10 Simple Linear Regression and Correlation

10–1 Using Statistics 409

10–2 The Simple Linear Regression Model 411

10–3 Estimation: The Method of Least Squares 414

10–4 Error Variance and the Standard Errors of Regression Estimators 424

10–5 Correlation 429

10–6 Hypothesis Tests about the Regression Relationship 434

10–7 How Good Is the Regression? 438

10–8 Analysis-of-Variance Table and an *F* Test of the Regression Model 443

10–9 Residual Analysis and Checking for Model Inadequacies 445

10–10 Use of the Regression Model for Prediction 454

10–11 Using the Computer 458

10–12 Summary and Review of Terms 464

Case 13 Firm Leverage and Shareholder Rights 466

Case 14 Risk and Return 467

LEARNING OBJECTIVES

After studying this chapter, you should be able to:

- Determine whether a regression experiment would be useful in a given instance.
- Formulate a regression model.
- Compute a regression equation.
- Compute the covariance and the correlation coefficient of two random variables.
- Compute confidence intervals for regression coefficients.
- Compute a prediction interval for a dependent variable.
- Test hypotheses about regression coefficients.
- Conduct an ANOVA experiment using regression results.
- Analyze residuals to check the validity of assumptions about the regression model.
- Solve regression problems using spreadsheet templates.
- Use the LINEST function to carry out a regression.

10–1 Using Statistics

In 1855, a 33-year-old Englishman settled down to a life of leisure in London after several years of travel throughout Europe and Africa. The boredom brought about by a comfortable life induced him to write, and his first book was, naturally, *The Art of Travel.* As his intellectual curiosity grew, he shifted his interests to science and many years later published a paper on heredity, “Natural Inheritance” (1889). He reported his discovery that sizes of seeds of sweet pea plants appeared to “revert,” or “regress,” to the mean size in successive generations. He also reported results of a study of the relationship between heights of fathers and the heights of their sons. A straight line was fit to the data pairs: height of son versus height of father. Here, too, he found a “regression to mediocrity”: The heights of the sons represented a movement away from their fathers, toward the average height. The man was Sir Francis Galton, a cousin of Charles Darwin. We credit him with the idea of statistical regression.

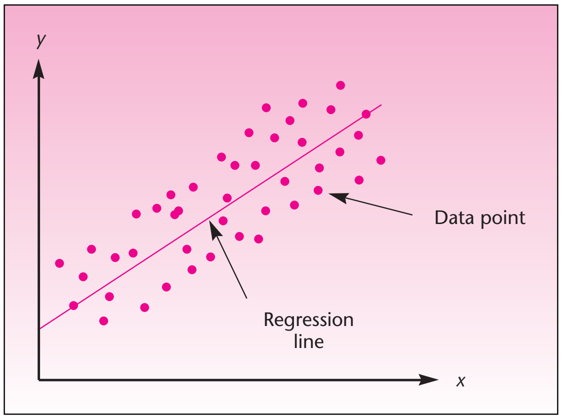
While most applications of regression analysis may have little to do with the “regression to the mean” discovered by Galton, the term **regression** remains. It now refers to the statistical technique of modeling the relationship between variables. In this chapter on **simple linear regression,** we model the relationship between two variables: a **dependent variable,** denoted by Y, and an **independent variable,** denoted by X. The model we use is a *straight-line relationship* between X and Y. When we model the relationship between the dependent variable Y and a set of several independent variables, or when the assumed relationship between Y and X is curved and requires the use of more terms in the model, we use a technique called *multiple regression.* This technique will be discussed in the next chapter.

Figure 10–1 is a general example of simple linear regression: fitting a straight line to describe the relationship between two variables X and Y. The points on the graph are randomly chosen observations of the two variables X and Y, and the straight line describes the general *movement* in the data–an increase in Y corresponding to an increase in X. An inverse straight-line relationship is also possible, consisting of a general decrease in Y as X increases (in such cases, the slope of the line is negative).

Regression analysis is one of the most important and widely used statistical techniques and has many applications in business and economics. A firm may be interested in estimating the relationship between advertising and sales (one of the most important topics of research in the field of marketing). Over a short range of values–when advertising is not yet overdone, giving diminishing returns–the relationship between advertising and sales may be well approximated by a straight line. The X variable in Figure 10–1 could denote advertising expenditure, and the Y variable could stand for the resulting sales for the same period. The data points in this case would be pairs of observations of the form $x_1 = \$75{,}570$, $y_1 = 134{,}679$ units; $x_2 = \$83{,}090$, $y_2 = 151{,}664$ units; etc. That is, the first month the firm spent $75,570 on advertising, and sales for the month were 134,679 units; the second month the company spent $83,090 on advertising, with resulting sales of 151,664 units for that month; and so on for the entire set of available data.

The data pairs, values of X paired with corresponding values of Y, are the points shown in a sketch of the data (such as Figure 10–1). A sketch of data on two variables is called a **scatter plot.** In addition to the scatter plot, Figure 10–1 shows the straight line believed to best show how the general trend of increasing sales corresponds, in this example, to increasing advertising expenditures. This chapter will teach you how to find the best line to fit a data set and how to use the line once you have found it.

FIGURE 10–1 **Simple Linear Regression**



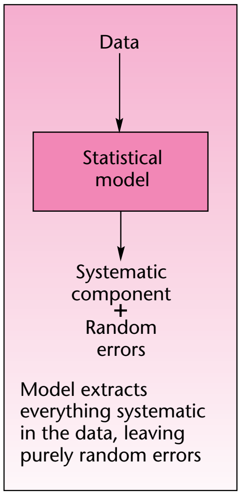
Although, in reality, our sample may consist of all available information on the two variables under study, we always assume that our data set constitutes a random sample of observations from a population of possible pairs of values of *X* and *Y*. Incidentally, in our hypothetical advertising sales example, we assume no carryover effect of advertising from month to month; every month's sales depend only on that month's level of advertising. Other common examples of the use of simple linear regression in business and economics are the modeling of the relationship between job performance (the dependent variable *Y*) and extent of training (the independent variable *X*); the relationship between returns on a stock (*Y*) and the riskiness of the stock (*X*); and the relationship between company profits (*Y*) and the state of the economy (*X*).

Model Building

Like the analysis of variance, both simple linear regression and multiple regression are *statistical models*. Recall that a statistical model is a set of mathematical formulas and assumptions that describe a real-world situation. We would like our model to explain as much as possible about the process underlying our data. However, due to the uncertainty inherent in all real-world situations, our model will probably not explain everything, and we will always have some remaining errors. The errors are due to unknown outside factors that affect the process generating our data.

A good statistical model is *parsimonious,* which means that it uses as few mathematical terms as possible to describe the real situation. The model captures the systematic behavior of the data, leaving out the factors that are nonsystematic and cannot be foreseen or predicted–the errors. The idea of a good statistical model is illustrated in Figure 10–2. The errors, denoted by ϵ, constitute the random component in the model. In a sense, the statistical model breaks down the data into a nonrandom, systematic component, which can be described by a formula, and a purely random component.

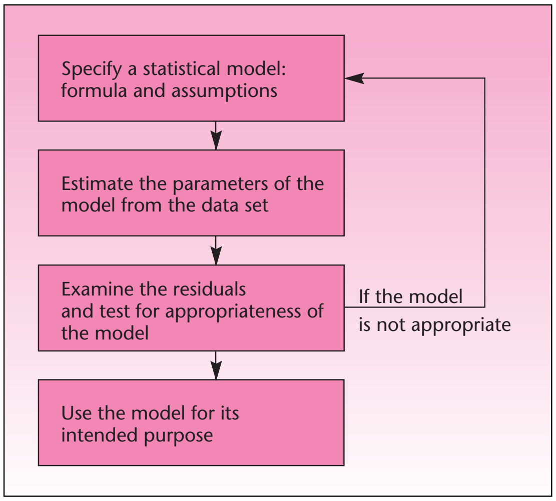
FIGURE 10–2
A Statistical Model



Model extracts everything systematic in the data, leaving purely random errors

How do we deal with the errors? This is where probability theory comes in. Since our model, we hope, captures everything systematic in the data, the remaining random errors are probably due to a large number of minor factors that we cannot trace. We assume that the random errors ϵ are *normally distributed.* If we have a properly constructed model, the resulting observed errors will have an average of zero (although few, if any, will actually equal zero), and they should also be *independent* of one another. We note that the assumption of a normal distribution of the errors is not absolutely necessary in the regression model. The assumption is made so that we can carry out statistical hypothesis tests using the *F* and *t* distributions. The only necessary assumption is that the errors ϵ have mean zero and a constant variance σ^2 and that they be uncorrelated with one another. In the

FIGURE 10–3 **Steps in Building a Statistical Model**


Specify a statistical model: formula and assumptions
Estimate the parameters of the model from the data set
Examine the residuals and test for appropriateness of the model
If the model is not appropriate
Use the model for its intended purpose

next section, we describe the simple linear regression model. We now present a general model-building methodology.

First, we propose a particular model to describe a given situation. For example, we may propose a simple linear regression model for describing the relationship between two variables. Then we estimate the model parameters from the random sample of data we have. The next step is to consider the observed errors resulting from the fit of the model to the data. These observed errors, called **residuals,** represent the information in the data not explained by the model. For example, in the ANOVA model discussed in Chapter 9, the within-group variation (leading to SSE and MSE) is due to the residuals. If the residuals are found to contain some nonrandom, *systematic* component, we reevaluate our proposed model and, if possible, adjust it to incorporate the systematic component found in the residuals; or we may have to discard the model and try another. When we believe that model residuals contain nothing more than pure randomness, we use the model for its intended purpose: *prediction* of a variable, *control* of a variable, or the *explanation* of the relationships among variables.

In the advertising sales example, once the regression model has been estimated and found to be appropriate, the firm may be able to use the model for predicting sales for a given level of advertising within the range of values studied. Using the model, the firm may be able to control its sales by setting the level of advertising expenditure. The model may help explain the effect of advertising on sales within the range of values studied. Figure 10–3 shows the usual steps of building a statistical model.

10–2 The Simple Linear Regression Model

Recall from algebra that the equation of a straight line is $Y = A + BX$, where A is the Y intercept and B is the slope of the line. In simple linear regression, we model the relationship between two variables X and Y as a straight line. Therefore, our model must contain two parameters: an intercept parameter and a slope parameter. The usual notation for the **population intercept** is β_0, and the notation for the **population slope** is β_1. If we include the error term ϵ, the population regression model is given in equation 10–1.

CHAPTER 15

The population simple linear regression model is

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{10–1}$$

where Y is the dependent variable, the variable we wish to explain or predict; X is the independent variable, also called the *predictor* variable; and ϵ is the error term, the only random component in the model and thus the only source of randomness in Y.

The model parameters are as follows:

β_0 is the Y intercept of the straight line given by $Y = \beta_0 + \beta_1 X$ (the line does not contain the error term).

β_1 is the slope of the line $Y = \beta_0 + \beta_1 X$.

FIGURE 10–4
Simple Linear Regression Model

$$Y = \underbrace{\beta_0 + \beta_1 X}_{\text{Nonrandom component: straight line}} + \underbrace{\epsilon}_{\text{Random error}}$$

The simple linear regression model of equation 10–1 is composed of two components: a nonrandom component, which is the line itself, and a purely random component—the error term ϵ. This is shown in Figure 10–4. The nonrandom part of the model, the straight line, is the equation for the *mean of Y, given X*. We denote the conditional mean of Y, given X, by $E(Y \mid X)$. Thus, if the model is correct, the *average* value of Y for a given value of X falls right *on* the regression line. The equation for the mean of Y, given X, is given as equation 10–2.

The conditional mean of Y is

$$E(Y \mid X) = \beta_0 + \beta_1 X \tag{10–2}$$

Comparing equations 10–1 and 10–2, we see that our model says that each value of Y comprises the average Y for the given value of X (this is the straight line), plus a random error. We will sometimes use the simplified notation $E(Y)$ for the line, remembering that this is the *conditional* mean of Y for a given value of X. As X increases, the average population value of Y also increases, assuming a positive slope of the line (or decreases, if the slope is negative). The *actual* population value of Y is equal to the average Y conditional on X, plus a random error ϵ. We thus have, for a given value of X,

$$Y = \text{Average } Y \text{ for given } X + \text{Error}$$

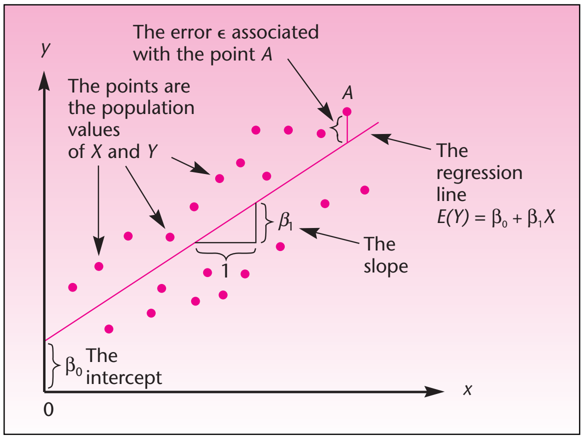
Figure 10–5 shows the population regression model.

We now state the assumptions of the simple linear regression model.

Model assumptions:

1. The relationship between X and Y is a straight-line relationship.
2. The values of the independent variable X are assumed fixed (not random); the only randomness in the values of Y comes from the error term ϵ.

FIGURE 10–5 **Population Regression Line**


y
The error ϵ associated with the point A
A
The points are the population values of X and Y
The regression line E(Y) = β0 + β1X
β1
1
The slope
β0 The intercept
0
x

3. The errors ϵ are normally distributed with mean 0 and a constant variance σ^2. The errors are uncorrelated (not related) with one another in successive observations.[1] In symbols:

$$\epsilon \sim N(0, \sigma^2) \qquad (10\text{–}3)$$

FIGURE 10–6 **Distributional Assumptions of the Linear Regression Model**

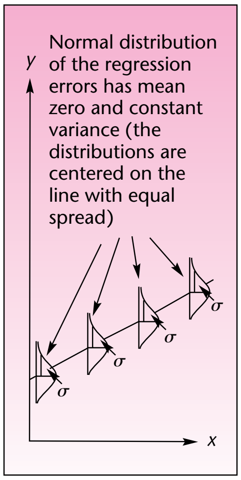

y
Normal distribution of the regression errors has mean zero and constant variance (the distributions are centered on the line with equal spread)
σ
σ
σ
σ
x

Figure 10–6 shows the distributional assumptions of the errors of the simple linear regression model. The population regression errors are normally distributed about the population regression line, with mean zero and equal variance. (The errors are equally spread about the regression line; the error variance does not increase or decrease as *X* increases.)

The simple linear regression model applies only if the true relationship between the two variables *X* and *Y* is a straight-line relationship. If the relationship is curved (*curvilinear*), then we need to use the more involved methods of the next chapter. In Figure 10–7, we show various relationships between two variables. Some are straight-line relationships that can be modeled by simple linear regression, and others are not.

So far, we have described the population model, that is, the assumed true relationship between the two variables *X* and *Y*. Our interest is focused on this unknown population relationship, and we want to *estimate* it, using sample information. We obtain a random sample of observations on the two variables, and we estimate the regression model parameters β_0 and β_1 from this sample. This is done by the *method of least squares,* which is discussed in the next section.

PROBLEMS

10–1. What is a statistical model?

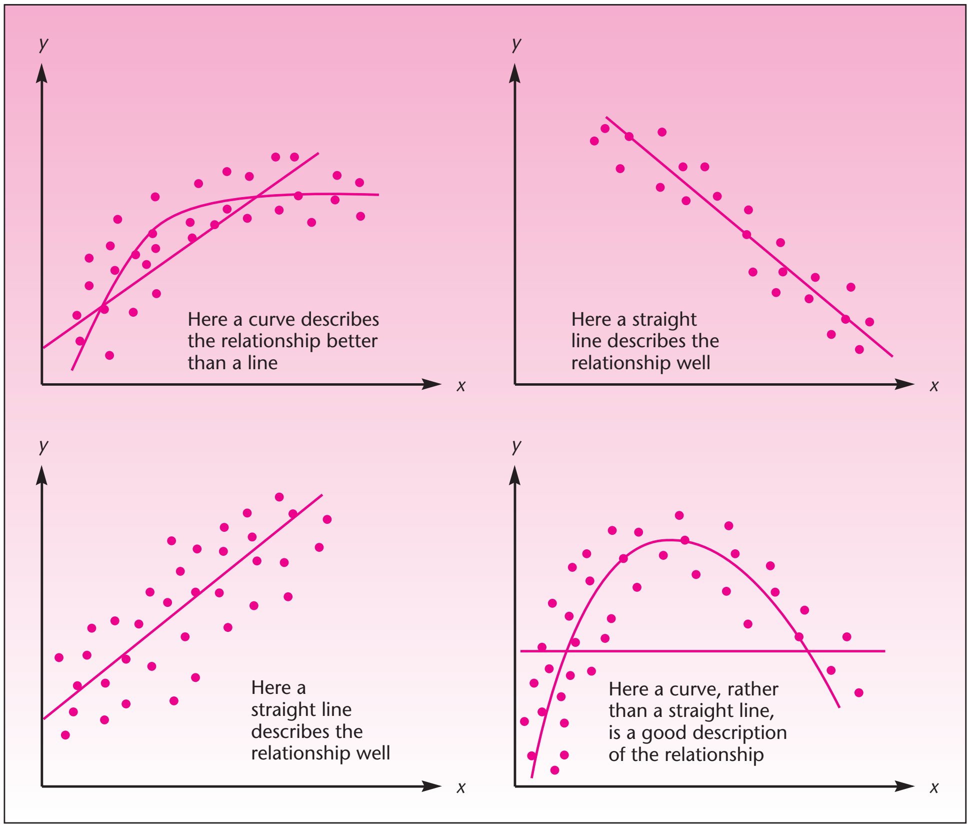
10–2. What are the steps of statistical model building?

10–3. What are the assumptions of the simple linear regression model?

10–4. Define the parameters of the simple linear regression model.

[1] The idea of statistical *correlation* will be discussed in detail in Section 10–5. In the case of the regression errors, we assume that successive errors $\epsilon_1, \epsilon_2, \epsilon_3, \ldots$ are uncorrelated: they are not related with one another; there is no trend, no joint movement in successive errors. Incidentally, the assumption of zero correlation together with the assumption of a normal distribution of the errors implies the assumption that the errors are independent of one another. Independence implies noncorrelation, but noncorrelation does not imply independence, except in the case of a normal distribution (this is a technical point).

FIGURE 10–7 Some Possible Relationships between *X* and *Y*



10–5. What is the conditional mean of Y, given X?

10–6. What are the uses of a regression model?

10–7. What are the purpose and meaning of the error term in regression?

10–8. A simple linear regression model was used for predicting the success of private-label products, which, according to the authors of the study, now account for 20% of global grocery sales, and the per capita gross domestic product for the country at which the private-label product is sold.[2] The regression equation is given as

$$\text{PLS} = \beta\ \text{GDPC} + \epsilon$$

where PLS = private label success, GDPC = per capita gross domestic product, β = regression slope, and ϵ = error term. What kind of regression model is this?

