Estimating Demand Functions

CHAPTER

5

When Sergio Zyman was the marketing chief of Coca-Cola, he once indicated that his company, one of the world's biggest advertisers, would put less emphasis on traditional newspaper, magazine, and TV ads and more emphasis on new marketing techniques like special programs on cable TV and product tie-ins with movies. All firms, not just Coke, must constantly reevaluate and adjust their marketing strategies. As stressed repeatedly in previous chapters, an effective manager must have a good working knowledge of the demand function for his or her firm's products.

The previous two chapters were concerned with the theory of demand; now we learn how to estimate a product's demand function. Consumer surveys and market experiments can be useful in providing such information, but the technique most frequently used to estimate demand functions is regression analysis.

While managers use some or all of these techniques (we mentioned the use of focus groups by Dell Computer in Chapter 4), the technique most frequently used to estimate demand functions is regression analysis (even much of the data gathered by questionnaire and focus group is analyzed by regression). In Chapter 3, we showed how Amtrak estimated its demand function with regression analysis. Since regression analysis is used repeatedly in subsequent chapters to

estimate production functions and cost functions and for forecasting, we devote considerable attention to this basic technique in this chapter.

The Identification Problem

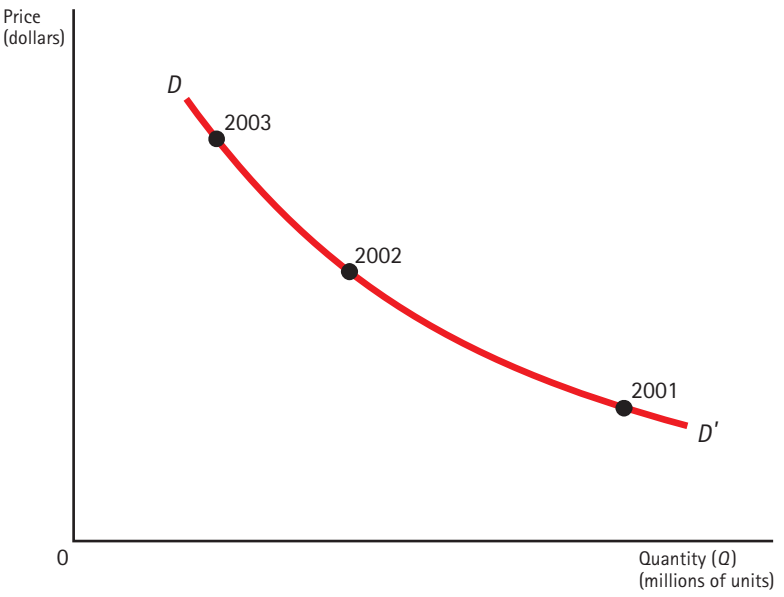
While it is very important that managers have reasonably accurate estimates of the demand functions for their own (and other) products, this does not mean that it is always easy to obtain such estimates. One problem that may arise in estimating demand curves should be recognized at the outset. Given the task of estimating the demand curve for a particular product, you might be inclined to plot the quantity demanded of the product in 2003 versus its 2003 price, the quantity demanded in 2002 versus its 2002 price, and so forth. If the resulting plot of points for 2001 to 2003 were as shown in Figure 5.1, you might be tempted to conclude that the demand curve is *DD′*.

Unfortunately, things are not so simple. Price, as we saw in Chapter 1, is determined by both the demand and supply curves for this product if the

FIGURE 5.1

Price Plotted against Quantity, 2001–2003

The curve *DD′* is unlikely to be a good estimate of the demand curve.



market is competitive. Specifically, the equilibrium value of price is at the level where the demand and supply curves intersect. The important point to note is that the demand and supply curves for this product may have been different each year. So, as shown in Figure 5.2, the supply curve may have shifted (from S_{01} in 2001 to S_{02} in 2002 to S_{03} in 2003), and the demand curve may have shifted (from D_{01} in 2001 to D_{02} in 2002 to D_{03} in 2003). As indicated in Figure 5.2, DD' is not even close to being a good approximation to the demand curve for this product in any of these three years.

In the situation in Figure 5.2, if you were to conclude that DD' was the demand curve, you would underestimate (in absolute value) the price elasticity of demand for this product in 2003 and 2002 and overestimate it (in absolute value) in 2001. In 2003, you would think that, if price were lowered from $30 to $28, the quantity demanded would increase from 10 to 12 million units per year. In fact, as shown in Figure 5.2, such a price reduction would result in an increase of the quantity demanded to 18, not 12, million units per year. This is a mammoth error in anyone's book.

FIGURE 5.2

Estimated Demand Curve Contrasted with Actual Demand Curves

The estimated demand curve DD' is not at all similar to the actual demand curves.

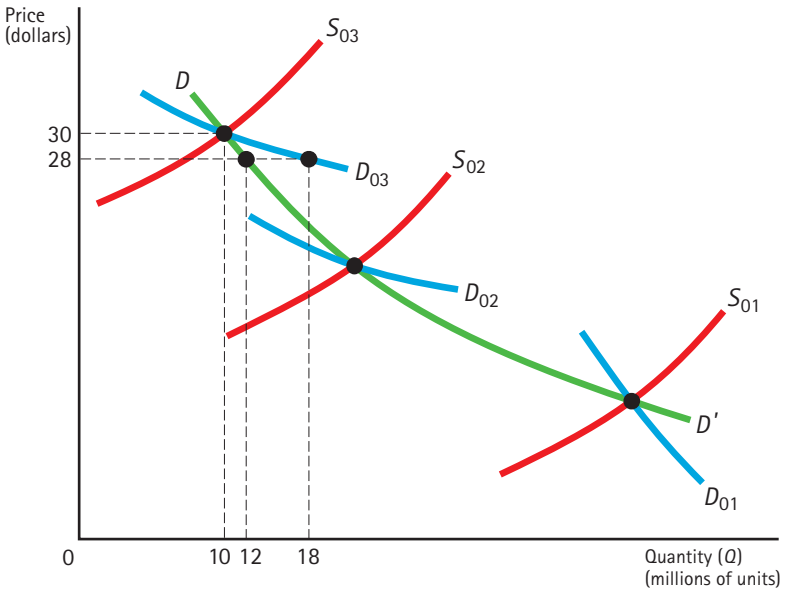
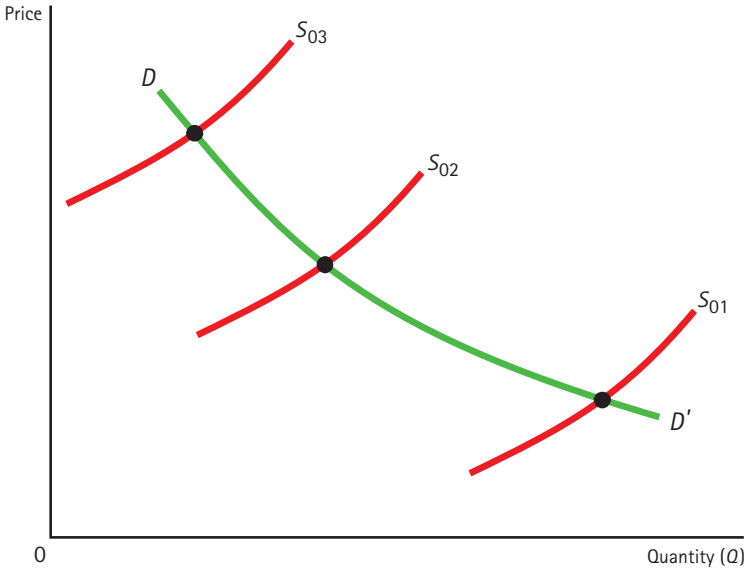


FIGURE 5.3

Fixed Demand Curve and Shifting Supply Curve

In this special case, *DD'* does represent the actual demand curve.



The point is that, because we are not holding constant a variety of nonprice variables like consumer tastes, incomes, the prices of other goods, and advertising, we cannot be sure that the demand curve was fixed during the period when the measurements were made. If the demand curve was fixed and only the supply curve changed during the period, we could be confident that the plot of points in Figure 5.1 represents the demand curve. As shown in Figure 5.3, the shifts in the supply curve trace out various points on the demand curve we want to measure.

How can we estimate a demand curve if it has not remained fixed in the past? There are many ways, some simple, some very complex. Econometric techniques recognize that price and quantity are related by both the supply curve and the demand curve and both these curves shift in response to nonprice variables. Some basic econometric techniques, such as regression analysis, are presented later in this chapter; others are too complex to be taken up here.[1]

[1]See J. Johnston, *Econometric Methods* (3d ed.; New York: McGraw-Hill, 1984); J. Kmenta, *Elements of Econometrics* (2d ed.; New York: Macmillan Co., 1986); or E. Berndt, *The Practice of Econometrics* (Reading, MA: Addison-Wesley, 1991).

Consumer interviews and market experiments are also widely used, as indicated in the next three sections.

Consumer Interviews

To obtain information concerning the demand function for a particular product, firms frequently interview consumers and administer questionnaires concerning their buying habits, motives, and intentions. Firms may also run focus groups in an attempt to discern consumers' tastes. For example, a firm might ask a random sample of consumers how much more gasoline they would purchase if its price were reduced by 5 percent. Or, a market researcher might ask a sample of consumers whether they liked a new type of perfume better than a leading existing brand, and if so, how much more they would be willing to pay for it (than for the existing brand).

Unfortunately, consumer surveys of this sort have many well-known limitations. The direct approach of simply asking people how much they would buy of a particular commodity at particular prices often does not seem to work very well. Frequently, the answers provided by consumers to such a hypothetical question are not very accurate. However, more subtle approaches can be useful. Interviews indicated that most buyers of a particular baby food selected it on their doctor's recommendation and that most of them knew very little about prices of substitutes. This information, together with other data, suggested that the price elasticity of demand was quite low in absolute value.[2]

Despite the limitations of consumer interviews and questionnaires, many managers believe that such surveys can reveal a great deal about how their firms can serve the market better. For example, the Campbell Soup Company's researchers contacted close to 110,000 people to talk about the taste, preparation, and nutritional value of food. On the basis of these interviews, Campbell changed the seasonings in five Le Menu dinners and introduced a line of low-salt soups (called Special Request). Some of the factors influencing the quality of survey results can be quite subtle. For example, according to research findings, there are sometimes advantages in respondents' keypunching answers, rather than verbalizing them, because the respondents tend to answer emotional questions more honestly this way.[3]

[2]J. Dean, "Estimating the Price Elasticity of Demand," in E. Mansfield, ed., *Managerial Economics and Operations Research* (4th ed.; New York: Norton, 1980).

[3]*New York Times*, November 8, 1987, p. 4F. Also, see W. Baumol, "The Empirical Determination of Demand Relationships," in *Managerial Economics and Operations Research*, ed. Mansfield.

Market Experiments

Another method of estimating the demand curve for a particular commodity is to carry out direct market experiments. The idea is to vary the price of the product while attempting to keep other market conditions fairly stable (or to take changes in other market conditions into account). For example, a manufacturer of ink conducted an experiment some years ago to determine the price elasticity of demand for its product. It raised the price from 15 cents to 25 cents in four cities and found that demand was quite inelastic. Attempts were made to estimate the cross elasticity of demand with other brands as well.

Controlled laboratory experiments can sometimes be carried out. Consumers are given money and told to shop in a simulated store. The experimenter can vary the prices, packaging, and location of particular products, and see the effects on the consumers' purchasing decisions. While this technique is useful, it suffers from the fact that consumers participating in such an experiment know that their actions are being monitored. For that reason, their behavior may depart from what it normally would be.

Before carrying out a market experiment, weigh the costs against the benefits. Direct experimentation can be expensive or risky because customers may be lost and profits cut by the experiment. For example, if the price of a product is raised as part of an experiment, potential buyers may be driven away. Also, since they are seldom really controlled experiments and since they are often of relatively brief duration and the number of observations is small, experiments often cannot produce all the information that is needed. Nonetheless, market experiments can be of considerable value, as illustrated by the following actual case.

