares estimator to hold. Second, as we will see, if samples are relatively large, it is no longer a necessary assumption for hypothesis testing and interval estimation.

Other Assumptions

In the more advanced material in Section 2.10, we considered stronger sets of assumptions for the simple regression model that are relevant for some data generating processes—nonrandom x's, random and independent x, and random sampling, as well as the random and strictly exogenous x case considered here. The properties and characteristics of our inference procedures—estimation and hypothesis testing—established for the random and strictly exogenous x case carry over to cases where stronger assumptions are applicable.

[4] In a similar way to the assumption about conditional homoskedasticity, we can show that $\text{cov}\left(e_i, e_j|\mathbf{X}\right) = 0$ implies $\text{cov}\left(y_i, y_j|\mathbf{X}\right) = 0$ and $\text{cov}\left(e_i, e_j\right) = 0$, but the unconditional covariance $\text{cov}\left(y_i, y_j\right)$ may not be zero.

5.2 Estimating the Parameters of the Multiple Regression Model

In this section, we consider the problem of using the least squares principle to estimate the unknown parameters of the multiple regression model. We will discuss estimation in the context of the model in (5.6), which we repeat here for convenience, with *i* denoting the *i*th observation.

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i$$

This model is simpler than the full model, yet all the results we present carry over to the general case with only minor modifications.

5.2.1 Least Squares Estimation Procedure

To find an estimator for estimating the unknown parameters we follow the least squares procedure that was first introduced in Chapter 2 for the simple regression model. With the least squares principle, we find those values of $(\beta_1, \beta_2, \beta_3)$ that minimize the sum of squared differences between the observed values of y_i and their expected values $E(y_i|\mathbf{X}) = \beta_1 + x_{i2}\beta_2 + x_{i3}\beta_3$. Mathematically, we minimize the sum of squares function $S(\beta_1, \beta_2, \beta_3)$, which is a function of the unknown parameters, given the data

$$\begin{aligned} S(\beta_1, \beta_2, \beta_3) &= \sum_{i=1}^{N} \left(y_i - E(y_i|\mathbf{X})\right)^2 \\ &= \sum_{i=1}^{N} \left(y_i - \beta_1 - \beta_2 x_{i2} - \beta_3 x_{i3}\right)^2 \end{aligned} \tag{5.9}$$

Given the sample observations y_i, and $\mathbf{x}_i$, minimizing the sum of squares function is a straightforward exercise in calculus. Details of this exercise are given in Appendix 5A. The solutions give us formulas for the least squares estimators for the β coefficients in a multiple regression model with two explanatory variables. They are extensions of those given in (2.7) and (2.8) for the simple regression model with one explanatory variable. There are three reasons for relegating these formulas to Appendix 5A instead of inflicting them on you here. First, they are complicated formulas that we do not expect you to memorize. Second, we never use these formulas explicitly; computer software uses the formulas to calculate least squares estimates. Third, we frequently have models with more than two explanatory variables, in which case the formulas become even more complicated. If you proceed with more advanced study in econometrics, you will discover that there is one relatively simple matrix algebra expression for the least squares estimator that can be used for all models, irrespective of the number of explanatory variables.

Although we always get the computer to do the work for us, it is important to understand the least squares principle and the difference between least squares estimators and least squares estimates. Looked at as a general way to use sample data, formulas for b_1, b_2, and b_3, obtained by minimizing (5.9), are estimation procedures, which are called the **least squares estimators** of the unknown parameters. In general, since their values are not known until the data are observed and the estimates calculated, the least squares estimators are random variables. Computer software applies the formulas to a specific sample of data producing **least squares estimates**, which are numeric values. These least squares estimators and estimates are also referred to as **ordinary** least squares estimators and estimates, abbreviated OLS, to distinguish them from other estimators and estimates such as weighted least squares and two-stage least squares that we encounter later in the book. To avoid too much notation, we use b_1, b_2, and b_3 to denote both the estimators and the estimates.

EXAMPLE 5.2 | OLS Estimates for Hamburger Chain Data

Table 5.2 contains the least squares results for the sales equation for Big Andy's Burger Barn. The least squares estimates are

$$b_1 = 118.91 \quad b_2 = -7.908 \quad b_3 = 1.863$$

Following Example 4.3, these estimates along with their standard errors and the equation's R^2 are typically reported in equation format as

$$\begin{aligned} \widehat{SALES} &= \underset{(6.35)}{118.91} - \underset{(1.096)}{7.908}\,PRICE + \underset{(0.683)}{1.863}\,ADVERT \\ &\text{(se)} \\ R^2 &= 0.448 \end{aligned} \tag{5.10}$$

From the information in this equation, one can readily construct interval estimates or test hypotheses for each of the β_k in a manner similar to that described in Chapter 3, but with a change in the number of degrees of freedom for the t-distribution. Like before, the t-values and p-values in Table 5.2 relate to testing $H_0\!:\beta_k = 0$ against the alternative $H_1\!:\beta_k \neq 0$ for $k = 1, 2, 3$.

We proceed by first interpreting the estimates in (5.10). Then, to explain the degrees of freedom change that arises from having more than one explanatory variable, and to reinforce earlier material, we go over the sampling properties of the least squares estimator, followed by interval estimation and hypothesis testing.

What can we say about the coefficient estimates in (5.10)?

1. The negative coefficient on *PRICE* suggests that demand is price elastic; we estimate that, with advertising held constant, an increase in price of $1 will lead to a fall in mean monthly revenue of $7908. Or, expressed differently, a reduction in price of $1 will lead to an increase in mean revenue of $7908. If such is the case, a strategy of price reduction through the offering of specials would be successful in increasing sales revenue. We do need to consider carefully the magnitude of the price change, however. A $1 change in price is a relatively large change. The sample mean of price is 5.69 and its standard deviation is 0.52. A 10-cent change is more realistic, in which case we estimate the mean revenue change to be $791.
2. The coefficient on advertising is positive; we estimate that with price held constant, an increase in advertising expenditure of $1000 will lead to an increase in mean sales revenue of $1863. We can use this information, along with the costs of producing the additional hamburgers, to determine whether an increase in advertising expenditures will increase profit.
3. The estimated intercept implies that if both price and advertising expenditure were zero the sales revenue would be $118,914. Clearly, this outcome is not possible; a zero price implies zero sales revenue. In this model, as in many others, it is important to recognize that the model is an approximation to reality in the region for which we have data. Including an intercept improves this approximation even when it is not directly interpretable.

In giving the above interpretations, we had to be careful to recognize the units of measurement for each of the variables. What would happen if we measured *PRICE* in cents instead of dollars and *SALES* in dollars instead of thousands of dollars? To discover the outcome, define the new variables measured in terms of the new units as $PRICE^* = 100 \times PRICE$ and $SALES^* = 1000 \times SALES$. Substituting for *PRICE* and *SALES*, our new fitted equation becomes

$$\frac{\widehat{SALES^*}}{1000} = 118.91 - 7.908\frac{PRICE^*}{100} + 1.863ADVERT$$

Multiplying through by 1000, we obtain

$$\widehat{SALES^*} = 118{,}910 - 79.08PRICE^* + 1863ADVERT$$

This is the estimated model that we would obtain if we applied least squares to the variables expressed in terms of the new units of measurement. The standard errors would change in the same way, but the R^2 will stay the same. In this form, a more direct interpretation of the coefficients is possible. A one cent increase in *PRICE* leads to a decline in mean *SALES* of $79.08. An increase in *ADVERT* of $1000 leads to an increase in mean sales revenue of $1863.

TABLE 5.2 Least Squares Estimates for Sales Equation for Big Andy's Burger Barn

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	118.9136	6.3516	18.7217	0.0000
PRICE	−7.9079	1.0960	−7.2152	0.0000
ADVERT	1.8626	0.6832	2.7263	0.0080
$R^2 = 0.4483$	$SSE = 1718.943$	$\hat{\sigma} = 4.8861$	$s_y = 6.48854$	

In addition to providing information about how sales change when price or advertising change, the estimated equation can be used for prediction. Suppose Big Andy is interested in predicting sales revenue for a price of \$5.50 and an advertising expenditure of \$1200. Including extra decimal places to get an accurate hand calculation, this prediction is

$$\begin{aligned} SALES &= 118.91 - 7.908PRICE + 1.863ADVERT \\ &= 118.914 - 7.9079 \times 5.5 + 1.8626 \times 1.2 \\ &= 77.656 \end{aligned}$$

The predicted value of sales revenue for $PRICE = 5.5$ and $ADVERT = 1.2$ is \$77,656.

Remark

A word of caution is in order about interpreting regression results: The negative sign attached to price implies that reducing the price will increase sales revenue. If taken literally, why should we not keep reducing the price to zero? Obviously that would not keep increasing total revenue. This makes the following important point: Estimated regression models describe the relationship between the economic variables for values similar to those found in the sample data. Extrapolating the results to extreme values is generally not a good idea. Predicting the value of the dependent variable for values of the explanatory variables far from the sample values invites disaster. Refer to Figure 4.2 and the surrounding discussion.

5.2.2 Estimating the Error Variance σ^2

There is one remaining parameter to estimate—the variance of the error term. For this parameter, we follow the same steps that were outlined in Section 2.7. Under assumptions MR1, MR2, and MR3, we know that

$$\sigma^2 = \text{var}(e_i|\mathbf{X}) = \text{var}(e_i) = E(e_i^2|\mathbf{X}) = E(e_i^2)$$

Thus, we can think of σ^2 as the expectation or population mean of the squared errors e_i^2. A natural estimator of this population mean is the sample mean $\hat{\sigma}^2 = \sum e_i^2/N$. However, the squared errors e_i^2 are unobservable, so we develop an estimator for σ^2 that is based on their counterpart, the squares of the least squares residuals. For the model in (5.6), these residuals are

$$\hat{e}_i = y_i - \hat{y}_i = y_i - (b_1 + b_2x_{i2} + b_3x_{i3})$$

An estimator for σ^2 that uses the information from $\hat{e}_i^2$ and has good statistical properties is

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^N \hat{e}_i^2}{N - K} \tag{5.11}$$

where K is the number of β parameters being estimated in the multiple regression model. We can think of $\hat{\sigma}^2$ as an average of $\hat{e}_i^2$ with the denominator in the averaging process being $N - K$ instead of N. It can be shown that replacing e_i^2 by $\hat{e}_i^2$ requires the use of $N - K$ instead of N for $\hat{\sigma}^2$ to be unbiased. Note that in equation (2.19), where there was only one explanatory variable and two coefficients, we had $K = 2$.

To appreciate further why $\hat{e}_i$ provide information about σ^2, recall that σ^2 measures the variation in e_i or, equivalently, the variation in y_i around the mean function $\beta_1 + \beta_2x_{i2} + \beta_3x_{i3}$. Since $\hat{e}_i$ are estimates of e_i, big values of $\hat{e}_i$ suggest σ^2 is large while small $\hat{e}_i$ suggest σ^2 is small. When we refer to "big" values of $\hat{e}_i$, we mean big positive ones or big negative ones. Using the squares of the residuals $\hat{e}_i^2$ means that positive values do not cancel with negative ones; thus, $\hat{e}_i^2$ provide information about the parameter σ^2.

EXAMPLE 5.3 | Error Variance Estimate for Hamburger Chain Data

In the hamburger chain example, we have $K = 3$. The estimate for our sample of data in Table 5.1 is

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{75} \hat{e}_i^2}{N - K} = \frac{1718.943}{75 - 3} = 23.874$$

Go back and have a look at Table 5.2. There are two quantities in this table that relate to the above calculation. The first is the **sum of squared errors**

$$SSE = \sum_{i=1}^{N} \hat{e}_i^2 = 1718.943$$

The second is the square root of $\hat{\sigma}^2$, given by

$$\hat{\sigma} = \sqrt{23.874} = 4.8861$$

Both these quantities typically appear in the output from your computer software. Different software refer to it in different ways. Sometimes $\hat{\sigma}$ is referred to as the **standard error of the regression**. Sometimes it is called the **root mse** (short for the square root of mean squared error).

5.2.3 Measuring Goodness-of-Fit

For the simple regression model studied in Chapter 4, we introduced R^2 as a measure of the proportion of variation in the dependent variable that is explained by variation in the explanatory variable. In the multiple regression model the same measure is relevant, and the same formulas are valid, but now we talk of the proportion of variation in the dependent variable explained by *all* the explanatory variables included in the model. The coefficient of determination is

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{N} \left(\hat{y}_i - \bar{y}\right)^2}{\sum_{i=1}^{N} \left(y_i - \bar{y}\right)^2} = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{N} \hat{e}_i^2}{\sum_{i=1}^{N} \left(y_i - \bar{y}\right)^2} \tag{5.12}$$

where SSR is the variation in y "explained" by the model (**sum of squares due to regression**), SST is the total variation in y about its mean (sum of squares, total), and SSE is the sum of squared least squares residuals (errors) and is that part of the variation in y that is not explained by the model.

The notation $\hat{y}_i$ refers to the predicted value of y for each of the sample values of the explanatory variables, that is,

$$\hat{y}_i = b_1 + b_2 x_{i2} + b_3 x_{i3} + \cdots + b_K x_{iK}$$

The sample mean $\bar{y}$ is both the mean of the y_i and the mean of the $\hat{y}_i$, providing the model that includes an intercept $\left(\beta_1 \text{ in this case}\right)$.

The value for SSE will be reported by almost all computer software, but sometimes SST is not reported. Recall, however, that the sample standard deviation for y, which is readily computed by most software, is given by

$$s_y = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} \left(y_i - \bar{y}\right)^2} = \sqrt{\frac{SST}{N-1}}$$

and so

$$SST = (N - 1) s_y^2$$

EXAMPLE 5.4 | R^2 for Hamburger Chain Data

Using the results for Big Andy's Burger Barn in Table 5.2, we find that $SST = 74 \times 6.48854^2 = 3115.485$ and $SSE = 1718.943$. Using these sums of squares, we have

$$R^2 = 1 - \frac{\sum_{i=1}^{N} \hat{e}_i^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} = 1 - \frac{1718.943}{3115.485} = 0.448$$

The interpretation of R^2 is that 44.8% of the variation in sales revenue about its mean is explained by the variation in price and the variation in the level of advertising expenditure. It means that, *in our sample*, 55.2% of the variation in revenue is left unexplained and is due to variation in the error term or variation in other variables that implicitly form part of the error term.

As mentioned in Section 4.2.2, the coefficient of determination is also viewed as a measure of the predictive ability of the model over the sample period, or as a measure of how well the estimated regression fits the data. The value of R^2 is equal to the squared sample correlation coefficient between $\hat{y}_i$ and y_i. Since the sample correlation measures the linear association between two variables, if the R^2 is high, that means there is a close association between the values of y_i and the values predicted by the model, $\hat{y}_i$. In this case, the model is said to "fit" the data well. If R^2 is low, there is not a close association between the values of y_i and the values predicted by the model, $\hat{y}_i$, and the model does not fit the data well.

One final note is in order. The intercept parameter β_1 is the y-intercept of the regression "plane," as shown in Figure 5.1. If, for theoretical reasons, you are *certain* that the regression plane passes through the origin, then $\beta_1 = 0$ and it can be omitted from the model. While this is not a common practice, it does occur, and regression software includes an option that removes the intercept from the model. If the model does not contain an intercept parameter, then the measure R^2 given in (5.12) is no longer appropriate. The reason it is no longer appropriate is that, without an intercept term in the model,

$$\sum_{i=1}^{N}(y_i - \bar{y})^2 \neq \sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{N}\hat{e}_i^2$$

or, $SST \neq SSR + SSE$. To understand why, go back and check the proof in Appendix 4B. In the sum of squares decomposition the cross-product term $\sum_{i=1}^{N}(\hat{y}_i - \bar{y})\hat{e}_i$ no longer disappears. Under these circumstances, it does not make sense to talk of the proportion of total variation that is explained by the regression. Thus, when your model does not contain a constant, it is better not to report R^2, even if your computer displays one.

5.2.4 Frisch–Waugh–Lovell (FWL) Theorem

The Frisch–Waugh–Lovell (FWL) Theorem[5] is a useful and somewhat surprising result that we use a number of times in the remainder of the book. It also helps understand in a multiple regression the interpretation of a coefficient estimate, *all other variables held constant*. To illustrate[6]

[5] Also known as the Frisch–Waugh Theorem or the decomposition theorem.

[6] An illustration is not a proof. For a nonmatrix algebra proof, see Michael C. Lovell (2008) "A Simple Proof of the FWL Theorem," *Journal of Economic Education*, Winter 2008, 88–91. A proof using matrix algebra is presented in William H. Greene (2018) *Econometric Analysis, Eighth Edition*, Boston: Prentice-Hall, 36–38.

this result, we use the sales equation $SALES_i = \beta_1 + \beta_2 PRICE_i + \beta_3 ADVERT_i + e_i$ and carry out the following steps:

1. Estimate the simple regression $SALES_i = a_1 + a_2 PRICE_i + error$ using the least squares estimator and save the least squares residuals.

$$\widetilde{SALES}_i = SALES_i - \left(\hat{a}_1 + \hat{a}_2 PRICE_i\right) = SALES_i - \left(121.9002 - 7.8291 PRICE_i\right)$$

2. Estimate the simple regression $ADVERT_i = c_1 + c_2 PRICE_i + error$ using the least squares estimator and save the least squares residuals.

$$\widetilde{ADVERT}_i = ADVERT_i - \left(\hat{c}_1 + \hat{c}_2 PRICE_i\right) = ADVERT_i - \left(1.6035 + 0.0423 PRICE_i\right)$$

3. Estimate the simple regression $\widetilde{SALES}_i = \beta_3 \widetilde{ADVERT}_i + \tilde{e}_i$ with no constant term. The estimate of β_3 is $b_3 = 1.8626$. This estimate is the same as that reported from the full regression in Table 5.2.
4. Compute the least squares residuals from step 3, $\hat{\tilde{e}}_i = \widetilde{SALES}_i - b_3 \widetilde{ADVERT}_i$. Compare these residuals to those from the complete model.

$$\hat{e}_i = SALES_i - \left(b_1 + b_2 PRICE_i + b_3 ADVERT_i\right)$$

You will find that the two sets of residuals $\hat{\tilde{e}}_i$ and $\hat{e}_i$ are identical. Consequently, the sums of squared residuals are also the same, $\sum \hat{e}_i^2 = \sum \hat{\tilde{e}}_i^2 = 1718.943$.

What have we shown?

- In steps 1 and 2, we removed (or "purged" or "partialled out") the linear influence of *PRICE* (and a constant term) from both *SALES* and *ADVERT* by estimating least squares regressions and computing the least squares residuals $\widetilde{SALES}$ and $\widetilde{ADVERT}$. These residual variables are *SALES* and *ADVERT* after removing, or "partialling out," the linear influence of *PRICE* and a constant.
- In step 3, we illustrate the first important result of the **FWL theorem**: the coefficient estimate for β_3 from the regression using the partialled-out variables $\widetilde{SALES}_i = \beta_3 \widetilde{ADVERT}_i + \tilde{e}_i$ is exactly the same as that from the full regression $SALES_i = \beta_1 + \beta_2 PRICE_i + \beta_3 ADVERT_i + e_i$. We have explained β_3 as "the change in monthly sales *SALES* (\$1000) when advertising expenditure *ADVERT* is increased by one unit (\$1000), and the price index *PRICE* **is held constant**." The FWL result gives a precise meaning to "is held constant." It means that β_3 is the effect of advertising expenditure on sales after the linear influence of price and a constant term have been removed from both.
- In step 4, we note the second important result of the FWL theorem: the least squares residuals and their sum of squares are identical when calculated from the full regression or the "partialled-out" model.

A few cautions are in order. First, pay attention to the constant term. Here we have included it with *PRICE* as a variable to be partialled out in steps 1 and 2. Consequently, a constant is not included in step 3. Second, estimating the partialled-out regression is not *completely* equivalent to estimating the original, complete model. When estimating $\widetilde{SALES}_i = \beta_3 \widetilde{ADVERT}_i + \tilde{e}_i$, your software will see only one parameter to estimate, β_3. Consequently, when computing the estimate of σ^2, software will use the degrees of freedom $N - 1 = 74$. This means that the reported estimated error variance will be too small. It is $\tilde{\sigma}^2 = \sum \hat{\tilde{e}}_i^2/(N-1) = 1718.943/74 = 23.2290$ compared to the estimate from the previous section that uses divisor $N - K = 75 - 3$, $\hat{\sigma}^2 = \sum \hat{e}_i^2/(N-3) = 1718.943/72 = 23.8742$.[7] Third, for illustration we have used estimates that are rounded to four decimals. In practice, your software will use more significant digits. The results of the theorem may suffer from rounding error if insufficient significant digits are used. The estimate in step 3 is

[7] This smaller error variance estimate means that the standard errors of the regression coefficients discussed in Section 5.3.1 will be too small.

accurate to four decimals in this example, but the least squares residuals in step 4 are off without using more significant digits.

The Frisch–Waugh–Lovell Theorem also applies in the multiple regression model $y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \cdots + \beta_K x_{iK} + e_i$. Partition the explanatory variables into two groups. The theorem works for any partition, but generally the variables that are not the primary focus of the analysis are partialled out. This group is sometimes called the collection of **control variables** as they are included for a proper specification of the regression model and "control for" the variables that are not of primary interest. For example, suppose that x_2 and x_3 are the variables of primary interest. Then the two groups are $g_1 = \left(x_{i2}, x_{i3}\right)$ and $g_2 = \left(x_{i1} = 1, x_{i4}, x_{i5}, \ldots, x_{iK}\right)$. Note that we have included the constant term in group two but not group one. Each variable must go into one group or the other but not both. The FWL theorem is then applied in the following steps:

1. Estimate the least squares regression with dependent variable y and the explanatory variables $g_2 = \left(x_{i1} = 1, x_{i4}, x_{i5}, \ldots, x_{iK}\right)$. Compute the least squares residuals, $\tilde{y}$.
2. Estimate the least squares regression for each variable in group one using explanatory variables $g_2 = \left(x_{i1} = 1, x_{i4}, x_{i5}, \ldots, x_{iK}\right)$ and compute the least squares residuals, $\tilde{x}_2$ and $\tilde{x}_3$.
3. Estimate the least squares regression using the partialled-out variables, $\tilde{y}_i = \beta_2 \tilde{x}_{i2} + \beta_3 \tilde{x}_{i3} + \tilde{e}_i$. The coefficient estimates b_2 and b_3 will be identical to the estimates from the full model.
4. The residuals from the partialled-out regression, $\hat{\tilde{e}}_i = \tilde{y}_i - \left(b_2 \tilde{x}_{i2} + b_3 \tilde{x}_{i3}\right)$, are identical to the residuals from the full model.

5.3 Finite Sample Properties of the Least Squares Estimator

In a general context, the least squares estimators $\left(b_1, b_2, b_3\right)$ are random variables; they take on different values in different samples and their values are unknown until a sample is collected and their values computed. The differences from sample to sample are called "sampling variation" and are unavoidable. The probability or **sampling distribution** of the OLS estimator describes how its estimates vary over all possible samples. The **sampling properties** of the OLS estimator refer to characteristics of this distribution. If the mean of the distribution of b_k is β_k, the estimator is unbiased. The variance of the distribution provides a basis for assessing the reliability of the estimates. If the variability of b_k across samples is relatively high, then it is hard to be confident that the values obtained in one realized sample will necessarily be close to the true parameters. On the other hand, if b_k is unbiased and its variability across samples is relatively low, we can be confident that an estimate from one sample will be reliable.

What we can say about the sampling distribution of the least squares estimator depends on what assumptions can realistically be made for the sample of data being used for estimation. For the simple regression model introduced in Chapter 2 we saw that, under the assumptions SR1 to SR5, the OLS estimator is best linear unbiased in the sense that there is no other linear unbiased estimator that has a lower variance. The same result holds for the general multiple regression model under assumptions MR1–MR5.

The Gauss–Markov Theorem: If assumptions MR1–MR5 hold, the least squares estimators are the **B**est **L**inear **U**nbiased **E**stimators (BLUE) of the parameters in the multiple regression model.[8]

[8] Similar remarks can be made about the properties of the least squares estimator in the multiple regression model under the more restrictive, but sometimes realistic, assumptions explored for the simple regression model in Section 2.10. Under the assumptions in that section, if all explanatory variables are statistically independent of all error terms, or if the observations on $\left(y_i, x_{i2}, x_{i3}, \ldots, x_{iK}\right)$ are collected via random sampling making them independent, the BLUE property still holds.

The implications of adding assumption MR6, that the errors are normally distributed, are also similar to those from the corresponding assumption made for the simple regression model. Conditional on **X**, the least squares estimator is normally distributed. Using this result, and the **error variance estimator** $\hat{\sigma}^2$, a *t*-statistic that follows a *t*-distribution can be constructed and used for interval estimation and hypothesis testing, along similar lines to the development in Chapter 3.

These various properties—BLUE and the use of the *t*-distribution for interval estimation and hypothesis testing—are **finite sample** properties. As long as $N > K$, they hold irrespective of the sample size N. We provide more details in the context of the multiple regression model in the remainder of this section and in Sections 5.4 and 5.5. There are, however, many circumstances where we are unable to rely on finite sample properties. Violation of some of the assumptions can mean that finite sample properties of the OLS estimator do not hold or are too difficult to derive. Also, as we travel through the book and encounter more complex models and assumptions designed for a variety of different types of sample data, an ability to use finite sample properties becomes the exception rather than the rule. To accommodate such situations we use what are called **large sample** or **asymptotic properties**. These properties refer to the behavior of the sampling distribution of an estimator as the sample size approaches infinity. Under less restrictive assumptions, or when faced with a more complex model, large sample properties can be easier to derive than finite sample properties. Of course, we never have infinite samples, but the idea is that if N is sufficiently large, then an estimator's properties as N becomes infinite will be a good approximation to that estimator's properties when N is large but finite. We discuss large sample properties and the circumstances under which they need to be invoked in Section 5.7. An example is the central limit theorem mentioned in Section 2.6. There we learnt that, if N is sufficiently large, the least squares estimator is approximately normally distributed even when assumption SR6, which specifies that the errors are normally distributed, is violated.