10–3 Estimation: The Method of Least Squares

CHAPTER 15

We want to find good estimates of the regression parameters β_0 and β_1. Remember the properties of good estimators, discussed in Chapter 5. Unbiasedness and efficiency are among these properties. A method that will give us good estimates of the regression

[2] Lien Lamey et al., "How Business Cycles Contribute to Private-Label Success: Evidence from the United States and Europe," *Journal of Marketing* 71 (January 2007), pp. 1–15.

coefficients is the **method of least squares.** The method of least squares gives us the *best linear unbiased estimators* (BLUE) of the regression parameters β_0 and β_1. These estimators both are unbiased and have the lowest variance of all possible unbiased estimators of the regression parameters. These properties of the least-squares estimators are specified by a well-known theorem, the *Gauss-Markov theorem*. We denote the least-squares estimators by b_0 and b_1.

The least-squares estimators are

$$b_0 \xrightarrow{\text{estimates}} \beta_0$$

$$b_1 \xrightarrow{\text{estimates}} \beta_1$$

The estimated regression equation is

$$Y = b_0 + b_1X + e \qquad (10\text{–}4)$$

where b_0 estimates β_0, b_1 estimates β_1, and e stands for the observed errors—the residuals from fitting the line $b_0 + b_1X$ to the data set of n points.

In terms of the data, equation 10–4 can be written with the subscript i to signify each particular data point:

$$y_i = b_0 + b_1x_i + e_i \qquad (10\text{–}5)$$

where $i = 1, 2, \ldots, n$. Then e_1 is the first residual, the distance from the first data point to the fitted regression line; e_2 is the distance from the second data point to the line; and so on to e_n, the nth error. The errors e_i are viewed as estimates of the true population errors ϵ_i. The equation of the regression line itself is as follows:

The regression line is

$$\hat{Y} = b_0 + b_1X \qquad (10\text{–}6)$$

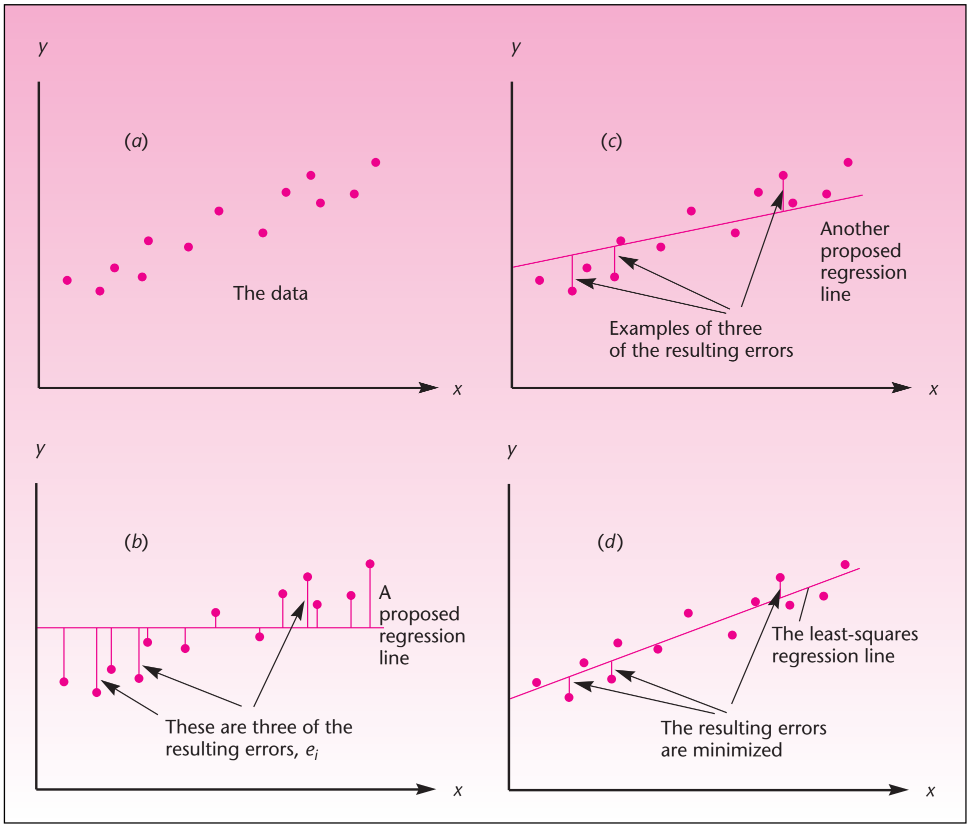
where $\hat{Y}$ (pronounced "Y hat") is the Y value *lying on the fitted regression line* for a given X.

Thus, $\hat{y}_1$ is the fitted value corresponding to x_1, that is, the value of y_1 without the error e_1, and so on for all $i = 1, 2, \ldots, n$. The fitted value Y is also called the *predicted value of* $\hat{Y}$ because if we do not know the actual value of Y, it is the value we would predict for a given value of X, using the estimated regression line.

Having defined the estimated regression equation, the errors, and the fitted values of Y, we will now demonstrate the principle of least squares, which gives us the BLUE regression parameters. Consider the data set shown in Figure 10–8(a). In parts (b), (c), and (d) of the figure, we show different lines passing through the data set and the resulting errors e_i.

As can be seen from Figure 10–8, the regression line proposed in part (b) results in very large errors. The errors corresponding to the line of part (c) are smaller than the ones of part (b), but the errors resulting from using the line proposed in part (d) are by far the smallest. The line in part (d) seems to move with the data and *minimize* the resulting errors. This should convince you that the line that best describes the trend in the data is the line that lies "inside" the set of

FIGURE 10–8 A Data Set of *X* and *Y* Pairs, and Different Proposed Straight Lines to Describe the Data



points; since some of the points lie above the fitted line and others below the line, some errors will be positive and others will be negative. If we want to minimize all the errors (both positive and negative ones), we should minimize the *sum of the squared errors* (SSE, as in ANOVA). Thus, we want to find the *least-squares* line–the line that minimizes SSE. We note that least squares is not the only method of fitting lines to data; other methods include minimizing the sum of the absolute errors. The method of least squares, however, is the most commonly used method to estimate a regression relationship. Figure 10–9 shows how the errors lead to the calculation of SSE.

We define the sum of squares for error in regression as

$$\text{SSE} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad (10\text{–}7)$$

Figure 10–10 shows different values of SSE corresponding to values of b_0 and b_1. The least-squares line is the particular line specified by values of b_0 and b_1 that minimize SSE, as shown in the figure.

FIGURE 10–9 Regression Errors Leading to SSE

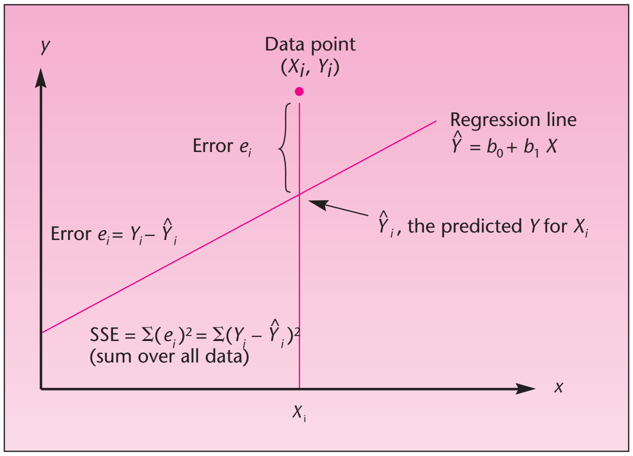
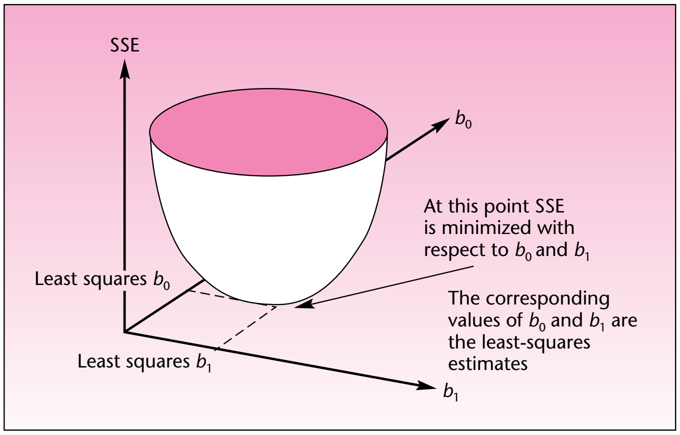

y
Data point
(X_i, Y_i)
Error e_i
Regression line
$\hat{Y} = b_0 + b_1 X$
Error $e_i = Y_i - \hat{Y}_i$
$\hat{Y}_i$, the predicted Y for X_i
SSE = $\Sigma(e_i)^2 = \Sigma(Y_i - \hat{Y}_i)^2$
(sum over all data)
x
X_i

FIGURE 10–10 The Particular Values b_0 and b_1 That Minimize SSE



Calculus is used in finding the expressions for b_0 and b_1 that minimize SSE. These expressions are called the *normal equations* and are given as equations 10–8.[3] This system of two equations with two unknowns is solved to give us the values of b_0 and b_1 that minimize SSE. The results are the least-squares estimators b_0 and b_1 of the simple linear regression parameters β_0 and β_1.

The **normal equations** are

$$\sum_{i=1}^{n} y_i = nb_0 + b_1 \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i y_i = b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 \qquad (10\text{–}8)$$

[3] We leave it as an exercise to the reader with background in calculus to derive the normal equations by taking the partial derivatives of SSE with respect to b_0 and b_1 and setting them to zero.

Before we present the solutions to the normal equations, we define the sums of squares SS_X and SS_Y and the sum of the cross-products SS_{XY}. These will be very useful in defining the least-squares estimates of the regression parameters, as well as in other regression formulas we will see later. The definitions are given in equations 10–9.

Definitions of sums of squares and cross-products useful in regression analysis:

$$SS_x = \sum(x - \bar{x})^2 = \sum x^2 - \frac{(\sum x)^2}{n}$$

$$SS_y = \sum(y - \bar{y})^2 = \sum y^2 - \frac{(\sum y)^2}{n}$$

$$SS_{xy} = \sum(x - \bar{x})(y - \bar{y}) = \sum xy - \frac{(\sum x)(\sum y)}{n} \qquad (10\text{–}9)$$

The first definition in each case is the conceptual one using squared distances from the mean; the second part is a computational definition. Summations are over all data.

We now give the solutions of the normal equations, the least-squares estimators b_0 and b_1.

Least-squares regression estimators include the slope

$$b_1 = \frac{SS_{xy}}{SS_x}$$

and the intercept

$$b_0 = \bar{y} - b_1\bar{x} \qquad (10\text{–}10)$$

The formula for the estimate of the intercept makes use of the fact that the *least-squares line always passes through the point* $(\bar{x}, \bar{y})$, the intersection of the mean of X and the mean of Y.

Remember that the obtained estimates b_0 and b_1 of the regression relationship are just realizations of *estimators* of the true regression parameters β_0 and β_1. As always, our estimators have standard deviations (and variances, which, by the Gauss-Markov theorem, are as small as possible). The estimates can be used, along with the assumption of normality, in the construction of confidence intervals for, and the conducting of hypothesis tests about, the true regression parameters β_0 and β_1. This will be done in the next section.

We demonstrate the process of estimating the parameters of a simple linear regression model in Example 10–1.

EXAMPLE 10–1

American Express Company has long believed that its cardholders tend to travel more extensively than others—both on business and for pleasure. As part of a comprehensive research effort undertaken by a New York market research firm on behalf of American Express, a study was conducted to determine the relationship between travel and charges on the American Express card. The research firm selected a random sample of 25 cardholders from the American Express computer file and recorded their total charges over a specified period. For the selected cardholders, information was also obtained, through a mailed questionnaire, on the total number of miles traveled by each cardholder during the same period. The data for this study are given in Table 10–1. Figure 10–11 is a scatter plot of the data.

Solution

As can be seen from the figure, it seems likely that a straight line will describe the trend of increase in dollar amount charged with increase in number of miles traveled. The least-squares line that fits these data is shown in Figure 10–12.

We will now show how the least-squares regression line in Figure 10–12 is obtained. Table 10–2 shows the necessary computations. From equations 10–9, using sums at the bottom of Table 10–2, we get

$$\text{SS}_X = \sum x^2 - \frac{(\sum x)^2}{n} = 293{,}426{,}946 - \frac{79{,}448^2}{25} = 40{,}947{,}557.84$$

and

$$\text{SS}_{XY} = \sum xy - \frac{(\sum x)(\sum y)}{n} = 390{,}185{,}014 - \frac{(79{,}448)(106{,}605)}{25} = 51{,}402{,}852.4$$

TABLE 10–1 American Express Study Data

Miles	Dollars
1,211	1,802
1,345	2,405
1,422	2,005
1,687	2,511
1,849	2,332
2,026	2,305
2,133	3,016
2,253	3,385
2,400	3,090
2,468	3,694
2,699	3,371
2,806	3,998
3,082	3,555
3,209	4,692
3,466	4,244
3,643	5,298
3,852	4,801
4,033	5,147
4,267	5,738
4,498	6,420
4,533	6,059
4,804	6,426
5,090	6,321
5,233	7,026
5,439	6,964

FIGURE 10–11 Data for the American Express Study

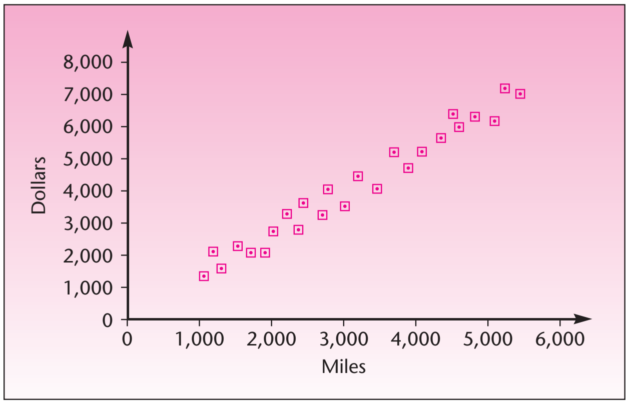


FIGURE 10–12 Least-Squares Line for the American Express Study

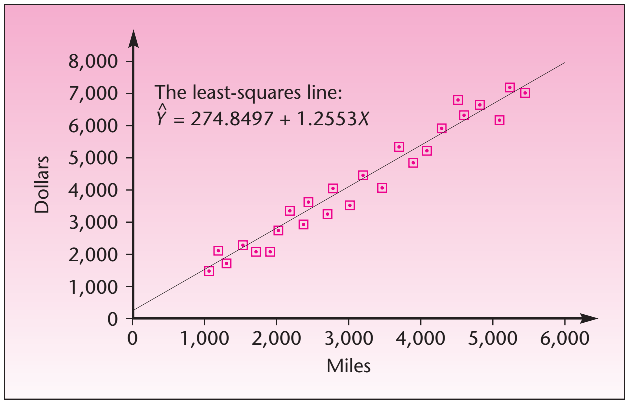


TABLE 10–2 The Computations Required for the American Express Study

Miles *X*	Dollars *Y*	X^2	Y^2	XY
1,211	1,802	1,466,521	3,247,204	2,182,222
1,345	2,405	1,809,025	5,784,025	3,234,725
1,422	2,005	2,022,084	4,020,025	2,851,110
1,687	2,511	2,845,969	6,305,121	4,236,057
1,849	2,332	3,418,801	5,438,224	4,311,868
2,026	2,305	4,104,676	5,313,025	4,669,930
2,133	3,016	4,549,689	9,096,256	6,433,128
2,253	3,385	5,076,009	11,458,225	7,626,405
2,400	3,090	5,760,000	9,548,100	7,416,000
2,468	3,694	6,091,024	13,645,636	9,116,792
2,699	3,371	7,284,601	11,363,641	9,098,329
2,806	3,998	7,873,636	15,984,004	11,218,388
3,082	3,555	9,498,724	12,638,025	10,956,510
3,209	4,692	10,297,681	22,014,864	15,056,628
3,466	4,244	12,013,156	18,011,536	14,709,704
3,643	5,298	13,271,449	28,068,804	19,300,614
3,852	4,801	14,837,904	23,049,601	18,493,452
4,033	5,147	16,265,089	26,491,609	20,757,851
4,267	5,738	18,207,289	32,924,644	24,484,046
4,498	6,420	20,232,004	41,216,400	28,877,160
4,533	6,059	20,548,089	36,711,481	27,465,447
4,804	6,426	23,078,416	41,293,476	30,870,504
5,090	6,321	25,908,100	39,955,041	32,173,890
5,233	7,026	27,384,289	49,364,676	36,767,058
5,439	6,964	29,582,721	48,497,296	37,877,196
79,448	106,605	293,426,946	521,440,939	390,185,014

Using equations 10–10 for the least-squares estimates of the slope and intercept parameters, we get

$$b_1 = \frac{SS_{XY}}{SS_X} = \frac{51{,}402{,}852.40}{40{,}947{,}557.84} = 1.255333776$$

and

$$b_0 = \bar{y} - b_1\bar{x} = \frac{106{,}605}{25} - 1.2553337776\left(\frac{79{,}448}{25}\right) = 274.8496866$$

Always carry out as many significant digits as you can in these computations. Here we carried out the computations by hand, for demonstration purposes. Usually, all computations are done by computer or by calculator. There are many hand calculators with a built-in routine for simple linear regression. From now on, we will present only the computed results, the least-squares estimates. The estimated least-squares relationship for Example 10–1 is reporting estimates to the second significant decimal:

$$Y = 274.85 + 1.26X + e \qquad (10\text{–}11)$$

The equation of the line itself, that is, the predicted value of Y for a given X, is

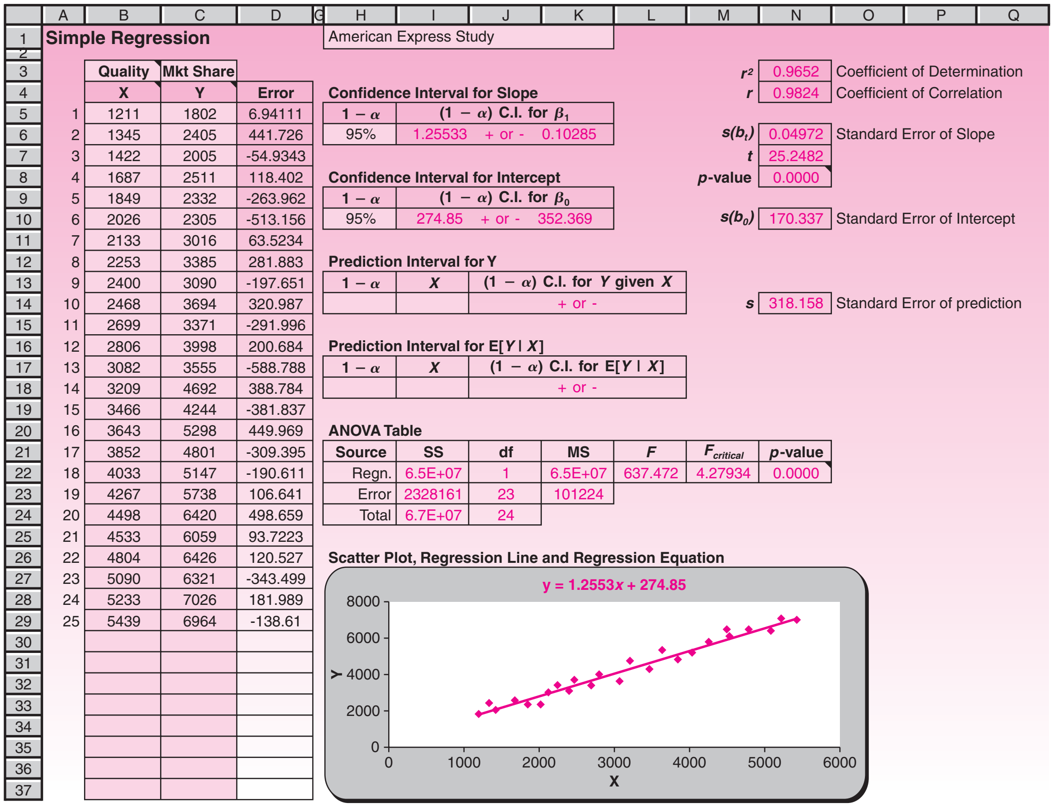
$$\hat{Y} = 274.85 + 1.26X \qquad (10\text{–}12)$$

The Template

Figure 10–13 shows the template that can be used to carry out a simple regression. The X and Y data are entered in columns B and C. The scatter plot at the bottom shows the regression equation and the regression line. Several additional statistics regarding the regression appear in the remaining parts of the template; these are explained in later sections. The error values appear in column D.