L'eggs: A Market Experiment

L'eggs Products, a subsidiary of the Hanes Corporation, markets L'eggs Pantyhose, the first major nationally branded and advertised hosiery product distributed through food and drug outlets. According to some estimates, it has been the largest-selling single brand in the hosiery industry. Jack Ward, group product manager of the firm, was interested in determining the effect on sales of four temporary promotion alternatives: a 40-cent price reduction for a package containing two pairs, a 25-cent price reduction for a package containing two pairs, a 20-cent price reduction per pair, and a coupon mailed to homes worth 25 cents off if a pair was purchased.[4]

[4]The material in this section is based on F. DeBruicker, J. Quelch, and S. Ward, *Cases in Consumer Behavior* (2d ed.; Englewood Cliffs, NJ: Prentice-Hall, 1986).

To test these four promotion alternatives, Jerry Clawson, director of marketing research, decided that each would be implemented in a carefully chosen test market, and the results would be compared with another market where no unusual promotion was carried out. Specifically, there was a 40-cent reduction (for two pairs) in Syracuse, New York; a 25-cent reduction (for two pairs) in Columbus, Ohio; a 20-cent reduction (for one pair) in Denver, Colorado; and a 25-cent coupon in Cincinnati, Ohio. The results in these markets were compared with those in Boise, Idaho, where no special promotion occurred.

According to the firm's sales research group, the results were as follows: "The two for 40¢-off promotion (Syracuse) was the most effective with a net short-term cumulative increase in sales of 53 percent felt over six weeks. The 20¢ price-off promotion (Denver) was the second most effective, with a net cumulative short-term increase of 20 percent felt over eight weeks. . . . The 25¢ coupon promotion (Cincinnati) was the least effective with a 3 percent short-term increase in sales felt over eight weeks."[5]

This is an example of how firms go about obtaining information concerning their market demand functions. In this case, the firm's managers were interested in the effects of both the form and size of the price cut, and they were concerned only with a temporary price cut. In other cases, firms are interested in the effects of more long-term price changes or of changes in product characteristics or advertising. But, regardless of these differences, marketing research of this sort can play an important role in providing data for the estimation of demand functions.

Regression Analysis

Although consumer interviews and direct market experiments are important sources of information concerning demand functions, they are not used as often as regression analysis. Suppose that a firm's demand function is

$$Y = A + B_1X + B_2P + B_3I + B_4P_r \tag{5.1}$$

where Y is the quantity demanded of the firm's product, X is the selling expense (such as advertising) of the firm, P is the price of its product, I is the disposable income of consumers, and P_r is the price of the competing product sold by its rival. What we want are estimates of the values of A, B_1, B_2, B_3, and B_4. Regression analysis enables us to obtain them from historical data concerning Y, X, P, I, and P_r.

[5]Ibid., p. 335. The validity of these results is discussed there also.

CONSULTANT'S CORNER

Marketing Plans at the Stafford Company

The Stafford Company developed a new type of electric drive. When the design engineering for this machine was finished, Stafford's managers began to make long-range plans concerning marketing this product. By means of field surveys and the analysis of published information, the firm's market research personnel estimated that about 10,000 electric drives of this general sort would be sold per year. The share of the total market that Stafford's new product would capture depended on its price. According to the firm's market research department, the relationship between price and market share was as follows:

Price	Market share
$ 800	11.0
900	10.2
1,000	9.2
1,100	8.4
1,200	7.5
1,300	6.6
1,400	5.6

Stafford's managers wanted advice in setting the price for their new drive, and to help determine the optimal price, they wanted a simple equation expressing the annual quantity demanded of the new product as a function of its price. They also wanted whatever information could readily be provided concerning the reliability of this equation. In particular, they were interested in whether they could safely use this equation to estimate the quantity demanded if price were set at $1,500 or $1,600.

Prepare a brief report supplying the information requested. (Note that the figures on market share in the table are expressed in percentage points. Thus, if the price of Stafford's new product is set at $800, it will capture 11.0 percent of the market for electric drives of this general sort, according to the market research department.)

Source: This section is based on an actual case, although the numbers and situation are disguised somewhat.

In the rest of this chapter, we describe the nature and application of regression analysis, a statistical technique that can be used to estimate many types of economic relationships, not just demand functions. We begin with the simple case in which the only factor influencing the quantity demanded is the firm's selling expense, then turn to the more complicated (and realistic) case in which the quantity demanded is affected by more than one factor, as it is in equation (5.1).

Regression analysis describes the way in which one variable is related to another. (As we see later in this chapter, regression techniques can handle more than two variables, but only two are considered at present.) Regression analysis derives an equation that can be used to estimate the unknown value of one variable on the basis of the known value of the other variable. For example, suppose that the Miller Pharmaceutical Company is scheduled to spend

TABLE 5.1

Selling Expense and Sales, Miller Pharmaceutical Company, Sample of Nine Years

Selling expense (millions of dollars)	Sales (millions of units)
1	4
2	6
4	8
8	14
6	12
5	10
8	16
9	16
7	12

$4 million next year on selling expense (for promotion, advertising, and related marketing activities) and it wants to estimate its next-year's sales, on the basis of the data in Table 5.1 regarding its sales and selling expense in the previous nine years. In this case, although the firm's selling expense next year is known, its next year's sales are unknown. Regression analysis describes the way in which the firm's sales are historically related to its selling expense.

Simple Regression Model

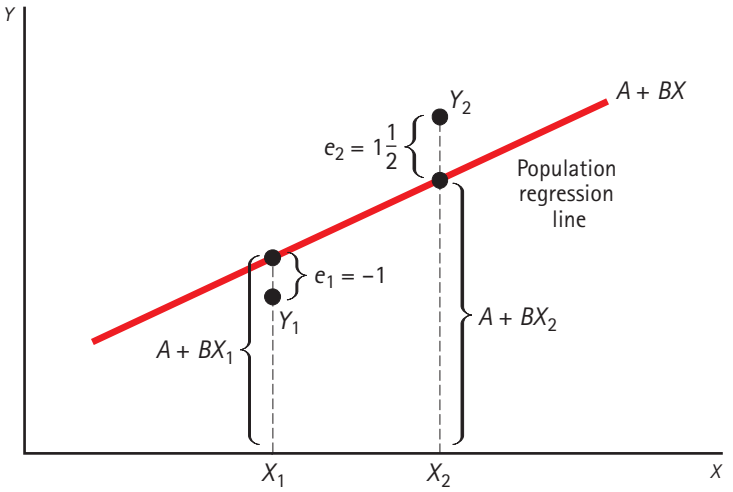
As you recall from Chapter 1, a **model** is a simplified or idealized representation of the real world. In this section, we describe the model—that is, the set of simplifying assumptions—on which regression analysis is based. We begin by visualizing a population of all relevant pairs of observations of the independent and dependent variables. For instance, in the case of the Miller Pharmaceutical Company, we visualize a population of pairs of observations concerning sales and selling expense. This population includes all the levels of sales corresponding to all the levels of selling expense in the history of the firm.

The mean of a variable equals the sum of its values divided by their number. Therefore, the mean of a variable that assumes four values, 3, 2, 1, and 0, is $(3 + 2 + 1 + 0)/4$, or 1.5. Regression analysis assumes that *the mean value of Y, given the value of X, is a linear function of X*. In other words, the mean value of the dependent variable is assumed to be a linear function of the independent variable, the equation of this being $A + BX$, as shown in Figure 5.4.

FIGURE 5.4

Regression Model

The mean value of Y, given the value of X, falls on the population regression line.



This straight line is called the **population regression line** or the **true regression line.**

Put differently, regression analysis assumes that

$$Y_i = A + BX_i + e_i \tag{5.2}$$

where Y_i is the ith observed value of the dependent variable and X_i is the ith observed value of the independent variable. Essentially, e_i is an **error term,** that is, a random amount that is added to $A + BX_i$ (or subtracted from it if e_i is negative). Because of the presence of this error term, the observed values of Y_i fall around the population regression line, not on it. Hence, as shown in Figure 5.4, if e_1 (the value of the error term for the first observation) is -1, Y_1 lies 1 below the population regression line. And if e_2 (the value of the error term for the second observation) is $+1.50$, Y_2 lies 1.50 above the population regression line. Regression analysis assumes that the values of e_i are independent and their mean value equals zero.[6]

[6]The values of e_1 and e_2 are independent if the probability distribution of e_1 does not depend on the value of e_2 and the probability distribution of e_2 does not depend on the value of e_1. Regression analysis also assumes that the variability of the values of e_i is the same, regardless of the value of X. Many of the tests described subsequently assume too that the values of e_i are normally distributed. For a description of the normal distribution, see Appendix B.

Although the assumptions underlying regression analysis are unlikely to be met completely, they are close enough to the truth in a sufficiently large number of cases that regression analysis is a powerful technique. Nonetheless, it is important to recognize at the start that, if these assumptions are not at least approximately valid, the results of a regression analysis can be misleading.

Sample Regression Line

The purpose of a regression analysis is to obtain the mathematical equation for a line that describes the average relationship between the dependent and independent variables. This line is calculated from the sample observations and is called the **sample** or **estimated regression line**. It should not be confused with the population regression line discussed in the previous section. Whereas the population regression line is based on the entire population, the sample regression line is based on only the sample.

The general expression for the sample regression line is

$$\hat{Y} = a + bX$$

where $\hat{Y}$ is the value of the dependent variable predicted by the regression line, and a and b are estimators of A and B, respectively. (An estimator is a function of the sample observations used to estimate an unknown parameter. For example, the sample mean is an estimator often used to estimate the population mean.) Since this equation implies that $\hat{Y} = a$ when $X = 0$, it follows that a is the value of $\hat{Y}$ at which the line intersects the Y axis. Therefore, a is often called the Y **intercept** of the regression line. And b, which clearly is the slope of the line, measures the change in the predicted value of Y associated with a one-unit increase in X.

Figure 5.5 shows the estimated regression line for the data concerning sales and selling expense of the Miller Pharmaceutical Company. The equation for this regression line is

$$\hat{Y} = 2.536 + 1.504X$$

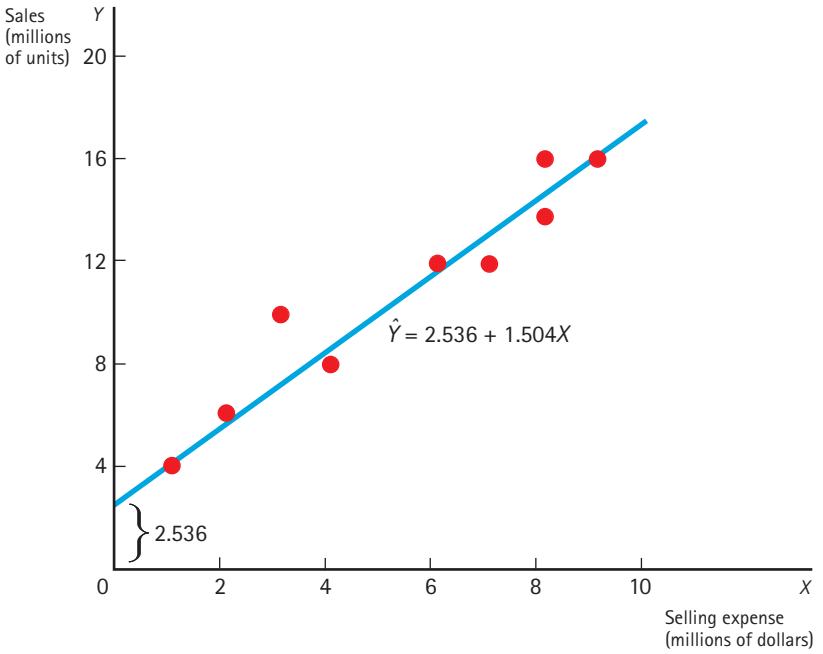
where $\hat{Y}$ is sales in millions of units and X is selling expense in millions of dollars. What is 2.536? It is the value of a, the estimator of A. What is 1.504? It is the value of b, the estimator of B. For the moment, we are not interested in how this equation was determined; what we want to consider is how it should be interpreted.