5.3.1 The Variances and Covariances of the Least Squares Estimators

The variances and covariances of the least squares estimators give us information about the reliability of the estimators b_1, b_2, and b_3. Since the least squares estimators are unbiased, the smaller their variances, the higher the probability that they will produce estimates "near" the true parameter values. For $K = 3$, we can express the conditional variances and covariances in an algebraic form that provides useful insights into the behavior of the least squares estimator. For example, we can show that

$$\operatorname{var}(b_2|\mathbf{X}) = \frac{\sigma^2}{\left(1 - r_{23}^2\right)\sum_{i=1}^{N}\left(x_{i2} - \bar{x}_2\right)^2} \tag{5.13}$$

where r_{23} is the sample correlation coefficient between the values of x_2 and x_3; see Section 4.2.1. Its formula is given by

$$r_{23} = \frac{\sum\left(x_{i2} - \bar{x}_2\right)\left(x_{i3} - \bar{x}_3\right)}{\sqrt{\sum\left(x_{i2} - \bar{x}_2\right)^2 \sum\left(x_{i3} - \bar{x}_3\right)^2}}$$

For the other variances and covariances, there are formulas of a similar nature. It is important to understand the factors affecting the variance of b_2:

1. Larger error variances σ^2 lead to larger variances of the least squares estimators. This is to be expected, since σ^2 measures the overall uncertainty in the model specification. If σ^2 is large, then data values may be widely spread about the regression function $E\left(y_i|\mathbf{X}\right) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3}$, and there is less information in the data about the parameter values. If σ^2 is small, then data values are compactly spread about the regression function $E\left(y_i|\mathbf{X}\right) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3}$, and there is more information about what the parameter values might be.

2. Larger sample sizes N imply smaller variances of the least squares estimators. A larger value of N means a larger value of the summation $\sum\left(x_{i2}-\bar{x}_2\right)^2$. Since this term appears in the denominator of (5.13), when it is large, $\operatorname{var}\left(b_2\right)$ is small. This outcome is also an intuitive one; more observations yield more precise parameter estimation.
3. More variation in an explanatory variable around its mean, measured in this case by $\sum\left(x_{i2}-\bar{x}_2\right)^2$, leads to a smaller variance of the least squares estimator. To estimate β_2 precisely, we prefer a large amount of variation in x_{i2}. The intuition here is that if the variation or change in x_2 is small, it is difficult to measure the effect of that change. This difficulty will be reflected in a large variance for b_2.
4. A larger correlation between x_2 and x_3 leads to a larger variance of b_2. Note that $1-r_{23}^2$ appears in the denominator of (5.13). A value of $|r_{23}|$ close to 1 means $1-r_{23}^2$ will be small, which in turn means $\operatorname{var}\left(b_2\right)$ will be large. The reason for this fact is that variation in x_{i2} about its mean adds most to the precision of estimation when it is not connected to variation in the other explanatory variables. When the variation in one explanatory variable is connected to variation in another explanatory variable, it is difficult to disentangle their separate effects. In Chapter 6, we discuss "collinearity," which is the situation when the explanatory variables are correlated with one another. Collinearity leads to increased variances of the least squares estimators.

Although our discussion has been in terms of a model where $K=3$, these factors affect the variances of the least squares estimators in the same way in larger models.

It is customary to arrange the estimated variances and covariances of the least squares estimators in a square array, which is called a matrix. This matrix has variances on its diagonal and covariances in the off-diagonal positions. It is called a **variance–covariance matrix** or, more simply, a **covariance matrix**. When $K=3$, the arrangement of the variances and covariances in the covariance matrix is

$$\operatorname{cov}\left(b_1,b_2,b_3\right)=\begin{bmatrix}\operatorname{var}\left(b_1\right) & \operatorname{cov}\left(b_1,b_2\right) & \operatorname{cov}\left(b_1,b_3\right)\\ \operatorname{cov}\left(b_1,b_2\right) & \operatorname{var}\left(b_2\right) & \operatorname{cov}\left(b_2,b_3\right)\\ \operatorname{cov}\left(b_1,b_3\right) & \operatorname{cov}\left(b_2,b_3\right) & \operatorname{var}\left(b_3\right)\end{bmatrix}$$

Before discussing estimation of this matrix, it is useful to distinguish between the covariance matrix conditional on the observed explanatory variables $\operatorname{cov}\left(b_1,b_2,b_3|\mathbf{X}\right)$, and the unconditional covariance matrix $\operatorname{cov}\left(b_1,b_2,b_3\right)$ that recognizes that most data generation is such that both y and $\mathbf{X}$ are random variables. Given that the OLS estimator is both conditionally and unconditionally unbiased, that is, $E\left(b_k\right)=E\left(b_k|\mathbf{X}\right)=\beta_k$, the unconditional covariance matrix is given by

$$\operatorname{cov}\left(b_1,b_2,b_3\right)=E_{\mathbf{X}}\left[\operatorname{cov}\left(b_1,b_2,b_3|\mathbf{X}\right)\right]$$

Taking the variance of b_2 as an example of one of the elements in this matrix, we have

$$\operatorname{var}\left(b_2\right)=E_{\mathbf{X}}\left[\operatorname{var}\left(b_2|\mathbf{X}\right)\right]=\sigma^2E_{\mathbf{X}}\left[\frac{1}{(1-r_{23}^2)\sum_{i=1}^{N}\left(x_{i2}-\bar{x}_2\right)^2}\right]$$

We use the same quantity to estimate both $\operatorname{var}\left(b_2\right)$ and $\operatorname{var}\left(b_2|\mathbf{X}\right)$. That is,

$$\widehat{\operatorname{var}}\left(b_2\right)=\widehat{\operatorname{var}}\left(b_2|\mathbf{X}\right)=\frac{\hat{\sigma}^2}{(1-r_{23}^2)\sum_{i=1}^{N}\left(x_{i2}-\bar{x}_2\right)^2}$$

This quantity is an unbiased estimator for both $\operatorname{var}\left(b_2\right)$ and $\operatorname{var}\left(b_2|\mathbf{X}\right)$. For estimating $\operatorname{var}\left(b_2|\mathbf{X}\right)$, we replace σ^2 with $\hat{\sigma}^2$ in equation (5.13). For estimating $\operatorname{var}\left(b_2\right)$, we replace σ^2 with $\hat{\sigma}^2$ *and* the unknown expectation $E_{\mathbf{X}}\left\{\left[(1-r_{23}^2)\sum_{i=1}^{N}\left(x_{i2}-\bar{x}_2\right)^2\right]^{-1}\right\}$ with $\left[(1-r_{23}^2)\sum_{i=1}^{N}\left(x_{i2}-\bar{x}_2\right)^2\right]^{-1}$. Similar replacements are made for the other elements in the covariance matrix.

EXAMPLE 5.5 | Variances, Covariances, and Standard Errors for Hamburger Chain Data

Using the estimate $\hat{\sigma}^2 = 23.874$ and our computer software package, the estimated variances and covariances for b_1, b_2, b_3, in the Big Andy's Burger Barn example are

$$\widehat{\text{cov}}(b_1, b_2, b_3) = \begin{bmatrix} 40.343 & -6.795 & -0.7484 \\ -6.795 & 1.201 & -0.0197 \\ -0.7484 & -0.0197 & 0.4668 \end{bmatrix}$$

Thus, we have

$$\begin{aligned} \widehat{\text{var}}(b_1) &= 40.343 & \widehat{\text{cov}}(b_1, b_2) &= -6.795 \\ \widehat{\text{var}}(b_2) &= 1.201 & \widehat{\text{cov}}(b_1, b_3) &= -0.7484 \\ \widehat{\text{var}}(b_3) &= 0.4668 & \widehat{\text{cov}}(b_2, b_3) &= -0.0197 \end{aligned}$$

TABLE 5.3 Covariance Matrix for Coefficient Estimates

	C	*PRICE*	*ADVERT*
C	40.3433	−6.7951	−0.7484
PRICE	−6.7951	1.2012	−0.0197
ADVERT	−0.7484	−0.0197	0.4668

Table 5.3 shows how this information is typically reported in the output from computer software. Of particular relevance are the standard errors of b_1, b_2, and b_3; they are given by the square roots of the corresponding estimated variances. That is,

$$\begin{aligned} \text{se}(b_1) &= \sqrt{\widehat{\text{var}}(b_1)} = \sqrt{40.343} = 6.3516 \\ \text{se}(b_2) &= \sqrt{\widehat{\text{var}}(b_2)} = \sqrt{1.201} = 1.0960 \\ \text{se}(b_3) &= \sqrt{\widehat{\text{var}}(b_3)} = \sqrt{0.4668} = 0.6832 \end{aligned}$$

Again, it is time to go back and look at Table 5.2. Notice that these values appear in the standard error column.

These standard errors can be used to say something about the range of the least squares estimates if we were to obtain more samples of 75 Burger Barns from different cities. For example, the standard error of b_2 is approximately $\text{se}(b_2) = 1.1$. We know that the least squares estimator is unbiased, so its mean value is $E(b_2) = \beta_2$. Suppose b_2 is approximately normally distributed, then based on statistical theory we expect 95% of the estimates b_2, obtained by applying the least squares estimator to other samples, to be within approximately two standard deviations of the mean β_2. Given our sample, $2 \times \text{se}(b_2) = 2.2$, so we estimate that 95% of the b_2 values would lie within the interval $\beta_2 \pm 2.2$. It is in this sense that the estimated variance of b_2, or its corresponding standard error, tells us something about the reliability of the least squares estimates. If the difference between b_2 and β_2 can be large, b_2 is not reliable; if the difference between b_2 and β_2 is likely to be small, then b_2 is reliable. Whether a particular difference is "large" or "small" will depend on the context of the problem and the use to which the estimates are to be put. This issue is considered again in later sections when we use the estimated variances and covariances to test hypotheses about the parameters and to construct interval estimates.

5.3.2 The Distribution of the Least Squares Estimators

We have asserted that, under the multiple regression model assumptions MR1–MR5, listed in Section 5.1, the least squares estimator b_k is the best linear unbiased estimator of the parameter β_k in the model

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \cdots + \beta_K x_{iK} + e_i$$

If we add assumption MR6, that the random errors e_i have normal probability distributions, then, conditional on $\mathbf{X}$, the dependent variable y_i is normally distributed:

$$(y_i|\mathbf{X}) \sim N\left((\beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK}), \sigma^2\right) \quad \Longleftrightarrow \quad (e_i|\mathbf{X}) \sim N(0, \sigma^2)$$

For a given $\mathbf{X}$, the least squares estimators are linear functions of dependent variables, which means that the conditional distribution of the least squares estimators is also normal:

$$(b_k|\mathbf{X}) \sim N\left(\beta_k, \text{var}(b_k|\mathbf{X})\right)$$

That is, given $\mathbf{X}$, each b_k has a normal distribution with mean β_k and variance $\text{var}(b_k|\mathbf{X})$. By subtracting its mean and dividing by the square root of its variance, we can transform the normal random variable b_k into a standard normal variable Z with mean zero and a variance of one.

$$Z = \frac{b_k - \beta_k}{\sqrt{\operatorname{var}\left(b_k|\mathbf{X}\right)}} \sim N(0,1), \qquad \text{for } k = 1, 2, \ldots, K \tag{5.14}$$

What is particularly helpful about this result is that the distribution of Z does not depend on any unknown parameters or on $\mathbf{X}$. Although the unconditional distribution of b_k will almost certainly not be normal—it depends on the distributions of both e and $\mathbf{X}$—we can use the standard normal distribution to make probability statements about Z irrespective of whether the explanatory variables are treated as fixed or random. As mentioned in Chapter 3, statistics with this property are called **pivotal**.

There is one remaining problem, however. Before we can use (5.14) to construct interval estimates for β_k or test hypothesized values for β_k, we need to replace the unknown parameter σ^2 that is a component of var $\left(b_k|\mathbf{X}\right)$ with its estimator $\hat{\sigma}^2$. Doing so yields a t random variable given by

$$t = \frac{b_k - \beta_k}{\sqrt{\widehat{\operatorname{var}}\left(b_k|\mathbf{X}\right)}} = \frac{b_k - \beta_k}{\operatorname{se}\left(b_k\right)} \sim t_{(N-K)} \tag{5.15}$$

Like Z in equation (5.14), the distribution of this t-statistic does not depend on any unknown parameters or on $\mathbf{X}$. It is a generalization of the result in equation (3.2). A difference is the degrees of freedom of the t random variable. In Chapter 3, where there were two coefficients to be estimated, the number of degrees of freedom was $(N-2)$. In this chapter, there are K unknown coefficients in the general model and *the number of degrees of freedom for t-statistics is* $(N-K)$.

Linear Combinations of Parameters

The result in (5.15) extends to a linear combination of coefficients that was introduced in Section 3.6. Suppose that we are interested in estimating or testing hypotheses about a linear combination of coefficients that in the general case is given by

$$\lambda = c_1\beta_1 + c_2\beta_2 + \cdots + c_K\beta_K = \textstyle\sum_{k=1}^{K} c_k\beta_k$$

Then

$$t = \frac{\hat{\lambda} - \lambda}{\operatorname{se}\left(\hat{\lambda}\right)} = \frac{\sum c_k b_k - \sum c_k \beta_k}{\operatorname{se}\left(\sum c_k b_k\right)} \sim t_{(N-K)} \tag{5.16}$$

This expression is a little intimidating, mainly because we have included all coefficients to make it general, and because hand calculation of $\operatorname{se}\left(\sum c_k b_k\right)$ is onerous if more than two coefficients are involved. For example, if $K = 3$, then

$$\operatorname{se}\left(c_1b_1 + c_2b_2 + c_3b_3\right) = \sqrt{\widehat{\operatorname{var}}\left(c_1b_1 + c_2b_2 + c_3b_3|\mathbf{X}\right)}$$

where

$$\begin{aligned}\widehat{\operatorname{var}}\left(c_1b_1 + c_2b_2 + c_3b_3|\mathbf{X}\right) = {} & c_1^2\widehat{\operatorname{var}}\left(b_1|\mathbf{X}\right) + c_2^2\widehat{\operatorname{var}}\left(b_2|\mathbf{X}\right) + c_3^2\widehat{\operatorname{var}}\left(b_3|\mathbf{X}\right) + 2c_1c_2\widehat{\operatorname{cov}}\left(b_1, b_2|\mathbf{X}\right) \\ & + 2c_1c_3\widehat{\operatorname{cov}}\left(b_1, b_3|\mathbf{X}\right) + 2c_2c_3\widehat{\operatorname{cov}}\left(b_2, b_3|\mathbf{X}\right)\end{aligned}$$

In many instances some of the c_k will be zero, which can simplify the expressions and the calculations considerably. If one c_k is equal to one, and the rest are zero, (5.16) simplifies to (5.15).

What happens if the errors are not normally distributed? Then the least squares estimator will not be normally distributed and (5.14), (5.15), and (5.16) will not hold exactly. They will, however, be approximately true in large samples. Thus, having errors that are not normally distributed does not stop us from using (5.15) and (5.16), but it does mean we have to be cautious if the sample size is not large. A test for normally distributed errors was given in Section 4.3.5. An example of errors that are not normally distributed can be found in Appendix 5C.

We now examine how the results in (5.15) and (5.16) can be used for interval estimation and hypothesis testing. The procedures are identical to those described in Chapter 3, except that the degrees of freedom change.

5.4 Interval Estimation

5.4.1 Interval Estimation for a Single Coefficient

Suppose we are interested in finding a 95% **interval estimate** for β_2, the response of average sales revenue to a change in price at Big Andy's Burger Barn. Following the procedures described in Section 3.1, and noting that we have $N - K = 75 - 3 = 72$ degrees of freedom, the first step is to find a value from the $t_{(72)}$-distribution, call it t_c, such that

$$P\left(-t_c < t_{(72)} < t_c\right) = 0.95 \tag{5.17}$$

Using the notation introduced in Section 3.1, $t_c = t_{(0.975, N-K)}$ is the 97.5-percentile of the $t_{(N-K)}$-distribution (the area or probability to the left of t_c is 0.975), and $-t_c = t_{(0.025, N-K)}$ is the 2.5-percentile of the $t_{(N-K)}$-distribution (the area or probability to the left of $-t_c$ is 0.025). Consulting the t-table (Statistical Table 2), we discover there is no entry for 72 degrees of freedom, but, from the entries for 70 and 80 degrees of freedom, it is clear that, correct to two decimal places, $t_c = 1.99$. If greater accuracy is required, your computer software can be used to find $t_c = 1.993$. Using this value, and the result in (5.15) for the second coefficient ($k = 2$), we can rewrite (5.17) as

$$P\left(-1.993 \le \frac{b_2 - \beta_2}{\text{se}\left(b_2\right)} \le 1.993\right) = 0.95$$

Rearranging this expression, we obtain

$$P\left[b_2 - 1.993 \times \text{se}\left(b_2\right) \le \beta_2 \le b_2 + 1.993 \times \text{se}\left(b_2\right)\right] = 0.95$$

The interval endpoints

$$\left[b_2 - 1.993 \times \text{se}\left(b_2\right),\ \ b_2 + 1.993 \times \text{se}\left(b_2\right)\right] \tag{5.18}$$

define a 95% interval estimator of β_2. If this interval estimator is used in many samples from the population, then 95% of them will contain the true parameter β_2. We can establish this fact before any data are collected, based on the model assumptions alone. Before the data are collected, we have confidence in the **interval estimation procedure (estimator)** because of its performance over all possible samples.

EXAMPLE 5.6 | Interval Estimates for Coefficients in Hamburger Sales Equation

A 95% interval estimate for β_2 based on our particular sample is obtained from (5.18) by replacing b_2 and $\text{se}\left(b_2\right)$ by their values $b_2 = -7.908$ and $\text{se}\left(b_2\right) = 1.096$. Thus, our 95% interval estimate for β_2 is given by[9]

$$\begin{aligned}&(-7.9079 - 1.9335 \times 1.096,\ 7.9079 + 1.9335 \times 1.096)\\&\quad = (-10.093, -5.723)\end{aligned}$$

This interval estimate suggests that decreasing price by \$1 will lead to an increase in average revenue somewhere between \$5723 and \$10,093. Or, in terms of a price change whose magnitude is more realistic, a 10-cent price reduction will lead to an average revenue increase between \$572 and \$1009. Based on this information, and the cost of making and selling more burgers, Big Andy can decide whether to proceed with a price reduction.

[9] For this and the next calculation, we used more digits so that it would match the more accurate computer output. You may see us do this occasionally.

Following a similar procedure for β_3, the response of average sales revenue to advertising, we find a 95% interval estimate is given by

$$(1.8626 - 1.9935 \times 0.6832, 1.8626 + 1.9935 \times 0.6832)$$
$$= (0.501, 3.225)$$

We estimate that an increase in advertising expenditure of \$1000 leads to an increase in average sales revenue of between \$501 and \$3225. This interval is a relatively wide one; it implies that extra advertising expenditure could be unprofitable (the revenue increase is less than \$1000) or could lead to a revenue increase more than three times the cost of the advertising. Another way of describing this situation is to say that the point estimate $b_3 = 1.8626$ is not very reliable, as its standard error (which measures sampling variability) is relatively large.

In general, if an interval estimate is uninformative because it is too wide, there is nothing immediate that can be done. A wide interval for the parameter β_3 arises because the estimated sampling variability of the least squares estimator b_3 is large. In the computation of an interval estimate, a large sampling variability is reflected by a large standard error. A narrower interval can only be obtained by reducing the variance of the estimator. Based on the variance expression in (5.13), one solution is to obtain more and better data exhibiting more independent variation. Big Andy could collect data from other cities and set a wider range of price and advertising combinations. It might be expensive to do so, however, and so he would need to assess whether the extra information is worth the extra cost. This solution is generally not open to economists, who rarely use controlled experiments to obtain data. Alternatively, we might introduce some kind of nonsample information on the coefficients. The question of how to use both sample and nonsample information in the estimation process is taken up in Chapter 6.

We cannot say, in general, what constitutes an interval that is too wide, or too uninformative. It depends on the context of the problem being investigated, and on how the information is to be used.

To give a general expression for an interval estimate, we need to recognize that the **critical value** t_c will depend on the degree of confidence specified for the interval estimate and the number of degrees of freedom. We denote the degree of confidence by $1 - \alpha$; in the case of a 95% interval estimate $\alpha = 0.05$ and $1 - \alpha = 0.95$. The number of degrees of freedom is $N - K$; in Big Andy's Burger Barn example this value was $75 - 3 = 72$. The value t_c is the percentile value $t_{(1-\alpha/2, N-K)}$, which has the property that $P\left[t_{(N-K)} \le t_{(1-\alpha/2, N-K)}\right] = 1 - \alpha/2$. In the case of a 95% confidence interval, $1 - \alpha/2 = 0.975$; we use this value because we require 0.025 in each tail of the distribution. Thus, we write the general expression for a $100(1 - \alpha)\%$ confidence interval as

$$\left[b_k - t_{(1-\alpha/2,N-K)} \times \text{se}\left(b_k\right),\ b_k + t_{(1-\alpha/2,\ N-K)} \times \text{se}\left(b_k\right)\right]$$

5.4.2 Interval Estimation for a Linear Combination of Coefficients

The t-statistic in (5.16) can also be used to create interval estimates for a variety of linear combinations of parameters. Such combinations are of interest if we are considering the value of $E(y|\mathbf{X})$ for a particular setting of the explanatory variables, or the effect of changing two or more explanatory variables simultaneously. They become especially relevant if the effect of an explanatory variable depends on two or more parameters, a characteristic of many nonlinear relationships that we explore in Section 5.6.

EXAMPLE 5.7 | Interval Estimate for a Change in Sales

Big Andy wants to make next week a big sales week. He plans to increase advertising expenditure by \$800 and drop the price by 40 cents. If the prices before and after the changes are $PRICE_0$ and $PRICE_1$, respectively, and those for advertising expenditure are $ADVERT_0$ and $ADVERT_1$, then the change in expected sales from Andy's planned strategy is

$$\begin{aligned}\lambda &= E(SALES_1|PRICE_1, ADVERT_1)\\ &\quad - E(SALES_0|PRICE_0, ADVERT_0)\\ &= [\beta_1 + \beta_2 PRICE_1 + \beta_3 ADVERT_1]\\ &\quad - [\beta_1 + \beta_2 PRICE_0 + \beta_3 ADVERT_0]\\ &= [\beta_1 + \beta_2(PRICE_0 - 0.4) + \beta_3(ADVERT_0 + 0.8)]\\ &\quad - [\beta_1 + \beta_2 PRICE_0 + \beta_3 ADVERT_0]\\ &= -0.4\beta_2 + 0.8\beta_3\end{aligned}$$

Andy would like a point estimate and a 90% interval estimate for λ.

A point estimate is given by

$$\begin{aligned}\hat{\lambda} &= -0.4b_2 + 0.8b_3 = -0.4 \times (-7.9079) + 0.8 \times 1.8626\\ &= 4.6532\end{aligned}$$

Our estimate of the expected increase in sales from Big Andy's strategy is \$4653.

From (5.16), we can derive a 90% interval estimate for $\lambda = -0.4\beta_2 + 0.8\beta_3$ as

$$\begin{aligned}&\left[\hat{\lambda} - t_c \times \text{se}(\hat{\lambda}), \hat{\lambda} + t_c \times \text{se}(\hat{\lambda})\right]\\ &= \left[(-0.4b_2 + 0.8b_3) - t_c \times \text{se}(-0.4b_2 + 0.8b_3),\right.\\ &\quad \left.(-0.4b_2 + 0.8b_3) + t_c \times \text{se}(-0.4b_2 + 0.8b_3)\right]\end{aligned}$$

where $t_c = t_{(0.95, 72)} = 1.666$. Using the covariance matrix of the coefficient estimates in Table 5.3, and the result for the variance of a linear function of two random variables—see equation (3.8)—we can calculate the standard error $\text{se}(-0.4b_2 + 0.8b_3)$ as follows:

$$\begin{aligned}&\text{se}(-0.4b_2 + 0.8b_3)\\ &= \sqrt{\widehat{\text{var}}(-0.4b_2 + 0.8b_3|\mathbf{X})}\\ &= \left[(-0.4)^2\widehat{\text{var}}(b_2|\mathbf{X}) + (0.8)^2\widehat{\text{var}}(b_3|\mathbf{X})\right.\\ &\quad \left.-2 \times 0.4 \times 0.8 \times \widehat{\text{cov}}(b_2, b_3|\mathbf{X})\right]^{1/2}\\ &= [0.16 \times 1.2012 + 0.64 \times 0.4668 - 0.64 \times (-0.0197)]^{1/2}\\ &= 0.7096\end{aligned}$$

Thus, a 90% interval estimate is

$$\begin{aligned}&(4.6532 - 1.666 \times 0.7096, 4.6532 + 1.666 \times 0.7096)\\ &= (3.471, 5.835)\end{aligned}$$

We estimate, with 90% confidence, that the expected increase in sales from Big Andy's strategy will lie between \$3471 and \$5835.

5.5 Hypothesis Testing

As well as being useful for interval estimation, the *t*-distribution result in (5.15) provides the foundation for testing hypotheses about individual coefficients. As you discovered in Chapter 3, hypotheses of the form $H_0: \beta_2 = c$ versus $H_1: \beta_2 \neq c$, where c is a specified constant, are called two-tail tests. Hypotheses with inequalities such as $H_0: \beta_2 \leq c$ versus $H_1: \beta_2 > c$ are called one-tail tests. In this section, we consider examples of each type of hypothesis. For a **two-tail test**, we consider testing the significance of an individual coefficient; for one-tail tests, some hypotheses of economic interest are considered. Using the result in (5.16), one- and two-tail tests can also be used to test hypotheses about linear combinations of coefficients. An example of this type follows those for testing hypotheses about individual coefficients. We will follow the step-by-step procedure for testing hypotheses that was introduced in Section 3.4. To refresh your memory, here are the steps again:

Step-by-Step Procedure for Testing Hypotheses

1. Determine the null and alternative hypotheses.
2. Specify the test statistic and its distribution if the null hypothesis is true.
3. Select α and determine the rejection region.
4. Calculate the sample value of the test statistic and, if desired, the *p*-value.
5. State your conclusion.

At the time these steps were introduced, in Chapter 3, you had not discovered p-values. Knowing about p-values (see Section 3.5) means that steps 3–5 can be framed in terms of the test statistic and its value and/or the p-value. We will use both.