Below the scatter plot is a panel for residual analysis. Here you will find the Durbin-Watson statistic, the residual plot, and the normal probability plot. The Durbin-Watson statistic will be explained in the next chapter, and the normal probability plot will be explained later in this chapter. The residual plot shows that there is no relationship between X and the residuals. Figure 10–14 shows the panel.

FIGURE 10–13 The Simple Regression Template
[Simple Regression.xls; Sheet: Regression]



Simple Regression

American Express Study

	Quality X	Mkt Share Y	Error
1	1211	1802	6.94111
2	1345	2405	441.726
3	1422	2005	-54.9343
4	1687	2511	118.402
5	1849	2332	-263.962
6	2026	2305	-513.156
7	2133	3016	63.5234
8	2253	3385	281.883
9	2400	3090	-197.651
10	2468	3694	320.987
11	2699	3371	-291.996
12	2806	3998	200.684
13	3082	3555	-588.788
14	3209	4692	388.784
15	3466	4244	-381.837
16	3643	5298	449.969
17	3852	4801	-309.395
18	4033	5147	-190.611
19	4267	5738	106.641
20	4498	6420	498.659
21	4533	6059	93.7223
22	4804	6426	120.527
23	5090	6321	-343.499
24	5233	7026	181.989
25	5439	6964	-138.61

r^2	0.9652	Coefficient of Determination
r	0.9824	Coefficient of Correlation
$s(b_1)$	0.04972	Standard Error of Slope
t	25.2482	
p-value	0.0000	
$s(b_0)$	170.337	Standard Error of Intercept
s	318.158	Standard Error of prediction

Confidence Interval for Slope

$1-\alpha$	$(1-\alpha)$ C.I. for β_1
95%	1.25533 + or - 0.10285

Confidence Interval for Intercept

$1-\alpha$	$(1-\alpha)$ C.I. for β_0
95%	274.85 + or - 352.369

Prediction Interval for Y

$1-\alpha$	X	$(1-\alpha)$ C.I. for Y given X
		+ or -

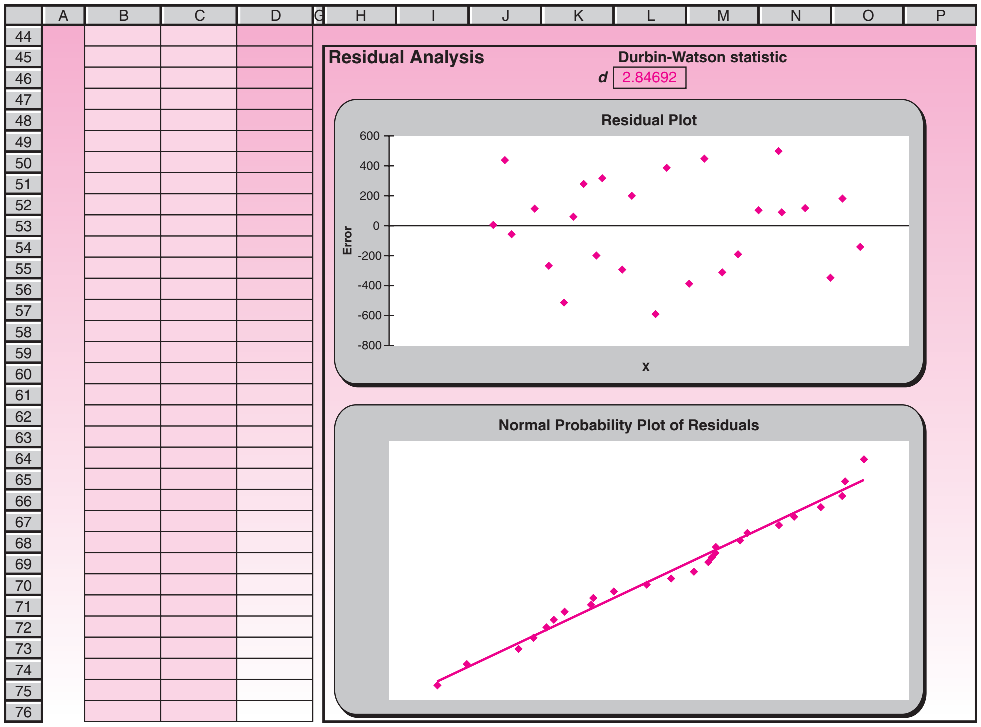
Prediction Interval for E[Y | X]

$1-\alpha$	X	$(1-\alpha)$ C.I. for E[Y | X]
		+ or -

ANOVA Table

Source	SS	df	MS	F	$F_{critical}$	p-value
Regn.	6.5E+07	1	6.5E+07	637.472	4.27934	0.0000
Error	2328161	23	101224			
Total	6.7E+07	24				

Scatter Plot, Regression Line and Regression Equation



Durbin-Watson statistic
d 2.84692

FIGURE 10–14 Residual Analysis in the Template [Simple Regression.xls; Sheet: Regression]

PROBLEMS

10–9. Explain the advantages of the least-squares procedure for fitting lines to data. Explain how the procedure works.

10–10. (A conceptually advanced problem) Can you think of a possible limitation of the least-squares procedure?

10–11. An article in the *Journal of Monetary Economics* assesses the relationship between percentage growth in wealth over a decade and a half of savings for baby boomers of age 40 to 55 with these people's income quartiles. The article presents a table showing five income quartiles, and for each quartile there is a reported percentage growth in wealth. The data are as follows.[4]

Income quartile:	1	2	3	4	5
Wealth growth (%):	17.3	23.6	40.2	45.8	56.8

Run a simple linear regression of these five pairs of numbers and estimate a linear relationship between income and percentage growth in wealth.

10-12. A financial analyst at Goldman Sachs ran a regression analysis of monthly returns on a certain investment (Y) versus returns for the same month on the Standard & Poor's index (X). The regression results included $\mathrm{SS}_X = 765.98$ and $\mathrm{SS}_{XY} = 934.49$. Give the least-squares estimate of the regression slope parameter.

[4] Edward N. Wolff, "The Retirement Wealth of the Baby Boom Generation," *Journal of Monetary Economics* 54 (January 2007), pp. 1–40.

10–13. Recently, research efforts have focused on the problem of predicting a manufacturer's market share by using information on the quality of its product. Suppose that the following data are available on market share, in percentage (Y), and product quality, on a scale of 0 to 100, determined by an objective evaluation procedure (X):

X:	27	39	73	66	33	43	47	55	60	68	70	75	82
Y:	2	3	10	9	4	6	5	8	7	9	10	13	12

Estimate the simple linear regression relationship between market share and product quality rating.

10–14. A pharmaceutical manufacturer wants to determine the concentration of a key component of cough medicine that may be used without the drug's causing adverse side effects. As part of the analysis, a random sample of 45 patients is administered doses of varying concentration (X), and the severity of side effects (Y) is measured. The results include $\bar{x} = 88.9$, $\bar{y} = 165.3$, $SS_X = 2{,}133.9$, $SS_{XY} = 4{,}502.53$, $SS_Y = 12{,}500$. Find the least-squares estimates of the regression parameters.

10–15. The following are data on annual inflation and stock returns. Run a regression analysis of the data and determine whether there is a linear relationship between inflation and total return on stocks for the periods under study.

Inflation (%)	Total Return on Stocks (%)
1	–3
2	36
12.6	12
–10.3	–8
0.51	53
2.03	–2
–1.8	18
5.79	32
5.87	24

10–16. An article in *Worth* discusses the immense success of one of the world's most prestigious cars, the Aston Martin Vanquish. This car is expected to keep its value as it ages. Although this model is new, the article reports resale values of earlier Aston Martin models over various decades.

Decade:	1960s	1970s	1980s	1990s	2000s
Present value of Aston Martin model (average):	\$180,000	\$40,000	\$60,000	\$160,000	\$200,000

Based on these limited data, is there a relationship between age and average price of an Aston Martin? What are the limitations of this analysis? Can you think of some hidden variables that could affect what you are seeing in the data?

10–17. For the data given below, regress one variable on the other. Is there an implication of causality, or are both variables affected by a third?

	Sample of Annual Transactions ($ millions)	
Year	Credit Card	Online Debit Card
2002	156	211
2003	204	280
2004	279	386
2005	472	551
2006	822	684
2007	1,213	905

10–18. (A problem requiring knowledge of calculus) Derive the normal equations (10–8) by taking the partial derivatives of SSE with respect to b_0 and b_1 and setting them to zero. [*Hint:* Set SSE $= \Sigma e^2 = \Sigma(y - \hat{y})^2 = \Sigma(y - b_0 - b_1 x)^2$, and take the derivatives of the last expression on the right.]

10–4 Error Variance and the Standard Errors of Regression Estimators

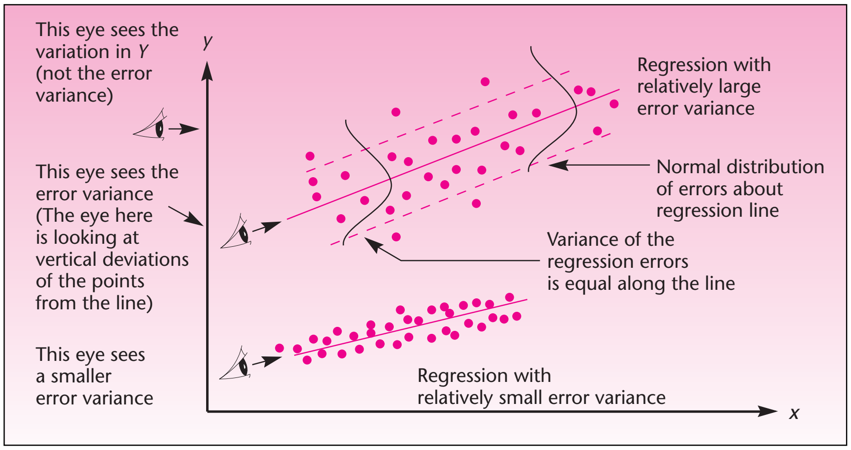
CHAPTER 15

Recall that σ^2 is the variance of the population regression errors ϵ and that this variance is assumed to be constant for all values of X in the range under study. The error variance is an important parameter in the context of regression analysis because it is a measure of the spread of the population elements about the regression line. Generally, the smaller the error variance, the more closely the population elements follow the regression line. The error variance is the variance of the dependent variable Y as “seen” by an eye looking in the direction of the regression line (the error variance is not the variance of Y). These properties are demonstrated in Figure 10–15.

The figure shows two regression lines. The top regression line in the figure has a larger error variance than the bottom regression line. The error variance for each regression is the variation in the data points as seen by the eye located at the base of the line, looking *in the direction of the regression line.* The variance of Y, on the other hand, is the variation in the Y values regardless of the regression line. That is, the variance of Y for each of the two data sets in the figure is the variation in the data as seen by an eye looking in a direction parallel to the X axis. Note also that the spread of the data is constant along the regression lines. This is in accordance with our assumption of equal error variance for all X.

Since σ^2 is usually unknown, we need to estimate it from our data. An unbiased estimator of σ^2, denoted by S^2, is the *mean square error* (*MSE*) of the regression. As you will soon see, sums of squares and mean squares in the context of regression analysis are very similar to those of ANOVA, presented in the preceding chapter. The degrees of freedom for error in the context of simple linear regression are $n - 2$ because we have n data points, from which two parameters, β_0 and β_1, are estimated (thus, two restrictions are imposed on the n points, leaving df $= n - 2$). The sum of squares for error (SSE) in regression analysis is defined as the sum of squared deviations of the data values Y from the fitted values $\hat{Y}$. The sum of squares for error may also be defined in terms of a computational formula using SS_X, SS_Y, and SS_{XY} as defined in equations 10–9. We state these relationships in equations 10–13.

FIGURE 10-15 **Two Examples of Regression Lines Showing the Error Variance**


This eye sees the variation in Y (not the error variance)
y
Regression with relatively large error variance
This eye sees the error variance (The eye here is looking at vertical deviations of the points from the line)
Normal distribution of errors about regression line
Variance of the regression errors is equal along the line
This eye sees a smaller error variance
Regression with relatively small error variance
x

$$\begin{aligned}\text{df(error)} &= n - 2\\ \text{SSE} &= \sum (Y - \hat{Y})^2\\ &= \text{SS}_Y - \frac{(\text{SS}_{XY})^2}{\text{SS}_X}\\ &= \text{SS}_Y - b_1\text{SS}_{XY} \qquad (10\text{–}13)\end{aligned}$$

An unbiased estimator of σ^2, denoted by S^2, is

$$\text{MSE} = \frac{\text{SSE}}{n - 2}$$

In Example 10–1, the sum of squares for error is

$$\begin{aligned}\text{SSE} &= \text{SS}_Y - b_1\text{SS}_{XY} = 66{,}855{,}898 - (1.255333776)(51{,}402{,}852.4)\\ &= 2{,}328{,}161.2\end{aligned}$$

and

$$\text{MSE} = \frac{\text{SSE}}{n - 2} = \frac{2{,}328{,}161.2}{23} = 101{,}224.4$$

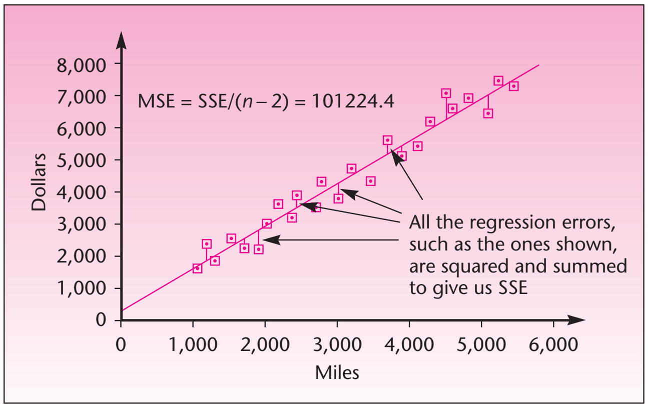
An estimate of the standard deviation of the regression errors σ is s, which is the square root of MSE. (The estimator S is not unbiased because the square root of an unbiased estimator, such as S^2, is not itself unbiased. The bias, however, is small, and the point is a technical one.) The estimate $s = \sqrt{\text{MSE}}$ of the standard deviation of the regression errors is sometimes referred to as *standard error of estimate.* In Example 10–1 we have

$$s = \sqrt{\text{MSE}} = \sqrt{101{,}224.4} = 318.1578225$$

The computation of SSE and MSE for Example 10–1 is demonstrated in Figure 10–16.

The standard deviation of the regression errors σ and its estimate s play an important role in the process of estimation of the values of the regression parameters β_0 and β_1.

FIGURE 10–16 **Computing SSE and MSE in the American Express Study**



This is so because σ is part of the expressions for the standard errors of both parameter estimators. The standard errors are defined next; they give us an idea of the accuracy of the least-squares estimates b_0 and b_1. *The standard error of* b_1 *is especially important because it is used in a test for the existence of a linear relationship between* X *and* Y. This will be seen in Section 10–6.

The standard error of b_0 is

$$s(b_0) = \frac{s\sqrt{\Sigma x^2}}{\sqrt{n\text{SS}_x}} \tag{10–14}$$

where $s = \sqrt{\text{MSE}}$.

The standard error of b_1 is very important, for the reason just mentioned. The true standard deviation of b_1 is $\sigma/\sqrt{\text{SS}_x}$, but since σ is not known, we use the estimated standard deviation of the errors, s.

The standard error of b_1 is

$$s(b_1) = \frac{s}{\sqrt{\text{SS}_X}} \tag{10–15}$$

Formulas such as equation 10–15 are nice to know, but you should not worry too much about having to use them. Regression analysis is usually done by computer, and the computer output will include the standard errors of the regression estimates. We will now show how the regression parameter estimates and their standard errors can be used in the construction of confidence intervals for the true regression parameters β_0 and β_1. In Section 10–6, as mentioned, we will use the standard error of b_1 for conducting the very important hypothesis test about the existence of a linear relationship between X and Y.

Confidence Intervals for the Regression Parameters

Confidence intervals for the true regression parameters β_0 and β_1 are easy to compute.

A $(1 - \alpha)$ 100% confidence interval for β_0 is

$$b_0 \pm t_{(\alpha/2,\, n-2)}s(b_0) \tag{10–16}$$

where $s(b_0)$ is as given in equation 10–14.

A $(1 - \alpha)$ 100% confidence interval for β_1 is

$$b_1 \pm t_{(\alpha/2,\, n-2)}s(b_1) \tag{10–17}$$

where $s(b_1)$ is as given in equation 10–15.

Let us construct 95% confidence intervals for β_0 and β_1 in the American Express example. Using equations 10–14 to 10–17, we get

$$s(b_0) = \frac{s\sqrt{\Sigma x^2}}{\sqrt{n\text{SS}_X}} = 318.16\frac{\sqrt{293{,}426{,}946}}{\sqrt{(25)(40{,}947{,}557.84)}} = 170.338 \tag{10–18a}$$

where the various quantities were computed earlier, including Σx^2, which is found at the bottom of Table 10–2.

A 95% confidence interval for β_0 is

$$b_0 \pm t_{(\alpha/2,\, n-2)}s(b_0) = 274.85 \pm 2.069(170.338) = [-77.58, 627.28] \qquad (10\text{–}18b)$$

where the value 2.069 is obtained from Appendix C, Table 3, for $1 - \alpha = 0.95$ and 23 degrees of freedom. We may be 95% confident that the true regression intercept is anywhere from −77.58 to 627.28. Again using equations 10–14 to 10–17, we get

$$s(b_1) = \frac{s}{\sqrt{SS_X}} = \frac{318.16}{\sqrt{40{,}947{,}557.84}} = 0.04972 \qquad (10\text{–}19a)$$

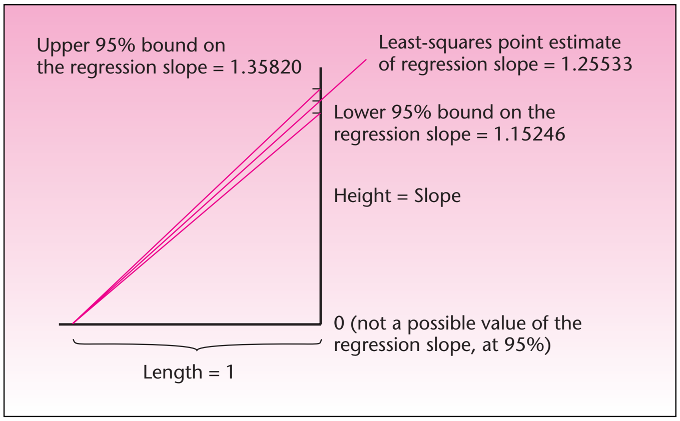
A 95% confidence interval for β_1 is

$$\begin{aligned} b_1 \pm t_{(\alpha/2,\, n-2)}s(b_1) &= 1.25533 \pm 2.069(0.04972) \\ &= [1.15246, 1.35820] \end{aligned} \qquad (10\text{–}19b)$$

From the confidence interval given in equation 10–19b, we may be 95% confident that the *true* slope of the (*population*) regression line is anywhere from 1.15246 to 1.3582. This range of values is far from zero, and so we may be quite confident that the true regression slope is not zero. This conclusion is very important, as we will see in the following sections. Figure 10–17 demonstrates the meaning of the confidence interval given in equation 10–19b.