At the outset, note the difference between Y and $\hat{Y}$. Whereas Y denotes an *observed* value of sales, $\hat{Y}$ denotes the *computed* or *estimated* value of sales,

FIGURE 5.5

Sample Regression Line

This line is an estimate of the population regression line.



based on the regression line. For example, the first row of Table 5.1 shows that, in the first year, the actual value of sales was 4 million units when selling expense was \$1 million. Therefore, Y = 4.0 millions of units when $X = 1$. In contrast, the regression line indicates that $\hat{Y} = 2.536 + 1.504(1)$, or 4.039 millions of units when $X = 1$. In other words, while the regression line predicts that sales will equal 4.039 millions of units when selling expense is \$1 million, the actual sales figure under these circumstances (in the first year) was 4 million units.

It is essential to be able to identify and interpret the Y intercept and slope of a regression line. What is the Y intercept of the regression line in the case of the Miller Pharmaceutical Company? It is 2.536 millions of units. This means that, if the firm's selling expense is zero, the estimated sales would be 2.536 millions of units. (As shown in Figure 5.5, 2.536 millions of units is the value of the dependent variable at which the regression line intersects the vertical axis.) What is the slope of the regression line in this case? It is 1.504. This means that the estimated sales go up by 1.504 millions of units when selling expense increases by \$1 million.

Method of Least Squares

The method used to determine the values of a and b is the so-called method of least squares. Since the deviation of the ith observed value of Y from the regression line equals $\hat{Y}_i - Y_i$, the sum of these squared deviations equals

$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - a - bX_i)^2 \tag{5.3}$$

where n is the sample size.[7] Using the minimization technique presented in Chapter 2, we can find the values of a and b that minimize the expression in equation (5.3) by differentiating this expression with respect to a and b and setting these partial derivatives equal to zero:

$$\frac{\partial \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{\partial a} = -2 \sum_{i=1}^{n} (Y_i - a - bX_i) = 0 \tag{5.4}$$

$$\frac{\partial \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{\partial b} = -2 \sum_{i=1}^{n} X_i (Y_i - a - bX_i) = 0 \tag{5.5}$$

Solving equations (5.4) and (5.5) simultaneously and letting $\overline{X}$ equal the mean value of X in the sample and $\overline{Y}$ equal the mean value of Y, we find that

$$b = \frac{\sum_{i=1}^{n} (X_i - \overline{X}(Y_i - \overline{Y})}{\sum_{i=1}^{n} (X_i - \overline{X})^2} \tag{5.6}$$

$$a = \overline{Y} - b\overline{X} \tag{5.7}$$

The value of b in equation (5.6) is often called the **estimated regression coefficient.**

[7]As pointed out in Chapter 1, Σ is the mathematical summation sign. What does ΣX_i mean? It means that the numbers to the right of the summation sign (that is, the values of X_i) should be summed from the lower limit on i (which is given below the Σ sign) to the upper limit on i (which is given above the Σ sign):

$$\sum_{i=1}^{n} X_i$$

means the same thing as $X_1 + X_2 + \cdots + X_n$.

TABLE 5.2

Computation of ΣX_i, ΣY_i, ΣX_i^2, ΣY_i^2, and $\Sigma X_i Y_i$

	X_i	Y_i	X_i^2	Y_i^2	$X_i Y_i$
	1	4	1	16	4
	2	6	4	36	12
	4	8	16	64	32
	8	14	64	196	112
	6	12	36	144	72
	5	10	25	100	50
	8	16	64	256	128
	9	16	81	256	144
	7	12	49	144	84
Total	50	98	340	1,212	638
$\overline{X} = 50/9 = 5.556$					
$\overline{Y} = 98/9 = 10.889$					

From a computational point of view, it frequently is easier to use a somewhat different formula for b than the one given in equation (5.6). This alternative formula, which yields the same answer as equation (5.6), is

$$b = \frac{n\sum_{i=1}^{n} X_i Y_i - \left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2}$$

In the case of the Miller Pharmaceutical Company, Table 5.2 shows the calculation of $\Sigma X_i Y_i$, ΣX_i^2, ΣX_i, and ΣY_i. Based on these calculations,

$$b = \frac{9(638) - (50)(98)}{9(340) - 50^2} = 1.504$$

Therefore, the value of b, the least-squares estimator of B, is 1.504, which is the result given in the previous section. In other words, an increase in selling expense of \$1 million results in an increase in estimated sales of about 1.504 millions of units.

Having calculated b, we can readily determine the value of a, the least-squares estimator of A. According to equation (5.7),

$$a = \overline{Y} - b\overline{X}$$

where $\overline{Y}$ is the mean of the values of Y, and $\overline{X}$ is the mean of the values of X. Since, as shown in Table 5.2, $\overline{Y} = 10.889$ and $\overline{X} = 5.556$, it follows that

$$\begin{aligned} a &= 10.889 - 1.504(5.556) \\ &= 2.536 \end{aligned}$$

Therefore, the least-squares estimate of A is 2.536 millions of units, which is the result given in the previous section.

Having obtained a and b, it is a simple matter to specify the average relationship in the sample between sales and selling expense for the Miller Pharmaceutical Company. This relationship is

$$\hat{Y} = 2.536 + 1.504X \tag{5.8}$$

where $\hat{Y}$ is measured in millions of units and X is measured in millions of dollars. As we know, this line is often called the *sample regression line*, or the *regression of Y on X*. It is the line presented in the previous section and plotted in Figure 5.5. Now, we see how this line is derived. (However, a computer usually does the calculations.)

To illustrate how a regression line of this sort can be used, suppose that the managers of the firm want to predict the firm's sales if they decide to devote \$4 million to selling expense. Using equation (5.8), they would predict that its sales would be

$$2.536 + 1.504(4) = 8.55. \tag{5.9}$$

Since sales are measured in millions of units, this means that sales would be expected to be 8.55 million units.

Coefficient of Determination

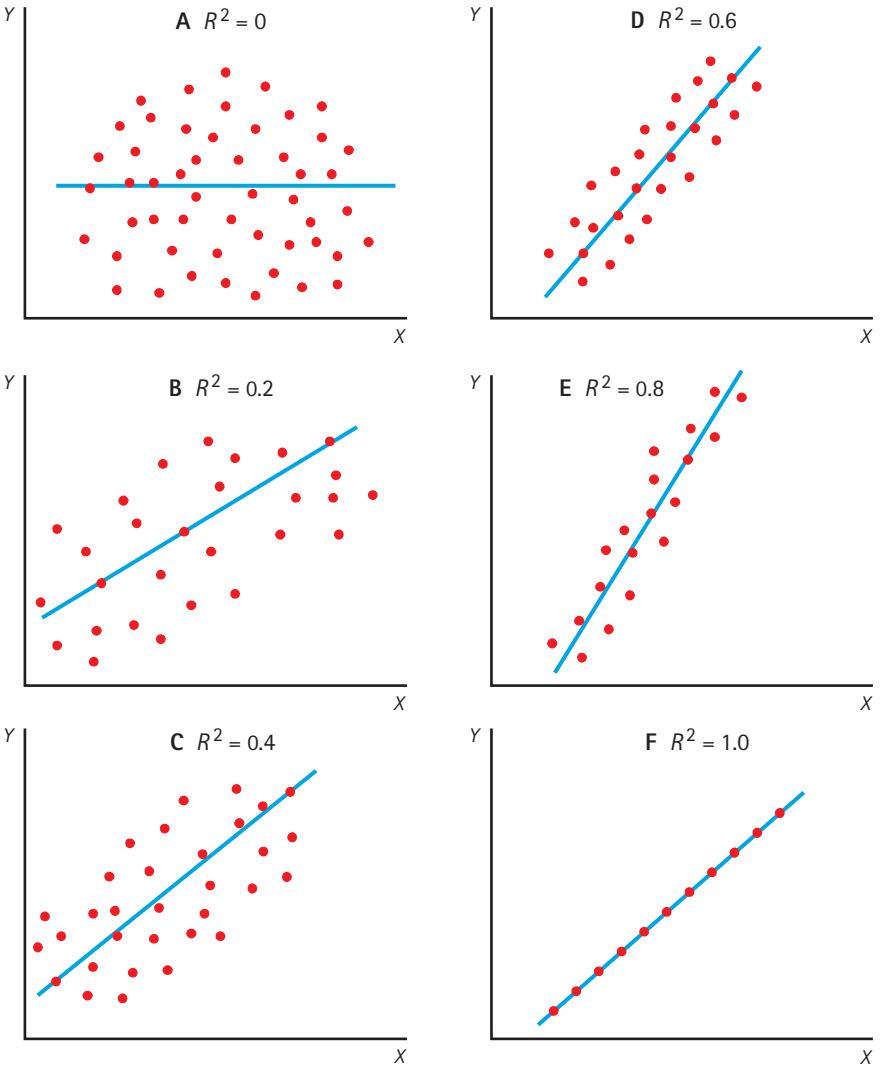
Once the regression line has been calculated, we want to know how well this line fits the data. There can be huge differences in how well a regression line fits a set of data, as shown in Figure 5.6. Clearly, the regression line in panel F of Figure 5.6 provides a much better fit than the regression line in panel B of the same figure. How can we measure how well a regression line fits the data?

The most commonly used measure of the goodness of fit of a regression line is the coefficient of determination. For present purposes, it is not necessary to know the formula for the coefficient of determination, because it is seldom calculated by hand. It is a particular item, often designated by R^2, or R-sq on a computer printout, as we shall see in the section after next.

FIGURE 5.6

Six Regression Lines: Coefficient of Determination Equals 0, 0.2, 0.4, 0.6, 0.8, and 1.0

When there is only one independent variable, the coefficient of determination is often designated by r^2, rather than R^2, but computer printouts generally use R^2, regardless of the number of independent variables. We use R^2 here, even though there is only one independent variable. See footnote 8.



CONCEPTS IN CONTEXT

How the Japanese Motorcycle Makers Used the Coefficient of Determination

In late 1982, Harley-Davidson asked the International Trade Commission (ITC), a federal agency that investigates possible injuries to U.S. firms and workers from imports, for relief from Japanese imports of heavyweight motorcycles. According to Harley-Davidson, the Japanese were selling their motorcycles at prices too low for it to meet. On the basis of Section 201 of the 1974 Trade Act, the ITC can impose tariffs or quotas on imported goods to provide "additional time to permit a seriously injured domestic industry to become competitive." But to receive such tariff or quota relief, the industry must demonstrate that the injuries it suffers are due to increased imports, not some other cause such as bad management or a recession.

Harley-Davidson's petition to the ITC was contested by the major Japanese motorcycle makers: Honda, Kawasaki, Suzuki, and Yamaha. One of their arguments was that general economic conditions, not Japanese imports, were the principal cause of Harley-Davidson's declining share of the market. In other words, they attributed Harley-Davidson's problems to the recession of the early 1980s. They pointed out that heavyweight motorcycles, which cost about $7,000, were a "big-ticket luxury consumer product" and that their sales would be expected to fall in a recession.

To back up this argument, John Reilly of ICF, Inc., the Japanese firms' chief economic consultant, calculated a regression, where Harley-Davidson's sales were the dependent variable and the level of blue-collar employment (a measure of general economic conditions) was the independent variable. He showed that the coefficient of determination was about 0.73. Then, he calculated a regression where Harley-Davidson's sales were the dependent variable, and the level of sales of Japanese motorcycles was the independent variable. He showed that the coefficient of determination was only about 0.22. From this comparison of the two coefficients of determination, he concluded that Harley-Davidson's sales were much more closely related to general economic conditions than to the level of sales of Japanese motorcycles.

Of course, this analysis tells us nothing about the effects of the price of Japanese motorcycles on Harley-Davidson's sales and profits. From many points of view, what was needed was an estimate of the market demand function for Harley-Davidson's motorcycles. Such an analysis would have related Harley-Davidson's sales to the price of Harley-Davidson's motorcycles, the price of Japanese motorcycles, the level of disposable income, and other variables discussed in Chapter 3. In any event, despite the evidence cited, the Japanese motorcycle manufacturers did not prevail. On the contrary, the ITC supported Harley-Davidson's petition, and on April 1, 1983, President Ronald Reagan imposed a substantial tariff (almost 50 percent) on imported (large) motorcycles.*

*See "Revving up for Relief: Harley-Davidson at the ITC," a case in the Study Guide accompanying this textbook. For further discussion, see J. Gomez-Ibanez and J. Kalt, *Cases in Microeconomics* (Englewood Cliffs, NJ: Prentice-Hall, 1990); P.C. Reid, *Well Made in America;* Lessons from Harley-Davidson on Being the Best (New York: McGraw-Hill, 1989); and *New York Times,* July 20, 1997.