5.5.1 Testing the Significance of a Single Coefficient

When we set up a multiple regression model, we do so because we believe that the explanatory variables influence the dependent variable y. If we are to confirm this belief, we need to examine whether or not it is supported by the data. That is, we need to ask whether the data provide any evidence to suggest that y is related to each of the explanatory variables. If a given explanatory variable, say x_k, has no bearing on y, then $\beta_k = 0$. Testing this null hypothesis is sometimes called a test of significance for the explanatory variable x_k. Thus, to find whether the data contain any evidence suggesting y is related to x_k, we test the null hypothesis

$$H_0: \beta_k = 0$$

against the alternative hypothesis

$$H_1: \beta_k \neq 0$$

To carry out the test, we use the test statistic (5.15), which, if the null hypothesis is true, is

$$t = \frac{b_k}{\text{se}\left(b_k\right)} \sim t_{(N-K)}$$

For the alternative hypothesis "not equal to," we use a two-tail test, introduced in Section 3.3.3, and reject H_0 if the computed t-value is greater than or equal to t_c (the critical value from the right side of the distribution) or less than or equal to $-t_c$ (the critical value from the left side of the distribution). For a test with level of significance α, $t_c = t_{(1-\alpha/2, N-K)}$ and $-t_c = t_{(\alpha/2, N-K)}$. Alternatively, if we state the acceptance–rejection rule in terms of the p-value, we reject H_0 if $p \leq \alpha$ and do not reject H_0 if $p > \alpha$.

EXAMPLE 5.8 | Testing the Significance of Price

In the Big Andy's Burger Barn example, we test, following our standard testing format, whether sales revenue is related to price:

1. The null and alternative hypotheses are $H_0: \beta_2 = 0$ and $H_1: \beta_2 \neq 0$.
2. The test statistic, if the null hypothesis is true, is $t = b_2 / \text{se}(b_2) \sim t_{(N-K)}$.
3. Using a 5% significance level ($\alpha = 0.05$), and noting that there are 72 degrees of freedom, the critical values that lead to a probability of 0.025 in each tail of the distribution are $t_{(0.975, 72)} = 1.993$ and $t_{(0.025, 72)} = -1.993$. Thus, we reject the null hypothesis if the calculated value of t from step 2 is such that $t \geq 1.993$ or $t \leq -1.993$. If $-1.993 < t < 1.993$, we do not reject H_0. Stating the acceptance–rejection rule in terms of the p-value, we reject H_0 if $p \leq 0.05$ and do not reject H_0 if $p > 0.05$.
4. The computed value of the t-statistic is

$$t = \frac{-7.908}{1.096} = -7.215$$

From your computer software, the p-value in this case can be found as

$$P\left(t_{(72)} > 7.215\right) + P\left(t_{(72)} < -7.215\right) = 2 \times \left(2.2 \times 10^{-10}\right)$$
$$= 0.000$$

Correct to three decimal places the result is p-value $= 0.000$.

5. Since $-7.215 < -1.993$, we reject $H_0: \beta_2 = 0$ and conclude that there is evidence from the data to suggest that sales revenue depends on price. Using the p-value to perform the test, we reject H_0 because $0.000 < 0.05$.

EXAMPLE 5.9 | Testing the Significance of Advertising Expenditure

For testing whether sales revenue is related to advertising expenditure, we have

1. $H_0:\beta_3 = 0$ and $H_1:\beta_3 \neq 0$.
2. The test statistic, if the null hypothesis is true, is $t = b_3/\text{se}(b_3) \sim t_{(N-K)}$.
3. Using a 5% significance level, we reject the null hypothesis if $t \geq 1.993$ or $t \leq -1.993$. In terms of the p-value, we reject H_0 if $p \leq 0.05$. Otherwise, we do not reject H_0.
4. The value of the test statistic is

$$t = \frac{1.8626}{0.6832} = 2.726$$

The p-value is given by

$$P\left(t_{(72)} > 2.726\right) + P\left(t_{(72)} < -2.726\right) = 2 \times 0.004$$
$$= 0.008$$

5. Because $2.726 > 1.993$, we reject H_0; the data support the conjecture that revenue is related to advertising expenditure. The same test outcome can be obtained using the p-value. In this case, we reject H_0 because $0.008 < 0.05$.

Note that the t-values -7.215 (Example 5.8) and 2.726 and their corresponding p-values 0.000 and 0.008 were reported in Table 5.2 at the same time that we reported the original least squares estimates and their standard errors. Hypothesis tests of this kind are carried out routinely by computer software, and their outcomes can be read immediately from the computer output that will be similar to Table 5.2.

When we reject a hypothesis of the form $H_0:\beta_k = 0$, we say that the estimate b_k is significant. Significance of a coefficient estimate is desirable—it confirms an initial prior belief that a particular explanatory variable is a relevant variable to include in the model. However, we cannot be absolutely certain that $\beta_k \neq 0$. There is still a probability α that we have rejected a true null hypothesis. Also, as mentioned in Section 3.4, statistical significance of an estimated coefficient should not be confused with the economic importance of the corresponding explanatory variable. If the estimated response of sales revenue to advertising had been $b_3 = 0.01$ with a standard error of $\text{se}(b_3) = 0.005$, then we would have concluded that b_3 is significantly different from zero; but, since the estimate implies increasing advertising by $1000 increases revenue by only $10, we would not conclude that advertising is important. We should also be cautious about concluding that statistical significance implies precise estimation. The advertising coefficient $b_3 = 1.8626$ was found to be significantly different from zero, but we also concluded that the corresponding 95% interval estimate (0.501, 3224) was too wide to be very informative. In other words, we were not able to get a precise estimate of β_3.

5.5.2 One-Tail Hypothesis Testing for a Single Coefficient

In Section 5.1, we noted that two important considerations for the management of Big Andy's Burger Barn were whether demand was price-elastic or price-inelastic and whether the additional sales revenue from additional advertising expenditure would cover the costs of the advertising. We are now in a position to state these questions as testable hypotheses, and to ask whether the hypotheses are compatible with the data.

EXAMPLE 5.10 | Testing for Elastic Demand

With respect to demand elasticity, we wish to know whether

- $\beta_2 \geq 0$: a decrease in price leads to a change in sales revenue that is zero or negative (demand is price-inelastic or has an elasticity of unity).
- $\beta_2 < 0$: a decrease in price leads to an increase in sales revenue (demand is price-elastic).

The fast food industry is very competitive with many substitutes for Andy's burgers. We anticipate elastic demand and put this conjecture as the alternative hypothesis. Following our standard testing format, we first state the null and alternative hypotheses:

1. $H_0: \beta_2 \geq 0$ (demand is unit-elastic or inelastic)
   $H_1: \beta_2 < 0$ (demand is elastic)
2. To create a test statistic, we act as if the null hypothesis is the equality $\beta_2 = 0$. Doing so is valid because if we reject H_0 for $\beta_2 = 0$, we also reject it for any $\beta_2 > 0$. Then, assuming that $H_0: \beta_2 = 0$ is true, from (5.15) the test statistic is $t = b_2 / \text{se}(b_2) \sim t_{(N-K)}$.
3. The rejection region consists of values from the t-distribution that are unlikely to occur if the null hypothesis is true. If we define "unlikely" in terms of a 5% significance level, then unlikely values of t are those less than the critical value $t_{(0.05,\, 72)} = -1.666$. Thus, we reject H_0 if $t \leq -1.666$ or if the p-value ≤ 0.05.
4. The value of the test statistic is

$$t = \frac{b_2}{\text{se}(b_2)} = \frac{-7.908}{1.096} = -7.215$$

   The corresponding p-value is $P\left(t_{(72)} < -7.215\right) = 0.000$.
5. Since $-7.215 < -1.666$, we reject $H_0: \beta_2 \geq 0$ and conclude that $H_1: \beta_2 < 0$ (demand is elastic) is more compatible with the data. The sample evidence supports the proposition that a reduction in price will bring about an increase in sales revenue. Since $0.000 < 0.05$, the same conclusion is reached using the p-value.

Note the similarities and differences between this test and the two-tail test of significance performed in Section 5.5.1. The calculated t-values are the same, but the critical t-values are different. Not only are the values themselves different, but with a two-tail test there are also two critical values, one from each side of the distribution. With a one-tail test there is only one critical value, from one side of the distribution. Also, the p-value from the one-tail test is usually, but not always, half that of the two-tail test, although this fact is hard to appreciate from this example because both p-values are essentially zero.

EXAMPLE 5.11 | Testing Advertising Effectiveness

The other hypothesis of interest is whether an increase in advertising expenditure will bring an increase in sales revenue that is sufficient to cover the increased cost of advertising. We want proof that our advertising is profitable. If not, we may change advertising firms. Since advertising will be profitable if $\beta_3 > 1$, we set up the hypotheses:

1. $H_0: \beta_3 \leq 1$ and $H_1: \beta_3 > 1$.
2. Treating the null hypothesis as the equality $H_0: \beta_3 = 1$, the test statistic that has the t-distribution when H_0 is true is, from (5.15),

$$t = \frac{b_3 - 1}{\text{se}(b_3)} \sim t_{(N-K)}$$

3. Choosing $\alpha = 0.05$ as our level of significance, the relevant critical value is $t_{(0.95,\, 72)} = 1.666$. We reject H_0 if $t \geq 1.666$ or if the p-value ≤ 0.05.
4. The value of the test statistic is

$$t = \frac{b_3 - \beta_3}{\text{se}(b_3)} = \frac{1.8626 - 1}{0.6832} = 1.263$$

   The p-value of the test is $P\left(t_{(72)} > 1.263\right) = 0.105$.
5. Since $1.263 < 1.666$, we do not reject H_0. There is insufficient evidence in our sample to conclude that advertising will be cost-effective. Using the p-value to perform the test, we again conclude that H_0 cannot be rejected, because $0.105 > 0.05$. Another way of thinking about the test outcome is as follows: Because the estimate $b_3 = 1.8626$ is greater than one, this estimate by itself suggests that advertising will be effective. However, when we take into account the precision of estimation, measured by the standard error, we find that $b_3 = 1.8626$ is not significantly greater than one. In the context of our hypothesis-testing framework, we cannot conclude with a sufficient degree of certainty that $\beta_3 > 1$.

5.5.3 Hypothesis Testing for a Linear Combination of Coefficients

We are often interested in testing hypotheses about linear combinations of coefficients. Will particular settings of the explanatory variables lead to a mean value of the dependent variable above a certain threshold? Will changes in the values of two or more explanatory variables lead to a mean dependent variable change that exceeds a predefined goal? The *t*-statistic in (5.16) can be used to answer these questions.

EXAMPLE 5.12 | Testing the Effect of Changes in Price and Advertising

Big Andy's marketing adviser claims that dropping the price by 20 cents will be more effective for increasing sales revenue than increasing advertising expenditure by \$500. In other words, she claims that $-0.2\beta_2 > 0.5\beta_3$. Andy does not wish to accept this proposition unless it can be verified by past data. He knows that the estimated change in expected sales from the price fall is $-0.2b_2 = -0.2 \times (-7.9079) = 1.5816$, and that the estimated change in expected sales from the extra advertising is $0.5b_3 = 0.5 \times 1.8626 = 0.9319$, so the marketer's claim appears to be correct. However, he wants to establish whether the difference $1.5816 - 0.9319$ could be attributable to sampling error, or whether it constitutes proof, at a 5% significance level, that $-0.2\beta_2 > 0.5\beta_3$. This constitutes a test about a linear combination of coefficients. Since $-0.2\beta_2 > 0.5\beta_3$ can be written as $-0.2\beta_2 - 0.5\beta_3 > 0$, we are testing a hypothesis about the linear combination $-0.2\beta_2 - 0.5\beta_3$.

Following our hypothesis testing steps, we have

1. $H_0: -0.2\beta_2 - 0.5\beta_3 \le 0$ (the marketer's claim is not correct)
   $H_1: -0.2\beta_2 - 0.5\beta_3 > 0$ (the marketer's claim is correct)
2. Using (5.16) with $c_2 = -0.2$, $c_3 = -0.5$ and all other c_k's equal to zero, and assuming that the equality in H_0 holds $(-0.2\beta_2 - 0.5\beta_3 = 0)$, the test statistic and its distribution when H_0 is true are
$$t = \frac{-0.2b_2 - 0.5b_3}{\text{se}(-0.2b_2 - 0.5b_3)} \sim t_{(72)}$$
3. For a one-tail test and a 5% significance level, the critical value is $t_{(0.95,\,72)} = 1.666$. We reject H_0 if $t \ge 1.666$ or if the *p*-value ≤ 0.05.
4. To find the value of the test statistic, we first compute
$$\begin{aligned}
&\text{se}(-0.2b_2 - 0.5b_3)\\
&= \sqrt{\widehat{\text{var}}(-0.2b_2 - 0.5b_3|\mathbf{X})}\\
&= \Big[(-0.2)^2\widehat{\text{var}}(b_2|\mathbf{X}) + (-0.5)^2\widehat{\text{var}}(b_3|\mathbf{X})\\
&\qquad + 2 \times (-0.2) \times (-0.5) \times \widehat{\text{cov}}(b_2, b_3|\mathbf{X})\Big]^{1/2}\\
&= [0.04 \times 1.2012 + 0.25 \times 0.4668 + 0.2 \times (-0.0197)]^{1/2}\\
&= 0.4010
\end{aligned}$$
   Then, the value of the test statistic is
$$t = \frac{-0.2b_2 - 0.5b_3}{\text{se}(-0.2b_2 - 0.5b_3)} = \frac{1.58158 - 0.9319}{0.4010} = 1.622$$
   The corresponding *p*-value is $P(t_{(72)} > 1.622) = 0.055$.
5. Since $1.622 < 1.666$, we do not reject H_0. At a 5% significance level, there is not enough evidence to support the marketer's claim. Alternatively, we reach the same conclusion using the *p*-value, because $0.055 > 0.05$.

5.6 Nonlinear Relationships

The multiple regression model that we have studied so far has the form

$$y = \beta_1 + \beta_2 x_2 + \cdots + \beta_K x_K + e \tag{5.19}$$

It is a linear function of variables (the *x*'s) and of the coefficients (the β's) and *e*. However, (5.19) is much more flexible than it at first appears. Although the assumptions of the multiple regression model require us to retain the property of linearity in the β's, many different nonlinear functions of variables can be specified by defining the *x*'s and/or *y* as transformations of original variables. Several examples of such transformations have already been encountered for the simple regression model. In Chapter 2, the quadratic model $y = \alpha_1 + \alpha_2 x^2 + e$ and the log-linear model $\ln(y) = \gamma_1 + \gamma_2 x + e$ were estimated. A detailed analysis of these and other nonlinear simple regression models—a linear-log model, a log-log model, and a cubic model—was given in Chapter 4. The same kind of variable transformations and interpretations of their coefficients carry over to multiple regression models. One class of models is that of **polynomial** equations such as the quadratic $y = \beta_1 + \beta_2 x + \beta_3 x^2 + e$ or the cubic $y = \alpha_1 + \alpha_2 x + \alpha_3 x^2 + \alpha_4 x^3 + e$. When we studied these models as examples of the simple regression model, we were constrained by the need to have only one right-hand-side variable, such as $y = \beta_1 + \beta_3 x^2 + e$ or $y = \alpha_1 + \alpha_4 x^3 + e$. Now that we are working within the framework of the multiple regression model, we can consider unconstrained polynomials with all their terms included. Another generalization is to include "cross-product" or "interaction" terms leading to a model such as $y = \gamma_1 + \gamma_2 x_2 + \gamma_3 x_3 + \gamma_4 x_2 x_3 + e$. In this section, we explore a few of the many options that are available for modeling nonlinear relationships. We begin with some examples of polynomial functions from economics. Polynomials are a rich class of functions that can parsimoniously describe relationships that are curved, with one or more peaks and valleys.

EXAMPLE 5.13 | Cost and Product Curves

In microeconomics, you studied "cost" curves and "product" curves that describe a firm. Total cost and total product curves are mirror images of each other, taking the standard "cubic" shapes shown in Figure 5.2. Average and marginal cost curves, and their mirror images, average and marginal product curves, take quadratic shapes, usually represented as shown in Figure 5.3.

The slopes of these relationships are not constant and cannot be represented by regression models that are "linear in the variables." However, these shapes are easily represented by polynomials. For example, if we consider the average cost relationship in Figure 5.3(a), a suitable regression model is

$$AC = \beta_1 + \beta_2 Q + \beta_3 Q^2 + e \qquad (5.20)$$

This quadratic function can take the "U" shape we associate with average cost functions. For the total cost curve in Figure 5.2(a), a cubic polynomial is in order,

$$TC = \alpha_1 + \alpha_2 Q + \alpha_3 Q^2 + \alpha_4 Q^3 + e \qquad (5.21)$$

These functional forms, which represent nonlinear shapes, can still be estimated using the least squares methods we have

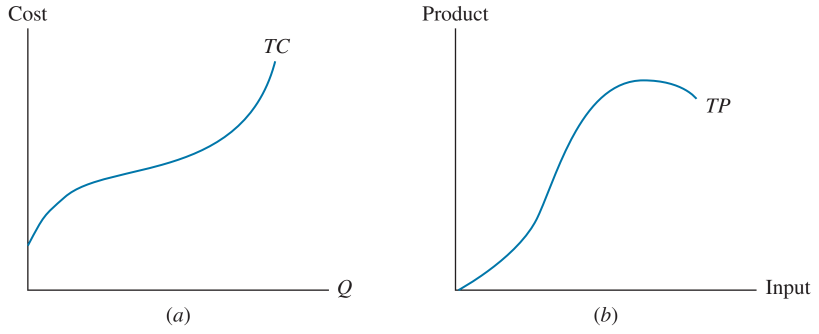

FIGURE 5.2 (*a*) Total cost curve and (*b*) total product curve.

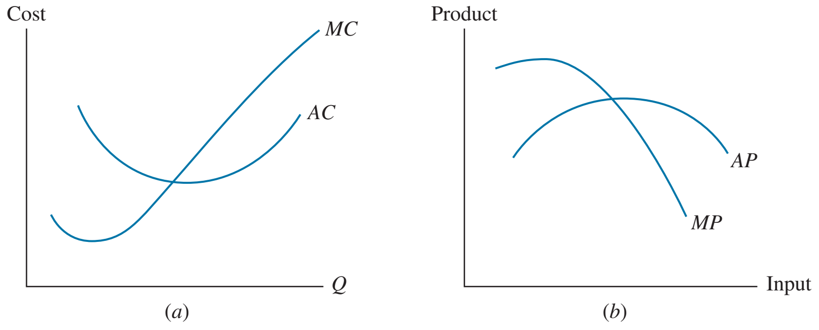

FIGURE 5.3 Average and marginal (*a*) cost curves and (*b*) product curves.

studied. The variables Q^2 and Q^3 are explanatory variables that are treated no differently from any others.

A difference in models of nonlinear relationships is in the interpretation of the parameters, which are not themselves slopes. To investigate the slopes, and to interpret the parameters, we need a little calculus. For the general polynomial function,

$$y = a_0 + a_1x + a_2x^2 + a_3x^3 + \cdots + a_px^p$$

the slope or derivative of the curve is

$$\frac{dy}{dx} = a_1 + 2a_2x + 3a_3x^2 + \cdots + pa_px^{p-1} \tag{5.22}$$

This slope changes depending on the value of x. Evaluated at a particular value, $x = x_0$, the slope is

$$\left.\frac{dy}{dx}\right|_{x=x_0} = a_1 + 2a_2x_0 + 3a_3x_0^2 + \cdots + pa_px_0^{p-1}$$

For more on rules of derivatives, see Appendix A.3.1.

Using the general rule in (5.22), the slope of the average cost curve (5.20) is

$$\frac{dE(AC)}{dQ} = \beta_2 + 2\beta_3Q$$

The slope of the average cost curve changes for every value of Q and depends on the parameters β_2 and β_3. For this U-shaped curve, we expect $\beta_2 < 0$ and $\beta_3 > 0$. The slope of the total cost curve (5.21), which is the marginal cost, is

$$\frac{dE(TC)}{dQ} = \alpha_2 + 2\alpha_3Q + 3\alpha_4Q^2$$

The slope is a quadratic function of Q, involving the parameters α_2, α_3, and α_4. For a U-shaped marginal cost curve, we expect the parameter signs to be $\alpha_2 > 0$, $\alpha_3 < 0$, and $\alpha_4 > 0$.

Using polynomial terms is an easy and flexible way to capture nonlinear relationships between variables. As we have shown, care must be taken when interpreting the parameters of models that contain polynomial terms. Their inclusion does not complicate **least squares estimation**—with one exception. It is sometimes true that having a variable and its square or cube in the same model causes **collinearity** problems. (See Section 6.4.)

EXAMPLE 5.14 | Extending the Model for Burger Barn Sales

In the Burger Barn model $SALES = \beta_1 + \beta_2PRICE + \beta_3ADVERT + e$, it is worth questioning whether the *linear* relationship between sales revenue, price, and advertising expenditure is a good approximation of reality. Having a linear model implies that increasing advertising expenditure will continue to increase sales revenue at the same rate irrespective of the existing levels of sales revenue and advertising expenditure—that is, that the coefficient β_3, which measures the response of $E(SALES|PRICE, ADVERT)$ to a change in $ADVERT$, is constant; it does not depend on the level of $ADVERT$. In reality, as the level of advertising expenditure increases, we would expect diminishing returns to set in. To illustrate what is meant by diminishing returns, consider the relationship between sales and advertising (assuming a fixed price) graphed in Figure 5.4. The figure shows the effect on sales of an increase of \$200 in advertising expenditure when the original level of advertising is (a) \$600 and (b) \$1,600. Note that the units in the graph are thousands

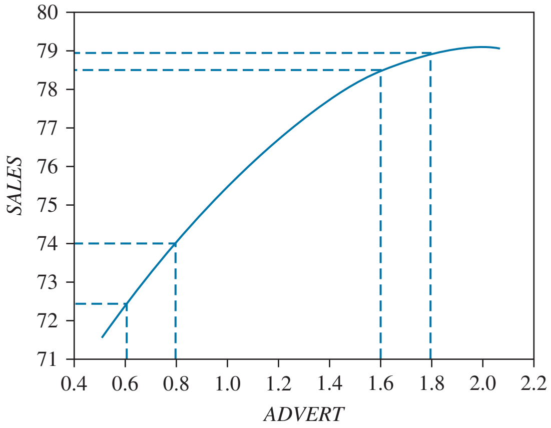

FIGURE 5.4 **A model where sales exhibits diminishing returns to advertising expenditure.**

of dollars, so these points appear as 0.6 and 1.6. At the smaller level of advertising, sales increase from \$72,400 to \$74,000, whereas at the higher level of advertising, the increase is a much smaller one, from \$78,500 to \$79,000. The linear model with the constant slope β_3 does not capture the diminishing returns.

What is required is a model where the slope changes as the level of *ADVERT* increases. One such model having this characteristic is obtained by including the squared value of advertising as another explanatory variable, making the new model

$$SALES = \beta_1 + \beta_2 PRICE + \beta_3 ADVERT + \beta_4 ADVERT^2 + e \tag{5.23}$$

Adding the term $\beta_4 ADVERT^2$ to our original specification yields a model in which the response of expected revenue to a change in advertising expenditure depends on the level of advertising. Specifically, by applying the polynomial derivative rule in (5.22), and holding *PRICE* constant, the response of $E(SALES|PRICE, ADVERT)$ to a change in *ADVERT* is

$$\left.\frac{\Delta E(SALES|PRICE, ADVERT)}{\Delta ADVERT}\right|_{(PRICE \text{ held constant})}$$
$$= \frac{\partial E(SALES|PRICE, ADVERT)}{\partial ADVERT} = \beta_3 + 2\beta_4 ADVERT \tag{5.24}$$

The partial derivative sign "∂" is used in place of the derivative sign "d" that we used in (5.22) because *SALES* depends on two variables, *PRICE* and *ADVERT*, and we are holding *PRICE* constant. See Appendix A.3.5 for further details about partial derivatives.

We refer to $\partial E(SALES|PRICE, ADVERT)/\partial ADVERT$ in (5.24) as the **marginal effect** of advertising on sales. In linear functions, the slope or marginal effect is constant. In nonlinear functions, it varies with one or more of the variables. To find the expected signs for β_3 and β_4, note that we expect the response of sales revenue to a change in advertising to be positive when $ADVERT = 0$. That is, we expect $\beta_3 > 0$. Also, to achieve diminishing returns, the response must decline as *ADVERT* increases. That is, we expect $\beta_4 < 0$.

Using least squares to estimate (5.23) yields

$$\begin{aligned} \widehat{SALES} &= \underset{(6.80)}{109.72} - \underset{(1.046)}{7.640}PRICE + \underset{(3.556)}{12.151}ADVERT \\ \text{(se)} \quad & \quad - \underset{(0.941)}{2.768}ADVERT^2 \end{aligned} \tag{5.25}$$

What can we say about the addition of $ADVERT^2$ to the equation? Its coefficient has the expected negative sign and is significantly different from zero at a 5% significance level. Moreover, the coefficient of *ADVERT* has retained its positive sign and continues to be significant. The estimated response of sales to advertising is

$$\frac{\partial \widehat{SALES}}{\partial ADVERT} = 12.151 - 5.536 ADVERT$$

Substituting into this expression we find that when advertising is at its minimum value in the sample of \$500 ($ADVERT = 0.5$), the marginal effect of advertising on sales is 9.383. When advertising is at a level of \$2000 ($ADVERT = 2$), the marginal effect is 1.079. Thus, allowing for diminishing returns to advertising expenditure has improved our model both statistically and in terms of meeting our expectations about how sales will respond to changes in advertising.