In the next chapter, we will discuss *joint* confidence intervals for both regression parameters β_0 and β_1, an advanced topic of secondary importance. (Since the two estimates are related, a joint interval will give us greater accuracy and a more meaningful, single confidence coefficient $1 - \alpha$. This topic is somewhat similar to the Tukey analysis of Chapter 9.) Again, we want to deemphasize the importance of inference about β_0, even though information about the standard error of the estimator of this parameter is reported in computer regression output. It is the inference about β_1 that is of interest to us. Inference about β_1 has implications for the existence of a linear relationship between X and Y; inference about β_0 has no such implications. In addition, you may be tempted to use the results of the inference about β_0 to "force" this parameter to equal

FIGURE 10–17 Interpretation of the Slope Estimation for Example 10–1


Upper 95% bound on the regression slope = 1.35820
Least-squares point estimate of regression slope = 1.25533
Lower 95% bound on the regression slope = 1.15246
Height = Slope
0 (not a possible value of the regression slope, at 95%)
Length = 1

zero or another number. Such temptation should be resisted for reasons that will be explained in a later section; therefore, we deemphasize inference about β_0.

EXAMPLE 10–2

The data below are international sales versus U.S. sales for the McDonald's chain for 10 years.

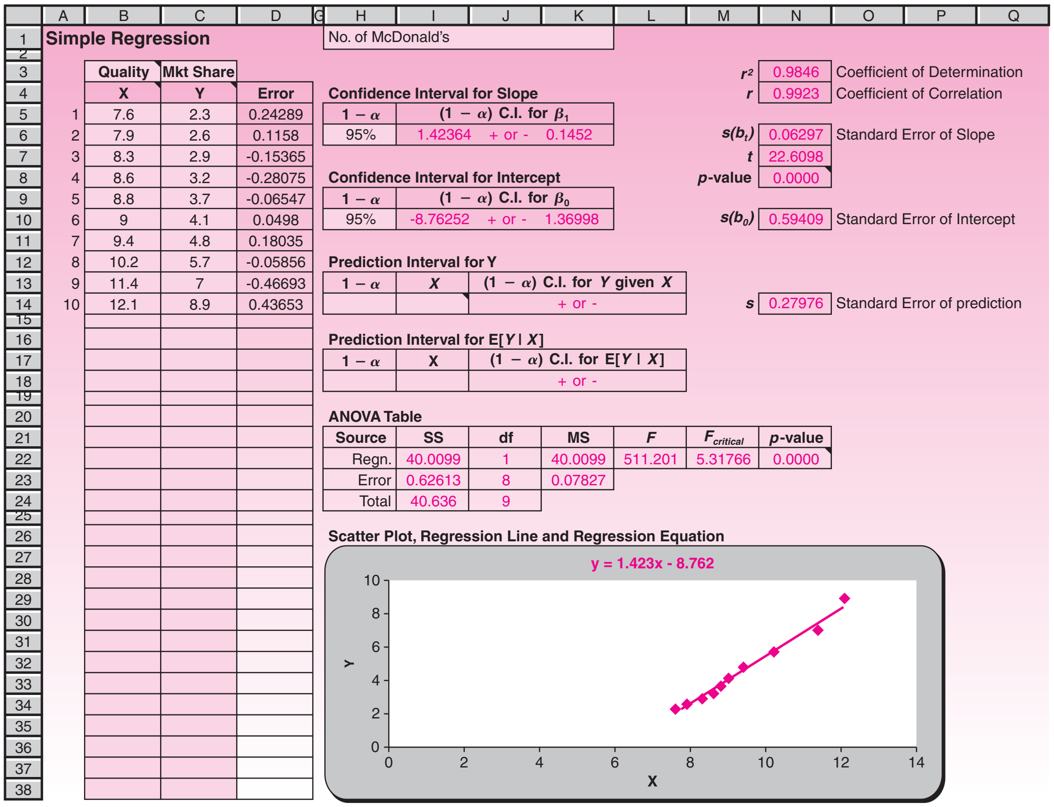
Sales for McDonald's at Year End (in billions)

U.S. Sales	International Sales
7.6	2.3
7.9	2.6
8.3	2.9
8.6	3.2
8.8	3.7
9.0	4.1
9.4	4.8
10.2	5.7
11.4	7.0
12.1	8.9

Use the template to regress McDonald's international sales, then answer the following questions:

1. What is the regression equation?
2. What is the 95% confidence interval for the slope?
3. What is the standard error of estimate?

Solution



Simple Regression

No. of McDonald's

	Quality X	Mkt Share Y	Error
1	7.6	2.3	0.24289
2	7.9	2.6	0.1158
3	8.3	2.9	-0.15365
4	8.6	3.2	-0.28075
5	8.8	3.7	-0.06547
6	9	4.1	0.0498
7	9.4	4.8	0.18035
8	10.2	5.7	-0.05856
9	11.4	7	-0.46693
10	12.1	8.9	0.43653

Confidence Interval for Slope

$1-\alpha$	$(1-\alpha)$ C.I. for β_1
95%	1.42364 + or - 0.1452

Confidence Interval for Intercept

$1-\alpha$	$(1-\alpha)$ C.I. for β_0
95%	-8.76252 + or - 1.36998

Prediction Interval for Y

$1-\alpha$	X	$(1-\alpha)$ C.I. for Y given X
		+ or -

Prediction Interval for E[Y | X]

$1-\alpha$	X	$(1-\alpha)$ C.I. for E[Y | X]
		+ or -

r^2	0.9846	Coefficient of Determination
r	0.9923	Coefficient of Correlation
$s(b_1)$	0.06297	Standard Error of Slope
t	22.6098	
p-value	0.0000	
$s(b_0)$	0.59409	Standard Error of Intercept
s	0.27976	Standard Error of prediction

ANOVA Table

Source	SS	df	MS	F	$F_{critical}$	p-value
Regn.	40.0099	1	40.0099	511.201	5.31766	0.0000
Error	0.62613	8	0.07827			
Total	40.636	9				

Scatter Plot, Regression Line and Regression Equation

1. From the template, the regression equation is $\hat{Y} = 1.4326X - 8.7625$.
2. The 95% confidence interval for the slope is 1.4236 ± 0.1452.
3. The standard error of estimate is 0.2798.

PROBLEMS

10–19. Give a 99% confidence interval for the slope parameter in Example 10–1. Is zero a credible value for the true regression slope?

10–20. Give an unbiased estimate for the error variance in the situation of problem 10–11. In this problem and others, you may either use a computer or do the computations by hand.

10–21. Find the standard errors of the regression parameter estimates for problem 10–11.

10–22. Give 95% confidence intervals for the regression slope and the regression intercept parameters for the situation of problem 10–11.

10–23. For the situation of problem 10–13, find the standard errors of the estimates of the regression parameters; give an estimate of the variance of the regression errors. Also give a 95% confidence interval for the true regression slope. Is zero a plausible value for the true regression slope at the 95% level of confidence?

10–24. Repeat problem 10–23 for the situation in problem 10–17. Comment on your results.

10–25. In addition to its role in the formulas of the standard errors of the regression estimates, what is the significance of s^2?

10–5 Correlation

CHAPTER 14

We now digress from regression analysis to discuss an important related concept: statistical *correlation*. Recall that one of the assumptions of the regression model is that the independent variable X is fixed rather than random and that the only randomness in the values of Y comes from the error term ϵ. Let us now relax this assumption and *assume that both* X *and* Y *are random variables*. In this new context, the study of the relationship between two variables is called *correlation analysis*.

In correlation analysis, we adopt a symmetric approach: We make no distinction between an independent variable and a dependent one. The correlation between two variables is a measure of the linear relationship between them. The correlation gives an indication of how well the two variables move together in a straight-line fashion. The correlation between X and Y is the same as the correlation between Y and X. We now define correlation more formally.

> The **correlation** between two random variables *X* and *Y* is a measure of the *degree of linear association* between the two variables.

Two variables are highly correlated if they move well together. Correlation is indicated by the **correlation coefficient.**

> The population correlation coefficient is denoted by ρ. The coefficient ρ can take on any value from -1, through 0, to 1.

The possible values of ρ and their interpretations are given below.

1. When ρ is equal to zero, there is no correlation. That is, there is no linear relationship between the two random variables.

2. When $\rho = 1$, there is a perfect, positive, linear relationship between the two variables. That is, whenever one of the variables, X or Y, increases, the other variable also increases; and whenever one of the variables decreases, the other one must also decrease.
3. When $\rho = -1$, there is a perfect negative linear relationship between X and Y. When X or Y increases, the other variable decreases; and when one decreases, the other one must increase.
4. When the value of ρ is between 0 and 1 in absolute value, it reflects the relative strength of the linear relationship between the two variables. For example, a correlation of 0.90 implies a relatively strong positive relationship between the two variables. A correlation of -0.70 implies a weaker, negative (as indicated by the minus sign), linear relationship. A correlation $\rho = 0.30$ implies a relatively weak (positive) linear relationship between X and Y.

A few sets of data on two variables, and their corresponding population correlation coefficients, are shown in Figure 10–18.

How do we arrive at the concept of correlation? Consider the pair of random variables X and Y. In correlation analysis, *we will assume that both* X *and* Y *are normally distributed random variables with means* μ_X *and* μ_Y *and standard deviations* σ_X *and* σ_Y, *respectively*. We define the *covariance* of X and Y as follows:

The **covariance** of two random variables X and Y is

$$\text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] \tag{10–20}$$

where μ_X is the (population) mean of X and μ_Y is the (population) mean of Y.

The covariance of X and Y is thus the expected value of the product of the deviation of X from its mean and the deviation of Y from its mean. The covariance is positive when the two random variables move together in the same direction, it is negative when the two random variables move in opposite directions, and it is zero when the two variables are not linearly related. Other than this, the covariance does not convey much. Its magnitude cannot be interpreted as an indication of the *degree* of linear association between the two variables, because the covariance's magnitude depends on the magnitudes of the standard deviations of X and Y. But if we divide the covariance by these standard deviations, we get a measure that is constrained to the range of values -1 to 1 and conveys information about the relative strength of the linear relationship between the two variables. This measure is the population correlation coefficient ρ.

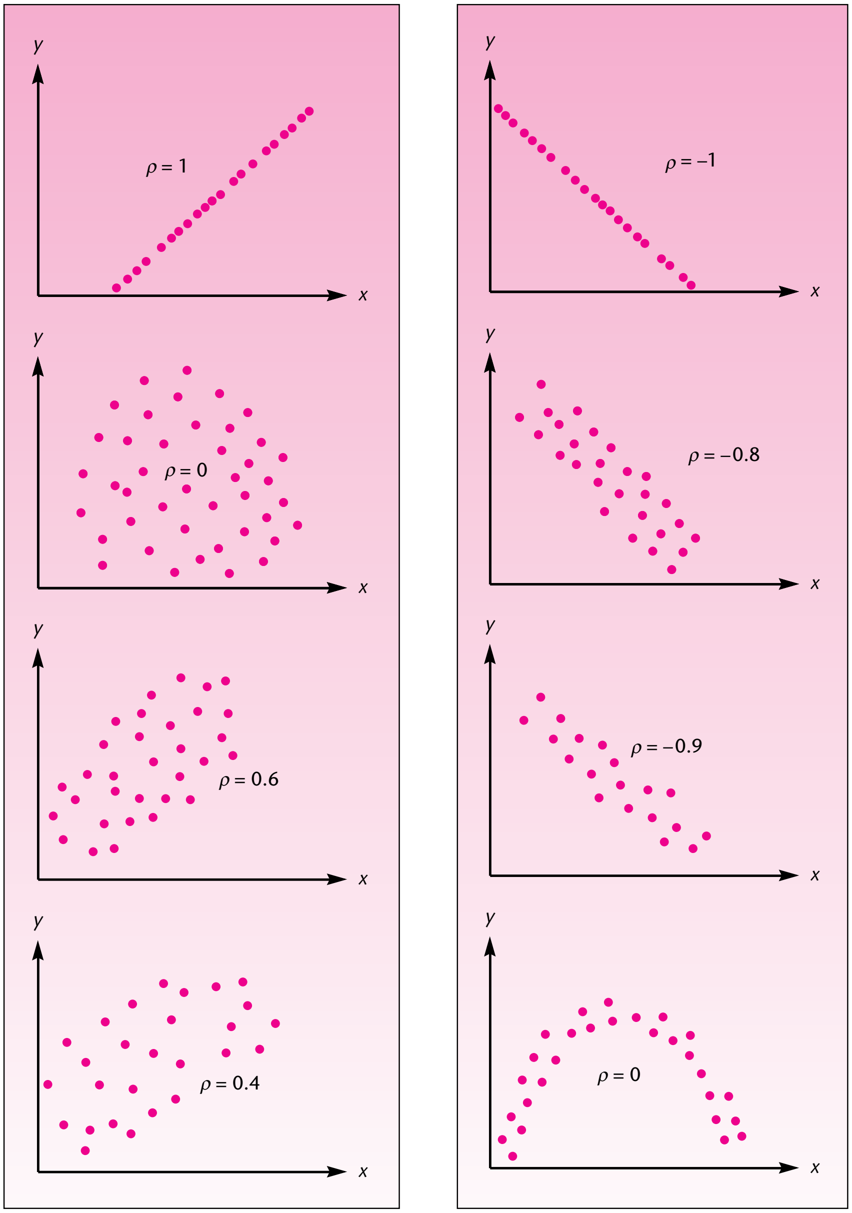
The **population correlation coefficient** is

$$\rho = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} \tag{10–21}$$

Figure 10–18 gives an idea of what data from populations with different values of ρ may look like.

Like all population parameters, the value of ρ is not known to us, and we need to estimate it from our random sample of (X, Y) observation pairs. It turns out that a sample estimator of $\text{Cov}(X, Y)$ is $SS_{XY}/(n - 1)$; an estimator of σ_X is $\sqrt{SS_X/(n - 1)}$; and an estimator of σ_Y is $\sqrt{SS_Y/(n - 1)}$. Substituting these estimators for their population counterparts in equation 10–21, and noting that the term $n - 1$ cancels, we get the *sample correlation coefficient,* denoted by r. This estimate of ρ, also referred to as the *Pearson product-moment correlation coefficient,* is given in equation 10–22.

FIGURE 10–18 **Several Possible Correlations between Two Variables**



The **sample correlation coefficient** is

$$r = \frac{SS_{XY}}{\sqrt{SS_X SS_Y}} \qquad (10\text{–}22)$$

In regression analysis, the square of the sample correlation coefficient, or r^2, has a special meaning and importance. This will be seen in Section 10–7.

We often use the sample correlation coefficient for descriptive purposes as a point estimator of the population correlation coefficient ρ. When r is large and positive (closer to $+1$), we say that the two variables are highly correlated in a positive way; when r is large and negative (toward -1), we say that the two variables are highly correlated in an inverse direction, and so on. That is, we view r as if it were the parameter ρ, which r estimates. However, r can be used as an estimator in testing hypotheses about the true correlation coefficient ρ. When such hypotheses are tested, the assumption of normal distributions of the two variables is required.

The most common test is a test of whether two random variables X and Y are correlated. The hypothesis test is

$$\begin{aligned} &H_0\colon \rho = 0 \\ &H_1\colon \rho \neq 0 \end{aligned} \qquad (10\text{–}23)$$

The test statistic for this particular test is

$$t_{(n-2)} = \frac{r}{\sqrt{(1 - r^2)/(n - 2)}} \qquad (10\text{–}24)$$

This test statistic may also be used for carrying out a one-tailed test for the existence of a positive only, or a negative only, correlation between X and Y. These would be one-tailed tests instead of the two-tailed test of equation 10–23, and the only difference is that the critical points for t would be the appropriate one-tailed values for a given α. The test statistic, however, is good *only* for tests where the null hypothesis assumes a zero correlation. When the true correlation between the two variables is anything but zero, the t distribution in equation 10–24 does not apply; in such cases the distribution is more complicated.[5] The test in equation 10–23 is the most common hypothesis test about the population correlation coefficient because it is a test for the existence of a linear relationship between two variables. We demonstrate this test with the following example.

EXAMPLE 10–3

A study was carried out to determine whether there is a linear relationship between the time spent in negotiating a sale and the resulting profits. A random sample of 27 market transactions was collected, and the time taken to conclude the sale as well as the resulting profit were recorded for each transaction. The sample correlation coefficient was computed: $r = 0.424$. Is there a linear relationship between the length of negotiations and transaction profits?

Solution

We want to conduct the hypothesis test H_0: $\rho = 0$ versus H_1: $\rho \neq 0$. Using the test statistic in equation 10–24, we get

$$t_{(25)} = \frac{r}{\sqrt{(1 - r^2)/(n - 2)}} = \frac{0.424}{\sqrt{(1 - 0.424^2)/25}} = 2.34$$

[5] In cases where we want to test H_0: $\rho = a$ versus H_1: $\rho \neq a$, where a is some number other than zero, we may do so by using the Fisher transformation: $z' = (1/2) \log [(1 + r)/(1 - r)]$, where z' is approximately normally distributed with mean $\mu' = (1/2) \log [(1 + \rho)/(1 - \rho)]$ and standard deviation $\sigma' = 1/\sqrt{n - 3}$. (Here *log* is taken to mean *natural logarithm*.) Such tests are less common, and a more complete description may be found in advanced texts. As an exercise, the interested reader may try this test on some data. [You need to transform z' to an approximate standard normal $z = (z' - \mu')/\sigma'$; use the null-hypothesis value of ρ in the formula for μ'.]

From Appendix C, Table 3, we find that the critical points for a *t* distribution with 25 degrees of freedom and $\alpha = 0.05$ are ± 2.060. Therefore, we reject the null hypothesis of no correlation in favor of the alternative that the two variables are linearly related. Since the critical points for $\alpha = 0.01$ are ± 2.787, and $2.787 > 2.34$, we are unable to reject the null hypothesis of no correlation between the two variables if we want to use the 0.01 level of significance. If we wanted to test (before looking at our data) only for the existence of a positive correlation between the two variables, our test would have been H_0: $\rho \leq 0$ versus H_1: $\rho > 0$ and we would have used only the right tail of the *t* distribution. At $\alpha = 0.05$, the critical point of *t* with 25 degrees of freedom is 1.708, and at $\alpha = 0.01$ it is 2.485. The null hypothesis would, again, be rejected at the 0.05 level but not at the 0.01 level of significance.