The value of the coefficient of determination varies between 0 and 1. *The closer it is to 1, the better the fit; the closer it is to 0, the poorer the fit.* In the case of the Miller Pharmaceutical Company, the coefficient of determination between sales and selling expense is 0.97, which indicates a very good fit. To get a feel for what a particular value of the coefficient of determination means, look at the six panels of Figure 5.6. Panel A shows that, if the coefficient of determination is 0, there is no relationship at all between the independent and dependent variables. Panel B shows that, if the coefficient of determination is 0.2, the regression line fits the data rather poorly. Panel C shows that, if it is 0.4, the regression line fits better but not very well. Panel D shows that, if it is 0.6, the fit is reasonably good. Panel E shows that, if it is 0.8, the fit is good. Finally, panel F shows that, if it is 1.0, the fit is perfect.[8] (A fuller discussion of the coefficient of determination is provided in the appendix to this chapter.

Multiple Regression

In previous sections of this chapter, we discussed regression techniques in the case in which there is only one independent variable. In practical applications of regression techniques, it frequently is necessary and desirable to include two or more independent variables. Now, we extend the treatment of regression to the case in which there is more than one independent variable.

Whereas a **simple regression** includes only one independent variable, a **multiple regression** includes two or more independent variables. Multiple regressions ordinarily are carried out on computers with the aid of statistical software packages like Minitab, SAS, or SPSS. So, there is no reason for you to learn how to do them by hand. The first step in multiple regression analysis is to identify the independent variables and specify the mathematical form of the equation relating the mean value of the dependent variable to these independent variables.

[8]If one is doing the calculations by hand, a convenient formula for the coefficient of determination is

$$r^2 = \frac{\left[n\sum_{i=1}^{n} X_i Y_i - \left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)\right]^2}{\left[n\sum_{i=1}^{n} X_1^2 - \left(\sum_{i=1}^{n} X_i\right)^2\right]\left[n\sum_{i=1}^{n} Y_i^2 - \left(\sum_{i=1}^{n} Y_i\right)^2\right]}$$

Table 5.2 contains the quantities to be inserted in this formula.

Note too that the square root of r^2, called the **correlation coefficient**, is also used to measure how well a simple regression equation fits the data. (The sign of the square root is the same as that of b.)

As pointed out in the note to Figure 5.6, computer printouts generally refer to the coefficient of determination as R^2, although statisticians often call it r^2 when there is only one independent variable.

In the case of the Miller Pharmaceutical Company, suppose that the firm's executives feel that its sales depend on its price, as well on its selling expense. More specifically, they assume that

$$Y_i = A + B_1X_i + B_2P_i + e_i \tag{5.10}$$

where X_i is the selling expense (in millions of dollars) of the firm during the ith year and P_i is the price (in dollars) of the firm's product during the ith year (measured as a deviation from \$10, the current price). Of course, B_2 is assumed to be negative. This is a different model from that in equation (5.2). Here, we assume that Y_i (the firm's sales in the ith year) depends on two independent variables, not one. Of course, there is no reason why more independent variables cannot be added, so long as data are available concerning their values and there is good reason to expect them to affect Y_i. But, to keep matters simple, we assume that the firm's executives believe that only selling expense and price should be included as independent variables.[9]

The object of multiple regression analysis is to estimate the unknown constants A, B_1, and B_2 in equation (5.10). Just as in the case of simple regression, these constants are estimated by finding the value of each that minimizes the sum of the squared deviations of the observed values of the dependent variable from the values of the dependent variable predicted by the regression equation. Suppose that a is an estimator of A, b_1 is an estimator of B_1, and b_2 is an estimator of B_2. Then, the value of the dependent variable Y_i predicted by the estimated regression equation is

$$\hat{Y}_i = a + b_1X_i + b_2P_i$$

and the deviation of this predicted value from the actual value of the dependent variable is

$$Y_i - \hat{Y}_i = Y_i - a - b_1X_i - b_2P_i$$

If these deviations are squared and summed, the result is

$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - a - b_1X_i - b_2P_i)^2 \tag{5.11}$$

where n is the number of observations in the sample. As pointed out earlier, we choose the values of a, b_1, and b_2 that minimize the expression in equation (5.11). These estimates are called *least-squares estimates*, as in the case of simple regression.

[9]As in the case of simple regression, it is assumed that the mean value of e_i is zero and that the values of e_i are statistically independent (recall footnote 6).

TABLE 5.3

Sales, Selling Expense, and Price, Miller Pharmaceutical Company, Sample of Nine Years

Selling expense (millions of dollars)	Sales (millions of units)	Price (less $10)
2	6	0
1	4	1
8	16	2
5	10	3
6	12	4
4	8	5
7	12	6
9	16	7
8	14	8

Computer programs, described in the following section, are available to calculate these least-squares estimates. Based on the data in Table 5.3, the computer output shows that $b_1 = 1.758$, $b_2 = -0.352$, and $a = 2.529$. Consequently, the estimated regression equation is

$$Y_i = 2.529 + 1.758X_i - 0.352P_i \tag{5.12}$$

The estimated value of B_1 is 1.758, as contrasted with our earlier estimate of B, which was 1.504. In other words, a $1 million increase in selling expense results in an increase in estimated sales of 1.758 million units, as contrasted with 1.504 million units in the simple regression in equation (5.8). The reason these estimates differ is that the present estimate of the effect of selling expense on sales holds constant the price, whereas the earlier estimate did not hold this factor constant. Since this factor affects sales, the earlier estimate is likely to be a biased estimate of the effect of selling expense on sales.[10]

[10]Of course, this regression is supposed to be appropriate only when X_i and P_i vary in a certain limited range. If P_i is large and X_i is small, the regression would predict a negative value of sales, which obviously is inadmissible. But, as long as the regression is not used to make predictions for values of X_i and P_i outside the range of the data given in Table 5.3, this is no problem. For simplicity, we assume in equation (5.10) that the effect of price on the mean value of sales (holding selling expense constant) can be regarded as linear in the relevant range. Alternatively, we could have assumed that it was quadratic or the constant-elasticity demand function discussed in Chapter 3 might have been used.

CONCEPTS IN CONTEXT

Color Balance and Shelf-Life Performance of Polaroid Film

In 1947, the prototype of the instant camera was demonstrated to the Optical Society of America. A year later, the Polaroid made the first instant camera and film available to the public. The single-step photographic process enabled pictures to be developed in 60 seconds. Unfortunately, Polaroid did not see the potential for the digital camera fast enough, and although they subsequently developed digital cameras, they were no longer the leader in the photography market. In addition, "one-hour" photo developing at the local drugstore, supermarket, or photo shop took away some of the advantage of Polaroid's "instant" pictures. In 2001, they voluntarily declared bankruptcy.

Sixty five percent of the assets (and trademark name) of the company were purchased by One Equity Partners (part of J.P. Morgan Chase) in 2002. Primary PDC (the interests of the old Polaroid Corporation own the other 35% of the new Polaroid Corporation. According to the new corporate description by Yahoo Finance, "the company makes instant film and camera, digital cameras, professional imaging equipment, and security ID-card systems." Its I-Zone instant camera is the nation's top-selling camera. So, while digital cameras are prevalent, many people take pictures using film cameras.

Regression analysis is important in many aspects of managerial economics, not just in estimating demand functions. For example, this technique helped the Polaroid Corporation, a leading manufacturer of cameras and film, to supply film at the peak of its usefulness. An extremely important consideration to Polaroid was how well films maintain their sensitivity, and whether they provided satisfactory photographic results and for how long. Information of this sort, together with

data concerning average elapsed time between the purchase and utilization of film, enabled Polaroid to make manufacturing adjustments to help consumers get good performance from Polaroid film.

One important characteristic of film is color balance—its ability to produce color. To see the effects of film age on color balance, Polaroid took 14 samples at monthly intervals, up to 13 months after manufacture. For each sample, the change in blue balance was measured. As shown in the graph, the color balance becomes bluer (that is, "cooler," not as "warm") as the film ages.

Using the techniques described in this chapter, Polaroid estimated the regression line:

$$\hat{Y} = 8.194 + 6.756X,$$

where Y is the change in blue balance and X is the age (in months) of the film. The coefficient of determination was 0.966, which indicates a close fit to the data.

According to Polaroid officials, this application of regression analysis was important. Together with data regarding consumer purchase and use patterns, it enabled "Polaroid to manufacture film that shifted those characteristics which determine picture quality to their optimum setting by the time the film was being used. In essence, Polaroid had the information to compensate in its manufacturing process for crucial alterations in film performance that happened as a result of the aging process."*

*D. Anderson, D. Sweeney, and T. Williams, *Statistics for Business and Economics* (3d ed.; St. Paul, MN: West, 1987), p. 523.

ANALYZING MANAGERIAL DECISIONS

How Good are *Ward's* Projections of Auto Output?

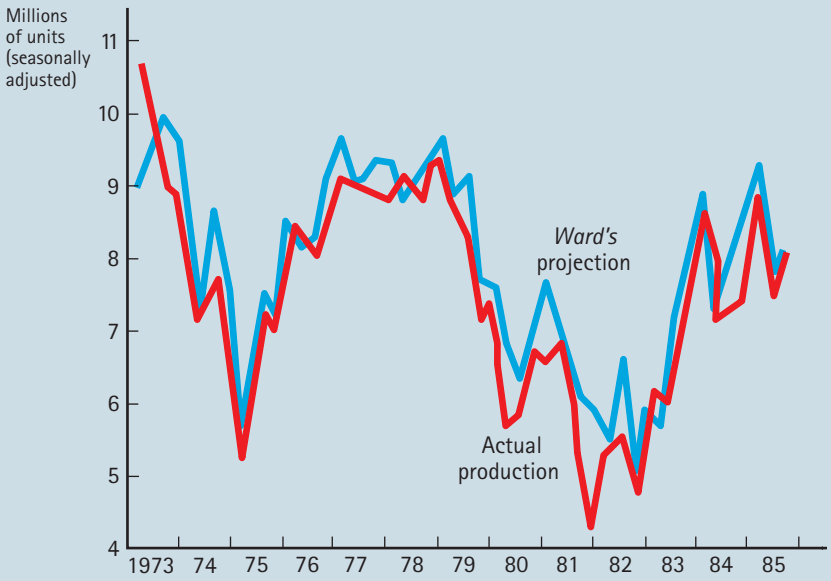
The automobile industry and its suppliers, as well as other industries and government agencies, try in a variety of ways to forecast auto output in the United States. Each month, *Ward's Automotive Reports* asks eight U.S. automakers to state their domestic production plans for the next three to eight months. The following figure shows actual domestic auto production and *Ward's* projections made at the beginning of each quarter. The average error is about a half-million cars per year, or about 6 percent.

To obtain a more precise estimate of the relationship between *Ward's* projections and actual output, Ethan Harris regressed actual output *(Y)* on *Ward's* projection *(X)* and the error in *Ward's* projection during the previous quarter *(E),* the result being

$$Y = 0.275 + 0.909X + 0.277E$$

The multiple coefficient of determination equals 0.838.

(a) If *Ward's* projection is 1 million cars higher in one quarter than in another, would you expect actual output to be 1 million cars higher? Why or why not? (b) If *Ward's* projection was 100,000 cars too high in the previous quarter, is it likely that actual output would be higher than if the projection had been 100,000 cars too low in the previous



quarter? (c) Does the regression provide a good or poor fit to the data?

SOLUTION (a) No. According to the equation, if *X* increases by 1 million, *Y* would be expected to increase by 0.909 times 1 million, or 909,000 (if *E* remains the same). (b) Under these circumstances, it is likely that actual output would be higher than it would if the projection had been 100,000 cars too low in the previous quarter. To see this, note that the regression coefficient of *E* in the regression equation is positive. Therefore, increases in *E* tend to be associated with increases in *Y*. (c) The fact that the multiple coefficient of determination is about 0.8 indicates that the fit is good (about like that in panel E of Figure 5.6).*

*For further discussion, see E. Harris, "Forecasting Automobile Output," *Federal Reserve Bank of New York Quarterly Review,* Winter 1985–86, reprinted in *Managerial Economics and Operations Research,* ed. Mansfield.