EXAMPLE 5.15 | An Interaction Variable in a Wage Equation

In the last example, we saw how the inclusion of $ADVERT^2$ in the regression model for *SALES* has the effect of making the marginal effect of *ADVERT* on *SALES* depend on the level of *ADVERT*. What if the marginal effect of one variable depends on the level of another variable? How do we model it? To illustrate, consider a wage equation relating *WAGE* (\$ earnings per hour) to years of education (*EDUC*) and years of experience (*EXPER*) in the following way:

$$\begin{aligned} WAGE = \beta_1 + \beta_2 EDUC + \beta_3 EXPER \\ + \beta_4(EDUC \times EXPER) + e \end{aligned} \tag{5.26}$$

Here we are suggesting that the effect of another year's experience on wage may depend on a worker's level of education, and, similarly, the effect of another year of education may depend on the number of years of experience. Specifically,

$$\frac{\partial E(WAGE|EDUC, EXPER)}{\partial EXPER} = \beta_3 + \beta_4 EDUC$$

$$\frac{\partial E(WAGE|EDUC, EXPER)}{\partial EDUC} = \beta_2 + \beta_4 EXPER$$

Using the Current Population Survey data (*cps5_small*) to estimate (5.26), we obtain

$$\begin{aligned} \widehat{WAGE} &= -\underset{(4.162)}{18.759} + \underset{(0.2833)}{2.6557}EDUC + \underset{(0.1335)}{0.2384}EXPER \\ \text{(se)} \quad & \quad - \underset{(0.009400)}{0.002747}(EDUC \times EXPER) \end{aligned}$$

The negative estimate $b_4 = -0.002747$ suggests that the greater the number of years of education, the less valuable is an extra year of experience. Similarly, the greater the number of years of experience, the less valuable is an extra year of education. For a person with eight years of education, we estimate that an additional year of experience leads to an increase in average wages of $0.2384 - 0.002747 \times 8 = \$0.22$, whereas for a person with 16 years of education, the approximate increase in wages from an extra year of experience is $0.2384 - 0.002747 \times 16 = \$0.19$. For someone with no experience, the extra average wage from an extra year of education is \$2.66. The value of an extra year of education falls to $2.6557 - 0.002747 \times 20 = \$2.60$ for someone with 20 years of experience. These differences are not large. Perhaps there is no interaction effect—its estimated coefficient is not significantly different from zero—or perhaps we could improve the specification of the model.

EXAMPLE 5.16 | A Log-Quadratic Wage Equation

In equation (5.26), we used *WAGE* as the dependent variable whereas, when we previously studied a wage equation in Example 4.10, ln(*WAGE*) was chosen as the dependent variable. Labor economists tend to prefer ln(*WAGE*), believing that a change in years of education or experience is more likely to lead to a constant percentage change in *WAGE* than a constant absolute change. Also, a wage distribution will typically be heavily skewed to the right. Taking logarithms yields a distribution, which is shaped more like a normal distribution.

In the following example, we make two changes to the model in (5.26). We replace *WAGE* with ln(*WAGE*), and we add the variable *EXPER*². Adding *EXPER*² is designed to capture diminishing returns to extra years of experience. An extra year of experience for an old hand with many years of experience is likely to be less valuable than it would be for a rookie with limited or no experience. Thus, we specify the model

$$\begin{aligned} \ln(WAGE) &= \beta_1 + \beta_2 EDUC + \beta_3 EXPER \\ &\quad + \beta_4(EDUC \times EXPER) + \beta_5 EXPER^2 + e \end{aligned} \tag{5.27}$$

Here the two marginal effects which, when multiplied by 100 give the approximate percentage changes in wages from extra years of experience and education, respectively, are

$$\begin{aligned} &\frac{\partial E[\ln(WAGE)\,|EDUC, EXPER]}{\partial EXPER} \\ &\quad = \beta_3 + \beta_4 EDUC + 2\beta_5 EXPER \end{aligned} \tag{5.28}$$

$$\frac{\partial E[\ln(WAGE)\,|EDUC, EXPER]}{\partial EDUC} = \beta_2 + \beta_4 EXPER \tag{5.29}$$

Having both the interaction term and the square of *EXPER* in the equation means that the marginal effect for experience will depend on both the level of education and the number of years of experience. Estimating (5.27) using the data in *cps5_small* yields

$$\begin{array}{lllll} \widehat{\ln(WAGE)} = & 0.6792 & +\,0.1359EDUC & +\,0.04890EXPER \\ \text{(se)} & (0.1561) & (0.0101) & (0.00684) \\ & -\,0.001268(EDUC \times EXPER) & & \\ & \ \big(0.000342\big) & & \\ & -\,0.0004741EXPER^2 & & \\ & \ (0.0000760) & & \end{array}$$

In this case, all estimates are significantly different from zero. Estimates of the percentage changes in wages from extra years of experience and extra years of education, computed using (5.28) and (5.29) for $EDUC = 8$ and 16 and $EXPER = 0$ and 20, are presented in Table 5.4. As expected, the value of an extra year of experience is greatest for someone with 8 years of education and no experience and smallest for someone with 16 years of education and 20 years of experience. We estimate that the value of an extra year of education is $13.59 - 11.06 = 2.53$ percentage points less for someone with 20 years of experience relative to someone with no experience.

TABLE 5.4 Percentage Changes in Wage

		% Δ*WAGE*/Δ*EXPER*		%Δ*WAGE*/Δ*EDUC*
		Years of education		
		8	16	
Years of experience	0	3.88	2.86	13.59
	20	1.98	0.96	11.06

5.7 Large Sample Properties of the Least Squares Estimator

It is nice to be able to use the finite sample properties of the OLS estimator or, indeed, any other estimator, to make inferences about population parameters[10]. Provided our assumptions are correct, we can be confident that we are basing our conclusions on procedures that are exact, whatever the sample size. However, the assumptions we have considered so far are likely to be too restrictive for many data sets. To accommodate less restrictive assumptions, as well as carry out inference for general functions of parameters, we need to examine the properties of estimators as sample size approaches infinity. Properties as sample size approaches infinity provide a good guide to properties in large samples. They will always be an approximation, but it is an approximation that improves as sample size increases. Large sample approximate properties are known as **asymptotic properties**. A question students always ask and instructors always evade is "how large does the sample have to be?" Instructors are evasive because the answer depends on the model, the estimator, and the function of parameters that is of interest. Sometimes $N = 30$ is adequate; sometimes $N = 1000$ or larger could be necessary. Some illustrations are given in the Monte Carlo experiments in Appendix 5C. In Appendix 5D, we explain how bootstrapping can be used to check whether a sample size is large enough for asymptotic properties to hold.

In this section, we introduce some large sample (asymptotic) properties and then discuss some of the circumstances where they are necessary.

5.7.1 Consistency

When choosing econometric estimators, we do so with the objective in mind of obtaining an estimate that is close to the true but unknown parameter with high probability. Consider the simple linear regression model $y_i = \beta_1 + \beta_2 x_i + e_i, i = 1, \ldots, N$. Suppose that for decision-making purposes we consider that obtaining an estimate of β_2 within "epsilon" of the true value is satisfactory. The probability of obtaining an estimate "close" to β_2 is

$$P\left(\beta_2 - \varepsilon \leq b_2 \leq \beta_2 + \varepsilon\right) \tag{5.30}$$

An estimator is said to be **consistent** if this probability converges to 1 as the sample size $N \to \infty$. Or, using the concept of a limit, the estimator b_2 is consistent if

$$\lim_{N\to\infty} P\left(\beta_2 - \varepsilon \leq b_2 \leq \beta_2 + \varepsilon\right) = 1 \tag{5.31}$$

What does this mean? In Figure 5.5, we depict the probability density functions $f\left(b_{N_i}\right)$ for the least squares estimator b_2 based on samples sizes $N_4 > N_3 > N_2 > N_1$. As the sample size increases the probability density function (*pdf*) becomes narrower. Why is that so? First of all, the least squares estimator is unbiased if MR1–MR5 hold, so that $E\left(b_2\right) = \beta_2$. This property is true in samples of all sizes. As the sample size changes, the center of the *pdf*s remains at β_2. However, as the sample size N gets larger, the variance of the estimator b_2 becomes smaller. The center of the *pdf* remains fixed at $E\left(b_2\right) = \beta_2$, and the variance decreases, resulting in probability density functions like $f\left(b_{N_i}\right)$. The probability that b_2 falls in the interval $\beta_2 - \varepsilon \leq b_2 \leq \beta_2 + \varepsilon$ is the area under the *pdf* between these limits. As the sample size increases, the probability of b_2 falling within the limits increases toward 1. In large samples, we can say that the least squares estimator will provide an estimate close to the true parameter with high probability.

[10] This section contains advanced materials.

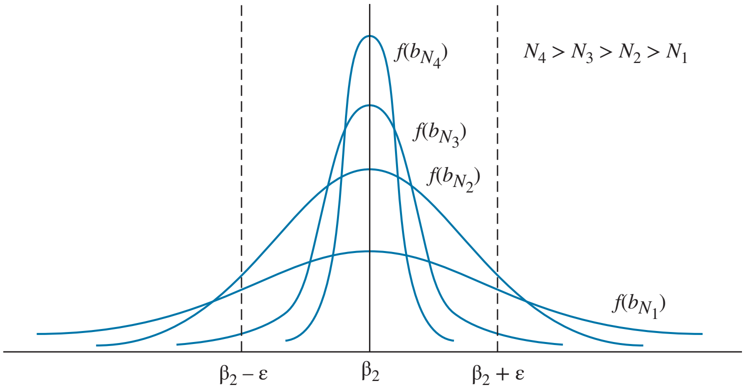

FIGURE 5.5 **An illustration of consistency.**

To appreciate why the variance decreases as N increases, consider the variance of the OLS estimator that we rewrite as follows:

$$\operatorname{var}(b_2) = \sigma^2 E\left(\frac{1}{\sum_{i=1}^{N}(x_i - \bar{x})^2}\right) = \frac{\sigma^2}{N} E\left(\frac{1}{\sum_{i=1}^{N}(x_i - \bar{x})^2 / N}\right) = \frac{\sigma^2}{N} E\left[\left(s_x^2\right)^{-1}\right] = \frac{\sigma^2}{N} C_x \quad (5.32)$$

Notice that the N's that we have introduced cancel out. This trick is used so that we can write the variance for b_2 in terms of the sample variance of x, $s_x^2 = \sum_{i=1}^{N}(x_i - \bar{x})^2 / N$.[11] Then, because $E\left[\left(s_x^2\right)^{-1}\right]$ is cumbersome, and a little intimidating, in the last equality we define the constant C_x as the expectation of the inverse of the sample variance. That is, $C_x = E\left[\left(s_x^2\right)^{-1}\right]$. The last result in (5.32) implies $\operatorname{var}(b_2) \to 0$ as $N \to \infty$.

The property of **consistency** applies to many estimators, even ones that are biased in finite samples. For example, the estimator $\hat{\beta}_2 = b_2 + 1/N$ is a biased estimator. The amount of the bias is

$$\operatorname{bias}\left(\hat{\beta}_2\right) = E\left(\hat{\beta}_2\right) - \beta_2 = \frac{1}{N}$$

For the estimator $\hat{\beta}_2$ the bias converges to zero as $N \to \infty$. That is,

$$\lim_{N \to \infty} \operatorname{bias}\left(\hat{\beta}_2\right) = \lim_{N \to \infty} \left[E\left(\hat{\beta}_2\right) - \beta_2\right] = 0 \quad (5.33)$$

In this case, the estimator is said to be **asymptotically unbiased**. Consistency for an estimator can be established by showing that the estimator is either unbiased or asymptotically unbiased, and that its variance converges to zero as $N \to \infty$,

$$\lim_{N \to \infty} \operatorname{var}\left(\hat{\beta}_2\right) = 0 \quad (5.34)$$

Conditions (5.33) and (5.34) are intuitive, and sufficient to establish an estimator to be consistent.

Because the probability density function of a consistent estimator collapses around the true parameter, and the probability that an estimator b_2 will be close to the true parameter β_2 approaches 1, the estimator b_2 is said to "converge in probability" to β_2, with the "in probability" part reminding us that it is the probability of being "close" in (5.31) that is the key factor. Several notations are used for this type of convergence. One is $b_2 \xrightarrow{p} \beta_2$, with the p over the arrow

[11] We have used N rather than $N - 1$ as the divisor for the sample variance. When dealing with properties as $N \to \infty$, it makes no difference which is used.

indicating "probability." A second is $\underset{N\to\infty}{\text{plim}}\left(b_2\right) = \beta_2$, with "plim" being short for "probability limit." Consistency is not just a large-sample alternative to unbiasedness; it is an important property in its own right. It is possible to find estimators that are unbiased but not consistent. The lack of consistency is considered undesirable even if an estimator is unbiased.

5.7.2 Asymptotic Normality

We mentioned earlier that the normal distribution assumption MR6: $\left(e_i|\mathbf{X}\right) \sim N\left(0, \sigma^2\right)$ is essential for the finite sample distribution of $\left(b_k|\mathbf{X}\right)$ to be normal and for t-statistics such as $t = \left(b_k - \beta_k\right)/\text{se}\left(b_k\right)$ to have an exact t-distribution for use in interval estimation and hypothesis testing. However, we then went on to say that all is not lost if the normality assumption does not hold because, from a central limit theorem, the distribution of b_k will be approximately normal and interval estimates and t-tests will be approximately valid in large samples. Large sample approximate distributions are called **asymptotic distributions**. The need to use asymptotic distributions will become more urgent as we examine more complex models and estimators.

To appreciate how asymptotic distributions work and to introduce some notation, consider the OLS estimator b_2 in the simple regression model $y_i = \beta_1 + \beta_2 x_i + e_i, i = 1, \ldots, N$. We argued that the consistency of b_2 implies that the *pdf* for b_2 collapses to the point β_2 as $N \to \infty$. How can we get an approximate large sample distribution for b_2 if its *pdf* collapses to a single point? We consider instead the distribution of $\sqrt{N}b_2$. Note that $E\left(b_2\right) = \beta_2$ and that, from (5.32), $\text{var}\left(b_2\right) = \sigma^2 C_x/N$. It follows that $E\left(\sqrt{N}b_2\right) = \sqrt{N}\beta_2$ and

$$\text{var}\left(\sqrt{N}b_2\right) = \left(\sqrt{N}\right)^2 \text{var}\left(b_2\right) = N\sigma^2 C_x/N = \sigma^2 C_x$$

That is,

$$\sqrt{N}b_2 \sim \left(\sqrt{N}\beta_2, \sigma^2 C_x\right) \tag{5.35}$$

Central limit theorems are concerned with the distribution of sums (or averages) of random variables as $N \to \infty$.[12] In Chapter 2—see equation (2.12)—we showed that $b_2 = \beta_2 + \left[\sum_{i=1}^N \left(x_i - \bar{x}\right)^2\right]^{-1} \sum_{i=1}^N \left(x_i - \bar{x}\right) e_i$ from which we can write

$$\sqrt{N}b_2 = \sqrt{N}\beta_2 + \left[s_x^2\right]^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^N \left(x_i - \bar{x}\right) e_i$$

Applying a central limit theorem to the sum $\sum_{i=1}^N \left(x_i - \bar{x}\right) e_i/\sqrt{N}$, and using $\left[s_x^2\right]^{-1} \xrightarrow{p} C_x$, it can be shown that the statistic obtained by normalizing (5.35) so that it has mean zero and variance one, will be approximately normally distributed. Specifically,

$$\frac{\sqrt{N}\left(b_2 - \beta_2\right)}{\sqrt{\sigma^2 C_x}} \overset{a}{\sim} N(0, 1)$$

The notation $\overset{a}{\sim}$ is used to denote the asymptotic or approximate distribution. Recognizing that $\text{var}\left(b_2\right) = \sigma^2 C_x/N$, we can rewrite the above result as

$$\frac{\left(b_2 - \beta_2\right)}{\sqrt{\text{var}\left(b_2\right)}} \overset{a}{\sim} N(0, 1)$$

[12] There are several central limit theorems designed to accommodate sums of random variables with different properties. Their treatment is relatively advanced. See, for example, William Greene, *Econometric Analysis 8e*, Pearson Prentice-Hall, 2018, online Appendix D.2.6, available at pages.stern.nyu.edu/~wgreene/Text/econometricanalysis.htm.

Going one step further, there is an important theorem that says replacing unknown quantities with consistent estimators does not change the asymptotic distribution of a statistic.[13] In this case, $\hat{\sigma}^2$ is a consistent estimator for σ^2 and $\left(s_x^2\right)^{-1}$ is a consistent estimator for C_x. Thus, we can write

$$t = \frac{\sqrt{N}\left(b_2 - \beta_2\right)}{\sqrt{\hat{\sigma}^2 / s_x^2}} = \frac{\left(b_2 - \beta_2\right)}{\sqrt{\widehat{\operatorname{var}}\left(b_2\right)}} = \frac{\left(b_2 - \beta_2\right)}{\operatorname{se}\left(b_2\right)} \overset{a}{\sim} N(0,\ 1) \tag{5.36}$$

This is precisely the t-statistic that we use for interval estimation and hypothesis testing. The result in (5.36) means that using it in large samples is justified when assumption MR6 is not satisfied. One difference is that we are now saying that the distribution of the statistic "t" is approximately "normal," not "t." However, the t-distribution approaches the normal as $N \to \infty$, and it is customary to use either the t or the normal distribution as the large sample approximation. Because use of (5.36) for interval estimation or hypothesis testing implies we are behaving as if b_2 is normally distributed with mean β_2 and variance $\widehat{\operatorname{var}}\left(b_2\right)$, this result is often written as

$$b_2 \overset{a}{\sim} N\left(\beta_2,\ \widehat{\operatorname{var}}\left(b_2\right)\right) \tag{5.37}$$

Finally, our exposition has been in terms of the distribution of b_2 in the simple regression model, but the result also holds for estimators of the coefficients in the multiple regression model. In Appendix 5C, we use Monte Carlo experiments to illustrate how the central limit theorem works and give examples of how large N needs to be for the normal approximation to be satisfactory.

5.7.3 Relaxing Assumptions

In the previous two sections we explained that, when assumptions MR1–MR5 hold, and MR6 is relaxed, the least squares estimator is consistent and asymptotically normal. In this section, we investigate what we can say about the properties of the least squares estimator when we modify the strict exogeneity assumption MR2: $E\left(e_i|\mathbf{X}\right) = 0$ to make it less restrictive.

Weakening Strict Exogeneity: Cross-Sectional Data It is convenient to consider modifications of $E\left(e_i|\mathbf{X}\right) = 0$ first for cross-sectional data and then for time-series data. For cross-sectional data, we return to the random sampling assumptions, explained in Section 2.2, and written more formally in Section 2.10. Generalizing these assumptions to the multiple regression model, random sampling implies the joint observations $\left(y_i, \mathbf{x}_i\right) = \left(y_i, x_{i1}, x_{i2}, \ldots, x_{iK}\right)$ are independent, and that the strict exogeneity assumption $E\left(e_i|\mathbf{X}\right) = 0$ reduces to $E\left(e_i|\mathbf{x}_i\right) = 0$. Under this and the remaining assumptions of the model under random sampling, the least squares estimator is best linear unbiased. We now examine the implications of replacing $E\left(e_i|\mathbf{x}_i\right) = 0$ with the weaker assumption

$$E\left(e_i\right) = 0 \quad \text{and} \quad \operatorname{cov}\left(e_i, x_{ik}\right) = 0 \quad \text{for } i = 1, 2, \ldots, N;\ k = 1, 2, \ldots, K \tag{5.38}$$

Why is (5.38) a weaker assumption? In Section 2.10, in the context of the simple regression model, we explained how $E\left(e_i|\mathbf{x}_i\right) = 0$ implies (5.38).[14] However, we cannot go back the other way. While $E\left(e_i|\mathbf{x}_i\right) = 0$ implies (5.38), (5.38) does not necessarily imply $E\left(e_i|\mathbf{x}_i\right) = 0$. Making the assumption $E\left(e_i|\mathbf{x}_i\right) = 0$ means that the best predictor for e_i is zero; there is no information in $\mathbf{x}_i$ that will help predict e_i. On the other hand, assuming $\operatorname{cov}\left(e_i, x_{ik}\right) = 0$ only implies there is no *linear* predictor for e_i that is better than zero. It does not rule out nonlinear functions of $\mathbf{x}_i$ that may help predict e_i.

Why is it useful to consider the weaker assumption in (5.38)? First, the weaker are the assumptions under which an estimator has desirable properties, the wider the applicability of

[13] For more precise details, see William Greene, *Econometric Analysis 8e*, Pearson Prentice-Hall, 2018, Theorem D.16, in online Appendix available at pages.stern.nyu.edu/~wgreene/Text/econometricanalysis.htm.

[14] A proof is given in Appendix 2G.

the estimator. Second, as we discover in Chapter 10, violation of the assumption $\text{cov}\left(e_i, x_{ik}\right) = 0$ provides a good framework for considering the problem of endogenous regressors.

The seemingly innocuous weaker assumption in (5.38) means we can no longer show that the least squares estimator is unbiased. Consider the least squares estimator for β_2 in the simple regression model $y_i = \beta_1 + \beta_2 x_i + e_i$. From (2.11) and (2.12),

$$b_2 = \beta_2 + \frac{\sum_{i=1}^{N}\left(x_i - \bar{x}\right)e_i}{\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2} \tag{5.39}$$

and

$$E\left(b_2\right) = \beta_2 + E\left(\frac{\sum_{i=1}^{N}\left(x_i - \bar{x}\right)e_i}{\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2}\right) \tag{5.40}$$

Now, $E\left(e_i\right) = 0$ and $\text{cov}\left(e_i, x_{ik}\right) = 0$ imply $E\left(x_i e_i\right) = 0$, but the last term in (5.39) is more complicated than that; it involves the covariance between e_i and a function of x_i. This covariance will not necessarily be zero, implying $E\left(b_2\right) \neq \beta_2$. We can show that b_2 is consistent, however. We can rewrite (5.39) as

$$b_2 = \beta_2 + \frac{\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)e_i}{\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2} = \beta_2 + \frac{\widehat{\text{cov}}\left(e_i, x_i\right)}{\widehat{\text{var}}\left(x_i\right)} \tag{5.41}$$

Because sample means, variances, and covariances computed from random samples are consistent estimators of their population counterparts,[15] we can say

$$\widehat{\text{cov}}\left(e_i, x_i\right) \xrightarrow{p} \text{cov}\left(e_i, x_i\right) = 0 \tag{5.42a}$$

$$\widehat{\text{var}}\left(x_i\right) \xrightarrow{p} \sigma_x^2 \tag{5.42b}$$

Thus, the second term in (5.41) converges in probability to zero, and $b_2 \xrightarrow{p} \beta_2$. It is also true that the asymptotic distribution of b_2 will be normal, as described in (5.36) and (5.37).

Weakening Strict Exogeneity: Time-Series Data

When we turn to time-series data, the observations $\left(y_t, \mathbf{x}_t\right)$, $t = 1, 2, \ldots, T$ are not collected via random sampling and so it is no longer reasonable to assume they are independent. The explanatory variables will almost certainly be correlated over time, and the likelihood of the assumption $E(e_t|\mathbf{X}) = 0$ being violated is very strong indeed. To see why, note that $E\left(e_t|\mathbf{X}\right) = 0$ implies

$$E\left(e_t\right) = 0 \quad \text{and} \quad \text{cov}\left(e_t, x_{sk}\right) = 0 \quad \text{for} \quad t, s = 1, 2, \ldots, T; \quad k = 1, 2, \ldots, K \tag{5.43}$$

This result says that the errors in every time period are uncorrelated with all the explanatory variables in every time period. In Section 2.10.2, three examples of how this assumption might be violated were described. Now would be a good time to check out those examples. To reinforce them, consider the simple regression model $y_t = \beta_1 + \beta_2 x_t + e_t$, which is being estimated with time-series observations in periods $t = 1, 2, \ldots, T$. If x_t is a policy variable whose settings depend on past outcomes $y_{t-1}, y_{t-2}, \ldots$, then x_t will be correlated with previous errors $e_{t-1}, e_{t-2}, \ldots$. This is evident from the equation for the previous period observation $y_{t-1} = \beta_1 + \beta_2 x_{t-1} + e_{t-1}$. If x_t is correlated with y_{t-1}, then it will also be correlated with e_{t-1} since y_{t-1} depends directly on e_{t-1}. Such a correlation is particularly evident if x_t is a lagged value of y_t. That is, $y_t = \beta_1 + \beta_2 y_{t-1} + e_t$. Models of this type are called autoregressive models; they are considered in Chapter 9.

The likely violation of $\text{cov}\left(e_t, x_{sk}\right) = 0$ for $s \neq t$ implies $E\left(e_t|\mathbf{X}\right) = 0$ will be violated, which in turn implies we cannot show that the least squares estimator is unbiased. It is possible to show

[15] This result follows from a law of large numbers. See Theorem D.4 and its corollary in the online Appendix to William Greene, *Econometric Analysis 8e*, Pearson Prentice-Hall, online at pages.stern.nyu.edu/~wgreene/Text/econometricanalysis.htm

it is consistent, however. To show consistency, we first assume that the errors and the explanatory variables in the same time period are uncorrelated. That is, we modify (5.43) to the less restrictive and more realistic assumption

$$E(e_t) = 0 \quad \text{and} \quad \text{cov}(e_t, x_{tk}) = 0 \quad \text{for} \quad t = 1, 2, \ldots, T;\ k = 1, 2, \ldots, K \tag{5.44}$$

Errors and the explanatory variables that satisfy (5.44) are said to be **contemporaneously uncorrelated**. We do not insist that $\text{cov}(e_t, x_{sk}) = 0$ for $t \neq s$. Now reconsider (5.41) written in terms of time-series observations

$$b_2 = \beta_2 + \frac{\frac{1}{T}\sum_{t=1}^{T}(x_t - \bar{x})e_t}{\frac{1}{T}\sum_{t=1}^{T}(x_t - \bar{x})^2} = \beta_2 + \frac{\widehat{\text{cov}}(e_t, x_t)}{\widehat{\text{var}}(x_t)} \tag{5.45}$$

Equation (5.45) is still valid, just as it was for cross-sectional observations. The question we need to ask to ensure consistency of b_2 is when the explanatory variables are not independent will it still be true that

$$\widehat{\text{cov}}(e_t, x_t) \xrightarrow{p} \text{cov}(e_t, x_t) = 0 \tag{5.46a}$$

$$\widehat{\text{var}}(x_t) \xrightarrow{p} \sigma_x^2 \tag{5.46b}$$

with σ_x^2 finite? The answer is "yes" as long as x is not "too dependent." If the correlation between the x_t's declines as they become further apart in time, then the results in (5.46) will hold. We reserve further discussion of the implications of the behavior of the explanatory variables in time-series regressions for Chapters 9 and 12. For the moment, we assume that their behavior is sufficiently cooperative for (5.46) to hold, so that the least squares estimator is consistent. At the same time, we recognize that, with time-series data, the least squares estimator is unlikely to be unbiased. **Asymptotic normality** can be shown by a central limit theorem, implying we can use (5.36) and (5.37) for interval estimation and hypothesis testing.