In regression analysis, the test for the existence of a linear relationship between *X* and *Y* is a test of whether the regression slope β_1 is equal to zero. The regression slope parameter is related to the correlation coefficient (as an exercise, compare the equations of the estimates *r* and b_1); when two random variables are uncorrelated, the population regression slope is zero.

We end this section with a word of caution. First, the existence of a correlation between two variables does not necessarily mean that one of the variables *causes* the other one. The determination of **causality** is a difficult question that cannot be directly answered in the context of correlation analysis or regression analysis. Also, the statistical determination that two variables are correlated may not always mean that they are correlated in any direct, meaningful way. For example, if we study any two population-related variables and find that both variables increase "together," this may merely be a reflection of the general increase in population rather than any direct correlation between the two variables. We should look for outside variables that may affect both variables under study.

PROBLEMS

10–26. What is the main difference between correlation analysis and regression analysis?

10–27. Compute the sample correlation coefficient for the data of problem 10–11.

10–28. Compute the sample correlation coefficient for the data of problem 10–13.

10–29. Using the data in problem 10–16, conduct the hypothesis test for the existence of a linear correlation between the two variables. Use $\alpha = 0.01$.

10–30. Is it possible that a sample correlation of 0.51 between two variables will not indicate that the two variables are really correlated, while a sample correlation of 0.04 between another pair of variables will be statistically significant? Explain.

10–31. The following data are indexed prices of gold and copper over a 10-year period. Assume that the indexed values constitute a random sample from the population of possible values. Test for the existence of a linear correlation between the indexed prices of the two metals.

Gold:	76, 62, 70, 59, 52, 53, 53, 56, 57, 56
Copper:	80, 68, 73, 63, 65, 68, 65, 63, 65, 66

Also, state one limitation of the data set.

10–32. Follow daily stock price quotations in the *Wall Street Journal* for a pair of stocks of your choice, and compute the sample correlation coefficient. Also, test for the existence of a nonzero linear correlation in the "population" of prices of the two stocks. For your sample, use as many daily prices as you can.

10–33. Again using the *Wall Street Journal* as a source of data, determine whether there is a linear correlation between morning and afternoon price quotations in London for an ounce of gold (for the same day). Any ideas?

10–34. A study was conducted to determine whether a correlation exists between consumers' perceptions of a television commercial (measured on a special scale) and their interest in purchasing the product (measured on a scale). The results are $n = 65$ and $r = 0.37$. Is there statistical evidence of a linear correlation between the two variables?

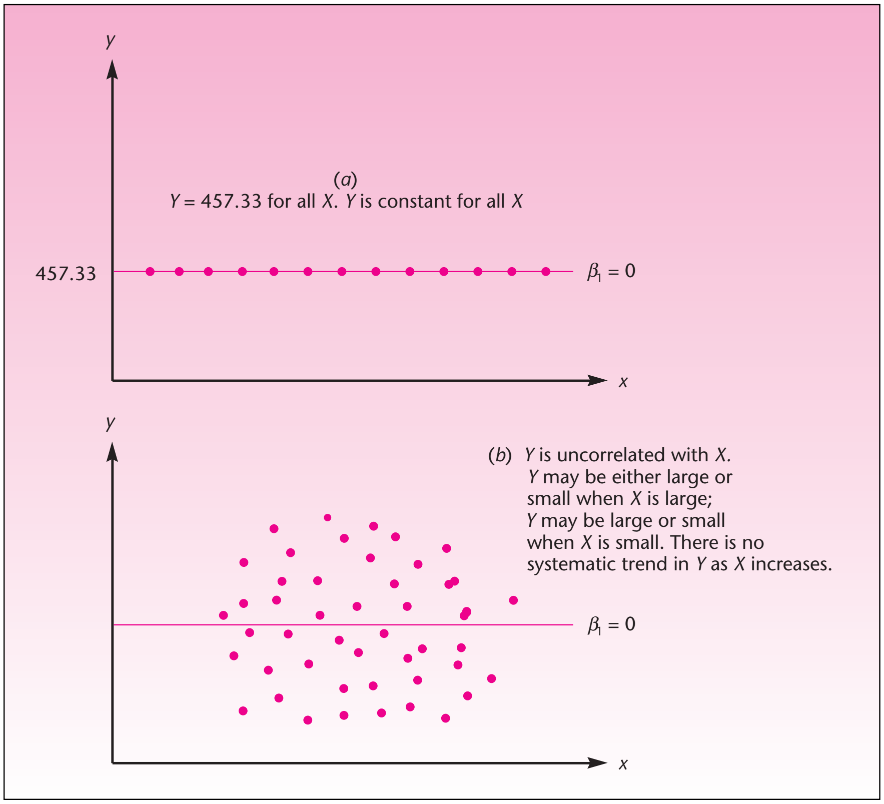
10–35. (Optional, advanced problem) Using the Fisher transformation (described in footnote 5), carry out a two-tailed test of the hypothesis that the population correlation coefficient for the situation of problem 10–34 is $\rho = 0.22$. Use $\alpha = 0.05$.

10–6 Hypothesis Tests about the Regression Relationship

When X and Y have no linear relationship, the population regression slope β_1 is equal to zero. Why? The population regression slope is equal to zero in either of two situations:

1. When Y is *constant* for all values of X. For example, $Y = 457.33$ for all X. This is shown in Figure 10–19(*a*). If Y is constant for all values of X, the slope of Y with respect to X, parameter β_1, is identically zero; there is no linear relationship between the two variables.

FIGURE 10–19 Two Possibilities Where the Population Regression Slope Is Zero



2. When the two variables are *uncorrelated*. When the correlation between X and Y is zero, as X increases Y may increase, or it may decrease, or it may remain constant. There is no *systematic* increase or decrease in the values of Y as X increases. This case is shown in Figure 10–19(*b*). As can be seen in the figure, data from this process are not "moving" in any pattern; thus, the line has no direction to follow. With no direction, the slope of the line is, again, zero.

Also, remember that the relationship may be curved, with no linear correlation, as was seen in the last part of Figure 10–18. In such cases, the slope may also be zero.

In all cases other than these, at least *some* linear relationship exists between the two variables X and Y; the slope of the line in all such cases would be either positive or negative, but not zero. Therefore, *the most important statistical test in simple linear regression is the test of whether the slope parameter* β_1 *is equal to zero*. If we conclude in any particular case that the true regression slope is equal to zero, this means that there is no linear relationship between the two variables: Either the dependent variable is constant, or—more commonly—the two variables are not linearly related. We thus have the following test for determining the existence of a linear relationship between two variables X and Y:

A hypothesis test for the existence of a linear relationship between X and Y is

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0 \qquad (10\text{–}25)$$

This test is, of course, a two-tailed test. Either the true regression slope is equal to zero, or it is not. If it is equal to zero, the two variables have no linear relationship; if the slope is not equal to zero, then it is either positive or negative (the two tails of rejection), in which case there is a linear relationship between the two variables. The test statistic for determining the rejection or nonrejection of the null hypothesis is given in equation 10–26. Given the assumption of normality of the regression errors, the test statistic possesses the t distribution with $n - 2$ degrees of freedom.

CHAPTER 15

The test statistic for the existence of a linear relationship between X and Y is

$$t_{(n-2)} = \frac{b_1}{s(b_1)} \qquad (10\text{–}26)$$

where b_1 is the least-squares estimate of the regression slope and $s(b_1)$ is the standard error of b_1. When the null hypothesis is true, the statistic has a t distribution with $n - 2$ degrees of freedom.

This test statistic is a special version of a general test statistic

$$t_{(n-2)} = \frac{b_1 - (\beta_1)_0}{s(b_1)} \qquad (10\text{–}27)$$

where $(\beta_1)_0$ is the value of β_1 under the null hypothesis. This statistic follows the format (Estimate − Hypothesized parameter value)/(Standard error of estimator). Since, in the test of equation 10–25, the hypothesized value of β_1 is zero, we have the simplified version of the test statistic, equation 10–26. One advantage of the simple form of our test statistic is that it allows us to conduct the test very quickly. Computer output for regression analysis usually contains a table similar to Table 10–3.

TABLE 10–3 An Example of a Part of the Computer Output for Regression

Variable	Estimate	Standard Error	*t* Ratio
Constant	5.22	0.5	10.44
X	4.88	0.1	48.80

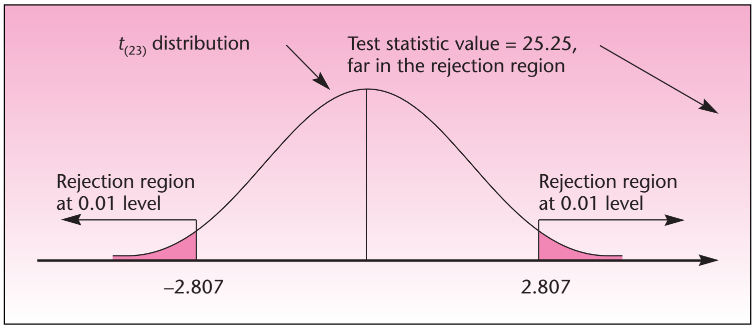
The estimate associated with X (or whatever name the user may have given to the independent variable in the computer program) is b_1. The standard error associated with X is $s(b_1)$. To conduct the test, all you need to do is to divide b_1 by $s(b_1)$. In the example of Table 10–3, $4.88/0.1 = 48.8$. The answer is reported in the table as the t ratio. The t ratio can now be compared with critical points of the t distribution with $n - 2$ degrees of freedom. Suppose that the sample size used was 100. Then the critical points for $\alpha = 0.05$, from the spreadsheet, are ± 1.98, and since $48.8 > 1.98$, we conclude that there is evidence of a linear relationship between X and Y in this hypothetical example. (Actually, the p-value is very small. Some computer programs will also report the p-value in an extra column on the right.) What about the first row in the table? The test suggested here is a test of whether the intercept β_0 (this is the constant) is equal to zero. The test statistic is the same as equation 10–26, but with subscripts 0 instead of 1. As we mentioned earlier, this test, although suggested by the output of computer routines, is usually not meaningful and should generally be avoided.

We now conduct the hypothesis test for the existence of a linear relationship between miles traveled and amount charged on the American Express card in Example 10–1. Our hypotheses are H_0: $\beta_1 = 0$ and H_1: $\beta_1 \neq 0$. Recall that for the American Express study, $b_1 = 1.25533$ and $s(b_1) = 0.04972$ (from equations 10–11 and 10–19a). We now compute the test statistic, using equation 10–26:

$$t = \frac{b_1}{s(b_1)} = \frac{1.25533}{0.04972} = 25.25$$

From the magnitude of the computed value of the statistic, we know that there is statistical evidence of a linear relationship between the variables, because 25.25 is certainly greater than any critical point of a t distribution with 23 degrees of freedom. We show the test in Figure 10–20. The critical points of t with 23 degrees of freedom and $\alpha = 0.01$ are obtained from Appendix C, Table 3. We conclude that there is evidence of a linear relationship between the two variables “miles traveled” and “dollars charged” in Example 10–1.

FIGURE 10–20 Test for a Linear Relationship for Example 10–1



Other Tests[6]

Although the test of whether the slope parameter is equal to zero is a very important test, because it is a test for the existence of a linear relationship between the two variables, other tests are possible in the context of regression. These tests serve secondary purposes. In financial analysis, for example, it is often important to determine from past performance data of a particular stock whether the stock generally moves with the market as a whole. If the stock does move with the stock market as a whole, the slope parameter of the regression of the stock's returns (Y) versus returns on the market as a whole (X) would be equal to 1.00. That is, $\beta_1 = 1$. We demonstrate this test with Example 10–4.

EXAMPLE 10–4

The *Market Sensitivity Report,* issued by Merrill Lynch, Inc., lists estimated beta coefficients of common stocks as well as their standard errors. *Beta* is the term used in the finance literature for the estimate b_1 of the regression of returns on a stock versus returns on the stock market as a whole. Returns on the stock market as a whole are taken by Merrill Lynch as returns on the Standard & Poor's 500 index. The report lists the following findings for common stock of Time, Inc.: beta = 1.24, standard error of beta = 0.21, $n = 60$. Is there statistical evidence to reject the claim that the Time stock moves, in general, with the market as a whole?

Solution

We want to carry out the special-purpose test H_0: $\beta_1 = 1$ versus H_1: $\beta_1 \neq 1$. We use the general test statistic of equation 10–27:

$$t_{(n-2)} = \frac{b_1 - (\beta_1)_0}{s(b_1)} = \frac{1.24 - 1}{0.21} = 1.14$$

Since $n - 2 = 58$, we use the standard normal distribution. The test statistic value is in the nonrejection region for any usual level α, and we conclude that there is no statistical evidence against the claim that Time moves with the market as a whole.

PROBLEMS

10–36. An interesting marketing research effort has recently been reported, which incorporates within the variables that predict consumer satisfaction from a product not only attributes of the product itself but also characteristics of the consumer who buys the product. In particular, a regression model was developed, and found successful, regressing consumer satisfaction S on a consumer's materialism M measured on a psychologically devised scale. For satisfaction with the purchase of sunglasses, the estimate of beta, the slope of S with respect to M, was $b = -2.20$. The reported t statistic was -2.53. The sample size was $n = 54$.[7] Is this regression statistically significant? Explain the findings.

10–37. A regression analysis was carried out of returns on stocks (Y) versus the ratio of book to market value (X). The resulting prediction equation is

$$Y = 1.21 + 3.1X(2.89)$$

where the number in parentheses is the standard error of the slope estimate. The sample size used is $n = 18$. Is there evidence of a linear relationship between returns and book to market value?

[6] This subsection may be skipped without loss of continuity.

[7] Jeff Wang and Melanie Wallendorf, "Materialism, Status Signaling, and Product Satisfaction," *Journal of the Academy of Marketing Science* 34, no. 4 (2006), pp. 494–505.

10–38. In the situation of problem 10–11, test for the existence of a linear relationship between the two variables.

10–39. In the situation of problem 10–13, test for the existence of a linear relationship between the two variables.

10–40. In the situation of problem 10–16, test for the existence of a linear relationship between the two variables.

10–41. For Example 10–4, test for the existence of a linear relationship between returns on the stock and returns on the market as a whole.

10–42. A regression analysis was carried out to determine whether wages increase for blue-collar workers depending on the extent to which firms that employ them engage in product exportation. The sample consisted of 585,692 German blue-collar workers. For each of these workers, the income was known as well as the percentage of the work that was related to exportation. The regression slope estimate was 0.009, and the t-statistic value was 1.51.[8] Carefully interpret and explain these findings.

10–43. An article in *Financial Analysts Journal* discusses results of a regression analysis of average price per share P on the independent variable X/k, where X/k is the contemporaneous earnings per share divided by firm-specific discount rate. The regression was run using a random sample of 213 firms listed in the *Value Line Investment Survey*. The reported results are

$$P = 16.67 + 0.68X/k(12.03)$$

where the number in parentheses is the standard error. Is there a linear relationship between the two variables?

10–44. A management recruiter wants to estimate a linear regression relationship between an executive's experience and the salary the executive may expect to earn after placement with an employer. From data on 28 executives, which are assumed to be a random sample from the population of executives that the recruiter places, the following regression results are obtained: $b_1 = 5.49$ and $s(b_1) = 1.21$. Is there a linear relationship between the experience and the salary of executives placed by the recruiter?

10–7 How Good Is the Regression?

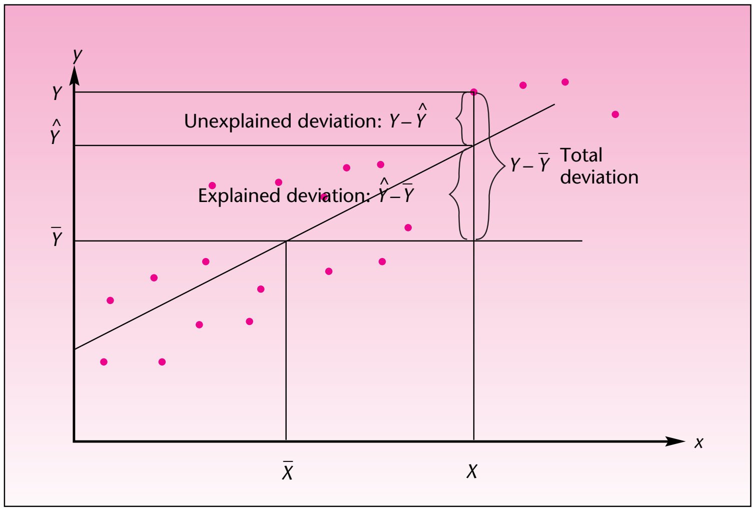
Once we have determined that a linear relationship exists between the two variables, the question is: How strong is the relationship? If the relationship is a strong one, prediction of the dependent variable can be relatively accurate, and other conclusions drawn from the analysis may be given a high degree of confidence.

We have already seen one measure of the regression fit: the mean square error. The MSE is an estimate of the variance of the true regression errors and is a measure of the variation of the data about the regression line. The MSE, however, depends on the nature of the data, and what may be a large error variation in one situation may not be considered large in another. What we need, therefore, is a *relative* measure of the degree of variation of the data about the regression line. Such a measure allows us to compare the fits of different models.

The relative measure we are looking for is a measure that compares the variation of Y about the regression line with the variation of Y without a regression line. This should remind you of analysis of variance, and we will soon see the relation of ANOVA to regression analysis. It turns out that the relative measure of regression fit

[8] Thorsten Schank, Claus Schnabel, and Joachim Wagner, "Do Exporters Really Pay Higher Wages? First Evidence from German Linked Employer–Employee Data," *Journal of International Economics* 72 (May 2007), pp. 52–74.

FIGURE 10–21 The Three Deviations Associated with a Data Point


y
Y
$\hat{Y}$
$\bar{Y}$
Unexplained deviation: $Y - \hat{Y}$
Explained deviation: $\hat{Y} - \bar{Y}$
$Y - \bar{Y}$ Total deviation
x
$\bar{X}$
X

we are looking for is the square of the estimated correlation coefficient r. It is called the *coefficient of determination*.

> The **coefficient of determination** r^2 is a descriptive measure of the strength of the regression relationship, a measure of how well the regression line fits the data.

The coefficient of determination r^2 is an estimator of the corresponding population parameter ρ^2, which is the square of the population coefficient of correlation between two variables X and Y. Usually, however, we use r^2 as a descriptive statistic—a relative measure of how well the regression line fits the data. Ordinarily, we do not use r^2 for inference about ρ^2.

CHAPTER 15

We will now see how the coefficient of determination is obtained directly from a decomposition of the variation in Y into a component due to error and a component due to the regression. Figure 10–21 shows the least-squares line that was fit to a data set. One of the data points (x, y) is highlighted. For this data point, the figure shows three kinds of deviations: the deviation of y from its mean $y - \bar{y}$, the deviation of y from its predicted value using the regression $y - \hat{y}$, and the deviation of the regression-predicted value of y from the mean of y, which is $\hat{y} - \bar{y}$. Note that the least-squares line passes through the point $(\bar{x}, \bar{y})$.

We will now follow exactly the same mathematical derivation we used in Chapter 9 when we derived the ANOVA relationships. There we looked at the deviation of a data point from its respective group mean–the error; here the error is the deviation of a data point from its regression-predicted value. In ANOVA, we also looked at the total deviation, the deviation of a data point from the grand mean; here we have the deviation of the data point from the mean of Y. Finally, in ANOVA we also considered the treatment deviation, the deviation of the group mean from the grand mean; here we have the *regression deviation*–the deviation of the predicted value from the mean of Y.