Software Packages and Computer Printouts

With few exceptions, regression analyses are carried out on computers, not by hand. Therefore, it is important that you know how to interpret computer printouts showing the results of regression calculations. Because there is a wide variety of "canned" programs for calculating regressions, no single format or list of items is printed out. However, the various kinds of printouts are sufficiently similar that it is worthwhile looking at two illustrations—Minitab and SAS—in some detail.

Figure 5.7 shows the Minitab printout from the multiple regression of the Miller Pharmaceutical Company's sales (designated as C1) on its selling expense (C2) and price (C3). According to this printout, the regression equation is

$$C1 = 2.529 + 1.758C2 - 0.352C3$$

The column headed "Coef" shows the estimated regression coefficient of each independent variable (called a "Predictor" on the printout). The intercept of the regression is the top figure in this vertical column (the figure in the horizontal row where the "Predictor" is "Constant"). The coefficient of determination (called R-sq) is shown in the middle of the printout. For a multiple regression, the coefficient of determination is often called the *multiple coefficient of determination.*[11]

Figure 5.8 shows the SAS printout for the same regression. To find the intercept of the equation, obtain the figure (2.529431) in the horizontal row labeled "INTERCEP" that is in the vertical column called "Parameter Estimate." To find the regression coefficient of selling expense, obtain the figure (1.758049) in the horizontal row labeled "C2" that is in the vertical column called "Parameter Estimate." To find the regression coefficient of price, obtain the figure (−0.351870) in the horizontal row labeled "C3" that is in the vertical column called "Parameter Estimate." The multiple coefficient of determination is the figure (0.9943) to the right of "R-square."

Interpreting the Computer Printout

The following additional statistics are also of considerable importance: the standard error of estimate, the *F* statistic, and the *t* statistic. Each is discussed briefly next. For more detailed discussions of each, see any business statistics textbook.[12]

The Standard Error of Estimate

A measure often used to indicate the accuracy of a regression model is the standard error of estimate, which is *a measure of the amount of scatter of individ-*

[11]The positive square root of the multiple coefficient of determination is called the *multiple correlation coefficient*, denoted *R*. It too is sometimes used to measure how well a multiple regression equation fits the data.

The *unadjusted* multiple coefficient of determination—R-sq in Figure 5.7—can never decrease as another independent variable is added; a related measure without this property is the *adjusted* multiple coefficient of determination—R-sq (adj.) in Figure 5.7. The latter is often denoted $\bar{R}^2$.

[12]For example, E. Mansfield, *Statistics for Business and Economics* (5th ed.; New York: Norton, 1994).

FIGURE 5.7

Minitab Printout of Results of Multiple Regression

```
MTB > regress c1 on 2 predictors in c2 and c3

The regression equation is
C1 = 2.53 + 1.76 C2 - 0.352 C3

Predictor              Coef             Stdev          t-ratio            p
Constant             2.5294            0.2884             8.77        0.000
C2                  1.75805           0.06937            25.34        0.000
C3                 -0.35187           0.07064            -4.98        0.002

s = 0.3702             R-sq = 99.4%     R-sq(adj) = 99.2%

Analysis of Variance

SOURCE           DF              SS              MS              F            p
Regression        2         144.067          72.033          525.72        0.000
Error             6           0.822           0.137
Total             8         144.889

SOURCE           DF          SEQ SS
C2                1         140.667
C3                1           3.399
```

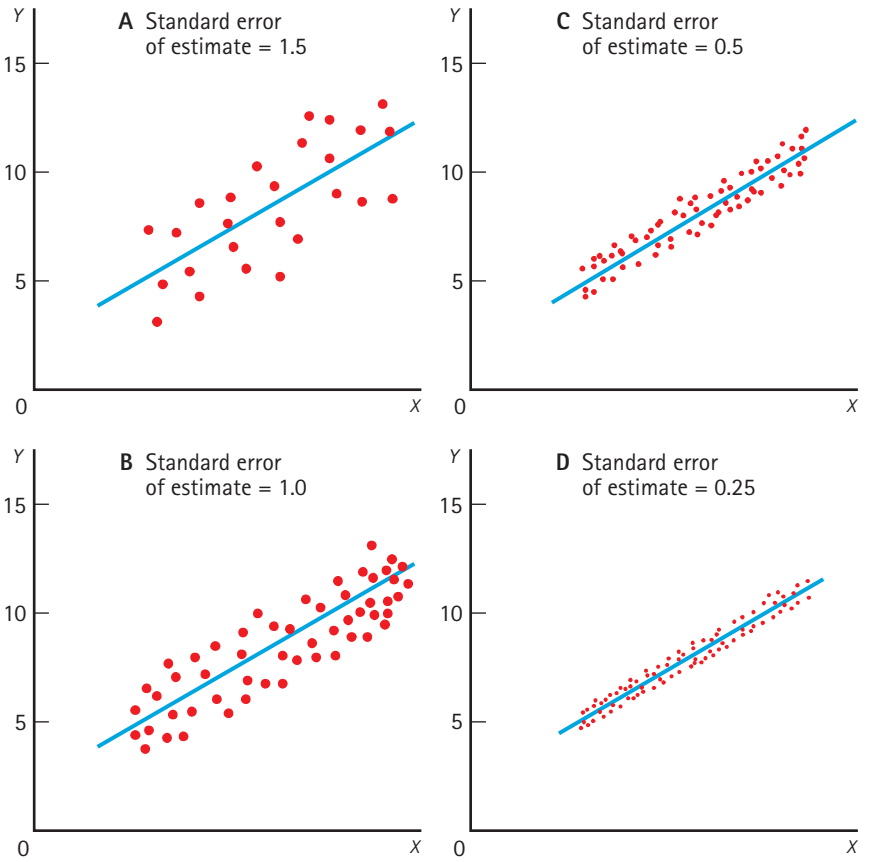
FIGURE 5.8

SAS Printout of Results of Multiple Regression

```
Dependent Variable: C1

Analysis of Variance

                                 Sum of          Mean
Source                 DF       Squares        Square       F Value     Prob>F

Model                   2      144.06678      72.03339        525.718     0.0001
Error                   6       0.82211       0.13702
C Total                 8      144.88889

   Root MSE       0.37016       R-square         0.9943
   Dep Mean      10.88889       Adj R-sq         0.9924
   C.V.           3.39944

Parameter Estimates

                    Parameter          Standard       T for HO:
Variable      DF     Estimate             Error     Parameter=0   Prob > |T|

INTERCEP       1     2.529431        0.28842968           8.770       0.0001
C2             1     1.758049        0.06937127          25.343       0.0001
C3             1    -0.351870        0.07064425          -4.981       0.0025
```

FIGURE 5.9

Four Regression Lines: Standard Error of Estimate Equals 1.5, 1.0, 0.5, and 0.25



ual observations about the regression line. The standard error of estimate is denoted by "s" in the Minitab printout in Figure 5.7 and by "ROOT MSE" in the SAS printout in Figure 5.8. A comparison of these printouts shows that, in the Miller Pharmaceutical multiple regression, the standard error is about 0.37 million units of sales. Of course, the answer is always the same, no matter which package we use.

To illustrate what the standard error of estimate measures, consider Figure 5.9. In panel A, the standard error of estimate is 1.5, which is much higher than in panel D, where it is 0.25. This is reflected in the much greater scatter in the points around the regression line in panel A than in panel D. As pointed out already, what the standard error of estimate measures is the amount of such

scatter. Clearly, the amount of scatter decreases as we move from panel A to panel B to panel C to panel D. Similarly, the standard error of estimate decreases as we move from panel A to panel B to panel C to panel D.

The standard error of estimate is useful in constructing prediction intervals, that is, intervals within which there is a specified probability that the dependent variable will lie. If this probability is set at 0.95, a very approximate prediction interval is

$$\hat{Y} \pm 2s_e \tag{5.13}$$

where $\hat{Y}$ is the predicted value of the dependent variable based on the sample regression and s_e is the standard error of estimate. For example, if the predicted value of the Miller Pharmaceutical Company's sales is 11 million units, the probability is about 0.95 that the firm's sales will be between 10.26 ($=11 - 2 \times 0.37$) million units and 11.74 ($=11 + 2 \times 0.37$) million units. However, it is important to note that equation (5.13) is a good approximation only if the independent variable is close to its mean; if this is not true, more complicated formulas must be used instead.[13]

The *F* Statistic

Frequently, the analyst wants to know whether any of the independent variables really influences the dependent variable. In the case of the Miller Pharmaceutical Company, the marketing director may ask whether the data indicate that either selling expense or price really influences the firm's sales. To answer such a question, one utilizes the F statistic, which is also included in the computer printout. The value of F is provided in the fifth horizontal row from the bottom of figures in the Minitab printout (Figure 5.7) and in the top horizon-

[13]The formula for the standard error of estimate is

$$\left[\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2/(n - k - 1)\right]^{0.5}$$

where k is the number of independent variables.

If the error term is normally distributed (see Appendix B for a description of the normal distribution), the exact prediction interval (with 0.95 probability) is

$$\hat{Y} \pm t_{0.025}s_e \left[\frac{n+1}{n} + \frac{(X^* - \bar{X})^2}{\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2/n}\right]^{0.5}$$

where $t_{0.025}$ is the value of a variable with the t distribution with $(n - 2)$ degrees of freedom that is exceeded with probability of 0.025, X^* is the value of the independent variable, and n is the sample size. (The t distribution is taken up in Appendix B.) This assumes that there is only one independent variable. For further discussion, see Mansfield, *Statistics for Business and Economics*.

tal row of figures in the SAS printout (Figure 5.8). Both printouts indicate that the value of F in the Miller Pharmaceutical case equals about 525.72.

Large values of F tend to imply that at least one of the independent variables has an effect on the dependent variable. Tables of the F distribution, a probability distribution named (or initialed) after the famous British statistician R. A. Fisher, are used to determine *the probability that an observed value of the F statistic could have arisen by chance, given that none of the independent variables has any effect on the dependent variable* (see Appendix B). This probability too is shown in the computer printout. It is denoted by "p" (immediately to the right of F) in the Minitab printout, and by "Prob.F" (immediately to the right of F VALUE) in the SAS printout. The value of this probability is 0.0001 (SAS) or 0.000 (Minitab); the difference is due to rounding.

Having this probability in hand, it is easy to answer the marketing director's question. Clearly, the probability is extremely small—only about 1 in 10,000—that one could have obtained such a strong relationship between the dependent and independent variables sheerly by chance. Therefore, the evidence certainly suggests that selling expense or price (or both) really influences the firm's sales.

The *t* Statistic

Managers and analysts often are interested in whether a particular independent variable influences the dependent variable. For example, the president of the Miller Pharmaceutical Company may want to determine whether the amount allocated to selling expense really affects the firm's sales. As we know from equation (5.12), the least-squares estimate of B_1 is 1.758, which suggests that selling expense has an effect on sales. But this least-squares estimate varies from one sample to another, and by chance it may be positive even if the true value of B_1 is zero.

To test whether the true value of B_1 is zero, we must look at the t statistic of B_1, which is presented in the printout. For Minitab, recall that B_1 is the regression coefficient of C2, since selling expense is denoted by C2. Therefore, to find the t statistic for B_1, we must locate the horizontal row of figures in the printout where the "Predictor" is C2 and obtain the figure in the vertical column called "t-ratio." If SAS is used, find the horizontal row of figures where the "Variable" is C2 and obtain the figure in the vertical column called "T for H0: Parameter = 0." If the error terms in the regression (that is, e_i) are normally distributed, the t statistic has a well-known probability distribution—the t distribution (see Appendix B).

All other things equal, the bigger is the value of the t statistic (in absolute terms), the smaller the probability that the true value of the regression coefficient in question really is zero. Based on the t distribution, it is possible to calculate *the probability, if the true value of the regression coefficient is zero, that the t statistic is as large (in absolute terms) as we observe*. This probability too is presented in the computer printout. For both Minitab and SAS, this proba-

bility is immediately to the right of the t statistic. For Minitab, it is in the vertical column labeled "p"; for SAS, it is in the vertical column labeled "Prob > T." Regardless of whether Minitab or SAS is used, this probability is shown to be about 0.0001 (see Figures 5.7 and 5.8).