5.7.4 Inference for a Nonlinear Function of Coefficients

The need for large sample or asymptotic distributions is not confined to situations where assumptions MR1–MR6 are relaxed. Even if these assumptions hold, we still need to use large sample theory if a quantity of interest involves a nonlinear function of coefficients. To introduce this problem, we return to Big Andy's Burger Barn and examine the optimal level of advertising.

EXAMPLE 5.17 | The Optimal Level of Advertising

Economic theory tells us to undertake all those actions for which the marginal benefit is greater than the marginal cost. This optimizing principle applies to Big Andy's Burger Barn as it attempts to choose the optimal level of advertising expenditure. Recalling that *SALES* denotes sales revenue or total revenue, the marginal benefit in this case is the marginal revenue from more advertising. From (5.24), the required marginal revenue is given by the marginal effect of more advertising $\beta_3 + 2\beta_4 ADVERT$. The marginal cost of \$1 of advertising is \$1 plus the cost of preparing the additional products sold due to effective advertising. If we ignore the latter costs, the marginal cost of \$1 of advertising expenditure is \$1. Thus, advertising should be increased to the point where

$$\beta_3 + 2\beta_4 ADVERT_0 = 1$$

with $ADVERT_0$ denoting the optimal level of advertising. Using the least squares estimates for β_3 and β_4 in (5.25), a point estimate for $ADVERT_0$ is

$$\widehat{ADVERT_0} = \frac{1 - b_3}{2b_4} = \frac{1 - 12.1512}{2 \times (-2.76796)} = 2.014$$

implying that the optimal monthly advertising expenditure is $2014.

To assess the reliability of this estimate, we need a standard error and an interval estimate for $(1-b_3)/2b_4$. This is a tricky problem, and one that requires the use of calculus to solve. What makes it more difficult than what we have done so far is the fact that it involves a **nonlinear function** of b_3 and b_4. Variances of nonlinear functions are hard to derive. Recall that the variance of a linear function, say, $c_3b_3 + c_4b_4$, is given by

$$\text{var}(c_3b_3 + c_4b_4) = c_3^2\text{var}(b_3) + c_4^2\text{var}(b_4) + 2c_3c_4\text{cov}(b_3, b_4) \tag{5.47}$$

Finding the variance of $(1-b_3)/2b_4$ is less straightforward. The best we can do is find an approximate expression that is valid in large samples. Suppose $\lambda = (1-\beta_3)/2\beta_4$ and $\hat{\lambda} = (1-b_3)/2b_4$; then, the approximate variance expression is

$$\text{var}(\hat{\lambda}) = \left(\frac{\partial\lambda}{\partial\beta_3}\right)^2 \text{var}(b_3) + \left(\frac{\partial\lambda}{\partial\beta_4}\right)^2 \text{var}(b_4) + 2\left(\frac{\partial\lambda}{\partial\beta_3}\right)\left(\frac{\partial\lambda}{\partial\beta_4}\right)\text{cov}(b_3, b_4) \tag{5.48}$$

This expression holds for all nonlinear functions of two estimators, not just $\hat{\lambda} = (1-b_3)/2b_4$. Also, note that for the linear case, where $\lambda = c_3\beta_3 + c_4\beta_4$ and $\hat{\lambda} = c_3b_3 + c_4b_4$, (5.48) reduces to (5.47). Using (5.48) to find an approximate expression for a variance is called the **delta method**. For further details, consult Appendix 5B.

We will use (5.48) to estimate the variance of $\hat{\lambda} = \widehat{ADVERT_0} = (1-b_3)/2b_4$, get its standard error, and use that to get an interval estimate for $\lambda = ADVERT_0 = (1-\beta_3)/2\beta_4$. If the use of calculus in (5.48) frightens you, take comfort in the fact that most software will automatically compute the standard error for you.

The required derivatives are

$$\frac{\partial\lambda}{\partial\beta_3} = -\frac{1}{2\beta_4}, \quad \frac{\partial\lambda}{\partial\beta_4} = -\frac{1-\beta_3}{2\beta_4^2}$$

To estimate $\text{var}(\hat{\lambda})$, we evaluate these derivatives at the least squares estimates b_3 and b_4.

Thus, for the estimated variance of the optimal level of advertising, we have

$$\begin{aligned}\widehat{\text{var}}(\hat{\lambda}) &= \left(-\frac{1}{2b_4}\right)^2 \widehat{\text{var}}(b_3) + \left(-\frac{1-b_3}{2b_4^2}\right)^2 \widehat{\text{var}}(b_4) \\ &\quad + 2\left(-\frac{1}{2b_4}\right)\left(-\frac{1-b_3}{2b_4^2}\right)\widehat{\text{cov}}(b_3, b_4) \\ &= \left(\frac{1}{2 \times 2.768}\right)^2 \times 12.646 \\ &\quad + \left(\frac{1-12.151}{2 \times 2.768^2}\right)^2 \times 0.88477 \\ &\quad + 2\left(\frac{1}{2 \times 2.768}\right)\left(\frac{1-12.151}{2 \times 2.768^2}\right) \times 3.2887 \\ &= 0.016567\end{aligned}$$

and

$$\text{se}(\hat{\lambda}) = \sqrt{0.016567} = 0.1287$$

We are now in a position to get a 95% interval estimate for $\lambda = ADVERT_0$. When dealing with a linear combination of coefficients in (5.16), and Section 5.4.2, we used the result $(\hat{\lambda} - \lambda)/\text{se}(\hat{\lambda}) \sim t_{(N-K)}$. In line with Section 5.7.2, this result can be used in exactly the same way for nonlinear functions, but a difference is that the result is only an approximate one for large samples, even when the errors are normally distributed. Thus, an approximate 95% interval estimate for $ADVERT_0$ is

$$\begin{aligned}&\left[\hat{\lambda} - t_{(0.975,71)}\text{se}(\hat{\lambda}),\ \hat{\lambda} + t_{(0.975,71)}\text{se}(\hat{\lambda})\right] \\ &= [2.014 - 1.994 \times 0.1287,\ 2.014 + 1.994 \times 0.1287] \\ &= [1.757,\ 2.271]\end{aligned}$$

We estimate with 95% confidence that the optimal level of advertising lies between $1757 and $2271.

EXAMPLE 5.18 | How Much Experience Maximizes Wages?

In Example 5.16, we estimated the wage equation

$$\begin{aligned}\ln(WAGE) &= \beta_1 + \beta_2 EDUC + \beta_3 EXPER \\ &\quad + \beta_4(EDUC \times EXPER) + \beta_5 EXPER^2 + e\end{aligned}$$

One of the implications of the quadratic function of experience is that, as a number of years of experience increases, wages will increase up to a point and then decline. Suppose we are interested in the number of years of experience, which maximizes $WAGE$. We can get this quantity by differentiating the wage equation with respect to $EXPER$, setting the first derivative equal to zero and solving for $EXPER$. It does not matter that the dependent variable is $\ln(WAGE)$ not $WAGE$; the value of $EXPER$ that maximizes $\ln(WAGE)$ will also maximize $WAGE$. Setting the first derivative in (5.28) equal to zero and solving for $EXPER$ yields

$$EXPER_0 = \frac{-\beta_3 - \beta_4 EDUC}{2\beta_5}$$

The maximizing value depends on the number of years of education. For someone with 16 years of education, it is

$$EXPER_0 = \frac{-\beta_3 - 16\beta_4}{2\beta_5}$$

Finding the standard error for an estimate of this function is tedious. It involves differentiating with respect to β_3, β_4, and β_5 and evaluating a variance expression involving three variances and three covariances—an extension of (5.48) to three coefficients. This is a problem better handled by your favorite econometric software. Taking this advice, we find $\widehat{EXPER}_0 = 30.17$ and $\text{se}\left(\widehat{EXPER}_0\right) = 1.7896$. Then, a 95% interval estimate of the number of years of experience that maximizes *WAGE* is

$$\left[\widehat{EXPER}_0 - t_{(0.975,1195)}\text{se}\left(\widehat{EXPER}_0\right),\right.$$
$$\left.\widehat{EXPER}_0 + t_{(0.975,1195)}\text{se}\left(\widehat{EXPER}_0\right)\right]$$

Inserting the relevant values yields

$$(30.17 - 1.962 \times 1.7896,\ 30.17 + 1.962 \times 1.7896)$$
$$= (26.7,\ 33.7)$$

We estimate that the number of years of experience that maximizes wages lies between 26.7 and 33.7 years.

5.8 Exercises

5.8.1 Problems

5.1 Consider the multiple regression model

$$y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + x_{i3}\beta_3 + e_i$$

with the seven observations on y_i, x_{i1}, x_{i2}, and x_{i3} given in Table 5.5.

TABLE 5.5 **Data for Exercise 5.1**

y_i	x_{i1}	x_{i2}	x_{i3}
1	1	0	1
1	1	1	−2
4	1	2	2
0	1	−2	1
1	1	1	−2
−2	1	−2	−1
2	1	0	1

Use a hand calculator or spreadsheet to answer the following questions:

a. Calculate the observations in terms of deviations from their means. That is, find $x_{i2}^* = x_{i2} - \bar{x}_2$, $x_{i3}^* = x_{i3} - \bar{x}_3$, and $y_i^* = y_i - \bar{y}$.

b. Calculate $\sum y_i^* x_{i2}^*$, $\sum x_{i2}^{*2}$, $\sum y_i^* x_{i3}^*$, $\sum x_{i2}^* x_{i3}^*$, and $\sum x_{i3}^{*2}$.

c. Use the expressions in Appendix 5A to find least squares estimates b_1, b_2, and b_3.

d. Find the least squares residuals $\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_7$.

e. Find the variance estimate $\hat{\sigma}^2$.

f. Find the sample correlation between x_2 and x_3.

g. Find the standard error for b_2.

h. Find *SSE*, *SST*, *SSR*, and R^2.

5.2 Use your answers to Exercise 5.1 to

a. Compute a 95% interval estimate for β_2.

b. Test the hypothesis $H_0\!:\beta_2 = 1.25$ against the alternative that $H_1\!:\beta_2 \neq 1.25$.

5.3 Consider the following model that relates the percentage of a household's budget spent on alcohol *WALC* to total expenditure *TOTEXP*, age of the household head *AGE*, and the number of children in the household *NK*.

$$WALC = \beta_1 + \beta_2 \ln(TOTEXP) + \beta_3 NK + \beta_4 AGE + e$$

This model was estimated using 1200 observations from London. An incomplete version of this output is provided in Table 5.6.

TABLE 5.6 **Output for Exercise 5.3**

Dependent Variable: *WALC*
Included observations: 1200

Variable	Coefficient	Std. Error	*t*-Statistic	Prob.
C	1.4515	2.2019		0.5099
ln(*TOTEXP*)	2.7648		5.7103	0.0000
NK		0.3695	−3.9376	0.0001
AGE	−0.1503	0.0235	−6.4019	0.0000
R-squared		Mean dependent var		6.19434
S.E. of regression		S.D. dependent var		6.39547
Sum squared resid	46221.62			

a. Fill in the following blank spaces that appear in this table.
   i. The *t*-statistic for b_1.
   ii. The standard error for b_2.
   iii. The estimate b_3.
   iv. R^2.
   v. $\hat{\sigma}$.

b. Interpret each of the estimates b_2, b_3, and b_4.

c. Compute a 95% interval estimate for β_4.What does this interval tell you?

d. Are each of the coefficient estimates significant at a 5% level? Why?

e. Test the hypothesis that the addition of an extra child decreases the mean budget share of alcohol by 2 percentage points against the alternative that the decrease is not equal to 2 percentage points. Use a 5% significance level.

5.4 Consider the following model that relates the percentage of a household's budget spent on alcohol, *WALC*, to total expenditure *TOTEXP*, age of the household head *AGE*, and the number of children in the household *NK*.

$$WALC = \beta_1 + \beta_2 \ln(TOTEXP) + \beta_3 NK + \beta_4 AGE + \beta_5 AGE^2 + e$$

Some output from estimating this model using 1200 observations from London is provided in Table 5.7. The covariance matrix relates to the coefficients b_3, b_4, and b_5.

a. Find a point estimate and a 95% interval estimate for the change in the mean budget percentage share for alcohol when a household has an extra child.

b. Find a point estimate and a 95% interval estimate for the marginal effect of *AGE* on the mean budget percentage share for alcohol when (i) $AGE = 25$, (ii) $AGE = 50$, and (iii) $AGE = 75$.

c. Find a point estimate and a 95% interval estimate for the age at which the mean budget percentage share for alcohol is at a minimum.

d. Summarize what you have discovered from the point and interval estimates in (a), (b), and (c).

e. Let **X** represent all the observations on all the explanatory variables. If $(e|\mathbf{X})$ is normally distributed, which of the above interval estimates are valid in finite samples? Which ones rely on a large sample approximation?

f. If $(e|\mathbf{X})$ is not normally distributed, which of the above interval estimates are valid in finite samples? Which ones rely on a large sample approximation?

TABLE 5.7 **Output for Exercise 5.4**

Variable	Coefficient
C	8.149
$\ln(TOTEXP)$	2.884
NK	−1.217
AGE	−0.5699
AGE^2	0.005515

	Covariance matrix		
	NK	AGE	AGE^2
NK	0.1462	−0.01774	0.0002347
AGE	−0.01774	0.03204	−0.0004138
AGE^2	0.0002347	−0.0004138	0.000005438

5.5 For each of the following two time-series regression models, and assuming MR1–MR6 hold, find $\text{var}\left(b_2|\mathbf{x}\right)$ and examine whether the least squares estimator is consistent by checking whether $\lim_{T\to\infty}\text{var}\left(b_2|\mathbf{x}\right) = 0$.

a. $y_t = \beta_1 + \beta_2 t + e_t, t = 1, 2, \ldots, T$. Note that $\mathbf{x} = (1, 2, \ldots, T)$, $\sum_{t=1}^{T}\left(t - \bar{t}\right)^2 = \sum_{t=1}^{T} t^2 - \left(\sum_{t=1}^{T} t\right)^2 \Big/ T$, $\sum_{t=1}^{T} t = T(T+1)/2$ and $\sum_{t=1}^{T} t^2 = T(T+1)(2T+1)/6$.

b. $y_t = \beta_1 + \beta_2 (0.5)^t + e_t, t = 1, 2, \ldots, T$. Here, $\mathbf{x} = \left(0.5, 0.5^2, \ldots, 0.5^T\right)$. Note that the sum of a geometric progression with first term r and common ratio r is

$$S = r + r^2 + r^3 + \cdots + r^n = \frac{r(1-r^n)}{1-r}$$

c. Provide an intuitive explanation for these results.

5.6 Suppose that, from a sample of 63 observations, the least squares estimates and the corresponding estimated covariance matrix are given by

$$\begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \\ -1 \end{bmatrix} \qquad \widehat{\text{cov}}\left(b_1, b_2, b_3\right) = \begin{bmatrix} 3 & -2 & 1 \\ -2 & 4 & 0 \\ 1 & 0 & 3 \end{bmatrix}$$

Using a 5% significance level, and an alternative hypothesis that the equality does not hold, test each of the following null hypotheses:

a. $\beta_2 = 0$
b. $\beta_1 + 2\beta_2 = 5$
c. $\beta_1 - \beta_2 + \beta_3 = 4$

5.7 After estimating the model $y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + e$ with $N = 203$ observations, we obtain the following information: $\sum_{i=1}^{N}\left(x_{i2} - \bar{x}_2\right)^2 = 1780.7$, $\sum_{i=1}^{N}\left(x_{i3} - \bar{x}_3\right)^2 = 3453.3$, $b_2 = 0.7176$, $b_3 = 1.0516$, $SSE = 6800.0$, and $r_{23} = 0.7087$.

a. What are the standard errors of the least squares estimates b_2 and b_3?
b. Using a 5% significance level, test the hypothesis $H_0{:}\beta_2 = 0$ against the alternative $H_1{:}\beta_2 \neq 0$.
c. Using a 10% significance level, test the hypothesis $H_0{:}\beta_3 \leq 0.9$ against the alternative $H_1{:}\beta_3 > 0.9$.
d. Given that $\widehat{\text{cov}}\left(b_2, b_3\right) = -0.019521$, use a 1% significance level to test the hypothesis $H_0{:}\beta_2 = \beta_3$ against the alternative $H_1{:}\beta_2 \neq \beta_3$.

5.8 There were 79 countries who competed in the 1996 Olympics and won at least one medal. For each of these countries, let *MEDALS* be the total number of medals won, *POPM* be population in millions, and *GDPB* be GDP in billions of 1995 dollars. Using these data we estimate the regression model $MEDALS = \beta_1 + \beta_2 POPM + \beta_3 GDPB + e$ to obtain

$$\begin{array}{llll}\widehat{MEDALS} = 5.917 + 0.01813\,POPM + 0.01026\,GDPB & & & R^2 = 0.4879 \\ \text{(se)} \quad (1.510) \quad (0.00819) \qquad\qquad (0.00136) & & & \end{array}$$

a. Given assumptions MR1–MR6 hold, interpret the coefficient estimates for β_2 and β_3.

b. Interpret R^2.

c. Using a 1% significance level, test the hypothesis that there is no relationship between the number of medals won and GDP against the alternative that there is a positive relationship. What happens if you change the significance level to 5%?

d. Using a 1% significance level, test the hypothesis that there is no relationship between the number of medals won and population against the alternative that there is a positive relationship. What happens if you change the significance level to 5%?

e. Test the following hypotheses using a 5% significance level:

i. $H_0: \beta_2 = 0.01$ against the alternative $H_1: \beta_2 \neq 0.01$

ii. $H_0: \beta_2 = 0.02$ against the alternative $H_1: \beta_2 \neq 0.02$

iii. $H_0: \beta_2 = 0.03$ against the alternative $H_1: \beta_2 \neq 0.03$

iv. $H_0: \beta_2 = 0.04$ against the alternative $H_1: \beta_2 \neq 0.04$

Are these test results contradictory? Why or why not?

f. Find a 95% interval estimate for β_2 and comment on it.

5.9 There were 64 countries who competed in the 1992 Olympics and won at least one medal. For each of these countries, let *MEDALS* be the total number of medals won, *POPM* be population in millions, and *GDPB* be GDP in billions of 1995 dollars. Excluding the United Kingdom, and using $N = 63$ observations, the model $MEDALS = \beta_1 + \beta_2 \ln(POPM) + \beta_3 \ln(GDPB) + e$ was estimated as

$$\begin{array}{ll}\widehat{MEDALS} = -13.153 + 2.764\ln(POPM) + 4.270\ln(GDPB) & \quad R^2 = 0.275 \\ \text{(se)} \qquad (5.974) \quad (2.070) \qquad\qquad (1.718) & \end{array}$$

a. Given assumptions MR1–MR6 hold, interpret the coefficient estimates for β_2 and β_3.

b. Interpret R^2.

c. Using a 10% significance level, test the hypothesis that there is no relationship between the number of medals won and GDP against the alternative that there is a positive relationship. What happens if you change the significance level to 5%?

d. Using a 10% significance level, test the hypothesis that there is no relationship between the number of medals won and population against the alternative that there is a positive relationship. What happens if you change the significance level to 5%?

e. Use the model to find point and 95% interval estimates for the expected number of medals won by the United Kingdom whose population and GDP in 1992 were 58 million and $1010 billion, respectively. [The standard error for $b_1 + \ln(58) \times b_2 + \ln(1010) \times b_3$ is 4.22196.]

f. The United Kingdom won 20 medals in 1992. Is the model a good one for predicting the mean number of medals for the United Kingdom? What is an approximate *p*-value for a test of $H_0: \beta_1 + \ln(58) \times \beta_2 + \ln(1010) \times \beta_3 = 20$ versus $H_1: \beta_1 + \ln(58) \times \beta_2 + \ln(1010) \times \beta_3 \neq 20$?

g. Without doing any of the calculations, write down the expression that is used to compute the standard error given in part (e).

5.10 Using data from 1950 to 1996 ($T = 47$ observations), the following equation for explaining wheat yield in the Mullewa Shire of Western Australia was estimated as

$$\begin{array}{l}\widehat{YIELD}_t = 0.1717 + 0.01117t + 0.05238\,RAIN_t \\ \text{(se)} \qquad (0.1537) \ (0.00262) \quad (0.01367)\end{array}$$

where $YIELD_t$ = wheat yield in tonnes per hectare in year t;

$TREND_t$ is a trend variable designed to capture technological change, with observations $t = 1, 2, \ldots, 47$;

$RAIN_t$ is total rainfall in inches from May to October (the growing season) in year t. The sample mean and standard deviation for *RAIN* are $\bar{x}_{RAIN} = 10.059$ and $s_{RAIN} = 2.624$.

a. Given assumptions MR1–MR5 hold, interpret the estimates for the coefficients of t and $RAIN$.

b. Using a 5% significance level, test the null hypothesis that technological change increases mean yield by no more than 0.01 tonnes per hectare per year against the alternative that the mean yield increase is greater than 0.01.

c. Using a 5% significance level, test the null hypothesis that an extra inch of rainfall increases mean yield by 0.03 tonnes per hectare against the alternative that the increase is not equal to 0.03.

d. Adding $RAIN^2$ to the equation and reestimating yields

$$\begin{array}{lllll} \widehat{YIELD}_t = & -0.6759 & +\ 0.011671t & +\ 0.2229RAIN_t & -\ 0.008155\,RAIN_t^2 \\ \text{(se)} & (0.3875) & (0.00250) & (0.0734) & (0.003453) \end{array}$$

What is the rationale for including $RAIN^2$? Does it have the expected sign?

e. Repeat part (b) using the model estimated in (d).

f. Repeat part (c) using the model estimated in (d), testing the hypothesis at the mean value of rainfall. (The estimated covariance between b_3 and b_4 (the coefficients of $RAIN$ and $RAIN^2$) is $\widehat{\text{cov}}(b_3, b_4) = -0.0002493$.)

g. Use the model in (d) to forecast yield in 1997, when the rainfall was 9.48 inches.

h. Suppose that you wanted to forecast 1997 yield before the rainfall was observed. What would be your forecast from the model in (a)? What would it be from the model in (d)?

5.11 When estimating wage equations, we expect that young, inexperienced workers will have relatively low wages; with additional experience their wages will rise, but then begin to decline after middle age, as the worker nears retirement. This life-cycle pattern of wages can be captured by introducing experience and experience squared to explain the level of wages. If we also include years of education, we have the equation

$$WAGE = \beta_1 + \beta_2 EDUC + \beta_3 EXPER + \beta_4 EXPER^2 + e$$

a. What is the marginal effect of experience on the mean wage?

b. What signs do you expect for each of the coefficients β_2, β_3, and β_4? Why?

c. After how many years of experience does the mean wage start to decline? (Express your answer in terms of β's.)

d. Estimating this equation using 600 observations yields

$$\begin{array}{lllll} \widehat{WAGE} = & -16.308 + & 2.329EDUC + & 0.5240EXPER - & 0.007582EXPER^2 \\ \text{(se)} & (2.745) & (0.163) & (0.1263) & (0.002532) \end{array}$$

The estimated covariance between b_3 and b_4 is $\widehat{\text{cov}}(b_3, b_4) = -0.00030526$. Find 95% interval estimates for the following:

i. The marginal effect of education on mean wage

ii. The marginal effect of experience on mean wage when $EXPER = 4$

iii. The marginal effect of experience on mean wage when $EXPER = 25$

iv. The number of years of experience after which the mean wage declines

5.12 This exercise uses data on 850 houses sold in Baton Rouge, Louisiana during mid-2005. We will be concerned with the selling price in thousands of dollars ($PRICE$), the size of the house in hundreds of square feet ($SQFT$), and the age of the house in years (AGE). The following two regression models were estimated:

$$PRICE = \alpha_1 + \alpha_2 AGE + v \quad \text{and} \quad SQFT = \delta_1 + \delta_2 AGE + u$$

The sums of squares and sums of cross products of the residuals from estimating these two equations are $\sum_{i=1}^{850} \hat{v}_i^2 = 10377817$, $\sum_{i=1}^{850} \hat{u}_i^2 = 75773.4$, $\sum_{i=1}^{850} \hat{u}_i \hat{v}_i = 688318$.

a. Find the least-squares estimate of β_2 in the model $PRICE = \beta_1 + \beta_2 SQFT + \beta_3 AGE + e$.

b. Let $\hat{e}_i = \hat{v}_i - b_2\hat{u}_i$. Show that $\sum_{i=1}^{850} \hat{e}_i^2 = \sum_{i=1}^{850} \hat{v}_i^2 - b_2 \sum_{i=1}^{850} \hat{v}_i \hat{u}_i$ where b_2 is the least-squares estimate for β_2.

c. Find an estimate of $\sigma^2 = \text{var}(e_i)$.

d. Find the standard error for b_2.

e. What is an approximate p-value for testing $H_0 : \beta_2 \geq 9.5$ against the alternative $H_1 : \beta_2 < 9.5$? What do you conclude from this p-value?

5.13 A concept used in macroeconomics is Okun's Law, which states that the change in unemployment from one period to the next depends on the rate of growth of the economy relative to a "normal" growth rate:

$$U_t - U_{t-1} = -\gamma\left(G_t - G_N\right)$$

where U_t is the unemployment rate in period t, G_t is the growth rate in period t, the "normal" growth rate G_N is that which is required to maintain a constant rate of unemployment, and $0 < \gamma < 1$ is an adjustment coefficient.

a. Show that the model can be written as $DU_t = \beta_1 + \beta_2 G_t$, where $DU_t = U_t - U_{t-1}$ is the change in the unemployment rate, $\beta_1 = \gamma G_N$, and $\beta_2 = -\gamma$.

b. Estimating this model with quarterly seasonally adjusted U.S. data from 1970 Q1 to 2014 Q4 yields

$$\widehat{DU_t} = 0.1989 - 0.2713G_t \qquad \hat{\sigma} = 0.2749$$

$$\operatorname{cov}\left(b_1, b_2\right) = \begin{pmatrix} 0.0007212 & -0.0004277 \\ -0.0004277 & 0.0006113 \end{pmatrix}$$

Use the estimates b_1 and b_2 to find estimates $\hat{\gamma}$ and $\hat{G}_N$.

c. Find standard errors for b_1, b_2, $\hat{\gamma}$, and $\hat{G}_N$. Are all these estimates significantly different from zero at a 5% level?

d. Using a 5% significance level test the null hypothesis that the natural growth rate is 0.8% per quarter against the alternative it is not equal to 0.8%.

e. Find a 95% interval estimate for the adjustment coefficient.

f. Find a 95% interval estimate for $E\left(U_{2015Q1} | U_{2014Q4} = 5.7991, G_{2015Q1} = 0.062\right)$.

g. Find a 95% prediction interval for U_{2015Q1} given $U_{2014Q4} = 5.7991$ and $G_{2015Q1} = 0.062$. Explain the difference between this interval and that in (f).