The error is also called the *unexplained deviation* because it is a deviation that cannot be explained by the regression relationship; the regression deviation is also called the *explained deviation* because it is that part of the deviation of a data point from the mean that can be explained by the regression relationship between X and Y. We *explain* why the Y value of a particular data point is above the mean of Y by the fact that its X component happens to be above the mean of X and by the fact that X and Y are linearly (and positively) related. As can be seen from Figure 10–21, and by simple arithmetic, we have

$$\begin{array}{ccccc} y - \bar{y} & = & y - \hat{y} & + & \hat{y} - \bar{y} \\ \text{Total deviation} & = & \text{Unexplained deviation (error)} & + & \text{Explained deviation (regression)} \end{array} \qquad (10\text{–}28)$$

As in the analysis of variance, we square all three deviations for each one of our data points, and we sum over all n points. Here, again, cross-terms drop out, and we are left with the following important relationship for the sums of squares:[9]

$$\begin{array}{ccccc} \sum_{i=1}^{n}(y_i - \bar{y})^2 & = & \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 & + & \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \\ \text{SST} & = & \text{SSE} & + & \text{SSR} \\ \text{(Total sum of squares)} & = & \text{(Sum of squares for error)} & + & \text{(Sum of squares for regression)} \end{array} \qquad (10\text{–}29)$$

The term SSR is also called the *explained variation;* it is the part of the variation in Y that is explained by the relationship of Y with the explanatory variable X. Similarly, SSE is the *unexplained variation,* due to error; the sum of the two is the *total variation* in Y.

We define the coefficient of determination as the sum of squares due to the regression divided by the total sum of squares. Since by equation 10–29 SSE and SSR add to SST, the coefficient of determination is equal to 1 minus SSE/SST. We have

$$r^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}} \qquad (10\text{–}30)$$

The coefficient of determination can be interpreted as *the proportion of the variation in* Y *that is explained by the regression relationship of* Y *with* X.

Recall that the correlation coefficient r can be between -1 and 1. Its square, r^2, can therefore be anywhere from 0 to 1. This is in accordance with the interpretation of r^2 as the *percentage of the variation in* Y *explained by the regression.* The coefficient is a measure of how closely the regression line fits the data; it is a measure of how much the variation in the values of Y is reduced once we regress Y on variable X. When $r^2 = 1$, we know that 100% of the variation in Y is explained by X. This means that the data all lie right on the regression line, and no errors result (because, from equation 10–30, SSE must be equal to zero). Since r^2 cannot be negative, we do not know whether the line slopes upward or downward (the direction can be found from b_1 or r), but we know that the line gives a *perfect fit* to the data. Such cases do not occur in business or economics. In fact, when there are no errors, no natural variation, there is no need for statistics.

At the other extreme is the case where the regression line explains nothing. Here the errors account for everything, and SSR is zero. In this case, we see from equation 10–30 that $r^2 = 0$. In such cases, X and Y have no linear relationship, and the true regression slope is probably zero (we say *probably* because r^2 is only an estimator, given to chance variation; it could possibly be estimating a nonzero ρ^2). Between the two cases $r^2 = 0$ and $r^2 = 1$ are values of r^2 that give an indication of the *relative fit* of the regression model to the data. *The higher* r^2 *is, the better the fit and the higher our confidence*

[9] The proof of the relation is left as an exercise for the mathematically interested reader.

in the regression. Be wary, however, of situations where the reported r^2 is exceptionally high, such as 0.99 or 0.999. In such cases, something may be wrong. We will see an example of this in the next chapter. Incidentally, in the context of multiple regression, discussed in the next chapter, we will use the notation R^2 for the coefficient of determination to indicate that the relationship is based on several explanatory X variables.

How high should the coefficient of determination be before we can conclude that a regression model fits the data well enough to use the regression with confidence? This question has no clear-cut answer. The answer depends on the intended use of the regression model. If we intend to use the regression for *prediction,* the higher the r^2, the more accurate will be our predictions.

An r^2 value of 0.9 or more is very good, a value greater than 0.8 is good, and a value of 0.6 or more may be satisfactory in some applications, although we must be aware of the fact that, in such cases, errors in prediction may be relatively high. When the r^2 value is 0.5 or less, the regression explains only 50% or less of the variation in the data; therefore, predictions may be poor. If we are interested only in understanding the relationship between the variables, lower values of r^2 may be acceptable, as long as we realize that the model does not explain much.

Figure 10–22 shows several regressions and their corresponding r^2 values. If you think of the total sum of squared deviations as being in a box, then r^2 is the proportion of the box that is filled with the explained sum of squares, the remaining part being the squared errors. This is shown for each regression in the figure.

Computing r^2 is easy if we express SSR, SSE, and SST in terms of the computational sums of squares and cross-products (equations 10–9):

$$\text{SST} = \text{SS}_Y \qquad \text{SSR} = b_1\text{SS}_{XY} \qquad \text{SSE} = \text{SS}_Y - b_1\text{SS}_{XY} \qquad (10\text{–}31)$$

We will now use equation 10–31 in computing the coefficient of determination for Example 10–1. For this example, we have

$$\begin{aligned}\text{SST} &= \text{SS}_Y = 66{,}855{,}898\\ \text{SSR} &= b_1\text{SS}_{XY} = (1.255333776)(51{,}402{,}852.4) = 64{,}527{,}736.8\end{aligned}$$

and

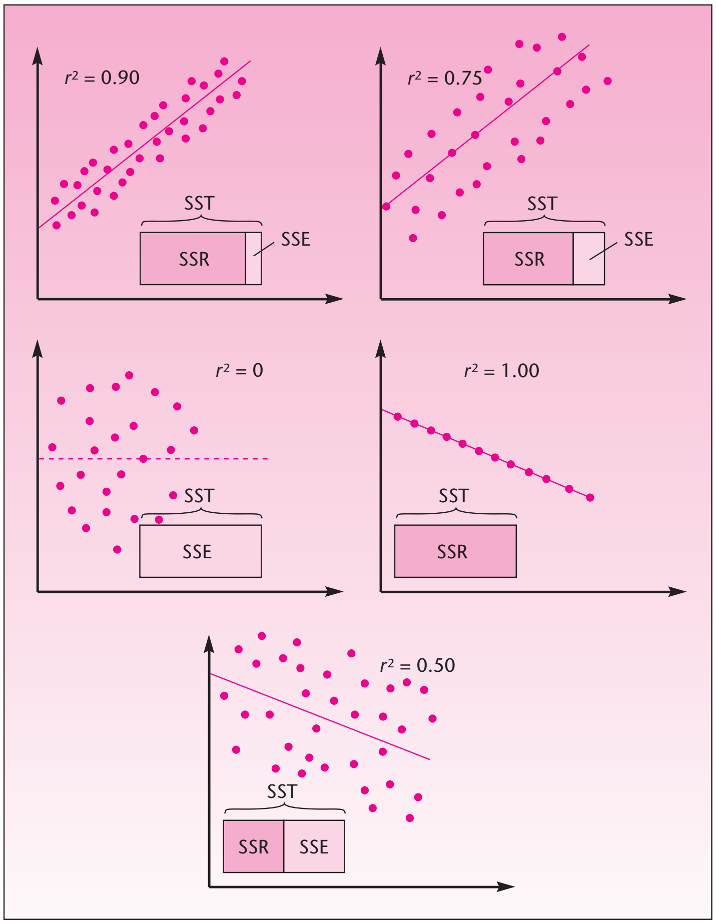
$$\text{SSE} = \text{SST} - \text{SSR} = 2{,}328{,}161.2$$

(These were computed when we found the MSE for this example.) We now compute r^2 as

$$r^2 = \frac{\text{SSR}}{\text{SST}} = \frac{64{,}527{,}736.8}{66{,}855{,}898} = 0.96518$$

The r^2 in this example is very high. The interpretation is that over 96.5% of the variation in charges on the American Express card can be explained by the relationship between charges on the card and extent of travel (miles). Again we note that while the computational formulas are easy to use, r^2 is always reported in a prominent place in regression computer output.

FIGURE 10–22 Value of the Coefficient of Determination in Different Regressions



In the next section, we will see how the sums of squares, along with the corresponding degrees of freedom, lead to mean squares—and to an analysis of variance in the context of regression. In closing this section, we note that in Chapter 11, we will introduce an adjusted coefficient of determination that accounts for degrees of freedom.

PROBLEMS

10–45. In problem 10–36, the coefficient of determination was found to be $r^2 = 0.09$.[10] What can you say about this regression, as far as its power to predict customer satisfaction with sunglasses using information on a customer's materialism score?

[10] Jeff Wang and Melanie Wallendorf, "Materialism, Status Signaling, and Product Satisfaction," *Journal of the Academy of Marketing* 34, no. 4 (2006), pp. 494–505.

11 MULTIPLE REGRESSION

11–1 Using Statistics 469
11–2 The *k*-Variable Multiple Regression Model 469
11–3 The *F* Test of a Multiple Regression Model 473
11–4 How Good Is the Regression? 477
11–5 Tests of the Significance of Individual Regression Parameters 482
11–6 Testing the Validity of the Regression Model 494
11–7 Using the Multiple Regression Model for Prediction 500
11–8 Qualitative Independent Variables 503
11–9 Polynomial Regression 513
11–10 Nonlinear Models and Transformations 521
11–11 Multicollinearity 531
11–12 Residual Autocorrelation and the Durbin-Watson Test 539
11–13 Partial *F* Tests and Variable Selection Methods 542
11–14 Using the Computer 548
11–15 Summary and Review of Terms 554
Case 15 Return on Capital for Four Different Sectors 556

LEARNING OBJECTIVES

After studying this chapter, you should be able to:

- Determine whether multiple regression would be applicable to a given instance.
- Formulate a multiple regression model.
- Carry out a multiple regression using the spreadsheet template.
- Test the validity of a multiple regression by analyzing residuals.
- Carry out hypothesis tests about the regression coefficients.
- Compute a prediction interval for the dependent variable.
- Use indicator variables in a multiple regression.
- Carry out a polynomial regression.
- Conduct a Durbin-Watson test for autocorrelation in residuals.
- Conduct a partial *F* test.
- Determine which independent variables are to be included in a multiple regression model.
- Solve multiple regression problems using the Solver macro.

11–1 Using Statistics

People often think that if something is good, then more of it is even better. In the case of the simple linear regression, explained in Chapter 10, this turns out to be true–as long as some rules are followed. Thus, if one X variable can help predict the value of Y, then several X variables may do an even better job–as long as they contain more information about Y.

A survey of the research literature in all areas of business reveals an overwhelmingly wide use of an extension of the method of Chapter 10, a model called multiple regression, which uses several independent variables in predicting a variable of interest.

11–2 The *k*-Variable Multiple Regression Model

The population regression model of a dependent variable Y on a set of k independent variables $X_1, X_2, \ldots, X_k$ is given by

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon \tag{11–1}$$

where β_0 is the Y intercept of the regression surface and each β_i, $i = 1, \ldots, k$, is the slope of the regression surface—sometimes called the **response surface**—with respect to variable X_i.

As with the simple linear regression model, we have some assumptions.

Model assumptions:

1. For each observation, the error term ϵ is normally distributed with mean zero and standard deviation σ and is independent of the error terms associated with all other observations. That is,

$$\epsilon_j \sim \mathrm{N}(0, \sigma^2) \qquad \text{for all } j = 1, 2, \ldots, n \tag{11–2}$$

independent of other errors.[1]

2. In the context of regression analysis, the variables X_j are considered *fixed quantities,* although in the context of correlational analysis, they are random variables. In any case, X_j *are independent of the error term* ϵ. When we assume that X_j are fixed quantities, we are assuming that we have realizations of k variables X_j and that the only randomness in Y comes from the error term ϵ.

For a case with $k = 2$ variables, the response surface is a plane in three dimensions (the dimensions are Y, X_1, and X_2). The plane is the surface of average response $E(Y)$ for any combination of the two variables X_1 and X_2. The response surface is given by the equation for $E(Y)$, which is the expected value of equation 11–1 with two independent variables. Taking the expected value of Y gives the value 0 to the

[1] The multiple regression model is valid under less restrictive assumptions than these. The assumptions of normality of the errors allows us to perform t tests and F tests of model validity. Also, all we need is that the errors be *uncorrelated* with one another. However, normal distribution + noncorrelation = independence.

error term ϵ. The equations for Y and $E(Y)$ in the case of regression with two independent variables are

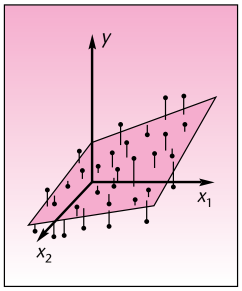
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \qquad (11\text{–}3)$$

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \qquad (11\text{–}4)$$

These are equations analogous to the case of simple linear regression. Here, instead of a regression line, we have a regression plane. Some values of Y (i.e., combinations of the X_i variables times their coefficients β_i, and the errors ϵ) are shown in Figure 11–1. The figure shows the response surface, the plane corresponding to equation 11–4.

FIGURE 11–1

A Two-Dimensional Response Surface $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ and Some Points



We estimate the regression parameters of equation 11–3 by the method of least squares. This is an extension of the procedure used in simple linear regression. In the case of two independent variables where the population model is equation 11–3, we need to estimate an equation of a plane that will minimize the sum of the squared errors $(Y - \hat{Y})^2$ over the entire data set of n points. The method is extendable to any k independent variables. In the case of $k = 2$, there are three equations, and their solutions are the least-squares estimates b_0, b_1, and b_2. These are estimates of the Y intercept, the slope of the plane with respect to X_1, and the slope of the plane with respect to X_2. The normal equations for $k = 2$ follow.

When the various sums Σy, Σx_1, and the other sums and products are entered into these equations, it is possible to solve the three equations for the three unknowns b_0, b_1, and b_2. These computations are always done by computer. We will, however, demonstrate the solution of equations 11–5 with a simple example.

The normal equations for the case of two independent variables:

$$\Sigma y = nb_0 + b_1 \Sigma x_1 + b_2 \Sigma x_2$$

$$\Sigma x_1 y = b_0 \Sigma x_1 + b_1 \Sigma x_1^2 + b_2 \Sigma x_1 x_2$$

$$\Sigma x_2 y = b_0 \Sigma x_2 + b_1 \Sigma x_1 x_2 + b_2 \Sigma x_2^2 \qquad (11\text{–}5)$$

EXAMPLE 11–1

Alka-Seltzer recently embarked on an in-store promotional campaign, with displays of its antacid featured prominently in supermarkets. The company also ran its usual radio and television commercials. Over a period of 10 weeks, the company kept track of its expenditure on radio and television advertising, variable X_1, as well as its spending on in-store displays, variable X_2. The resulting sales for each week in the area studied were recorded as the dependent variable Y. The company analyst conducting the study hypothesized a linear regression model of the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

linking sales volume with the two independent variables, advertising and in-store promotions. The analyst wanted to use the available data, considered a random sample of 10 weekly observations, to estimate the parameters of the regression relationship.

Solution Table 11–1 gives the data for this study in terms of Y, X_1, and X_2, all in thousands of dollars. The table also gives additional columns of products and squares of data

TABLE 11–1 **Various Quantities Needed for the Solution of the Normal Equations for Example 11–1 (numbers are in thousands of dollars)**

Y	X_1	X_2	X_1X_2	x_1^2	x_2^2	X_1Y	X_2Y
72	12	5	60	144	25	864	360
76	11	8	88	121	64	836	608
78	15	6	90	225	36	1,170	468
70	10	5	50	100	25	700	350
68	11	3	33	121	9	748	204
80	16	9	144	256	81	1,280	720
82	14	12	168	196	144	1,148	984
65	8	4	32	64	16	520	260
62	8	3	24	64	9	496	186
90	18	10	180	324	100	1,620	900
743	123	65	869	1,615	509	9,382	5,040

values needed for the solution of the normal equations. These columns are X_1X_2, X_1^2, X_2^2, X_1Y, and X_2Y. The sums of these columns are then substituted into equations 11–5, which are solved for the estimates b_0, b_1, and b_2 of the regression parameters.

From Table 11–1, the sums needed for the solution of the normal equations are $\Sigma y = 743$, $\Sigma x_1 = 123$, $\Sigma x_2 = 65$, $\Sigma x_1 y = 9{,}382$, $\Sigma x_2 y = 5{,}040$, $\Sigma x_1 x_2 = 869$, $\Sigma x_1^2 = 1{,}615$, and $\Sigma x_2^2 = 509$. When these sums are substituted into equations 11–5, we get the resulting normal equations:

$$\begin{aligned} 743 &= 10b_0 + 123b_1 + 65b_2 \\ 9{,}382 &= 123b_0 + 1{,}615b_1 + 869b_2 \\ 5{,}040 &= 65b_0 + 869b_1 + 509b_2 \end{aligned}$$

Solution of this system of equations by substitution, or by any other method of solution, gives

$$b_0 = 47.164942 \qquad b_1 = 1.5990404 \qquad b_2 = 1.1487479$$

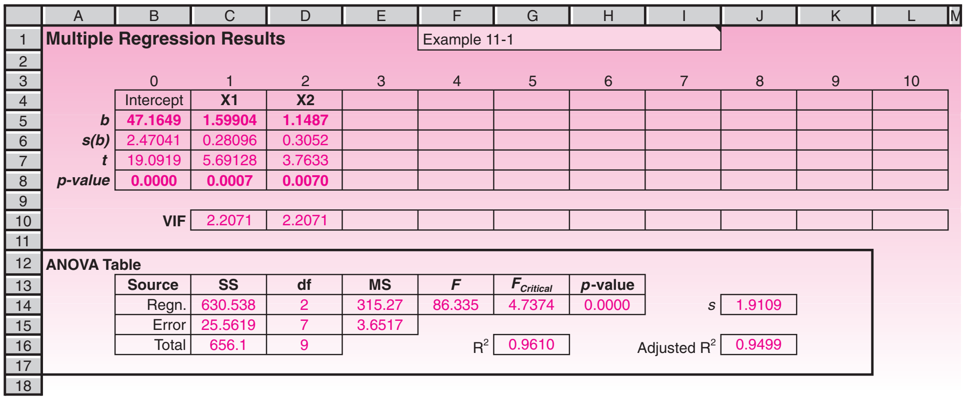
These are the *least-squares estimates* of the true regression parameters β_0, β_1, and β_2. Recall that the normal equations (equations 11–5) are originally obtained by calculus methods. (They are the results of differentiating the sum of squared errors with respect to the regression coefficients and setting the results to zero.)

Figure 11–2 shows the results page of the template on which the same problem has been solved. The template is described later.

The meaning of the estimates b_0, b_1, and b_2 as the Y intercept, the slope with respect to X_1, and the slope with respect to X_2, respectively, of the estimated regression surface is illustrated in Figure 11–3.