Given this probability, we can readily answer the question put forth by the president of the Miller Pharmaceutical Company. Recall that the president wanted to know whether the amount allocated to selling expense really affects the firm's sales. Given the results obtained in the previous paragraph, it seems extremely likely that the amount allocated to selling expense really does affect sales. After all, according to the previous paragraph, the probability is only about 1 in 10,000 that chance alone would have resulted in as large a t statistic (in absolute terms) as we found, based on the firm's previous experience.[14]

Multicollinearity

One important problem that can arise in multiple regression studies is **multicollinearity**, a situation in which two or more independent variables are very highly correlated. In the case of the Miller Pharmaceutical Company, suppose that there had been a perfect linear relationship in the past between the firm's selling expense and its price. In a case of this sort, it is impossible to estimate the regression coefficients of both independent variables (X and P) because the data provide no information concerning the effect of one independent variable, holding the other independent variable constant. All that can be observed is the

[14]Note that this is a *two-tailed test* of the hypothesis that selling expense has no effect on sales. That is, it is a test of this hypothesis against the alternative hypothesis that the true regression coefficient of selling expense is either positive or negative. In many cases, a *one-tailed test*—for example, in which the alternative hypothesis states that the true regression coefficient is positive only—may be more appropriate.

Frequently, a manager would like to obtain an interval estimate for the true value of a regression coefficient. In other words, he or she wants an interval that has a particular probability of including the true value of this regression coefficient. To find an interval that has a probability equal to $(1 - \alpha)$ of including this true value, you can calculate

$$b_1 \pm t_{\alpha/2} s_{b1} \tag{5.14}$$

where s_{b1} is the standard error of b_1 (in the horizontal row labeled "C2" and the vertical column labeled "Stdev" in the Minitab printout, or in the horizontal row labeled "C2" and the vertical column labeled "Standard Error" in the SAS printout) and where $t_{\alpha/2}$ is the $\alpha/2$ point on the t distribution with $(n - k - 1)$ degrees of freedom (see Appendix B). If α is set equal to 0.05, you obtain an interval that has a 95 percent probability of including B_1. In the case of the Miller Pharmaceutical Company, since, $B_1 = 1.758$, $s_{b1} = 0.069$, and $t_{0.025} = 2.447$ it follows that a 95 percent confidence interval for B_1 is

$$1.758 \pm 2.447\ (0.069)$$

or 1.589 to 1.927. For further discussion, see any business statistics textbook.

effect of both independent variables together, given that they both move together in the way they have in previous years.

Regression analysis estimates the effect of each independent variable by seeing how much effect this one independent variable has on the dependent variable when other independent variables are held constant. If two independent variables move together in a rigid, lockstep fashion, there is no way to tell how much effect each has separately; all we can observe is the effect of both combined. If there is good reason to believe that the independent variables will continue to move in lockstep in the future as they have in the past, multicollinearity does not prevent us from using regression analysis to predict the dependent variable. Since the two independent variables are perfectly correlated, one of them in effect stands for both and we therefore need use only one in the regression analysis. However, if the independent variables cannot be counted on to continue to move in lockstep, this procedure is dangerous, since it ignores the effect of the excluded independent variable.

In reality, you seldom encounter cases in which independent variables are perfectly correlated, but you often encounter cases in which independent variables are so highly correlated that, although it is possible to estimate the regression coefficient of each variable, these regression coefficients cannot be estimated at all accurately. To cope with such situations, it sometimes is possible to alter the independent variables in such a way as to reduce multicollinearity. Suppose that a managerial economist wants to estimate a regression equation where the quantity demanded per year of a certain good is the dependent variable and the average price of this good and disposable income of U.S. consumers are the independent variables. If disposable income is measured in money terms (that is, without adjustment for changes in the price level), there may be a high correlation between the independent variables. But if disposable income is measured in real terms (that is, with adjustment for changes in the price level), this correlation may be reduced considerably. Therefore, the managerial economist may decide to measure disposable income in real rather than money terms to reduce multicollinearity.

If techniques of this sort cannot reduce multicollinearity, there may be no alternative but to acquire new data that do not contain the high correlation among the independent variables. Whether you (or your board of directors) like it or not, there may be no way to estimate accurately the regression coefficient of a particular independent variable that is very highly correlated with some other independent variable.

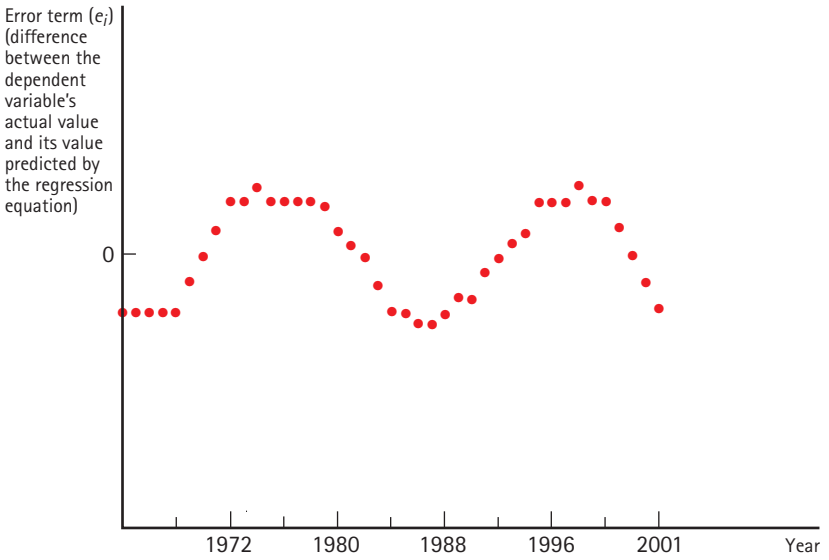
Serial Correlation

In addition to multicollinearity, another important problem that can occur in regression analysis is that the error terms (the values of e_i) are not indepen-

FIGURE 5.10

Serial Correlation of Error Terms

If the error term in one year is positive, the error term in the next year is almost always positive. If the error term in one year is negative, the error term in the next year is almost always negative.



dent; instead, they are serially correlated. For example, Figure 5.10 shows a case in which, if the error term in one period is positive, the error term in the next period is almost always positive. Similarly, if the error term in one period is negative, the error term in the next period almost always is negative. In such a situation, we say that the errors are *serially correlated* (or *autocorrelated*, which is another term for the same thing).[15] Because this violates the assumptions underlying regression analysis, it is important that we be able to detect its occurrence. (Recall that regression analysis assumes that the values of e_i are independent.)

To see whether serial correlation is present in the error terms in a regression, we can use the Durbin-Watson test. Let $\hat{e}_i$ be the difference between Y_i

[15]This is a case of positive serial correlation. (It is the sort of situation frequently encountered in managerial economics.) If the error term in one period tends to be positive (negative) and if the error term in the previous period is negative (positive), this is a case of negative serial correlation. More is said about this subsequently.

and $\hat{Y}_i$, the value of Y_i predicted by the sample regression. To apply the Durbin-Watson test, we (or in most cases, the computer) must calculate

$$d = \frac{\sum_{i=2}^{n} (\hat{e}_i - \hat{e}_{i-1})^2}{\sum_{i=1}^{n} \hat{e}_i^2} \tag{5.15}$$

Durbin and Watson provided tables that show whether d is so high or so low that the hypothesis that there is no serial correlation should be rejected. (Note that d is often called the Durbin-Watson statistic.)

Suppose we want to test this hypothesis against the alternative hypothesis that there is **positive** serial correlation. (Positive serial correlation would mean that e_i is directly related to e_{i-1}, as in Figure 5.10.) If so, we should reject the hypothesis of no serial correlation if $d < d_L$ and accept this hypothesis if $d > d_U$. If $d_L \leq d \leq d_U$, the test is inconclusive. The values of d_L and d_U are shown in Appendix Table 7. (Note that these values depend on the sample size n and on k, the number of independent variables in the regression.) On the other hand, suppose the alternative hypothesis is that there is **negative** serial correlation. (Negative serial correlation means that e_i is inversely related to e_{I-1}.) If so, we should reject the hypothesis of no serial correlation if $d > 4 - d_L$ and accept this hypothesis if $d < 4 - d_U$. If $4 - d_U \leq d \leq 4 - d_L$, the test is inconclusive.[16]

One way to deal with the problem of serial correlation, if it exists, is to take first differences of all the independent and dependent variables in the regression. For example, in the case of the Miller Pharmaceutical Company, we might use the change in sales relative to the previous year (rather than the level of sales) as the dependent variable. And the change in selling expense relative to the previous year (rather than the level of selling expense) and the change in price relative to the previous year (rather than the level of price) might be used as the independent variables in the regression.[17]

[16]For a two-tailed test of both positive and negative serial correlation, reject the hypothesis of no serial correlation if $d < d_L$ or if $d > 4 - d_L$, and accept this hypothesis if $d_U < d < 4 - d_U$. Otherwise, the test is inconclusive. For a two-tailed test, the significance level is double the significance level shown in Appendix Table 7.

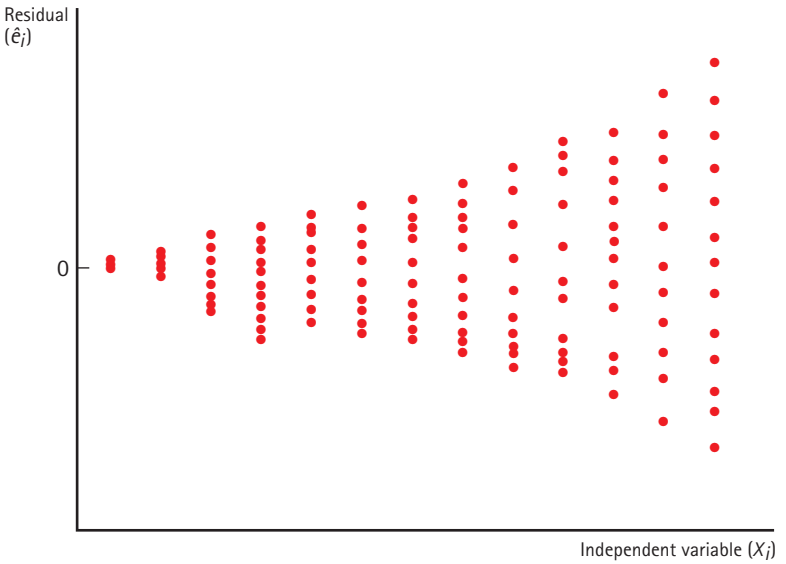
[17]The use of first differences, while useful in some cases, is not always appropriate. For further discussion, see Johnston, *Econometric Methods*.

It is also important to avoid specification errors, which result when one or more significant explanatory variables is not included in the regression. If specification errors arise, the estimated regression coefficients may be biased and the regression equation may not predict very well. Also, problems can arise if the independent variables in a regression contain substantial measurement errors, since the regression coefficients of these variables often tend to be biased toward zero.

FIGURE 5.11

Residuals Indicating That the Variation in the Error Terms Is Not Constant

As you can see, the residuals vary less when X is small than when it is large.