5.14 Consider the regression model $y_i = \beta_1 + \beta_2 x_i + e_i$ where the pairs $\left(y_i, x_i\right)$, $i = 1, 2, \ldots, N$, are random independent draws from a population.

a. Suppose the marginal distribution of x_i is log-normal. To appreciate the nature of the log-normal distribution, consider a normal random variable $W \sim N\left(\mu_W, \sigma_W^2\right)$. Then, $X = e^W$ has a log-normal distribution with mean $\mu_X = \exp\left(\mu_W + \sigma_W^2/2\right)$ and variance $\sigma_X^2 = \left(\exp\left(\sigma_W^2\right) - 1\right)\mu_X^2$. Assume that $\left(e_i | x_i\right) \sim N\left(0, \sigma_e^2\right)$.

i. Will the least squares estimator $\left(b_1, b_2\right)$ for the parameters $\left(\beta_1, \beta_2\right)$ be unbiased?

ii. Will it be consistent?

iii. Will it be normally distributed conditional on $\mathbf{x} = \left(x_1, x_2, \ldots, x_N\right)$?

iv. Will the marginal distribution of $\left(b_1, b_2\right)$ (not conditional on $\mathbf{x}$) be normally distributed?

v. Will t-tests for β_1 and β_2 be justified in finite samples or are they only large sample approximations?

vi. Suppose $\mu_w = 0$, $\sigma_w^2 = 1$, and $x_i = \exp\left(w_i\right)$. What is the asymptotic variance of the least squares estimator for β_2? (Express in terms of σ_e^2 and N.)

b. Suppose now that $x_i \sim N(0, 1)$ and that $\left(e_i | x_i\right)$ has a log-normal distribution with mean and variance $\mu_e = \exp\left(\mu_v + \sigma_v^2/2\right)$ and $\sigma_e^2 = \left(\exp\left(\sigma_v^2\right) - 1\right)\mu_e^2$, where $v = \ln(e) \sim N\left(\mu_v, \sigma_v^2\right)$.

i. Show that we can rewrite the model as $y_i = \beta_1^* + \beta_2 x_i + e_i^*$ where

$$\beta_1^* = \beta_1 + \exp\left(\mu_v + \sigma_v^2/2\right) \text{ and } e_i^* = e_i - \exp\left(\mu_v + \sigma_v^2/2\right)$$

ii. Show that $E\left(e_i^* | x_i\right) = 0$ and $\operatorname{var}\left(e_i^* | x_i\right) = \sigma_e^2$.

iii. Will the least squares estimator b_2 for the parameter β_2 be unbiased?

iv. Will it be consistent?

v. Will it be normally distributed conditional on $\mathbf{x} = \left(x_1, x_2, \ldots, x_N\right)$?

vi. Will the marginal distribution of b_2 (not conditional on $\mathbf{x}$) be normally distributed?

vii. Will t-tests for β_2 be justified in finite samples or are they only large sample approximations?

viii. What is the asymptotic variance of the least squares estimator for β_2? (Express in terms of σ_e^2 and N.)

5.15 Consider the regression model $y_i = \beta_1 + \beta_2 x_i + e_i$ where the pairs $\left(y_i, x_i\right)$, $i = 1, 2, \ldots, N$, are random independent draws from a population, $x_i \sim N(0, 1)$, and $E\left(e_i | x_i\right) = c\left(x_i^2 - 1\right)$ where c is a constant.

a. Show that $E\left(e_i\right) = 0$.

b. Using the result $\operatorname{cov}\left(e_i, x_i\right) = E_x\left[\left(x_i - \mu_x\right) E\left(e_i | x_i\right)\right]$, show that $\operatorname{cov}\left(e_i, x_i\right) = 0$.

c. Will the least squares estimator for β_2 be (i) unbiased? (ii) consistent?

5.16 Consider a log-linear regression for the weekly sales of a national brand of canned tuna (brand *A*), expressed as thousands of cans, *CANS*, as a function of the prices of two competing brands (brands *B* and *C*), expressed as percentages of the price of brand *A*. That is,

$$\ln(CANS) = \beta_1 + \beta_2 RPRCE\_B + \beta_3 RPRCE\_C + e$$

where $RPRCE\_B = \left(PRICE_B/PRICE_A\right) \times 100$ and $RPRCE\_C = \left(PRICE_C/PRICE_A\right) \times 100$.

a. Given assumptions MR1–MR5 hold, how do you interpret β_2 and β_3? What signs do you expect for these coefficients? Why?

Using $N = 52$ weekly observations, the least squares estimated equation is

$$\begin{array}{llll} \widehat{\ln(CANS)} = -2.724 + 0.0146 RPRCE\_B + 0.02649\, RPRCE\_C & & \hat{\sigma} = 0.5663 \\ \text{(se)} \qquad (0.582) \quad (0.00548) \qquad\qquad (0.00544) & & \widehat{\text{cov}}\left(b_2, b_3\right) = -0.0000143 \end{array}$$

b. Using a 10% significance level, test the null hypothesis that an increase in *RPRCE_B* of one percentage point leads to a 2.5% increase in the mean number of cans sold against the alternative that the increase is not 2.5%.

c. Using a 10% significance level, test the null hypothesis that an increase in *RPRCE_C* of one percentage point leads to a 2.5% increase in the mean number of cans sold against the alternative that the increase is not 2.5%.

d. Using a 10% significance level, test $H_0 : \beta_2 = \beta_3$ against the alternative $H_1 : \beta_2 \neq \beta_3$. Does the outcome of this test contradict your findings from parts (b) and (c)?

e. Which brand do you think is the closer substitute for brand *A*, brand *B*, or brand *C*? Why?

f. Use the corrected predictor introduced in Section 4.5.3 to estimate the expected number of brand *A* cans sold under the following scenarios:

i. $RPRCE\_B = 125$, $RPRCE\_C = 100$

ii. $RPRCE\_B = 111.11$, $RPRCE\_C = 88.89$

iii. $RPRCE\_B = 100$, $RPRCE\_C = 80$

g. The producers of brands *B* and *C* have set the prices of their cans of tuna to be $1 and 80 cents, respectively. The producer of brand *A* is considering three possible prices for her cans: 80 cents, 90 cents, or $1. Use the results from part (f) to find which of these three price settings will maximize revenue from sales.

5.8.2 Computer Exercises

5.17 Use econometric software to verify your answers to Exercise 5.1, parts (c), (e), (f), (g), and (h).

5.18 Consider the following two expenditure share equations where the budget share for food *WFOOD*, and the budget share for clothing *WCLOTH*, are expressed as functions of total expenditure *TOTEXP*.

$$WFOOD = \beta_1 + \beta_2 \ln(TOTEXP) + e_F \qquad \text{(XR5.18.1)}$$

$$WCLOTH = \alpha_1 + \alpha_2 \ln(TOTEXP) + e_C \qquad \text{(XR5.18.2)}$$

a. A commodity is regarded as a luxury if the coefficient of ln(*TOTEXP*) is positive and a necessity if it is negative. What signs would you expect for β_2 and α_2?

b. Using the data in the file *london5*, estimate the above equations using observations on households with one child. Comment on the estimates and their significance. Can you explain any possibly counterintuitive outcomes?

c. Using a 1% significance level, test $H_0 : \beta_2 \geq 0$ against the alternative $H_1 : \beta_2 < 0$. Why might you set up the hypotheses in this way?

d. Using a 1% significance level, test $H_0 : \alpha_2 \geq 0$ against the alternative $H_1 : \alpha_2 < 0$. Why might you set up the hypotheses in this way?

e. Estimate the two equations using observations on households with two children. Construct 95% interval estimates for β_2 and α_2 for both one- and two-child households. Based on these interval estimates, would you conjecture that the coefficients of ln(*TOTEXP*) are the same or different for one- and two-child households?

f. Use all observations to estimate the following two equations and test, at a 95% significance level, whether your conjectures in part (e) are correct. (NK = number of children in the household.)

$$WFOOD = \gamma_1 + \gamma_2 \ln(TOTEXP) + \gamma_3 NK + \gamma_4 NK \times \ln(TOTEXP) + e_F \quad \text{(XR5.18.3)}$$

$$WCLOTH = \delta_1 + \delta_2 \ln(TOTEXP) + \delta_3 NK + \delta_4 NK \times \ln(TOTEXP) + e_C \quad \text{(XR5.18.4)}$$

g. Compare the estimates for $\partial E(WFOOD|\mathbf{X})/\partial \ln(TOTEXP)$ from (XR5.18.1) for $NK = 1, 2$ with those from (XR5.18.3) for $NK = 1, 2$.

5.19 Consider the following expenditure share equation where the budget share for food *WFOOD* is expressed as a function of total expenditure *TOTEXP*.

$$WFOOD = \beta_1 + \beta_2 \ln(TOTEXP) + e_F \quad \text{(XR5.19.1)}$$

In Exercise 4.12, it was noted that the elasticity of expenditure on food with respect to total expenditure is given by

$$\varepsilon = 1 + \frac{\beta_2}{\beta_1 + \beta_2 \ln(TOTEXP)}$$

Also, in Exercise 5.18 it was indicated that a good is a necessity if $\beta_2 < 0$.

a. Show that $\beta_2 < 0$ if and only if $\varepsilon < 1$. That is, a good is a necessity if its expenditure elasticity is less than one (inelastic).

b. Use observations in the data file *london5* to estimate (XR5.19.1) and comment on the results.

c. Find point estimates and 95% interval estimates for the mean budget share for food, for total expenditure values (i) $TOTEXP = 50$ (the fifth percentile of *TOTEXP*), (ii) $TOTEXP = 90$ (the median), and (iii) $TOTEXP = 170$ (the 95th percentile).

d. Find point estimates and 95% interval estimates for the elasticity ε, for total expenditure values (i) $TOTEXP = 50$ (the fifth percentile), (ii) $TOTEXP = 90$ (the median), and (iii) $TOTEXP = 170$ (the 95th percentile).

e. Comment on how the mean budget share and the expenditure elasticity for food change as total expenditure changes. How does the reliability of estimation change as total expenditure changes?

5.20 A generalized version of the estimator for β_2 proposed in Exercise 2.9 by Professor I.M. Mean for the regression model $y_i = \beta_1 + \beta_2 x_i + e_i, i = 1, 2, \ldots, N$ is

$$\hat{\beta}_{2,mean} = \frac{\bar{y}_2 - \bar{y}_1}{\bar{x}_2 - \bar{x}_1}$$

where $(\bar{y}_1, \bar{x}_1)$ and $(\bar{y}_2, \bar{x}_2)$ are the sample means for the first and second halves of the sample observations, respectively, after ordering the observations according to increasing values of x. Given that assumptions MR1–MR6 hold:

a. Show that $\hat{\beta}_{2,mean}$ is unbiased.

b. Derive an expression for $\text{var}\left(\hat{\beta}_{2,mean}|\mathbf{x}\right)$.

c. Write down an expression for $\text{var}\left(\hat{\beta}_{2,mean}\right)$.

d. Under what conditions will $\hat{\beta}_{2,mean}$ be a consistent estimator for β_2?

e. Randomly generate observations on x from a uniform distribution on the interval (0,10) for sample sizes $N = 100, 500, 1000$, and, if your software permits, $N = 5000$. Assuming $\sigma^2 = 1000$, for each sample size, compute:

  i. $\text{var}(b_2|\mathbf{x})$ and $\text{var}\left(\hat{\beta}_{2,mean}|\mathbf{x}\right)$ where b_2 is the OLS estimator.

  ii. Estimates for $E\left[\left(s_x^2\right)^{-1}\right]$ and $E\left[4/\left(\bar{x}_2 - \bar{x}_1\right)^2\right]$ where s_x^2 is the sample standard deviation for x using N as a divisor.

f. Comment on the relative magnitudes of your answers in part (e), (i) and (ii) and how they change as sample size increases. Does it appear that $\hat{\beta}_{2,mean}$ is consistent?

g. Show that $E\left[\left(s_x^2\right)^{-1}\right] \xrightarrow{p} 0.12$ and $E\left[4/\left(\bar{x}_2 - \bar{x}_1\right)^2\right] \xrightarrow{p} 0.16$. [*Hint*: The variance of a uniform random variable defined on the interval (a, b) is $(b - a)^2/12$.]

h. Suppose that the observations on x were not ordered according to increasing magnitude but were randomly assigned to any position. Would the estimator $\hat{\beta}_{2,mean}$ be consistent? Why or why not?

5.21 Using the data in the file *toody5*, estimate the model

$$Y_t = \beta_1 + \beta_2 TREND_t + \beta_3 RAIN_t + \beta_4 RAIN_t^2 + \beta_5 (RAIN_t \times TREND_t) + e_t$$

where Y_t = wheat yield in tons per hectare in the Toodyay Shire of Western Australia in year t; $TREND_t$ is a trend variable designed to capture technological change, with observations 0, 0.1, 0.2, …, 4.7; 0 is for the year 1950, 0.1 is for the year 1951, and so on up to 4.7 for the year 1997; $RAIN_t$ is total rainfall in decimeters (dm) from May to October (the growing season) in year t (1 decimeter = 4 inches).

a. Report your estimates, standard errors, t-values, and p-values in a table.
b. Are each of your estimates significantly different from zero at a (i) 5% level, (ii) 10% level?
c. Do the coefficients have the expected signs? Why? (One of the objectives of technological improvements is the development of drought-resistant varieties of wheat.)
d. Find point and 95% interval estimates of the marginal effect of extra rainfall in (i) 1959 when the rainfall was 2.98 dm and (ii) 1995 when the rainfall was 4.797 dm. Discuss the results.
e. Find point and 95% interval estimates for the amount of rainfall that would maximize expected yield in (i) 1959 and (ii) 1995. Discuss the results.

5.22 Using the data in the file *toody5*, estimate the model

$$Y_t = \beta_1 + \beta_2 TREND_t + \beta_3 RAIN_t + \beta_4 RAIN_t^2 + \beta_5 (RAIN_t \times TREND_t) + e_t$$

where Y_t = wheat yield in tons per hectare in the Toodyay Shire of Western Australia in year t; $TREND_t$ is a trend variable designed to capture technological change, with observations 0, 0.1, 0.2, …, 4.7; 0 is for the year 1950, 0.1 is for the year 1951, and so on up to 4.7 for the year 1997; $RAIN_t$ is total rainfall in decimeters (dm) from May to October (the growing season) in year t (1 decimeter = 4 inches).

a. Report your estimates, standard errors, t-values, and p-values in a table.
b. For 1974, when $TREND = 2.4$ and $RAIN = 4.576$, use a 5% significance level to test the null hypothesis that extra rainfall will not increase expected yield against the alternative that it will increase expected yield.
c. Assuming rainfall is equal to its median value of 3.8355 dm, find point and 95% interval estimates of the expected improvement in wheat yield from technological change over the period 1960–1995.
d. There is concern that climate change is leading to a decline in rainfall over time. To test this hypothesis, estimate the equation $RAIN = \alpha_1 + \alpha_2 TREND + e$. Test, at a 5% significance level, the null hypothesis that mean rainfall is not declining over time against the alternative hypothesis that it is declining.
e. Using the estimated equation from part (d), estimate mean rainfall in 1960 and in 1995.
f. Suppose that $TREND_{1995} = TREND_{1960}$, implying there had been no technological change from 1960 to 1995. Use the estimates from part (e) to find an estimate of the decline in mean yield from 1960 to 1995 attributable to climate change.
g. Suppose that $E(RAIN_{1995}) = E(RAIN_{1960})$, implying there had been no rainfall change from 1960 to 1995. Find an estimate of the increase in mean yield from 1960 to 1995 attributable to technological change.
h. Compare the estimates you obtained in parts (c), (f), and (g).

5.23 The file *cocaine* contains 56 observations on variables related to sales of cocaine powder in northeastern California over the period 1984–1991. The data are a subset of those used in the study Caulkins, J. P. and R. Padman (1993), "Quantity Discounts and Quality Premia for Illicit Drugs," *Journal of the American Statistical Association*, 88, 748–757. The variables are

$PRICE$ = price per gram in dollars for a cocaine sale
$QUANT$ = number of grams of cocaine in a given sale
$QUAL$ = quality of the cocaine expressed as percentage purity
$TREND$ = a time variable with 1984 = 1 up to 1991 = 8
Consider the regression model

$$PRICE = \beta_1 + \beta_2 QUANT + \beta_3 QUAL + \beta_4 TREND + e$$

a. What signs would you expect on the coefficients β_2, β_3, and β_4?

b. Use your computer software to estimate the equation. Report the results and interpret the coefficient estimates. Have the signs turned out as you expected?

c. What proportion of variation in cocaine price is explained jointly by variation in quantity, quality, and time?

d. It is claimed that the greater the number of sales, the higher the risk of getting caught. Thus, sellers are willing to accept a lower price if they can make sales in larger quantities. Set up H_0 and H_1 that would be appropriate to test this hypothesis. Carry out the hypothesis test.

e. Test the hypothesis that the quality of cocaine has no influence on expected price against the alternative that a premium is paid for better-quality cocaine.

f. What is the average annual change in the cocaine price? Can you suggest why price might be changing in this direction?

5.24 The file *collegetown* contains data on 500 single-family houses sold in Baton Rouge, Louisiana during 2009–2013. We will be concerned with the selling price in thousands of dollars (*PRICE*), the size of the house in hundreds of square feet (*SQFT*), and the age of the house measured as a categorical variable (*AGE*), with 1 representing the newest and 11 the oldest. Let **X** denote all observations on *SQFT* and *AGE*. Use all observations to estimate the following regression model:

$$PRICE = \beta_1 + \beta_2 SQFT + \beta_3 (SQFT \times AGE) + e$$

a. Report the results. Are the estimated coefficients significantly different from zero?

b. Write down expressions for the marginal effects $\partial E(PRICE|\mathbf{X})/\partial SQFT$ and $\partial E(PRICE|\mathbf{X})/\partial AGE$. Interpret each of the coefficients. Given the categorical nature of the variable *AGE*, what assumptions are being made?

c. Find point and 95% interval estimates for the marginal effect $\partial E(PRICE|\mathbf{X})/\partial SQFT$ for houses that are (i) 5 years old, (ii) 20 years old, and (iii) 40 years old. How do these estimates change as *AGE* increases? (Refer to the file *collegetown.def* for the definition of *AGE*.)

d. As a house gets older and moves from one age category to the next, the expected price declines by $6000. Set up this statement as a null hypothesis for houses with (i) 1500 square feet, (ii) 3000 square feet, and (iii) 4500 square feet. Using a 5% significance level, test each of the null hypotheses against an alternative that the price decline is not $6000. Discuss the outcomes.

e. Find a 95% prediction interval for the price of a 60-year old house with 2500 square feet. In the data set there are four 60-year old houses with floor space between 2450 and 2550 square feet. What prices did they sell for? How many of these prices fall within your prediction interval? Is the model a good one for forecasting price?

5.25 The file *collegetown* contains data on 500 single-family houses sold in Baton Rouge, Louisiana during 2009–2013. We will be concerned with the selling price in thousands of dollars (*PRICE*), and the size of the house in hundreds of square feet (*SQFT*). Use all observations to estimate the following regression model:

$$\ln(PRICE) = \beta_1 + \beta_2 SQFT + \beta_3 SQFT^{1/2} + e$$

Suppose that assumptions MR1–MR6 all hold. In particular, $(e|SQFT) \sim N\left(0, \sigma^2\right)$.

a. Report the results. Are the estimated coefficients significantly different from zero?

b. Write down an expression for the marginal effect $\partial E\left[\ln(PRICE|SQFT)\right]/\partial SQFT$. Discuss the nature of this marginal effect and the expected signs for β_2 and β_3.

c. Find and interpret point and 95% interval estimates for $\partial E\left[\ln(PRICE|SQFT)\right]/\partial SQFT$ for houses with (i) 1500 square feet, (ii) 3000 square feet, and (iii) 4500 square feet.

d. Show that

$$\frac{\partial E[PRICE|SQFT]}{\partial SQFT} = \left(\beta_2 + \frac{\beta_3}{2SQFT^{1/2}}\right) \times \exp\left\{\beta_1 + \beta_2 SQFT + \beta_3 SQFT^{1/2} + \sigma^2/2\right\}$$

For future reference, we write this expression as $\partial E(PRICE|SQFT)/\partial SQFT = S \times C$ where

$$S = \left(\beta_2 + \frac{\beta_3}{2SQFT^{1/2}}\right) \times \exp\left\{\beta_1 + \beta_2 SQFT + \beta_3 SQFT^{1/2}\right\} \quad \text{and } C = \exp\left\{\sigma^2/2\right\}$$

Correspondingly, we let $\hat{S}$ and $\hat{C}$ denote estimates for S and C obtained by replacing unknown parameters by their estimates.

e. Estimate $\partial E(PRICE|SQFT)/\partial SQFT = S \times C$ for houses with (i) 1500 square feet, (ii) 3000 square feet, and (iii) 4500 square feet.

f. Finding the asymptotic standard errors for the estimates in (e) is tricky because of the product $\hat{S} \times \hat{C}$. To avoid such trickiness, find the standard errors for $\hat{S}$ for each type of house in (e).

g. For each type of house, and a 5% significance level, use the estimates from (e) and the standard errors from (f) to test the hypotheses

$$H_0: \frac{\partial E(PRICE|SQFT)}{\partial SQFT} = 9 \quad H_1: \frac{\partial E(PRICE|SQFT)}{\partial SQFT} \neq 9$$

What do you conclude?

h. (optional) To get the "correct" standard errors for $\hat{S} \times \hat{C}$, we proceed as follows. First, given $\text{var}\left(\hat{\sigma}^2\right) = 2\sigma^4/(N-K)$, find an estimate for $\text{var}\left(\hat{C}\right)$. It can be shown that $\hat{S}$ and $\hat{C}$ are independent. Using results on the product of independent random variables, an estimator for the variance of $\hat{S} \times \hat{C}$ is

$$\widehat{\text{var}}\left(\hat{S} \times \hat{C}\right) = \widehat{\text{var}}\left(\widehat{\frac{\partial E(PRICE|SQFT)}{\partial SQFT}}\right) = \hat{S}^2\,\widehat{\text{var}}\left(\hat{C}\right) + \hat{C}^2\,\widehat{\text{var}}\left(\hat{S}\right) + \widehat{\text{var}}\left(\hat{C}\right)\widehat{\text{var}}\left(\hat{S}\right)$$

Use this result to find standard errors for $\hat{S} \times \hat{C}$. How do they compare with the standard errors obtained in (f)? Are they likely to change the outcomes of the hypothesis tests in (g)?

5.26 Consider the presidential voting data (data file *fair5*) introduced in Exercise 2.23. Details of the variables can be found in that exercise.

a. Using all observations, estimate the regression model

$$VOTE = \beta_1 + \beta_2 GROWTH + \beta_3 INFLAT + e$$

Report the results. Are the estimates for β_2 and β_3 significantly different from zero at a 10% significance level? Did you use one- or two-tail tests? Why?

b. Assume the inflation rate is 3% and the Democrats are the incumbent party. Predict the percentage vote for both parties when the growth rate is (i) −2%, (ii) 0%, and (iii) 3%.

c. Assume the inflation rate is 3% and the Republicans are the incumbent party. Predict the percentage vote for both parties when the growth rate is (i) −2%, (ii) 0%, and (iii) 3%.

d. Based on your answers to parts (b) and (c), do you think the popular vote tends to be more biased toward the Democrats or the Republicans?

e. Consider the following two scenarios:

1. The Democrats are the incumbent party, the growth rate is 2% and the inflation rate is 2%.
2. The Republicans are the incumbent party, the growth rate is 2% and the inflation rate is 2%.

Using a 5% significance level, test the null hypothesis that the expected share of the Democratic vote under scenario 1 is equal to the expected share of the Republican vote under scenario 2.

5.27 In this exercise, we consider the auction market for art first introduced in Exercise 2.24. The variables in the data file *ashcan_small* that we will be concerned with are as follows:

$RHAMMER$ = the price at which a painting sold in thousands of dollars
$YEARS\_OLD$ = the time between completion of the painting and when it was sold
$INCHSQ$ = the size of the painting in square inches

Create a new variable $INCHSQ10 = INCHSQ/10$ to express size in terms of tens of square inches. Only consider observations where the art was sold ($SOLD = 1$).

a. Estimate the following equation and report the results:

$$RHAMMER = \beta_1 + \beta_2 YEARS\_OLD + \beta_3 INCHSQ10 + e$$

b. How much do paintings appreciate on a yearly basis? Find a 95% interval estimate for the expected yearly price increase.

c. How much more valuable are large paintings? Find a 95% interval estimate for the expected extra value from an extra 10 square inches.

d. Add the variable $INCHSQ10^2$ to the model and re-estimate. Report the results. Why would you consider adding this variable?

e. Does adding this variable have much impact on the interval estimate in part (b)?

f. Find 95% interval estimates for the expected extra value from an extra 10 square inches for art of the following sizes: (i) 50 square inches (sixth percentile), (ii) 250 square inches (approximately the median), and (iii) 900 square inches (97th percentile). Comment on how the value of an extra 10 square inches changes as the painting becomes larger.

g. Find a 95% interval estimate for the painting size that maximizes price.

h. Find a 95% interval estimate for the expected price of a 75-year-old, 100-square-inch painting.

i. How long would you have to keep a 100-square-inch painting for the expected price to become positive?