The general multiple regression model, equation 11–1, has one Y intercept parameter and k slope parameters. Each slope parameter β_i, $i = 1, \ldots, k$, represents the amount of increase (or decrease, in case it is negative) in $E(Y)$ for an increase of 1 unit in variable X_i when all other variables are kept constant. The regression coefficients β_i are therefore sometimes referred to as *net regression coefficients* because they represent the net change in $E(Y)$ for a change of 1 unit in the variable they represent, all else

FIGURE 11–2 The Results from the Template [Multiple Regression.xls; Sheet: Results]



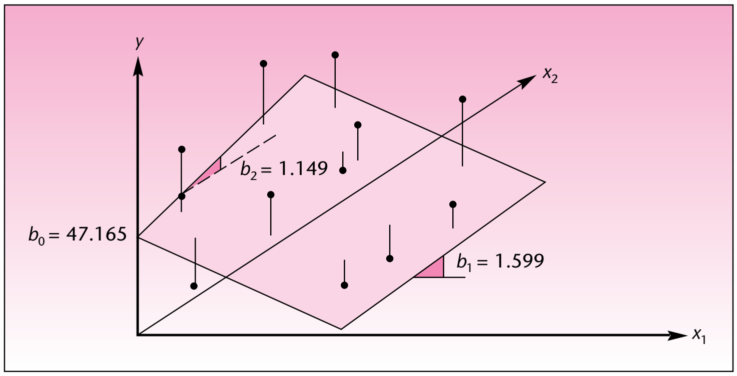
Multiple Regression Results Example 11-1

	0	1	2	3	4	5	6	7	8	9	10
	Intercept	**X1**	**X2**								
b	**47.1649**	**1.59904**	**1.1487**								
s(b)	2.47041	0.28096	0.3052								
t	19.0919	5.69128	3.7633								
p-value	**0.0000**	**0.0007**	**0.0070**								
VIF		2.2071	2.2071								

ANOVA Table

Source	SS	df	MS	*F*	$F_{Critical}$	*p*-value		
Regn.	630.538	2	315.27	86.335	4.7374	0.0000	*s*	1.9109
Error	25.5619	7	3.6517					
Total	656.1	9		R^2	0.9610		Adjusted R^2	0.9499

FIGURE 11–3 The Least-Squares Regression Surface for Example 11–1



remaining constant.[2] This is often difficult to achieve in multiple regression analysis since the explanatory variables are often interrelated in some way.

The Estimated Regression Relationship

CHAPTER 17

The **estimated regression relationship** is

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + \cdots + b_kX_k \qquad (11\text{–}6)$$

where $\hat{Y}$ is the predicted value of Y, the value lying *on* the estimated regression surface. The terms b_i, $i = 0, \ldots, k$, are the least-squares estimates of the population regression parameters β_i.

The least-squares estimators giving us the b_i are BLUEs (best linear unbiased estimators).

[2] For the reader with knowledge of calculus, we note that the coefficient β_i is the partial derivative of $E(Y)$ with respect to X_i: $\beta_i = \partial E(Y)/\partial X_i$.

The estimated regression relationship can also be written in a way that shows how each value of Y is expressed as a linear combination of the values of X_i plus an error term. This is given in equation 11–7.

$$y_j = b_0 + b_1 x_{1j} + b_2 x_{2j} + \cdots + b_k x_{kj} + e_j \qquad j = 1, \ldots, n \qquad (11\text{–}7)$$

In Example 11–1 the estimated regression relationship of sales volume Y on advertising X_1 and in-store promotions X_2 is given by

$$\hat{Y} = 47.164942 + 1.5990404X_1 + 1.1487479X_2$$

PROBLEMS

11–1. What are the assumptions underlying the multiple regression model? What is the purpose of the assumption of normality of the errors?

11–2. In a regression analysis of sales volume Y versus the explanatory variables advertising expenditure X_1 and promotional expenditures X_2, the estimated coefficient b_2 is equal to 1.34. Explain the meaning of this estimate in terms of the impact of promotional expenditure on sales volume.

11–3. In terms of model assumptions, what is the difference between a multiple regression model with k independent variables and a correlation analysis involving these variables?

11–4. What is a response surface? For a regression model with seven independent variables, what is the dimensionality of the response surface?

11–5. Again, for a multiple regression model with $k = 7$ independent variables, how many normal equations are there leading to the values of the estimates of the regression parameters?

11–6. What are the BLUEs of the regression parameters?

11–7. For a multiple regression model with two independent variables, results of the analysis include $\Sigma y = 852$, $\Sigma x_1 = 155$, $\Sigma x_2 = 88$, $\Sigma x_1 y = 11{,}423$, $\Sigma x_2 y = 8{,}320$, $\Sigma x_1 x_2 = 1{,}055$, $\Sigma x_1^2 = 2{,}125$, and $\Sigma x_2^2 = 768$, $n = 100$. Solve the normal equations for this regression model, and give the estimates of the parameters.

11–8. A realtor is interested in assessing the impact of size (in square feet) and distance from the center of town (in miles) on the value of homes (in thousands of dollars) in a certain area. Nine randomly chosen houses are selected; data are as follows.

Y (value):	345, 238, 452, 422, 328, 375, 660, 466, 290
X_1 (size):	1,650, 1,870, 2,230, 1,740, 1,900, 2,000, 3,200, 1,860, 1,230
X_2 (distance):	3.5, 0.5, 1.5, 4.5, 1.8, 0.1, 3.4, 3.0, 1.0

Compute the estimated regression coefficients, and explain their meaning.

11–9. The estimated regression coefficients in Example 11–1 are $b_0 = 47.165$, $b_1 = 1.599$, and $b_2 = 1.149$ (rounded to three decimal places). Explain the meaning of each of the three numbers in terms of the situation presented in the example.

11–3 The *F* Test of a Multiple Regression Model

The first statistical test we need to conduct in our evaluation of a multiple regression model is a test that will answer the basic question: Is there a linear regression relationship between the dependent variable Y and *any* of the explanatory, independent variables X_i suggested by the regression equation under consideration? If the proposed regression relationship is given in equation 11–1, a statistical test that can answer this important question is as follows.

A statistical hypothesis test for the existence of a linear relationship between Y and any of the X_i is

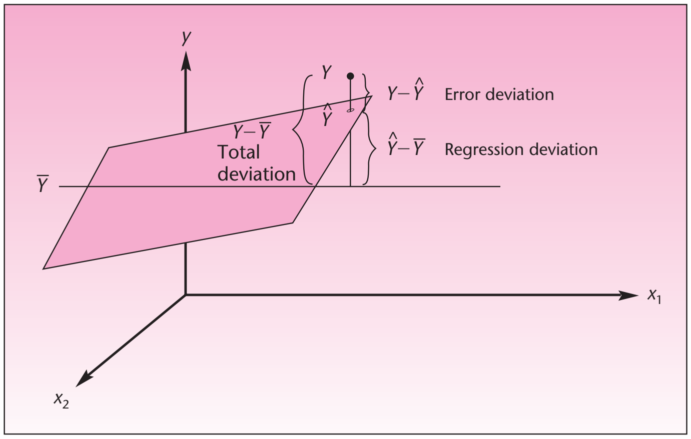
$$\begin{aligned} &H_0\text{: } \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_k = 0 \\ &H_1\text{: Not all the } \beta_i \ (i = 1, \ldots, k) \text{ are zero} \end{aligned} \qquad (11\text{–}8)$$

If the null hypothesis is true, no linear relationship exists between Y and any of the independent variables in the proposed regression equation. In such a case, there is nothing more to do. There is no regression. If, on the other hand, we reject the null hypothesis, there is statistical evidence to conclude that a regression relationship exists between Y and at least one of the independent variables proposed in the regression model.

To carry out the important test in equation 11–8, we will perform an analysis of variance. The ANOVA is the same as the one given in Chapter 10 for simple linear regression, except that here we have k independent variables instead of just 1. Therefore, the F test of the analysis of variance is not equivalent to the t test for the significance of the slope parameter, as was the case in Chapter 10. Since in multiple regression there are k slope parameters, we have k different t tests to follow the ANOVA.

Figure 11–4 is an extension of Figure 10–21 to the case of $k = 2$ independent variables–to a regression plane instead of a regression line. The figure shows a particular data point y, the predicted point $\hat{y}$ which lies on the estimated regression surface, and the mean of the dependent variable $\bar{y}$. The figure shows the three deviations associated with the data point: the error deviation $y - \hat{y}$, the regression deviation $\hat{y} - \bar{y}$, and the total deviation $y - \bar{y}$. As seen from the figure, the three deviations satisfy the relation: Total deviation = Regression deviation + Error deviation. As in the case of simple linear regression, when we square the deviations and sum them over all n data points, we get the following relation for the sums of squares. The sums of

FIGURE 11–4 Decomposition of the Total Deviation in Multiple Regression Analysis



squares are denoted by SST for the total sum of squares, SSR for the regression sum of squares, and SSE for the error sum of squares.

$$\text{SST} = \text{SSR} + \text{SSE} \tag{11–9}$$

This is the same as equation 10–29. The difference lies in the degrees of freedom. In simple linear regression, the degrees of freedom for error were $n - 2$ because two parameters, an intercept and a slope, were estimated from a data set of n points. In multiple regression, we estimate k slope parameters and an intercept from a data set of n points. Therefore, the degrees of freedom for error are $n - (k + 1)$. The degrees of freedom for the regression are k, and the total degrees of freedom are $n - 1$. Again, the degrees of freedom are additive. Table 11–2 is the ANOVA table for a multiple regression model with k independent variables.

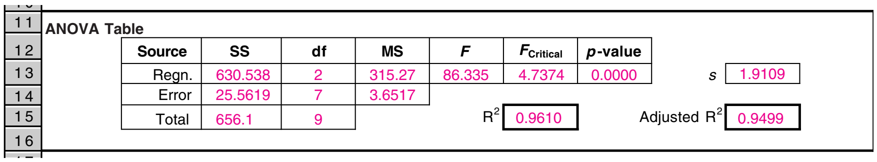
For Example 11–1, we present the ANOVA table computed by using the template. The results are shown in Table 11–3. Since the p-value is small, we reject the null hypothesis that both slope parameters β_1 and β_2 are zero (equation 11–8), in favor of the alternative that the slope parameters are not both zero. We conclude that there is evidence of a linear regression relationship between sales and at least one of the two variables, advertising or in-store promotions (or both). The F test is shown in Figure 11–5.

Note that since Example 11–1 has two independent variables, we do not yet know whether there is a regression relationship between sales and both advertising and in-store promotions, or whether the relationship exists between sales and one of the two variables only–and if so, which one. All we know is that our data present statistical evidence to conclude that a relationship exists between sales and at least one of the two independent variables. This is, of course, true for all cases with two or more independent variables. The F test only tells us that there is evidence of a relationship between the dependent variable and at least one of the independent variables in the full regression equation under consideration. Once we conclude that a relationship exists, we need to conduct separate tests to determine which of the slope parameters β_i, where $i = 1, \ldots, k$, are different from zero. Therefore, k further tests are needed.

TABLE 11–2 ANOVA Table for Multiple Regression

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	*F* Ratio
Regression	SSR	k	$\text{MSR} = \dfrac{\text{SSR}}{k}$	$F = \dfrac{\text{MSR}}{\text{MSE}}$
Error	SSE	$n - (k + 1)$	$\text{MSE} = \dfrac{\text{SSE}}{n - (k + 1)}$	
Total	SST	$n - 1$		

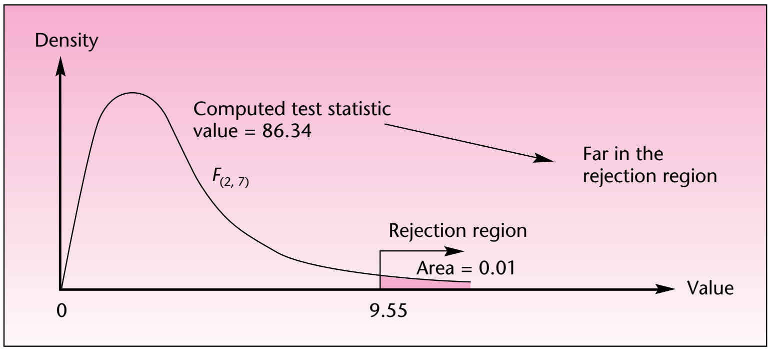
TABLE 11–3 ANOVA Table Produced by the Template [Multiple Regression.xls; Sheet: Results]



ANOVA Table

Source	SS	df	MS	*F*	F_{Critical}	*p*-value		
Regn.	630.538	2	315.27	86.335	4.7374	0.0000	*s*	1.9109
Error	25.5619	7	3.6517					
Total	656.1	9		R^2	0.9610		Adjusted R^2	0.9499

FIGURE 11–5 Regression *F* Test for Example 11–1



Compare the use of ANOVA tables in multiple regression with the analysis of variance discussed in Chapter 9. Once we rejected the null hypothesis that all *r* population means are equal, we required further analysis (the Tukey procedure or an alternative technique) to determine where the differences existed. In multiple regression, the further tests necessary for determining which variables are important are *t* tests. These tests tell us which variables help explain the variation in the values of the dependent variable and which variables have no explanatory power and should be eliminated from the regression model. Before we get to the separate tests of multiple regression parameters, we want to be able to evaluate how good the regression relationship is as a whole.

PROBLEMS

11–10. Explain what is tested by the hypothesis test in equation 11–8. What conclusion should be reached if the null hypothesis is not rejected? What conclusion should be reached if the null hypothesis is rejected?

11–11. In a multiple regression model with 12 independent variables, what are the degrees of freedom for error? Explain.

11–12. A study was reported about the effects of the number of hours worked, on average, and the average hourly income on unemployment in different countries.[3] Suppose that the regression analysis resulted in SSE = 8,650, SSR = 988, and the sample size was 82 observations. Is there a regression relationship between the unemployment rate and at least one of the explanatory variables?

11–13. Avis is interested in estimating weekly costs of maintenance of its rental cars of a certain size based on these variables: number of miles driven during the week, number of renters during the week, the car's total mileage, and the car's age. A regression analysis is carried out, and the results include $n = 45$ cars (each car selected randomly, during a randomly selected week of operation), SSR = 7,768, and SST = 15,673. Construct a complete ANOVA table for this problem, and test for the existence of a linear regression relationship between weekly maintenance costs and any of the four independent variables considered.

[3] Christopher A. Pissarides, "Unemployment and Hours of Work," *International Economic Review,* February 2007, pp. 1–36.

11–14. Nissan Motor Company wanted to find leverage factors for marketing the Maxima model in the United States. The company hired a market research firm in New York City to carry out an analysis of the factors that make people favor the model in question. As part of the analysis, the market research firm selected a random sample of 17 people and asked them to fill out a questionnaire about the importance of three automobile characteristics: prestige, comfort, and economy. Each respondent reported the importance he or she gave to each of the three attributes on a 0–100 scale. Each respondent then spent some time becoming acquainted with the car's features and drove it on a test run. Finally, each of the respondents gave an overall appeal score for the model on a 0–100 scale. The appeal score was considered the dependent variable, and the three attribute scores were considered independent variables. A multiple regression analysis was carried out, and the results included the following ANOVA table. Complete the table. Based on the results, is there a regression relationship between the appeal score and at least one of the attribute variables? Explain.

```
Analysis of Variance
  SOURCE      DF      SS      MS
Regression          7474.0
Error
Total               8146.5
```

11–4 How Good Is the Regression?

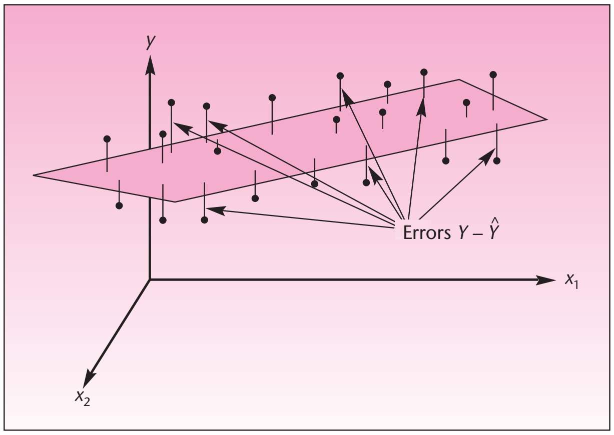
The mean square error MSE is an unbiased estimator of the variance of the population errors ϵ, which we denote by σ^2. The mean square error is defined in equation 11–10.

The **mean square error** is

$$\text{MSE} = \frac{\text{SSE}}{n - (k + 1)} = \frac{\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}{n - (k + 1)} \tag{11–10}$$

The errors resulting from the fit of a regression surface to our set of n data points are shown in Figure 11–6. The smaller the errors, the better the fit of the regression

FIGURE 11–6 **Errors in a Multiple Regression Model (shown for $k = 2$)**



model. Since the mean square error is the average squared error, where averaging is done by dividing by the degrees of freedom, MSE is a measure of how well the regression fits the data. The square root of MSE is an estimator of the standard deviation of the population regression errors σ. (Note that a square root of an unbiased estimator is not unbiased; therefore, $\sqrt{\text{MSE}}$ is not an unbiased estimator of σ, but is still a good estimator.) The square root of MSE is usually denoted by s and is referred to as the *standard error of estimate.*

The **standard error of estimate** is

$$s = \sqrt{\text{MSE}} \qquad (11\text{–}11)$$

This statistic is usually reported in computer output of multiple regression analysis. The mean square error and its square root are measures of the size of the errors in regression and give no indication about the *explained* component of the regression fit (see Figure 11–4, showing the breakdown of the total deviation of any data point to the error and regression components). A measure of regression fit that does incorporate the explained as well as the unexplained components is the *multiple coefficient of determination,* denoted by R^2. This measure is an extension to multiple regression of the coefficient of determination in simple linear regression, denoted by r^2.

The **multiple coefficient of determination R^2** measures the proportion of the variation in the dependent variable that is explained by the combination of the independent variables in the multiple regression model:

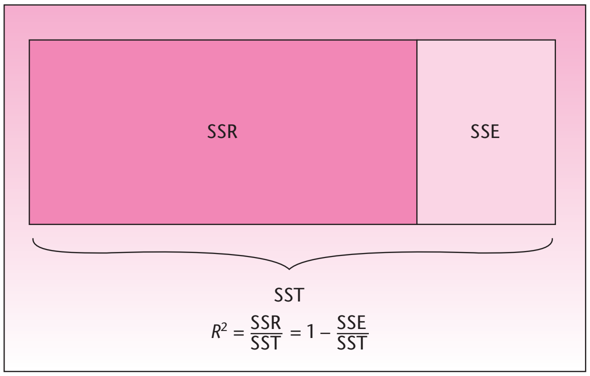
$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}} \qquad (11\text{–}12)$$

Note that R^2 is also equal to SSR/SST because SST = SSR + SSE. We prefer the definition in equation 11–12 for consistency with another measure of how well the regression model fits our data, the *adjusted* multiple coefficient of determination, which will be introduced shortly.

The measures SSE, SSR, and SST are reported in the ANOVA table for multiple regression. Because of the importance of R^2, however, it is reported separately in computer output of multiple regression analysis. The square root of the multiple coefficient of determination, $R = \sqrt{R^2}$, is the **multiple correlation coefficient.** In the context of multiple regression analysis (rather than correlation analysis), the multiple coefficient of determination R^2 is the important measure, not R. The coefficient of determination measures the percentage of variation in Y explained by the X variables; thus, it is an important measure of how well the regression model fits the data. In correlation analysis, where the X_i variables as well as Y are assumed to be random variables, the multiple correlation coefficient R measures the strength of the linear relationship between Y and the k variables X_i.

Figure 11–7 shows the breakdown of the total sum of squares (the sum of squared deviations of all n data points from the mean of Y; see Figure 11–6) into the sum of squares due to the regression (the explained variation) and the sum of squares due to error (the unexplained variation). The interpretation of R^2 is the same as that of r^2 in simple linear regression. The difference is that here the regression errors are measured as deviations from a regression surface that has higher dimensionality than a regression line. The multiple coefficient of determination R^2 is a very useful measure of performance of a multiple regression model. It does, however, have some limitations.