Further Analysis of the Residuals

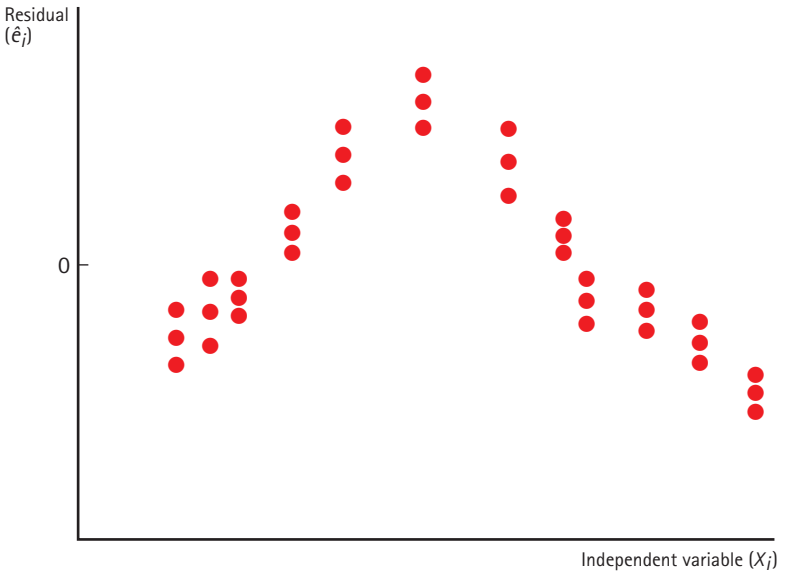
In the previous section, we used $\hat{e}_i$ (the difference between the actual value of Y_i and its value predicted by the sample regression) to test for serial correlation. Since it is a measure of the extent to which Y_i *cannot* be explained by the regression, $\hat{e}_i$ is often called the **residual** for the ith observation. Now we describe additional ways in which the residuals—that is, the values of $\hat{e}_i$—can be used to test whether the assumptions underlying regression analysis are met. We begin by plotting the value of each residual against the value of the independent variable. (For simplicity, we suppose only one independent variable.) That is, we plot $\hat{e}_i$ against X_i, which is the independent variable.

Suppose that the plot is as shown in Figure 5.11. As you can see, the values of the residuals are much more variable when X_i is large than when it is small. In other words, the variation in $\hat{e}_i$ increases as X_i increases. Since regression analysis assumes that *the variation in the error terms is the same, regardless of the value of the independent variable*, the plot in Figure 5.11 indicates that this assumption is violated. Two ways to remedy this situation are to use a

FIGURE 5.12

Residuals Indicating That the Relationship between the Dependent and Independent Variables Is Nonlinear, Not Linear

The residuals are negative when X is very small or very large and positive when X is of medium size.



weighted least-squares regression or to change the form of the dependent variable. For example, we might use log Y rather than Y as the dependent variable.[18]

If the plot of $\hat{e}_i$ against X_i looks like Figure 5.12, this is an indication that the relationship between the dependent and independent variables is not linear. When X is very low and very high, the linear regression *overestimates* the dependent variable, as shown by the fact that the residuals tend to be negative. When X is of medium size, the linear regression underestimates the dependent variable, as shown by the fact that the residuals tend to be positive. It appears that a quadratic relationship fits the data better than a linear one. So, rather than assume that equation (5.2) holds, we should assume that

$$Y_i = A + B_1X_i - B_2X_i^2 + e_i$$

Using the multiple regression techniques described previously, the values of A, B_1, and B_2 can be estimated.

[18]For further details, see Johnston, *Econometric Methods*.

Summary

1. An identification problem may occur if price in various periods is plotted against quantity demanded and the resulting relationship is used to estimate the demand curve. Because nonprice variables are not held constant, the demand curve may have shifted over time. Nonetheless, sophisticated econometric methods may be used to estimate the demand function. Also, market experiments and consumer interviews may be of value. For example, firms sometimes vary price from one city or region to another, to see what the effects are on quantity demanded. An actual illustration of this sort was the evaluation of the four promotion alternatives by L'eggs Products.
2. Regression analysis is useful in estimating demand functions and other economic relationships. The regression line shows the average relationship between the dependent variable and the independent variable. The method of least squares is the standard technique used to fit a regression line to a set of data. If the regression line is $\hat{Y} = a + bX$ and if a and b are calculated by least squares,

$$b = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (X_i - \overline{X})^2}$$

and

$$a = \overline{Y} - b\overline{X}$$

This value of b is often called the *estimated regression coefficient.*
3. Whereas a simple regression includes only one independent variable, a multiple regression includes more than one independent variable. An advantage of multiple regression over a simple regression is that you frequently can predict the dependent variable more accurately if more than one independent variable is used. Also, if the dependent variable is influenced by more than one independent variable, a simple regression of the dependent variable on a single independent variable may result in a biased estimate of the effect of this independent variable on the dependent variable.
4. The first step in multiple regression analysis is to identify the independent variables and specify the mathematical form of the equation relating the mean value of the dependent variable to the independent variables. For example, if Y is the dependent variable and X and P are identified as the independent variables, one might specify that

$$Y_i = A + B_1X_i + B_2P_i + e_i$$

ANALYZING MANAGERIAL DECISIONS

How Fed Economists Forecast Auto Output

Since purchases by the auto industry account for more than half of the rubber and lead consumed in this country as well as a major portion of the steel, aluminum, and a variety of other materials, it is obvious that many firms and government agencies, as well as the auto firms themselves, are interested in forecasting auto output. The Federal Reserve Bank of New York has published an article describing how the regression techniques described in this chapter have been used for this purpose. According to the author, Ethan Harris, the quantity of autos produced quarterly depends on five variables: (1) real disposable income, (2) the ratio of retail auto inventories to sales, (3) the average price of new cars (relative to the overall consumer price index), (4) the price level for nonauto durable goods, and (5) the prime rate (the interest rate banks charge their best customers).

The regression results follow. The probability that the *t* statistic for each of the regression coefficients is as large (in absolute terms) as it is here, if the true value of the regression coefficient is zero, is less than 0.01, except for the case of the nonauto price.

The value of the adjusted multiple coefficient of determination is 0.862, the standard error of estimate is 532, and the Durbin-Watson statistic (*d*) is 2.26. According to Ethan Harris, this regression equation has predicted auto output with a mean (absolute) error of about 6.9 percent.

(a) Would you expect the regression coefficient of the inventory-sales ratio to be negative? If so, why? (b) Can we be reasonably sure that the true value of the regression coefficient of the inventory-sales ratio is not zero? Why or why not? (c) Is there evidence of positive serial correlation of the error terms? (d) Can we use this regression as an estimate of the demand curve for autos? Why or why not?

SOLUTION (a) Yes. If inventories are large relative to sales, one would expect auto firms to produce less than they would if inventories were small. (b) Yes. According to the preceding discussion, the probability that the *t* statistic for the regression coefficient of the inventory-sales ratio would be as great as 6.1 (in absolute terms) would be less than 0.01 if the true regression coefficient were zero. Hence, if this true regression coefficient were zero, it is exceedingly unlikely that the *t* statistic (in absolute terms) would equal its observed value or more. (c) No. Since the value of n is approximately 50 and $k = 5$, Appendix Table 7 shows that $d_L = 1.26$ and $d_U = 1.69$ if the significance level equals 0.025. The observed value of the Durbin-Watson statistic (2.26) is greater than d_U (1.69); this means that we should accept the hypothesis that there is no positive serial correlation. (d) No. One important indication that this is true is that the regression coefficient of the auto price is positive. Clearly, this regression equation cannot be used as an estimate of the demand curve for autos.

Variable	Regression coefficient	*t* statistic
Constant	−22,302	−4.5
Disposable income	12.9	6.6
Prime rate	−97.8	−3.2
Inventory-sales ratio	−19.9	−6.1
Auto price	230	5.0
Nonauto price	6.0	2.1

where e_i is an error term. To estimate B_1 and B_2 (called the *true regression coefficients* of X and P) as well as A (the intercept of this true regression equation), we use the values that minimize the sum of squared deviations of Y_i from $\hat{Y}_i$, the value of the dependent variable predicted by the estimated regression equation.

5. In a simple regression, the coefficient of determination is used to measure the closeness of fit of the regression line. In a multiple regression, the multiple coefficient of determination, R^2, plays the same role. The closer R^2 is to 0, the poorer the fit; the closer it is to 1, the better the fit.
6. The F statistic can be used to test whether any of the independent variables has an effect on the dependent variable. The standard error of estimate can help to indicate how well a regression model can predict the dependent variable. The t statistic for the regression coefficient of each independent variable can be used to test whether this independent variable has any effect on the dependent variable. Computer printouts show the probability that the t statistic is as big (in absolute terms) as we observed, given that this independent variable has no effect on the dependent variable.
7. A difficult problem that can occur in multiple regression is multicollinearity, a situation in which two or more of the independent variables are highly correlated. If multicollinearity exists, it may be impossible to estimate accurately the effect of particular independent variables on the dependent variable. Another frequently encountered problem arises when the error terms in a regression are serially correlated. The Durbin-Watson test can be carried out to determine whether this problem exists. Plots of the residuals can help to detect cases in which the variation of the error terms is not constant or where the relationship is nonlinear not linear.

Problems

1. The Klein Corporation's marketing department, using regression analysis, estimates the firm's demand function, the result being

$$Q = -104 - 2.1P + 3.2I + 1.5A + 1.6Z$$
$$R^2 = 0.89$$
$$\text{Standard error of estimate} = 108$$

where Q is the quantity demanded of the firm's product (in tons), P is the price of the firm's product (in dollars per ton), I is per capita income (in dollars), A is the firm's advertising expenditure (in thousands of dollars), and Z is the price (in dollars) of a competing product. The regression is based on 200 observations.

a. According to the computer printout, the probability is 0.005 that the t statistic for the regression coefficient of A would be as large (in absolute terms) as it is in this case if in fact A has no effect on Q. Interpret this result.
b. If $I = 5{,}000$, $A = 20$, and $Z = 1{,}000$, what is the Klein Corporation's demand curve?
c. If $P = 500$ (and the conditions in part b hold), estimate the quantity demanded of the Klein Corporation's product.
d. How well does this regression equation fit the data?

2. Since all the Hawkins Company's costs (other than advertising) are essentially fixed costs, it wants to maximize its total revenue (net of advertising expenses). According to a regression analysis (based on 124 observations) carried out by a consultant hired by the Hawkins Company,

$$Q = -23 - 4.1P + 4.2I + 3.1A$$

where Q is the quantity demanded of the firm's product (in dozens), P is the price of the firm's product (in dollars per dozen), I is per capita income (in dollars), and A is advertising expenditure (in dollars).
a. If the price of the product is \$10 per dozen, should the firm increase its advertising?
b. If the advertising budget is fixed at \$10,000, and per capita income equals \$8,000, what is the firm's marginal revenue curve?
c. If the advertising budget is fixed at \$10,000, and per capita income equals \$8,000, what price should the Hawkins Company charge?

3. The 1980 sales and profits of seven steel companies were as follows:

Firm	Sales ($ billions)	Profit ($ billions)
Armco	5.7	0.27
Bethlehem	6.7	0.12
Bundy	0.2	0.00
Carpenter	0.6	0.04
Republic	3.8	0.05
U.S. Steel (now USX)	12.5	0.46
Westran	0.5	0.00

a. Calculate the sample regression line, where profit is the dependent variable and sales is the independent variable.
b. Estimate the 1980 average profit of a steel firm with 1980 sales of \$2 billion.
c. Can this regression line be used to predict a steel firm's profit in 2006? Explain.

4. The Cherry Manufacturing Company's chief engineer examines a random sample of 10 spot welds of steel. In each case, the shear strength of the weld and the diameter of the weld are determined, the results being as follows:

Shear strength (pounds)	Weld diameter (thousandths of an inch)
680	190
800	200
780	209
885	215
975	215
1,025	215
1,100	230
1,030	250
1,175	265
1,300	250

   a. Does the relationship between these two variables seem to be direct or inverse? Does this accord with common sense? Why or why not? Does the relationship seem to be linear?
   b. Calculate the least-squares regression of shear strength on weld diameter.
   c. Plot the regression line. Use this regression line to predict the average shear strength of a weld $^1/_5$ inch in diameter. Use the regression line to predict the average shear strength of a weld $^1/_4$ inch in diameter.