5.28 In this exercise, we consider the auction market for art first introduced in Exercise 2.24. The variables in the data file *ashcan_small* that we will be concerned with are as follows:

$RHAMMER$ = the price at which a painting sold in thousands of dollars
$YEARS\_OLD$ = the time between completion of the painting and when it was sold
$INCHSQ$ = the size of the painting in square inches

Create a new variable $INCHSQ10 = INCHSQ/10$ to express size in terms of tens of square inches. Only consider observations where the art was sold ($SOLD = 1$).

a. Estimate the following log-linear equation and report the results:

$$\ln(RHAMMER) = \beta_1 + \beta_2 YEARS\_OLD + \beta_3 INCHSQ10 + e$$

b. How much do paintings appreciate on a yearly basis? Find a 95% interval estimate for the expected percentage price increase per year.

c. How much more valuable are large paintings? Using a 5% significance level, test the null hypothesis that painting an extra 10 square inches increases the value by 2% or less against the alternative that it increases the value by more than 2%.

d. Add the variable $INCHSQ10^2$ to the model and re-estimate. Report the results. Why would you consider adding this variable?

e. Does adding this variable have much impact on the interval estimate in part (b)?

f. Redo the hypothesis test in part (c) for art of the following sizes: (i) 50 square inches (sixth percentile), (ii) 250 square inches (approximately the median), and (iii) 900 square inches (97th percentile). What do you observe?

g. Find a 95% interval estimate for the painting size that maximizes price.

h. Find a 95% interval estimate for the expected price of a 75-year-old, 100-square-inch painting. (Use the estimator $\exp\left\{E\left[\ln(RHAMMER|YEARS\_OLD = 75, INCHSQ10 = 10)\right]\right\}$ and its standard error.)

5.29 What is the relationship between crime and punishment? This important question has been examined by Cornwell and Trumbull[16] using a panel of data from North Carolina. The cross sections are 90 counties, and the data are annual for the years 1981–1987. The data are in the file *crime*.

Using the data from 1986, estimate a regression relating the log of the crime rate $LCRMRTE$ to the probability of an arrest $PRBARR$ (the ratio of arrests to offenses), the probability of conviction $PRBCONV$ (the ratio of convictions to arrests), the probability of a prison sentence $PRBPRIS$ (the ratio of prison sentences to convictions), the number of police per capita $POLPC$, and the weekly wage in construction $WCON$. Write a report of your findings. In your report, explain what effect you would expect each of the variables to have on the crime rate and note whether the estimated coefficients have the expected signs and are significantly different from zero. What variables appear to be the most important for crime deterrence? Can you explain the sign for the coefficient of $POLPC$?

5.30 In Section 5.7.4, we discovered that the optimal level of advertising for Big Andy's Burger Barn, $ADVERT_0$, satisfies the equation $\beta_3 + 2\beta_4 ADVERT_0 = 1$. Using a 5% significance level, test whether each of the following levels of advertising could be optimal: (a) $ADVERT_0 = 1.75$, (b) $ADVERT_0 = 1.9$, and (c) $ADVERT_0 = 2.3$. What are the p-values for each of the tests?

5.31 Each morning between 6:30 AM and 8:00 AM Bill leaves the Melbourne suburb of Carnegie to drive to work at the University of Melbourne. The time it takes Bill to drive to work ($TIME$), depends on the departure time ($DEPART$), the number of red lights that he encounters ($REDS$), and the number of trains that he has to wait for at the Murrumbeena level crossing ($TRAINS$). Observations on these

[16] "Estimating the Economic Model of Crime with Panel Data," *Review of Economics and Statistics*, 76, 1994, 360–366. The data were kindly provided by the authors.

variables for the 249 working days in 2015 appear in the file *commute5*. *TIME* is measured in minutes. *DEPART* is the number of minutes after 6:30 AM that Bill departs.

a. Estimate the equation

$$TIME = \beta_1 + \beta_2 DEPART + \beta_3 REDS + \beta_4 TRAINS + e$$

Report the results and interpret each of the coefficient estimates, including the intercept β_1.

b. Find 95% interval estimates for each of the coefficients. Have you obtained precise estimates of each of the coefficients?

c. Using a 5% significance level, test the null hypothesis that Bill's expected delay from each red light is 2 minutes or more against the alternative that it is less than 2 minutes.

d. Using a 10% significance level, test the null hypothesis that the expected delay from each train is 3 minutes against the alternative that it is not 3 minutes.

e. Using a 5% significance level, test the null hypothesis that Bill can expect a trip to be at least 10 minutes longer if he leaves at 7:30 AM instead of 7:00 AM, against the alternative that it will not be 10 minutes longer. (Assume other things are equal.)

f. Using a 5% significance level, test the null hypothesis that the expected delay from a train is at least three times greater than the expected delay from a red light against the alternative that it is less than three times greater.

g. Suppose that Bill encounters six red lights and one train. Using a 5% significance level, test the null hypothesis that leaving Carnegie at 7:00 AM is early enough to get him to the university on or before 7:45 AM against the alternative that it is not. [Carry out the test in terms of the expected time $E(TIME|\mathbf{X})$ where **X** represents the observations on all explanatory variables.]

h. Suppose that, in part (g), it is imperative that Bill is not late for his 7:45 AM meeting. Have the null and alternative hypotheses been set up correctly? What happens if these hypotheses are reversed?

5.32 Reconsider the variables and model from Exercise 5.31

$$TIME = \beta_1 + \beta_2 DEPART + \beta_3 REDS + \beta_4 TRAINS + e$$

Suppose that Bill is mainly interested in the magnitude of the coefficient β_2. He has control over his departure time, but no control over the red lights or the trains.

a. Regress *DEPART* on the variables *REDS* and *TRAINS* and save the residuals. Which coefficient estimates are significantly different from zero at a 5% level? For the significant coefficient(s), do you think the relationship is causal?

b. Regress *TIME* on the variables *REDS* and *TRAINS* and save the residuals. Are the estimates for the coefficients of *REDS* and *TRAINS* very different from the estimates for β_3 and β_4 obtained by estimating the complete model with *DEPART* included?

c. Use the residuals from parts (a) and (b) to estimate the coefficient β_2 and adjust the output to obtain its correct standard error.

5.33 Use the observations in the data file *cps5_small* to estimate the following model:

$$\ln(WAGE) = \beta_1 + \beta_2 EDUC + \beta_3 EDUC^2 + \beta_4 EXPER + \beta_5 EXPER^2 + \beta_6(EDUC \times EXPER) + e$$

a. At what levels of significance are each of the coefficient estimates "significantly different from zero"?

b. Obtain an expression for the marginal effect $\partial E\left[\ln(WAGE)|EDUC, EXPER\right]/\partial EDUC$. Comment on how the estimate of this marginal effect changes as *EDUC* and *EXPER* increase.

c. Evaluate the marginal effect in part (b) for all observations in the sample and construct a histogram of these effects. What have you discovered? Find the median, 5th percentile, and 95th percentile of the marginal effects.

d. Obtain an expression for the marginal effect $\partial E\left[\ln(WAGE)|EDUC, EXPER\right]/\partial EXPER$. Comment on how the estimate of this marginal effect changes as *EDUC* and *EXPER* increase.

e. Evaluate the marginal effect in part (d) for all observations in the sample and construct a histogram of these effects. What have you discovered? Find the median, 5th percentile, and 95th percentile of the marginal effects.

f. David has 17 years of education and 8 years of experience, while Svetlana has 16 years of education and 18 years of experience. Using a 5% significance level, test the null hypothesis that Svetlana's expected log-wage is equal to or greater than David's expected log-wage, against the alternative that David's expected log-wage is greater. State the null and alternative hypotheses in terms of the model parameters.

g. After eight years have passed, when David and Svetlana have had eight more years of experience, but no more education, will the test result in (f) be the same? Explain this outcome?

h. Wendy has 12 years of education and 17 years of experience, while Jill has 16 years of education and 11 years of experience. Using a 5% significance level, test the null hypothesis that their marginal effects of extra experience are equal against the alternative that they are not. State the null and alternative hypotheses in terms of the model parameters.

i. How much longer will it be before the marginal effect of experience for Jill becomes negative? Find a 95% interval estimate for this quantity.

Appendix 5A Derivation of Least Squares Estimators

In Appendix 2A, we derived expressions for the least squares estimators b_1 and b_2 in the simple regression model. In this appendix, we proceed with a similar exercise for the multiple regression model; we describe how to obtain expressions for b_1, b_2, and b_3 in a model with two explanatory variables. Given sample observations on y, x_2, and x_3, the problem is to find values for β_1, β_2, and β_3 that minimize

$$S(\beta_1, \beta_2, \beta_3) = \sum_{i=1}^{N}(y_i - \beta_1 - \beta_2 x_{i2} - \beta_3 x_{i3})^2$$

The first step is to partially differentiate S with respect to β_1, β_2, and β_3 and to set the first-order partial derivatives to zero. This yields

$$\frac{\partial S}{\partial \beta_1} = 2N\beta_1 + 2\beta_2 \sum x_{i2} + 2\beta_3 \sum x_{i3} - 2\sum y_i$$

$$\frac{\partial S}{\partial \beta_2} = 2\beta_1 \sum x_{i2} + 2\beta_2 \sum x_{i2}^2 + 2\beta_3 \sum x_{i2}x_{i3} - 2\sum x_{i2}y_i$$

$$\frac{\partial S}{\partial \beta_3} = 2\beta_1 \sum x_{i3} + 2\beta_2 \sum x_{i2}x_{i3} + 2\beta_3 \sum x_{i3}^2 - 2\sum x_{i3}y_i$$

Setting these partial derivatives equal to zero, dividing by 2, and rearranging yields

$$\begin{aligned} Nb_1 + \sum x_{i2}b_2 + \sum x_{i3}b_3 &= \sum y_i \\ \sum x_{i2}b_1 + \sum x_{i2}^2 b_2 + \sum x_{i2}x_{i3}b_3 &= \sum x_{i2}y_i \\ \sum x_{i3}b_1 + \sum x_{i2}x_{i3}b_2 + \sum x_{i3}^2 b_3 &= \sum x_{i3}y_i \end{aligned} \tag{5A.1}$$

The least squares estimators for b_1, b_2, and b_3 are given by the solution of this set of three *simultaneous equations*, known as the **normal equations**. To write expressions for this solution, it is convenient to express the variables as deviations from their means. That is, let

$$y_i^* = y_i - \bar{y}, \quad x_{i2}^* = x_{i2} - \bar{x}_2, \quad x_{i3}^* = x_{i3} - \bar{x}_3$$

Then the least squares estimates b_1, b_2, and b_3 are

$$b_1 = \bar{y} - b_2\bar{x}_2 - b_3\bar{x}_3$$

$$b_2 = \frac{\left(\sum y_i^* x_{i2}^*\right)\left(\sum x_{i3}^{*2}\right) - \left(\sum y_i^* x_{i3}^*\right)\left(\sum x_{i2}^* x_{i3}^*\right)}{\left(\sum x_{i2}^{*2}\right)\left(\sum x_{i3}^{*2}\right) - \left(\sum x_{i2}^* x_{i3}^*\right)^2}$$

$$b_3 = \frac{\left(\sum y_i^* x_{i3}^*\right)\left(\sum x_{i2}^{*2}\right) - \left(\sum y_i^* x_{i2}^*\right)\left(\sum x_{i3}^* x_{i2}^*\right)}{\left(\sum x_{i2}^{*2}\right)\left(\sum x_{i3}^{*2}\right) - \left(\sum x_{i2}^* x_{i3}^*\right)^2}$$

For models with more than three parameters, the solutions become quite messy without using matrix algebra; we will not show them. Computer software used for multiple regression computations solves normal equations such as those in (5A.1) to obtain the least squares estimates.

Appendix 5B The Delta Method

In Sections 3.6, 5.3, 5.4, and 5.5, we discussed estimating and testing **linear combinations** of parameters. If the regression errors are normal, the results discussed there hold in finite samples. If the regression errors are not normal, then those results hold in large samples, as discussed in Section 5.7. We now turn to **nonlinear functions** of regression parameters that were considered in Section 5.7.4 and provide some background for the results given there. You will be surprised in the subsequent chapters how many times we become interested in **nonlinear functions** of regression parameters. For example, we may find ourselves interested in functions such as $g_1(\beta_2) = \exp(\beta_2/10)$ or $g_2(\beta_1, \beta_2) = \beta_1/\beta_2$. The first function $g_1(\beta_2)$ is a function of the single parameter β_2. Intuitively, we would estimate this function of β_2 using $g_1(b_2)$. The second function $g_2(\beta_1, \beta_2)$ is a function of two parameters and similarly $g_2(b_1, b_2)$ seems like a reasonable estimator. Working with nonlinear functions of the estimated parameters requires additional tools because, even if the regression errors are normal, nonlinear functions of them are not normally distributed in finite samples, and usual variance formulas do not apply.

5B.1 Nonlinear Function of a Single Parameter

The key to working with nonlinear functions of a single parameter is the Taylor series approximation discussed in Appendix A, Derivative Rule 9. It is stated there as

$$f(x) \cong f(a) + \left.\frac{df(x)}{dx}\right|_{x=a}(x-a) = f(a) + f'(a)(x-a)$$

The value of a function at x is approximately equal to the value of the function at $x = a$, plus the derivative of the function evaluated at $x = a$, times the difference $x - a$. This approximation works well when the function is smooth and the difference $x - a$ is not too large. We will apply this rule to $g_1(b_2)$ replacing x with b_2 and a with β_2

$$g_1(b_2) \cong g_1(\beta_2) + g_1'(\beta_2)(b_2 - \beta_2) \tag{5B.1}$$

This Taylor series expansion of $g_1(b_2)$ shows the following:

1. If $E(b_2) = \beta_2$, then $E[g_1(b_2)] \cong g_1(\beta_2)$.
2. If b_2 is a biased but consistent estimator, so that $b_2 \xrightarrow{p} \beta_2$, then $g_1(b_2) \xrightarrow{p} g_1(\beta_2)$.
3. The variance of $g_1(b_2)$ is given by $\text{var}[g_1(b_2)] \cong [g_1'(\beta_2)]^2 \text{var}(b_2)$, which is known as the **delta method**. The delta method follows from working with the Taylor series approximation

$$\begin{aligned}\text{var}[g_1(b_2)] &= \text{var}\,[g_1(\beta_2) + g_1'(\beta_2)(b_2 - \beta_2)]\\ &= \text{var}\,[g_1'(\beta_2)(b_2 - \beta_2)] \text{ because } g_1(\beta_2) \text{ is not random}\\ &= [g_1'(\beta_2)]^2 \text{var}(b_2 - \beta_2) \text{ because } g_1'(\beta_2) \text{ is not random}\\ &= [g_1'(\beta_2)]^2 \text{var}(b_2) \text{ because } \beta_2 \text{ is not random}\end{aligned}$$

4. The estimator $g_1(b_2)$ has an approximate normal distribution in large samples,

$$g_1(b_2) \overset{a}{\sim} N\left[g_1(\beta_2)\,,\, [g_1'(\beta_2)]^2 \text{var}(b_2)\right] \tag{5B.2}$$

The asymptotic normality of $g_1(b_2)$ means that we can test nonlinear hypotheses about β_2, such as $H_0{:}\,g_1(\beta_2) = c$, and we can construct interval estimates of $g_1(\beta_2)$ in the usual way. To implement the delta method, we replace β_2 by its estimate b_2 and the true variance $\text{var}(b_2)$ by its estimate $\widehat{\text{var}}(b_2)$ which, for the simple regression model, is given in equation (2.21).

EXAMPLE 5.19 | An Interval Estimate for $\exp(\beta_2/10)$

To illustrate the delta method calculations, we use one sample from the $N = 20$ simulation considered in Appendix 5C; it is stored as *mc20*. For these data values, the fitted regression is

$$\hat{y} = 87.44311 + 10.68456x$$
$$(\text{se})\ (33.8764) \qquad (2.1425\ )$$

The nonlinear function we consider is $g_1(\beta_2) = \exp(\beta_2/10)$. In the simulation we know the value $\beta_2 = 10$ and therefore the value of the function is $g_1(\beta_2) = \exp(\beta_2/10) = e^1 = 2.71828$. To apply the delta method, we need the derivative $g_1'(\beta_2) = \exp(\beta_2/10) \times (1/10)$ (see Appendix A, Derivative Rule 7), and the estimated covariance matrix in Table 5B.1.

The estimated value of the nonlinear function is

$$g_1(b_2) = \exp(b_2/10) = \exp(10.68456/10) = 2.91088$$

The estimated variance is

$$\widehat{\text{var}}[g_1(b_2)] = [g_1'(b_2)]^2 \widehat{\text{var}}(b_2) = [\exp(b_2/10) \times (1/10)]^2 \widehat{\text{var}}(b_2)$$
$$= [\exp(10.68456/10) \times (1/10)]^2 4.59045 = 0.38896$$

TABLE 5B.1 Estimated Covariance Matrix

	b_1	b_2
b_1	1147.61330	−68.85680
b_2	−68.85680	4.59045

and

$$\text{se}[g_1(b_2)] = 0.62367.$$

The 95% interval estimate is

$$g_1(b_2) \pm t_{(0.975,20-2)}\text{se}[g_1(b_2)] = 2.91088 \pm 2.10092 \times 0.62367$$
$$= (1.60061, 4.22116)$$

5B.2 Nonlinear Function of Two Parameters[17]

When working with functions of two (or more) parameters the approach is much the same, but the Taylor series approximation changes to a more general form. For a function of two parameters, the Taylor series approximation is

$$g_2(b_1, b_2) \cong g_2(\beta_1, \beta_2) + \frac{\partial g_2(\beta_1, \beta_2)}{\partial \beta_1}(b_1 - \beta_1) + \frac{\partial g_2(\beta_1, \beta_2)}{\partial \beta_2}(b_2 - \beta_2) \qquad (5B.3)$$

1. If $E(b_1) = \beta_1$ and $E(b_2) = \beta_2$, then $E[g_2(b_1, b_2)] \cong g_2(\beta_1, \beta_2)$.
2. If b_1 and b_2 are consistent estimators, so that $b_1 \xrightarrow{p} \beta_1$ and $b_2 \xrightarrow{p} \beta_2$, then $g_2(b_1, b_2) \xrightarrow{p} g_2(\beta_1, \beta_2)$.
3. The variance of $g_2(b_1, b_2)$ is given by the **delta method** as

$$\text{var}[g_2(b_1, b_2)] \cong \left[\frac{\partial g_2(\beta_1, \beta_2)}{\partial \beta_1}\right]^2 \text{var}(b_1) + \left[\frac{\partial g_2(\beta_1, \beta_2)}{\partial \beta_2}\right]^2 \text{var}(b_2)$$
$$+ 2\left[\frac{\partial g_2(\beta_1, \beta_2)}{\partial \beta_1}\right]\left[\frac{\partial g_2(\beta_1, \beta_2)}{\partial \beta_2}\right] \text{cov}(b_1, b_2) \qquad (5B.4)$$

4. The estimator $g_2(b_1, b_2)$ has an approximate normal distribution in large samples,

$$g_2(b_1, b_2) \overset{a}{\sim} N(g_2(\beta_1, \beta_2),\ \text{var}[g_2(b_1, b_2)]\ ) \qquad (5B.5)$$

The asymptotic normality of $g_2(b_1, b_2)$ means that we can test nonlinear hypotheses such as $H_0 : g_2(\beta_1, \beta_2) = c$, and we can construct interval estimates of $g_2(\beta_1, \beta_2)$ in the usual way.

[17] This section contains advanced material. The general case involving a function of more than two parameters requires matrix algebra. See William Greene, *Econometric Analysis 8e*, Pearson Prentice-Hall, 2018, Theorems D.21A and D.22 in online Appendix available at pages.stern.nyu.edu/~wgreene/text/econometricanalysis.htm.

In practice we evaluate the derivatives at the estimates b_1 and b_2, and the variances and covariances by their usual estimates from equations such as those for the simple regression model in (2.20)–(2.22).

EXAMPLE 5.20 | An Interval Estimate for β_1/β_2

The nonlinear function of two parameters that we consider is $g_2(\beta_1, \beta_2) = \beta_1/\beta_2$. To employ the delta method, we require the derivatives (see Appendix A, Derivative Rules 3 and 6)

$$\frac{\partial g_2(\beta_1,\beta_2)}{\partial \beta_1} = \frac{1}{\beta_2} \quad \text{and} \quad \frac{\partial g_2(\beta_1,\beta_2)}{\partial \beta_2} = -\frac{\beta_1}{\beta_2^2}$$

The estimate $g_2(b_1, b_2) = b_1/b_2 = 87.44311/10.68456 = 8.18406$ and its estimated variance is

$$\begin{aligned}\widehat{\text{var}}\left[g_2(b_1,b_2)\right] &= \left[\frac{1}{b_2}\right]^2 \widehat{\text{var}}(b_1) + \left[-\frac{b_1}{b_2^2}\right]^2 \widehat{\text{var}}(b_2) \\ &\quad + 2\left[\frac{1}{b_2}\right]\left[-\frac{b_1}{b_2^2}\right]\widehat{\text{cov}}(b_1,b_2) \\ &= 22.61857\end{aligned}$$

The delta method standard error is $\text{se}(b_1/b_2) = 4.75590$. The resulting 95% interval estimate for β_1/β_2 is $(-1.807712, 18.17583)$. While all this seems incredibly complicated, most software packages will compute at least the estimates and standard errors automatically. And now that you understand the calculations, you can be confident when you use the "canned" routines.

Appendix 5C Monte Carlo Simulation

In Appendices 2H and 3C, we introduced a Monte Carlo simulation to illustrate the repeated sampling properties of the least squares estimators. In this appendix, we use the same framework to illustrate the repeated sampling performances of interval estimators and hypothesis tests when the errors are not normally distributed.

Recall that the **data generation process** for the simple linear regression model is given by

$$y_i = E(y_i|x_i) + e_i = \beta_1 + \beta_2 x_i + e_i, \quad i = 1, \ldots, N$$

The Monte Carlo parameter values are $\beta_1 = 100$ and $\beta_2 = 10$. The value of x_i is 10 for the first $N/2$ observations and 20 for the remaining $N/2$ observations, so that the regression functions are

$$\begin{aligned} E(y_i|x_i = 10) &= 100 + 10x_i = 100 + 10 \times 10 = 200, \quad i = 1, \ldots, N/2 \\ E(y_i|x_i = 20) &= 100 + 10x_i = 100 + 10 \times 20 = 300, \quad i = (N/2) + 1, \ldots, N \end{aligned}$$

5C.1 Least Squares Estimation with Chi-Square Errors

In this appendix, we modify the simulation in an important way. The random errors are independently distributed but with normalized chi-square distributions. In Figure B.7, the *pdf*s of several chi-square distributions are shown. We will use the $\chi^2_{(4)}$ in this simulation, which is skewed with a long tail to the right. Let $v_i \sim \chi^2_{(4)}$. The expected value and variance of this random variable are $E(v_i) = 4$ and $\text{var}(v_i) = 8$, respectively, so that $z_i = (v_i - 4)/\sqrt{8}$ has mean zero and variance one. The random errors we employ are $e_i = 50z_i$ so that $\text{var}(e_i|x_i) = \sigma^2 = 2500$, as in earlier appendices.

As before, we use $M = 10{,}000$ Monte Carlo simulations, using the sample sizes $N = 20$, 40 (as before), 100, 200, 500, and 1000. Our objectives are to illustrate that the least squares estimators of β_1, β_2, and the estimator $\hat{\sigma}^2$ are unbiased, and to investigate whether hypothesis tests and interval estimates perform as they should, even though the errors are not normally distributed. As in Appendix 3C, we

- Test the null hypothesis $H_0 : \beta_2 = 10$ using the one-tail alternative $H_0 : \beta_2 > 10$. The critical value for the test is the 95th percentile of the t-distribution with $N - 2$ degrees of freedom, $t_{(0.95, N-2)}$. We report the percentage of rejections from this test (*REJECT*).
- Contruct a 95% interval estimate for β_2 and report the percentage of the estimates (*COVER*) that contain the true parameter, $\beta_2 = 10$.
- Compute the percentage of the time (*CLOSE*) that the estimates b_2 are in the interval $\beta_2 \pm 1$, or between 9 and 11. Based on our theory, this percentage should increase toward 1 as N increases.

The Monte Carlo simulation results are summarized in Table 5C.1.

The unbiasedness of the least squares estimators is verified by the average values of the estimates being very close to the true parameter values for all sample sizes. The percentage of estimates that are "close" to the true parameter value rises as the sample size N increases, verifying the consistency of the estimator. Because the rejection rates from the t-test are close to 0.05 and the coverage of the interval estimates is close to 95%, the approximate normality of the estimators is very good. To illustrate, in Figure 5C.1 we present the histogram of the estimates b_2 for $N = 40$.

TABLE 5C.1 **The Least Squares Estimators, Tests, and Interval Estimators**

N	$\overline{b}_1$	$\overline{b}_2$	$\overline{\hat{\sigma}^2}$	*REJECT*	*COVER*	*CLOSE*
20	99.4368	10.03317	2496.942	0.0512	0.9538	0.3505
40	100.0529	9.99295	2498.030	0.0524	0.9494	0.4824
100	99.7237	10.01928	2500.563	0.0518	0.9507	0.6890
200	99.8427	10.00905	2497.473	0.0521	0.9496	0.8442
500	100.0445	9.99649	2499.559	0.0464	0.9484	0.9746
1000	100.0237	9.99730	2498.028	0.0517	0.9465	0.9980

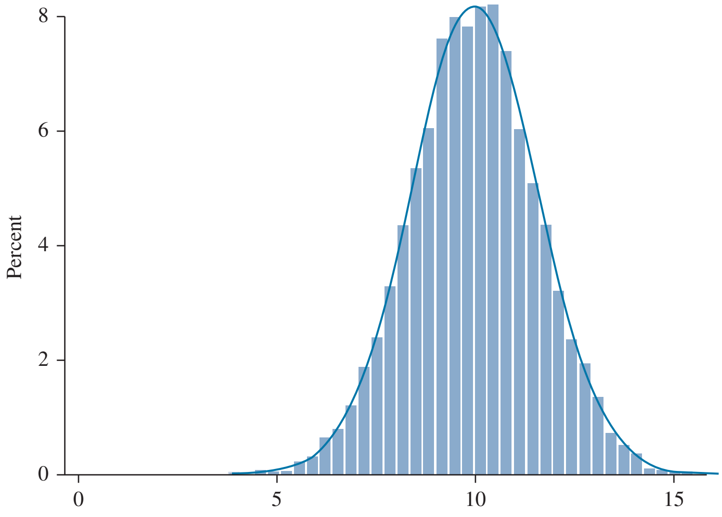

FIGURE 5C.1 **Histogram of the estimates b_2 for $N = 40$.**

It is very bell shaped, with the superimposed normal density function fitting it very well. The nonnormality of the errors does not invalidate inferences in this model, even with only $N = 40$ sample observations.