5. The Kramer Corporation's marketing manager calculates a regression, where the quantity demanded of the firm's product (designated as "C1") is the dependent variable and the price of the product (designated as "C2") and consumers' disposable income (designated as "C3") are independent variables. The Minitab printout for this regression follows:

```
MTB > regress c1 on 2 predictors in c2 and c3
The regression equation is
C1 = 40.8 - 1.02 C2 + 0.00667 C3

Predictor           Coef         Stdev    t-ratio           p
Constant          40.833         1.112      36.74       0.000
C2              -1.02500       0.06807     -15.06       0.000
C3              0.006667      0.005558       1.20       0.244

s=1.361                     R-sq=91.6%          R-sq(adj)=90.8%
Analysis of variance

SOURCE                DF            SS          MS           F           p
Regression             2        422.92      211.46      114.11       0.000
Error                 21         38.92        1.85
Total                 23        461.83

SOURCE                DF        SEQ SS
C2                     1        420.25
C3                     1          2.67
```

a. What is the intercept of the regression?
b. What is the estimated regression coefficient of the product's price?
c. What is the estimated regression coefficient of disposable income?
d. What is the multiple coefficient of determination?
e. What is the standard error of estimate?
f. What is the probability that the observed value of the F statistic could arise by chance, given that neither of the independent variables has any effect on the dependent variable?
g. What is the probability, if the true value of the regression coefficient of price is zero, that the t statistic is as large (in absolute terms) as we observe?
h. What is the probability, if the true value of the regression coefficient of disposable income is zero, that the t statistic is as large (in absolute terms) as we observe?
i. Describe briefly what this regression means.

6. Railroad executives must understand how the costs incurred in a freight yard are related to the output of the yard. The two most important services performed by a yard are switching and delivery, and it seems reasonable to use the number of cuts switched and the number of cars delivered during a particular period as a measure of output. (A cut is a group of cars that rolls as a unit onto the same classification track; it is often used as a unit of switching output.) A study of one of the nation's largest railroads assumed that

$$C_i = A + B_1 S_i + B_2 D_i + e_i$$

where C_i is the cost incurred in this freight yard on the ith day, S_i is the number of cuts switched in this yard on the ith day, D_i is the number of cars delivered in this yard on the ith day, and e_i is an error term. Data were obtained regarding C_i, S_i, and D_i for 61 days. On the basis of the procedures described in this chapter, these data were used to obtain estimates of A, B_1, and B_2. The resulting regression equation was

$$\hat{C}_i = 4{,}914 + 0.42 S_i + 2.44 D_i$$

where $\hat{C}_i$ is the cost (in dollars) predicted by the regression equation for the ith day.[19]

a. If you were asked to evaluate this study, what steps would you take to determine whether the principal assumptions underlying regression analysis were met?

[19]For a much more detailed account of this study, see E. Mansfield and H. Wein, "A Managerial Application of a Cost Function by a Railroad," a case in the Study Guide accompanying this textbook.

b. If you were satisfied that the underlying assumptions were met, of what use might this regression equation be to the railroad? Be specific.
c. Before using the study's regression equation, what additional statistics would you like to have? Why?
d. If the Durbin-Watson statistic equals 2.11, is there evidence of serial correlation in the residuals?

7. Mary Palmquist, a Wall Street securities analyst, wants to determine the relationship between the nation's gross domestic product (GDP) and the profits (after taxes) of the General Electric Company. She obtains the following data concerning each variable:

Year	Gross domestic product (billions of dollars)	General Electric's profits (millions of dollars)
1965	688	355
1966	753	339
1967	796	361
1968	868	357
1969	936	278
1970	982	363
1971	1,063	510
1972	1,171	573
1973	1,306	661
1974	1,407	705
1975	1,529	688
1976	1,706	931

a. What are the least-squares estimates of the intercept and slope of the true regression line, where GE's profits are the dependent variable and GDP is the independent variable?
b. On the average, what effect does a $1 increase in gross domestic product seem to have on the profits of GE?
c. If Ms. Palmquist feels that next year's GDP will be $2 trillion, what forecast of GE's profits will she make on the basis of the regression?
d. What is the coefficient of determination between the nation's gross domestic product and GE's profits?
e. Do the results obtained in previous parts of this problem prove that changes in GE's profits are caused by changes in the gross domestic product? Can we be sure that GE's profit is a linear function of the GDP? What other kinds of functions might be as good or better?
f. If you were the financial analyst, would you feel that this regression line was an adequate model to forecast GE's profits? Why or why not?

8. In the manufacture of cloth, the weft packages should not disintegrate unduly during weaving. A direct measure of the tendency to disintegrate exists, but it is laborious and uneconomical to carry out. In addition, there are indirect measures based on laboratory tests. The Brockway Textile Company would like to determine the extent to which one of these indirect measures is correlated with the direct measure. If the correlation is high enough, the firm believes that it may be able to use the indirect measure instead of the direct measure.

   An experiment was carried out in which both the direct and indirect measures of the tendency to disintegrate were calculated for 18 lots of packages. The results follow:

	Measure	
Lot	Direct	Indirect
1	31	6.2
2	31	6.2
3	21	10.1
4	21	8.4
5	57	2.9
6	80	2.9
7	35	7.4
8	10	7.3
9	0	11.1
10	0	10.7
11	35	4.1
12	63	3.5
13	10	5.0
14	51	4.5
15	24	9.5
16	15	8.5
17	80	2.6
18	90	2.9

   a. What is the coefficient of determination between the two measures?
   b. What linear regression line would you use to predict the value of the direct measure on the basis of knowledge of the indirect measure?
   c. On the basis of your findings, write a brief report indicating the factors to be weighed in deciding whether to substitute the indirect measure for the direct measure.

9. The Kingston Company hires a consultant to estimate the demand function for its product. Using regression analysis, the consultant estimates the demand function to be

$$\log Q = 2.01 - 0.148 \log P + 0.258 \log Z$$

where Q is the quantity demanded (in tons) of Kingston's product, P is the price (in dollars per ton) of Kingston's product, and Z is the price (in dollars per ton) of a rival product.

a. Calculate the price elasticity of demand for Kingston's product.
b. Calculate the cross elasticity of demand between Kingston's product and the rival product.
c. According to the consultant, $\overline{R}^2 = 0.98$ and the standard error of estimate is 0.001. If the number of observations is 94, comment on the goodness of fit of the regression.

10. During the 1960s, the Boston and Maine Railroad conducted an experiment in which it reduced fares by about 28 percent for approximately a year to estimate the price elasticity of demand. This large fare reduction resulted in essentially no change in the railroad's revenues.
    a. What problems exist in carrying out an experiment of this sort?
    b. Taken at face value, what seemed to be the price elasticity of demand?

11. Because of a shift in consumer tastes, the market demand curve for high-quality red wine has shifted steadily to the right. If the market supply curve has remained fixed (and is upward sloping to the right), there has been an increase over time in both the price of such wine and in the quantity sold.
    a. If one were to plot price against quantity sold, would the resulting relationship approximate the market demand curve?
    b. If not, what would this relationship approximate?

12. The Brennan Company uses regression analysis to obtain the following estimate of the demand function for its product:

$$\log Q = 2 - 1.2 \log P + 1.5 \log I$$

where Q is quantity demanded, P is price, and I is consumers' disposable income.

    a. Brennan's president is considering a 5 percent price reduction. He argues that these results indicate that such action will result in a 6 percent increase in the number of units sold by the firm. Do you agree? Why or why not?
    b. The firm's treasurer points out that, according to the computer printout, the probability that the t statistic of log P is as large (in absolute value) as it is, given that log P has no real effect on log Q, is about 0.5. He says that the estimate of the price elasticity is unreliable. Do you agree? Why or why not?
    c. How can the firm obtain a more accurate estimate of the price elasticity of demand?

Appendix: The Coefficient of Determination and the Concept of Explained Variation

In this appendix, we provide a fuller explanation of what the coefficient of determination is and how it can be interpreted. To begin with, we must discuss the concept of variation, which refers to a sum of squared deviations. The total variation in the dependent variable Y equals

$$\sum_{i=1}^{n} (Y_i - \overline{Y})^2 \tag{5.16}$$

In other words, the total variation equals the sum of the squared deviations of Y from its mean.

To measure how well a regression line fits the data, we divide the total variation in the dependent variable into two parts: the variation that *can* be explained by the regression line and the variation that *cannot* be explained by the regression line. To divide the total variation in this way, we must note that, for the ith observation,

$$(Y_i - \overline{Y}) = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \overline{Y}) \tag{5.17}$$

where $\hat{Y}_i$ is the value of Y_i that would be predicted on the basis of the regression line. In other words, as shown in Figure 5.13, the discrepancy between Y_i and the mean value of Y can be split into two parts: the discrepancy between Y_i and the point on the regression line directly below (or above) Y_i and the discrepancy between the point on the regression line directly below (or above) Y_i and $\overline{Y}$.

It can be shown that[20]

$$\sum_{i=1}^{n} (Y_i - \overline{Y})^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2 \tag{5.18}$$

[20]To derive this result, we square both sides of equation (5.17) and sum the result over all values of i. We find that

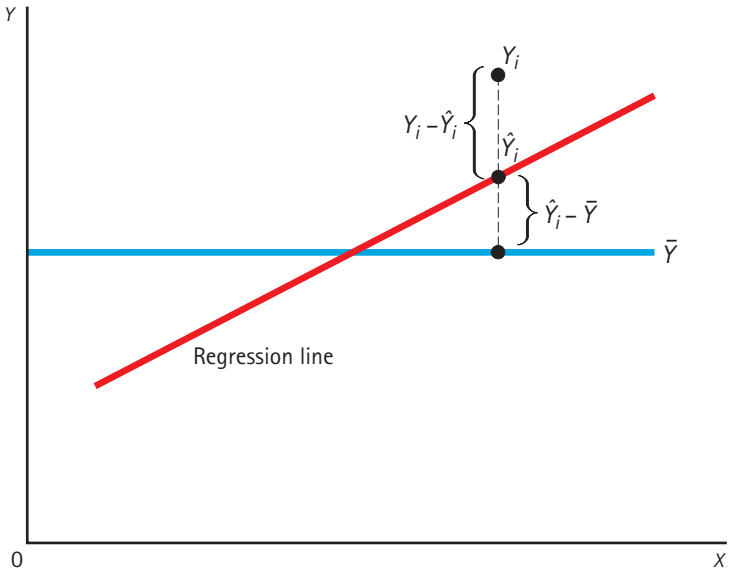
$$\begin{aligned}\sum_{i=1}^{n} (Y_i - \overline{Y})^2 &= \sum_{i=1}^{n} [(Y_i - \hat{Y}_i) + (\hat{Y}_i - \overline{Y})]^2 \\ &= \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2 + 2 \sum_{i=1}^{n} (Y_i - \hat{Y}_i)(\hat{Y}_i - \overline{Y})\end{aligned}$$

The last term on the right hand side equals zero, so equation (5.18) follows.

FIGURE 5.13

Division of $(Y_i - \bar{Y})$ into Two Parts: $(Y_i - \hat{Y}_i)$ and $(\hat{Y}_i - \bar{Y})$

This division is carried out to measure how well the regression line fits the data.



The term on the left-hand side of this equation shows the total variation in the dependent variable. The first term on the right-hand side measures the *variation in the dependent variable not explained by the regression*. This is a reasonable interpretation of this term, since it is the sum of squared deviations of the actual observations from the regression line. Clearly, the larger is the value of this term, the poorer the regression equation fits the data. The second term on the right-hand side of the equation measures the *variation in the dependent variable explained by the regression*. This is a reasonable interpretation of this term, since it shows how much the dependent variable would be expected to vary on the basis of the regression alone.

To measure the closeness of fit of a simple regression line, we use the **coefficient of determination**, which equals

$$1 - \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2} \tag{5.19}$$

In other words, the coefficient of determination equals

$$1 - \frac{\text{variation not explained by regression}}{\text{total variation}} = \frac{\text{variation explained by regression}}{\text{total variation}} \tag{5.20}$$

Clearly, the coefficient of determination is a reasonable measure of the closeness of fit of the regression line, since it equals the proportion of the total variation in the dependent variable explained by the regression line. The closer it is to 1, the better the fit; the closer it is to 0, the poorer the fit.

When a multiple regression is calculated, the multiple coefficient of determination is used to measure the goodness of fit of the regression. The multiple coefficient of determination is defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n} (Y_i - \overline{Y})^2} \tag{5.21}$$

where $\hat{Y}_i$ is the value of the dependent variable that is predicted from the regression equation. So, as in the case of the simple coefficient of determination covered earlier,

$$R^2 = \frac{\text{variation explained by regression}}{\text{total variation}} \tag{5.22}$$

This means that R^2 *measures the proportion of the total variation in the dependent variable explained by the regression equation.*