5C.2 Monte Carlo Simulation of the Delta Method

In this Monte Carlo simulation, again using 10,000 samples, we compute the value of the nonlinear function estimator $g_1(b_2) = \exp(b_2/10)$ for each sample, and we test the true null hypothesis $H_0: g_1(\beta_2) = \exp(\beta_2/10) = e^1 = 2.71828$ using a two-tail test at the 5% level of significance. We are interested in how well the estimator does in finite samples (recall that the random errors are not normally distributed and that the function is nonlinear), and how well the test performs. In Table 5C.2, we report the average of the parameter estimates for each sample size. Note that the mean estimate converges toward the true value as N becomes larger. The test at the 5% level of significance rejects the true null hypothesis about 5% of the time. The test statistic is

$$t = \frac{g_1(b_2) - 2.71828}{\mathrm{se}\left[g_1(b_2)\right]} \sim t_{(N-2)}$$

The fact that the t-test rejects the correct percentage of the time implies not only that the estimates are well behaved but also that the standard error in the denominator is correct, and that the distribution of the statistic is "close" to its limiting standard normal distribution. In Table 5C.2, $\overline{\mathrm{se}\left[\exp(b_2/10)\right]}$ is the average of the nominal standard errors calculated using the delta method, and std. dev. $\left[\exp(b_2/10)\right]$ is the standard deviation of the estimates that measures the actual, true variation in the Monte Carlo estimates. We see that for sample sizes $N = 20$ and $N = 40$, the average of the standard errors calculated using the delta method is smaller than the true standard deviation, meaning that on average, in this illustration, the delta method overstates the precision of the estimates $\exp(b_2/10)$. The average standard error calculated using the delta method is close to the true standard deviation for larger sample sizes. We are reminded that the delta method standard errors are valid in large samples, and in this illustration the sample size $N = 100$ seems adequate for the asymptotic result to hold. The histogram of the estimates for sample size $N = 40$ in Figure 5C.2 shows only the very slightest deviation from normality, which is why the t-test performs so well.

We now examine how well the delta method works at different sample sizes for estimating the function $g_2(\beta_1/\beta_2)$ and approximating its variance and asymptotic distribution. The mean estimates in Table 5C.3 show that there is some bias in the estimates for small samples sizes. However, the bias diminishes as the sample size increases and is close to the true value, 10, when $N = 100$. The average of the delta method standard errors, $\overline{\mathrm{se}(b_1/b_2)}$, is smaller than the actual, Monte Carlo, standard deviation of the estimates b_1/b_2 for sample sizes $N = 20$, 40, and 100. This illustrates the lesson that the more complicated the nonlinear function, or model, the larger the sample size that is required for asymptotic results to hold.

TABLE 5C.2 Simulation Results for $g_1(\beta_2) = \exp(\beta_2/10)$

N	$\overline{\exp(b_2/10)}$	$\overline{\mathrm{se}\left[\exp(b_2/10)\right]}$	Std. dev. $\left[\exp(b_2/10)\right]$	*REJECT*
20	2.79647	0.60738	0.63273	0.0556
40	2.75107	0.42828	0.44085	0.0541
100	2.73708	0.27208	0.27318	0.0485
200	2.72753	0.19219	0.19288	0.0503
500	2.72001	0.12148	0.12091	0.0522
1000	2.71894	0.08589	0.08712	0.0555

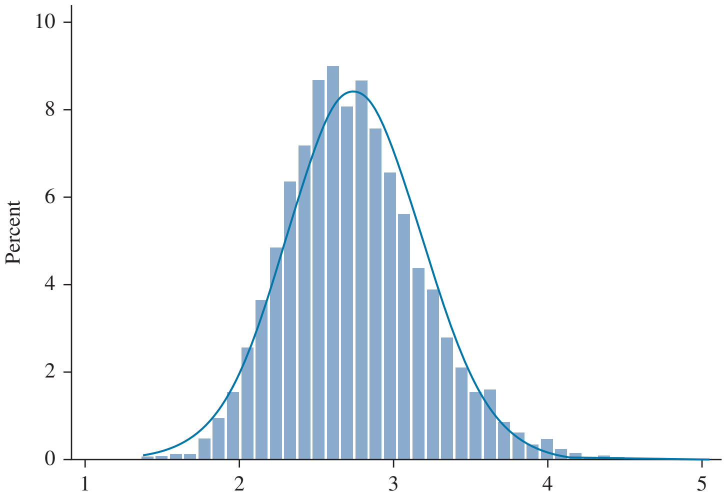

FIGURE 5C.2 Histogram of $g_1(b_2) = \exp(b_2/10)$.

TABLE 5C.3 Simulation Results for $g_2(b_1, b_2) = b_1/b_2$

N	$\overline{b_1/b_2}$	$\overline{\text{se}(b_1/b_2)}$	Std. dev. (b_1/b_2)
20	11.50533	7.18223	9.19427
40	10.71856	4.36064	4.71281
100	10.20997	2.60753	2.66815
200	10.10097	1.82085	1.82909
500	10.05755	1.14635	1.14123
1000	10.03070	0.80829	0.81664

The Monte Carlo simulated values of $g_2(b_1, b_2) = b_1/b_2$ are shown in Figures 5C.3(a) and (b) from the experiments with $N = 40$ and $N = 200$. With sample size $N = 40$, there is pronounced skewness. With $N = 200$, the distribution of the estimates is much more symmetric and bell shaped.

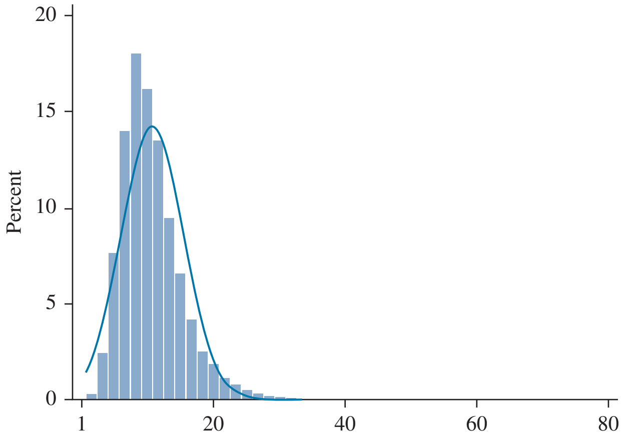

FIGURE 5C.3a Histogram of $g_2(b_1, b_2) = b_1/b_2$, $N = 40$.

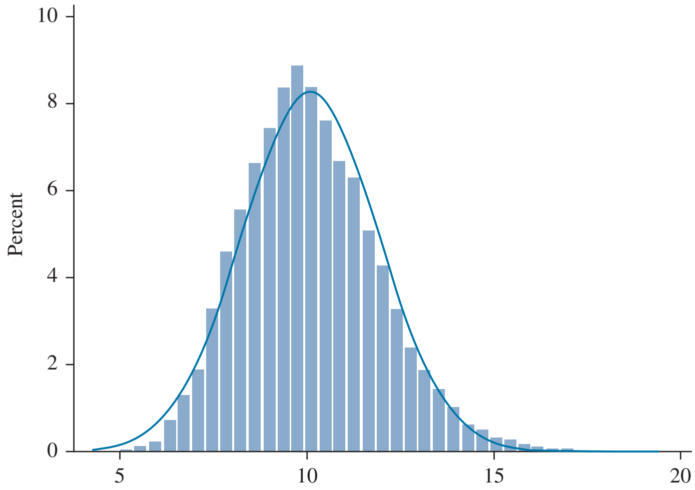

FIGURE 5C.3b **Histogram of $g_2(b_1, b_2) = b_1/b_2$, $N = 200$.**

Appendix 5D Bootstrapping

In Section 2.7.3, we discuss the interpretation of **standard errors** of estimators. Least squares estimates vary from sample to sample simply because the composition of the sample changes. This is called **sampling variability**. For the least squares estimators we have derived formulas for the variance of the least squares estimators. For example, in the simple regression model $y_i = \beta_1 + \beta_2 x_i + e_i$, the variance of the least squares estimator of the slope is $\text{var}(b_2|\mathbf{x}) = \sigma^2 / \sum(x_i - \bar{x})^2$ and the standard error is $\text{se}(b_2) = \left[\hat{\sigma}^2 / \sum(x_i - \bar{x})^2\right]^{1/2}$. We were able to derive this formula using the model assumptions and linear form of the least squares estimator.

However, there are estimators for whom no easy standard errors can be computed. The estimators may be based on complex multistep procedures, or they may be nonlinear functions. In many cases, we can show that the estimators are **consistent** and **asymptotically normal**. We discussed these properties in Section 5.7. For an estimator $\hat{\beta}$, these properties mean that $\hat{\beta} \overset{a}{\sim} N\left[\beta, \text{var}\left(\hat{\beta}\right)\right]$. In this expression, $\text{var}\left(\hat{\beta}\right)$ is an **asymptotic variance** that is appropriate in large samples. If the asymptotic variance is known, then the **nominal standard error**, that is valid in large samples, is $\text{se}\left(\hat{\beta}\right) = \left[\widehat{\text{var}}\left(\hat{\beta}\right)\right]^{1/2}$. Asymptotic variance formulas can be difficult to derive. We illustrated the **delta method**, in Appendices 5B and 5C.2, for finding asymptotic variances of nonlinear functions of the least squares estimators. Even in those simple cases, there are derivatives and tedious algebra.

The **bootstrap procedure** is an alternative and/or complement to the analytic derivation of asymptotic variances. **Bootstrapping** can be used to compute standard errors for complicated and nonlinear estimators. It uses the speed of modern computing and a technique called **resampling**. In this section, we explain the bootstrapping technique and several ways that it can be used. In particular, we can use bootstrapping to

1. Estimate the bias of the estimator $\hat{\beta}$.
2. Obtain a standard error $\text{se}\left(\hat{\beta}\right)$ that is valid in large samples.
3. Construct confidence intervals for β.
4. Find critical values for test statistics.

5D.1 Resampling

To illustrate resampling suppose we have N independent and identically distributed data pairs (y_i, x_i). This is the case if we collect random samples from a specific population.[18] To keep matters simple let $N = 5$. This is for illustration only. A hypothetical sample is given in Table 5D.1. Resampling means randomly select $N = 5$ rows **with replacement** to form a new sample. The phrase **with replacement** means that after randomly selecting one row, and adding it to a new data set, we return the selected row to the original data where it might be randomly selected again, or not.

Perhaps seeing an algorithm for doing this will help. It begins with the concept of a **uniform random number** on the zero to one interval, $u \sim \text{uniform}(0,1)$. Uniform random numbers are a core part of numerical methods for simulations. We discuss them in Appendix B.4.1. Roughly speaking, the uniformly distributed random value u is equally likely to take any value in the interval (0,1). Computer scientists have designed algorithms so that repeated draws using a **uniform random number generator** are independent of one another. These are built into every econometric software package, although the algorithms used may vary slightly from one to the next. To randomly pick a row of data,

1. Let $u^* = (5 \times u) + 1$. This value is greater than 1 but less than 6.
2. Drop the decimal portion to obtain a random integer b that is 1, 2, 3, 4, or 5.

Table 5D.2 illustrates the process for $N = 5$. These steps are automated by many software packages, so you will not have to do the programming yourself, but it is a good idea to know what is happening. The values j in Table 5D.2 are the rows from the original data set that will constitute the first **bootstrap sample**. The first bootstrap sample will contain observations 5, 1, 2, and the third observation twice, as shown in Table 5D.3.[19] This is perfectly OK. Resampling means that

TABLE 5D.1 The Sample

Observation	y	x
1	$y_1 = 6$	$x_1 = 0$
2	$y_2 = 2$	$x_2 = 1$
3	$y_3 = 3$	$x_3 = 2$
4	$y_4 = 1$	$x_4 = 3$
5	$y_5 = 0$	$x_5 = 4$

TABLE 5D.2 Random Integers

u	u^*	j
0.9120440	5.56022	5
0.0075452	1.037726	1
0.2808588	2.404294	2
0.4602787	3.301394	3
0.5601059	3.800529	3

[18] Bootstrap techniques for time-series data are much different, and we will not discuss them here.

[19] Random number generators use a “starting value,” called a **seed**. By choosing a seed the same sequence of random numbers can be obtained in subsequent runs. See Appendix B.4.1 for a discussion of how one class of random number generators work.

TABLE 5D.3 One Bootstrap Sample

Observation	y	x
5	$y_5 = 0$	$x_5 = 4$
1	$y_1 = 6$	$x_1 = 0$
2	$y_2 = 2$	$x_2 = 1$
3	$y_3 = 3$	$x_3 = 2$
3	$y_3 = 3$	$x_3 = 2$

some observations will be chosen multiple times, and others (such as observation 4 in this case) will not appear at all.

5D.2 Bootstrap Bias Estimate

The estimator $\hat{\beta}$ may be a biased estimator. Estimator bias is the difference between the estimator's expected value and the true parameter, or

$$\text{bias}\left(\hat{\beta}\right) = E\left(\hat{\beta}\right) - \beta$$

For a consistent estimator the bias disappears as $N \to \infty$, but we can estimate the bias given a sample of size N. Using the process described in the previous section, obtain bootstrap samples $b = 1, 2, \ldots, B$, each of size N. Using each bootstrap sample obtain an estimate $\hat{\beta}_b$. If $B = 200$, then we have 200 bootstrap sample estimates $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_{200}$. The average, or sample mean, of the B bootstrap sample estimates is

$$\bar{\hat{\beta}} = \frac{1}{B}\sum_{b=1}^{B}\hat{\beta}_b$$

The bootstrap estimate of the bias is

$$\text{bootstrap }\widehat{\text{bias}}\left(\hat{\beta}\right) = \bar{\hat{\beta}} - \hat{\beta}_O$$

where $\hat{\beta}_O$ is the estimate obtained using the original sample [the subscript is "oh" and not zero]. In this calculation, $\bar{\hat{\beta}}$ plays the role of $E\left(\hat{\beta}\right)$ and $\hat{\beta}_O$, the estimate from the original sample, plays the role of the true parameter β. A descriptive saying about bootstrapping is that that "$\hat{\beta}_O$ is true in the sample," emphasizing the role played by the original sample estimate, $\hat{\beta}_O$.

5D.3 Bootstrap Standard Error

Bootstrap standard error calculation requires B bootstrap samples of size N. For the purpose of computing standard errors, the number of bootstrap samples should be at least 50, and perhaps 200 or 400, depending on the complexity of your estimation problem.[20] The bootstrap standard error is the **sample standard deviation** of the B bootstrap estimates. The sample standard deviation is the square root of the sample variance. The bootstrap estimate of $\text{var}\left(\hat{\beta}\right)$ is the sample variance of the bootstrap estimates $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_B$,

$$\text{bootstrap var}\left(\hat{\beta}\right) = \sum_{b=1}^{B}\left(\hat{\beta}_b - \bar{\hat{\beta}}\right)^2\Big/(B-1)$$

[20] Try a number of bootstraps B. For standard errors $B = 200$ is a good starting value. Compute the bootstrap standard error. Change the random number seed a few times. If the bootstrap standard error changes little, then B is large enough. If there are substantial changes, increase B.

The bootstrap standard error is

$$\text{bootstrap se}\left(\hat{\beta}\right) = \sqrt{\text{bootstrap var}\left(\hat{\beta}\right)} = \sqrt{\sum_{b=1}^{B}\left(\hat{\beta}_b - \bar{\hat{\beta}}\right)^2 \Big/ (B-1)}$$

In large samples, the bootstrap standard error is no better, or worse, than the theoretically derived standard error. The advantage of the bootstrap standard error is that we need not derive the theoretical standard error, which can sometimes be very difficult. Even if the theoretical standard error can be obtained, the bootstrap standard error can be used as a check of the estimate based on a theoretical formula. If the bootstrap standard error is considerably different from the theory-based standard error, then either (i) the sample size N is not large enough to justify asymptotic theory, or (ii) the theoretical formula has an error. The theoretical standard error could be wrong if one of the model assumptions does not hold, or there is a math error, or there is an error in the software calculating the estimate based on the theoretical standard error (yes, that sometimes happens).

We can use the bootstrap standard error the same way as the usual standard error. An asymptotically justified $100(1-\alpha)\%$ interval estimator of β is

$$\hat{\beta} \pm t_c\left[\text{bootstrap se}\left(\hat{\beta}\right)\right]$$

where t_c is the $1-\alpha/2$ percentile of the t-distribution. In large samples, using $t_c = 1.96$ leads to a 95% interval estimate. This is sometimes called the **normal-based bootstrap confidence interval**.

For testing the null hypothesis $H_0 : \beta = c$ against $H_1 : \beta \neq c$, a valid test statistic is

$$t = \frac{\hat{\beta} - c}{\text{bootstrap se}\left(\hat{\beta}\right)}$$

If the null hypothesis is true, the test statistic has a standard normal distribution[21] in large samples. At the 5% level, we reject the null hypothesis if $t \geq 1.96$ or $t \leq -1.96$.

5D.4 Bootstrap Percentile Interval Estimate

A **percentile interval estimate**, or **percentile confidence interval**, does not use the approximate large sample normality of an estimator. Recall that in the simple regression model a 95% interval estimator is obtained from equation (3.5), which is

$$P\left[b_k - t_c\text{se}\left(b_k\right) \leq \beta_k \leq b_k + t_c\text{se}\left(b_k\right)\right] = 1 - \alpha$$

where $t_c = t_{(0.975,\, N-K)}$. The interval estimator $\left[b_k - t_c\text{se}\left(b_k\right), b_k + t_c\text{se}\left(b_k\right)\right]$ will contain the true parameter β_k in 95% of **repeated samples** from the same population. Another descriptive phrase used when discussing bootstrapping is that we "treat the sample as the population." This makes the point that by using bootstrapping, we are trying to learn about an estimator's **sampling properties**; or how the estimator performs in repeated samples. Bootstrapping treats each bootstrap sample as a "repeated sample." Using this logic, if we obtain many bootstrap samples, and many estimates (sorting the B bootstrap estimates from smallest to largest) a 95% percentile interval estimate is $\left[\hat{\beta}^*_{(0.025)},\ \hat{\beta}^*_{(0.975)}\right]$ where $\hat{\beta}^*_{(0.025)}$ is the 2.5%-percentile of the B bootstrap estimates, and $\hat{\beta}^*_{(0.975)}$ is the 97.5%-percentile of the B bootstrap estimates. Because of the way software programmers find percentiles, it is useful to choose B such that $\alpha(B+1)$ is a convenient integer. If $B = 999$, then the 2.5%-percentile is the 25th value and the 97.5%-percentile is the 975th value. If $B = 1999$, then the 2.5%-percentile is the 50th value and the 97.5%-percentile is the 1950th value. Calculating percentile interval estimates requires a larger number of bootstrap samples than calculating a standard error. Intervals calculated this way are not necessarily symmetrical.

[21] Because of its large sample justification, some software packages will call this statistic "z."

5D.5 Asymptotic Refinement

If it is possible to derive a theoretical expression for the variance of an estimator that is valid in large samples, then we can combine it with bootstrapping to improve upon standard asymptotic theory. Asymptotic refinement produces a test statistic critical value that leads to more accurate tests. What do we mean by that? A test of $H_0 : \beta = c$ against $H_1 : \beta \neq c$ uses an asymptotically valid nominal standard error and the t-statistic $t = \left(\hat{\beta} - c\right)/\text{se}\left(\hat{\beta}\right)$. If $\alpha = 0.05$, we reject the null hypothesis if $t \geq 1.96$ or $t \leq -1.96$. This test is called a **symmetrical two-tail test**. In finite (small) samples, the actual rejection probability is not $\alpha = 0.05$ but $P(\text{reject } H_0|H_0 \text{ is } true) = \alpha + error$. The *error* goes to zero as the sample size N approaches infinity. More precisely, $N \times error \leq N^*$ where N^* is some upper bound. In order for this to be true, as $N \to \infty$ the *error* must approach zero, $error \to 0$. Not only must $error \to 0$, but also it must approach zero at the same rate as $N \to \infty$, so that the two effects are offsetting, with product $N \times error$ staying a finite number. This is called convergence to zero at rate "N." Using a bootstrap critical value, t_c^*, instead of 1.96 it can be shown that $N^2 \times error \leq N^*$, so that the test size *error* converges to zero at rate N^2. We have a more accurate test because the *error* in the test size goes to zero faster using the bootstrap critical value.

The gain in accuracy is "easy" to obtain. Resample the data B times. In each bootstrap sample, compute

$$t_b = \frac{\hat{\beta}_b - \hat{\beta}_O}{\text{se}\left(\hat{\beta}_b\right)}$$

In this expression, $\hat{\beta}_b$ is the estimate in the bth bootstrap sample, $\hat{\beta}_O$ is the estimate based on the original sample, and $\text{se}\left(\hat{\beta}_b\right)$ is the nominal standard error, the usual theory-based standard error, calculated using the bth bootstrap sample. This is the bootstrap equivalent of equation (3.3). To find the bootstrap critical value t_c^* (i) compute $|t_b|$, (ii) sort them in ascending magnitude, then (iii) t_c^* is the $100(1 - \alpha)$-percentile of $|t_b|$. To test $H_0 : \beta = c$ against $H_1 : \beta \neq c$ use the t-statistic $t = \left(\hat{\beta} - c\right)/\text{se}\left(\hat{\beta}\right)$ computed with the original sample, and reject the null hypothesis if $t \geq t_c^*$ or $t \leq -t_c^*$. The $100(1 - \alpha)\%$ interval estimate $\hat{\beta} \pm t_c^* \text{se}\left(\hat{\beta}\right)$ is sometimes called a **percentile-*t*** interval estimate.

For a right-tail test, $H_0 : \beta \leq c$ against $H_1 : \beta > c$, t_c^* is the $100(1 - \alpha)$-percentile of t_b, dropping the absolute value operation. Reject the null hypothesis if $t \geq t_c^*$. For a left-tail test, $H_0 : \beta \geq c$ against $H_1 : \beta < c$, t_c^* is the 100α-percentile of t_b. Reject the null hypothesis if $t \leq t_c^*$.

EXAMPLE 5.21 | Bootstrapping for Nonlinear Functions $g_1\left(\beta_2\right) = \exp\left(\beta_2/10\right)$ and $g_2\left(\beta_1, \beta_2\right) = \beta_1/\beta_2$.

Clearly it is time for an example! Using the same Monte Carlo design as in Appendix 5C, we create one sample for $N = 20$, 40, 100, 200, 500, and 1000. They are in the data files *mc20*, *mc40*, *mc100*, *mc200*, *mc500*, and *mc1000*.

First we explore bootstrapping $g_1\left(\beta_2\right) = \exp\left(\beta_2/10\right)$. Table 5D.4a contains the estimates, delta method standard error, and an asymptotically justified 95% interval estimate

$$\exp\left(b_2/10\right) \pm \left\{1.96 \times \text{se}\left[\exp\left(b_2/10\right)\right]\right\}$$

Compare these to Table 5C.2 containing the Monte Carlo *averages* of the estimates, the nominal (delta method) standard errors, and the standard deviation of the estimates.

Because we will calculate percentile interval estimates and a bootstrap critical value, we use $B = 1999$ bootstrap samples as the basis for the estimates in Table 5D.4b. The bootstrap estimates of the bias diminish as the sample size increases, reflecting the consistency of the estimator. The bootstrap standard errors for $N = 20$, 40, and 100 are quite similar to the delta method standard errors for these sample sizes shown in Table 5D.4a. They are not as similar to the Monte Carlo average nominal standard error and standard deviation in Table 5C.2. However, once the sample size is $N = 200$ or more, the bootstrap standard errors are much closer to the results in Table 5C.2. In Table 5D.4b, we also

TABLE 5D.4a Delta Method $g_1(\beta_2) = \exp(\beta_2/10) = 2.71828$

N	$g_1(b_2) = \exp(b_2/10)$	$\text{se}[\exp(b_2/10)]$	95% Interval
20	2.91088	0.62367	[1.6885, 4.1332]
40	2.34835	0.37781	[1.6079, 3.0888]
100	2.98826	0.30302	[2.3945, 3.5822]
200	2.86925	0.20542	[2.4666, 3.2719]
500	2.63223	0.11241	[2.4119, 2.8526]
1000	2.78455	0.08422	[2.6195, 2.9496]

TABLE 5D.4b Bootstrapping $g_1(\beta_2) = \exp(\beta_2/10)$

N	Bootstrap Bias	Bootstrap se	PI	t_c^*
20	0.0683	0.6516	[2.0098, 4.5042]	3.0063
40	0.0271	0.3796	[1.7346, 3.2173]	2.2236
100	0.0091	0.3050	[2.4092, 3.6212]	2.0522
200	0.0120	0.2039	[2.4972, 3.3073]	1.9316
500	−0.0001	0.1130	[2.4080, 2.8567]	2.0161
1000	0.0025	0.0844	[2.6233, 2.9593]	1.9577

report the 95% **percentile interval** (PI) **estimate** for each sample size. Finally, we report the asymptotically refined critical value that would be used for a symmetrical two-tail test at the 5% level of significance, or when constructing a confidence interval. Based on these values, we judge that sample sizes $N = 20$ and 40 are not really sufficiently large to support asymptotic inferences in our specific samples, but if we do proceed, then the usual critical value 1.96 should not be used for t-tests or interval estimates. For sample sizes $N = 100$ or more, it appears that usual asymptotic procedures can be justified.

Table 5D.5 contains similar results for the function $g_2(\beta_1, \beta_2) = \beta_1/\beta_2$. The estimates, bootstrap bias, delta method standard error, and bootstrap standard error tell a similar story. For this nonlinear function, a ratio of two parameters, $N = 200$ or more would make us feel better about asymptotic inference. It is reassuring when the bootstrap and delta method standard errors are similar, although these are somewhat smaller than the average nominal standard error and standard deviations in Table 5C.3. Expressions containing ratios of parameters in one form or another often require larger samples for asymptotic inference to hold.

TABLE 5D.5 Bootstrapping $g_2(\beta_1, \beta_2) = \beta_1/\beta_2$

N	$g_2(b_1, b_2) = b_1/b_2$	Bootstrap Bias	$\text{se}(b_1/b_2)$	Bootstrap se
20	8.18406	0.7932	4.75590	4.4423
40	13.15905	1.0588	5.38959	6.0370
100	7.59037	0.2652	2.14324	2.3664
200	8.71779	0.0714	1.64641	1.6624
500	10.74195	0.0825	1.15712	1.2180
1000	9.44545	0.0120	0.73691	0.7412