FIGURE 11–7 **Decomposition of the Sum of Squares in Multiple Regression, and the Definition of R^2**



Recall the story at the beginning of this chapter about the student who wanted to predict the nation's economic future with a multiple regression model that had many variables. It turns out that, for any given data set of n points, as the number of variables in the regression model increases, so does R^2. You have already seen how this happens: The greater the number of variables in the regression equation, the more the regression surface "chases" the data until it overfits them. Since the fit of the regression model increases as we increase the number of variables, R^2 cannot decrease and approaches 1.00, or 100% explained variation in Y. This can be very deceptive, as the model–while appearing to fit the data very well–would produce poor predictions.

Therefore, a new measure of fit of a multiple regression model must be introduced: the *adjusted* (or corrected) *multiple coefficient of determination.* The adjusted multiple coefficient of determination, denoted $\bar{R}^2$, is the multiple coefficient of determination corrected for degrees of freedom. It accounts, therefore, not only for SSE and SST, but also for their appropriate degrees of freedom. This measure does not always increase as new variables are entered into our regression equation. When $\bar{R}^2$ does increase as a new variable is entered into the regression equation, including the variable in the equation may be worthwhile. The adjusted measure is defined as follows:

The **adjusted multiple coefficient of determination** is

$$\bar{R}^2 = 1 - \frac{\text{SSE}/[n - (k + 1)]}{\text{SST}/(n - 1)} \qquad (11\text{–}13)$$

The adjusted R^2 is the R^2 (defined in equation 11–12) where both SSE and SST are divided by their respective degrees of freedom. Since SSE/$[n - (k + 1)]$ is the MSE, we can say that, in a sense, $\bar{R}^2$ is a mixture of the two measures of the performance of a regression model: MSE and R^2. The denominator on the right-hand side of equation 11–13 would be *mean square total,* were we to define such a measure.

Computer output for multiple regression analysis usually includes the adjusted R^2. If it is not reported, we can get $\bar{R}^2$ from R^2 by a simple formula:

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n - 1}{n - (k + 1)} \qquad (11\text{–}14)$$

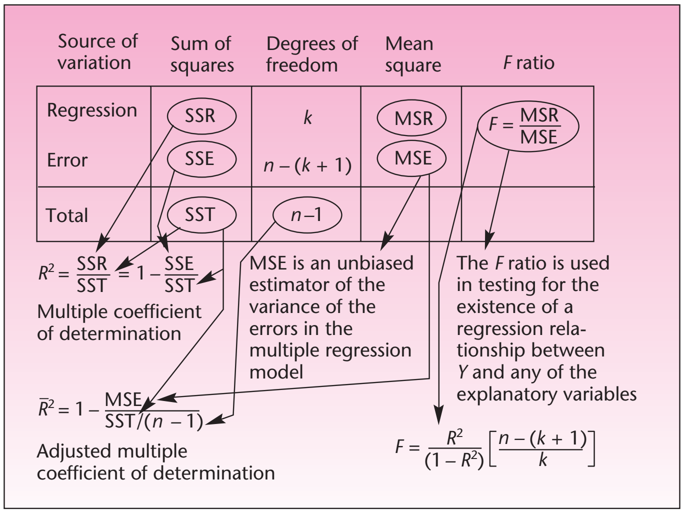
The proof of the relation between R^2 and $\bar{R}^2$ has instructional value and is left as an exercise. *Note:* Unless the number of variables is relatively large compared to the number of data points (as in the economics student's problem), R^2 and $\bar{R}^2$ are close to each other in value. Thus, in many situations, consideration of only the uncorrected measures R^2 is sufficient. We evaluate the fit of a multiple regression model based on this measure. When we are considering whether to include an independent variable in a regression model that already contains other independent variables, the increase in R^2 when the new variable is added must be weighed against the loss of 1 degree of freedom for error resulting from the addition of the variable (a new parameter would be added to the equation). With a relatively small data set and several independent variables in the model, adding a new variable if R^2 increases, say, from 0.85 to 0.86, may not be worthwhile. As mentioned earlier, in such cases, the adjusted measure $\bar{R}^2$ may be a good indicator of whether to include the new variable. We may decide to include the variable if $\bar{R}^2$ increases when the variable is added.

Of several possible multiple regression models with different independent variables, the model that minimizes MSE will also maximize $\bar{R}^2$. This should not surprise you, since MSE is related to the adjusted measure $\bar{R}^2$. The use of the two criteria MSE and $\bar{R}^2$ in selecting variables to be included in a regression model will be discussed in a later section.

We now return to the analysis of Example 11–1. Note that in Table 11–3 $R^2 = 0.961$, which means that 96.1% of the variation in sales volume is explained by the combination of the two independent variables, advertising and in-store promotions. Note also that the adjusted R^2 is 0.95, which is very close to the unadjusted measure. We conclude that the regression model fits the data very well since a high percentage of the variation in Y is explained by X_1, and/or X_2 (we do not yet know which of the two variables, if not both, is important). The standard error of estimate s is an estimate of σ, the standard deviation of the population regression errors. Note that R^2 is also a *statistic,* like s or MSE. It is a sample estimate of the population multiple coefficient of determination ρ^2, a measure of the proportion of the explained variation in Y in the entire population of Y and X_i values.

All three measures of the performance of a regression model—MSE (and its square root s), the coefficient of determination R^2, and the adjusted measure $\bar{R}^2$—are obtainable from quantities reported in the ANOVA table. This is shown in Figure 11–8, which demonstrates the relations among the different measures.

FIGURE 11–8 Measures of Performance of a Regression Model and the ANOVA Table


Source of variation
Sum of squares
Degrees of freedom
Mean square
F ratio
Regression
SSR
k
MSR
$F = \frac{\text{MSR}}{\text{MSE}}$
Error
SSE
$n - (k + 1)$
MSE
Total
SST
$n-1$
$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$
Multiple coefficient of determination
MSE is an unbiased estimator of the variance of the errors in the multiple regression model
The F ratio is used in testing for the existence of a regression relationship between Y and any of the explanatory variables
$\bar{R}^2 = 1 - \frac{\text{MSE}}{\text{SST}/(n-1)}$
Adjusted multiple coefficient of determination
$F = \frac{R^2}{(1 - R^2)}\left[\frac{n - (k + 1)}{k}\right]$

PROBLEMS

11–15. Under what conditions is it important to consider the adjusted multiple coefficient of determination?

11–16. Explain why the multiple coefficient of determination never decreases as variables are added to the multiple regression model.

11–17. Would it be useful to consider an adjusted coefficient of determination in a simple linear regression situation? Explain.

11–18. Prove equation 11–14.

11–19. Can you judge how well a regression model fits the data by considering the mean square error only? Explain.

11–20. A regression analysis was carried out of the stock return on the first day of an IPO (initial public offering) based on four variables: assessed benefit of the IPO, assessed improved market perception, assessed perception of market strength at the time of the IPO, and assessed growth potential due to patent or copyright ownership. The adjusted R^2 was 2.1%, and the F value was 2.27. The sample consisted of 438 responses from the chief financial officers of firms who issued IPOs from January 1, 1996, through June 15, 2002.[4] Analyze these results.

11–21. A portion of the regression output for the Nissan Motor Company study of problem 11–14 follows. Interpret the findings, and show how these results are obtainable from the ANOVA table results presented in problem 11–14. How good is the regression relationship between the overall appeal score for the automobile and the attribute-importance scores? Also, obtain the adjusted R^2 from the multiple coefficient of determination.

```
s = 7.192        R² = 91.7%        R² (ADJ) = 89.8%
```

11–22. A study of the market for mortgage-backed securities included a regression analysis of security effects and time effects on market prices as dependent variable. The sample size was 383 and the R^2 was 94%.[5] How good is this regression? Would you confidently predict market price based on security and time effects? Explain.

11–23. In the Nissan Motor Company situation in problem 11–21, suppose that a new variable is considered for inclusion in the equation and a new regression relationship is analyzed with the new variable included. Suppose that the resulting multiple coefficient of determination is $R^2 = 91.8\%$. Find the adjusted multiple coefficient of determination. Should the new variable be included in the final regression equation? Give your reasons for including or excluding the variable.

11–24. An article on pricing and competition in marketing reports the results of a regression analysis.[6] Information price was the dependent variable, and the independent variables were six marketing measures. The R^2 was 76.9%. Interpret the strength of this regression relationship. The number of data points was 242, and the F-test value was 44.8. Conduct the test and state your conclusions.

11–25. The following excerpt reports the results of a regression of excess stock returns on firm size and stock price, both variables being ranked on some scale. Explain, critique, and evaluate the reported results.

[4] James C. Brau, Patricia A. Ryan, and Irv DeGraw, "Initial Public Offerings: CFO Perceptions," *Financial Review* 41 (2006), pp. 483–511.

[5] Xavier Garbaix, Arvind Krishnamurthy, and Olivier Vigneron, "Limits of Arbitrage: Theory and Evidence from the Mortgage-Backed Securities Market," *Journal of Finance* 42, no. 2 (2007), pp. 557–595.

[6] Markus Christen and Miklos Sarvary, "Competitive Pricing of Information: A Longitudinal Experiment," *Journal of Marketing Research* 44 (February 2007), pp. 42–56.

Estimated Coefficient Value (*t* Statistic)

INTCPT	X1	X2	ADJUSTED-R^2
Ordinary Least-Squares Regression Results			
0.484	−0.030	−0.017	0.093
(5.71)***	(−2.91)***	(−1.66)*	

*Denotes significance at the 10% level.
**Denotes significance at the 5% level.
***Denotes significance at the 1% level.

11–26. A study of Dutch tourism behavior included a regression analysis using a sample of 713 respondents. The dependent variable, number of miles traveled on vacation, was regressed on the independent variables, family size and family income; and the multiple coefficient of determination was $R^2 = 0.72$. Find the adjusted multiple coefficient of determination $\overline{R}^2$. Is this a good regression model? Explain.

11–27. A regression analysis was carried out to assess sale prices of land in Uganda based on many variables that describe the owner of the land: age, educational level, number of males in the household, and more.[7] Suppose that there are eight independent variables, 500 data points, SSE = 6,179, and SST = 23,108. Construct an ANOVA table, conduct the F test, find R^2 and $\overline{R}^2$, and find the MSE.

11–5 Tests of the Significance of Individual Regression Parameters

Until now, we have discussed the multiple regression model in general. We saw how to test for the existence of a regression relationship between Y and at least one of a set of independent X_i variables by using an F test. We also saw how to evaluate the fit of the general regression model by using the multiple coefficient of determination and the adjusted multiple coefficient of determination. We have not yet seen, however, how to evaluate the significance of individual regression parameters β_i. A test for the significance of an individual parameter is important because it tells us whether the variable in question, X_h, has explanatory power with respect to the dependent variable. Such a test tells us whether the variable in question should be included in the regression equation.

In the last section, we saw that some indication about the benefit from inclusion of a particular variable in the regression equation is gained by comparing the adjusted coefficient of determination of a regression that includes the variable of interest with the value of this measure when the variable is not included. In this section, we will perform individual t tests for the significance of each slope parameter β_i. As we will see, however, we must use caution in interpreting the results of the individual t tests.

In Chapter 10 we saw that the hypothesis test

$$\boxed{\begin{aligned} &H_0: \beta_1 = 0 \\ &H_1: \beta_1 \neq 0 \end{aligned}}$$

can be carried out using either a t statistic $t = b_1/s(b_1)$ or an F statistic. Both tests were shown to be equivalent because F with 1 degree of freedom for the numerator is a squared t random variable with the same number of degrees of freedom as the denominator of F. A simple linear regression has only one slope, β_1, and if that slope is zero, there is no linear regression relationship. In multiple regression, where $k > 1$, the two

[7] J.M. Baland et al., "The Distributive Impact of Land Markets in Uganda," *Economic Development and Cultural Change* 55, no. 2 (2007), pp. 283–311.

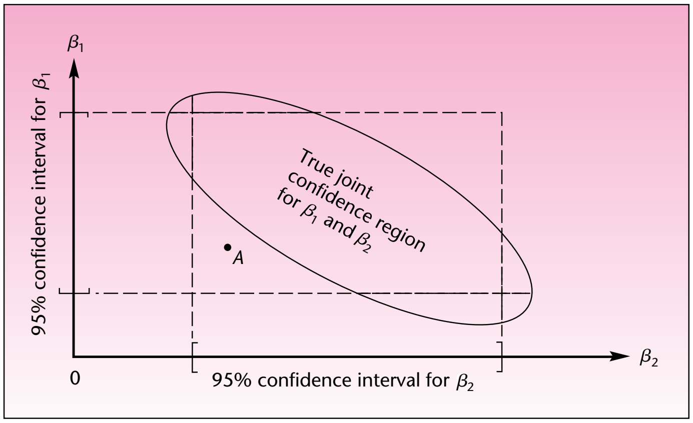
tests are not equivalent. The F test tells us whether a relationship exists between Y and at least one of the X_i, and the k ensuing t tests tell us which of the X_i variables are important and should be included in the regression equation. From the similarity of this situation with the situation of analysis of variance discussed in Chapter 9, you probably have guessed at least one of the potential problems: The individual t tests are each carried out at a single level of significance α, and we cannot determine the level of significance of the family of all k tests of the regression slopes jointly. The problem is further complicated by the fact that the tests are not independent of each other because the regression estimates come from the same data set.

Recall that hypothesis tests and confidence intervals are related. We may test hypotheses about regression slope parameters (in particular, the hypothesis that a slope parameter is equal to zero), or we may construct confidence intervals for the values of the slope parameters. If a 95% confidence interval for a slope parameter β_h contains the point zero, then the hypothesis test H_0: $\beta_h = 0$ carried out using $\alpha = 0.05$ would lead to nonrejection of the null hypothesis and thus to the conclusion that there is no evidence that the variable X_h has a linear relationship with Y.

We will demonstrate the interdependence of the separate tests of significance of the slope parameters with the use of confidence intervals for these parameters. When $k = 2$, there are two regression slope parameters: β_1 and β_2. (As in simple linear regression, usually there is no interest in testing hypotheses about the intercept parameter.) The sample estimators of the two regression parameters are b_1 and b_2. These estimators (and their standard errors) are correlated with each other (and assumed to be normally distributed). Therefore, the joint confidence region for the pair of parameters (β_1, β_2) is an *ellipse.* If we consider the estimators b_1 and b_2 separately, the joint confidence region will be a rectangle, with each side a separate confidence interval for a single parameter. This is demonstrated in Figure 11–9. A point inside the rectangle formed by the two separate confidence intervals for the parameters, such as point A in the figure, seems like a plausible value for the pair of regression slopes (β_1, β_2) but is not *jointly* plausible for the parameters. Only points inside the ellipse in the figure are jointly plausible for the pair of parameters.

Another problem that may arise in making inferences about individual regression slope coefficients is due to **multicollinearity**—the problem of correlations among the independent variables themselves. In multiple regression, we hope to have a strong correlation between each independent variable and the dependent

FIGURE 11–9 Joint Confidence Region and Individual Confidence Intervals for the Slope Parameters β_1 and β_2


β_1
95% confidence interval for β_1
True joint confidence region for β_1 and β_2
A
0
95% confidence interval for β_2
β_2

variable Y. Such correlations give the independent X_i variables predictive power with respect to Y. However, we do not want the independent variables to be correlated with one another. When the independent variables are correlated with one another, we have multicollinearity. When this happens, the independent variables rob one another of explanatory power. Many problems may then arise. One problem is that the standard errors of the individual slope estimators become unusually high, making the slope coefficients seem statistically not significant (not different from zero). For example, if we run a regression of job performance Y versus the variables age X_1 and experience X_2, we may encounter multicollinearity. Since, in general, as age increases so does experience, the two independent variables are not independent of each other; the two variables rob each other of explanatory power with respect to Y. If we run this regression, it is likely that—even though experience affects job performance—the individual test for significance of the slope parameter β_2 would lead to nonrejection of the null hypothesis that this slope parameter is equal to zero. Much will be said later about the problem of multicollinearity. Remember that in the presence of multicollinearity, the significance of any regression parameter depends on the other variables included in the regression equation. Multicollinearity may also cause the signs of some estimated regression parameters to be the opposite of what we expect.

Another problem that may affect the individual tests of significance of model parameters occurs when one of the model assumptions is violated. Recall from Section 11–2 that one of the assumptions of the regression model is that the error terms ϵ_j are uncorrelated with one another. When this condition does not hold, as may happen when our data are time series observations (observations ordered by time: yearly data, monthly data, etc.), we encounter the problem of autocorrelation of the errors. This causes the standard errors of the slope estimators to be unusually small, making some parameters seem more significant than they really are. This problem, too, should be considered, and we will discuss it in detail later.

Forewarned of problems that may arise, we now consider the tests of the individual regression parameters. In a regression model of Y versus k independent variables X_1, $X_2, \ldots, X_k$, we have k tests of significance of the slope parameters $\beta_1, \beta_2, \ldots, \beta_k$:

Hypothesis tests about individual regression slope parameters:

$$\begin{array}{ll}
(1) & H_0\colon \beta_1 = 0 \\
 & H_1\colon \beta_1 \neq 0 \\
(2) & H_0\colon \beta_2 = 0 \\
 & H_1\colon \beta_2 \neq 0 \\
\cdot & \quad\cdot \\
\cdot & \quad\cdot \\
\cdot & \quad\cdot \\
(k) & H_0\colon \beta_k = 0 \\
 & H_1\colon \beta_k \neq 0
\end{array} \qquad (11\text{–}15)$$

These tests are carried out by comparing each test statistic with a critical point of the distribution of the test statistic. The distribution of each test statistic, when the appropriate null hypothesis is true, is the t distribution with $n - (k + 1)$ degrees of freedom. The distribution depends on our assumption that the regression errors are normally distributed. The test statistic for each hypothesis test (i) in equations 11–15 (where $i = 1, 2, \ldots, k$) is the slope estimate b_i, divided by the standard error of the estimator $s(b_i)$. The estimates and the standard errors are reported in the computer output. Each $s(b_i)$ is an estimate of the population standard deviation of the estimator $\sigma(b_i)$, which is unknown to us.[8] The test statistics for the hypothesis tests (1) through (k) in equations 11–15 are as follows:

Test statistics for tests about individual regression slope parameters:

For test i ($i = 1, \ldots, k$):

$$t_{[n-(k+1)]} = \frac{b_i - 0}{s(b_i)} \qquad (11\text{–}16)$$

We write each test statistic as the estimate minus zero (the null-hypothesis value of β_i) to stress the fact that we may test the null hypothesis that β_i is equal to any number, not necessarily zero. Testing for equality to zero is most important because it tells us whether there is evidence that variable X_i has a linear relationship with Y. It tells us whether there is statistical evidence that variable X_i has explanatory power with respect to the dependent variable.

Let us look at a quick example. Suppose that a multiple regression analysis is carried out relating the dependent variable Y to five independent variables X_1, X_2, X_3, X_4, and X_5. In addition, suppose that the F test resulted in rejection of the null hypothesis that none of the predictor variables has any explanatory power with respect to Y; suppose also that R^2 of the regression is respectably high. As a result, we believe that the regression equation gives a good fit to the data and potentially may be used for prediction purposes. Our task now is to test the importance of each of the X_i variables separately. Suppose that the sample size used in this regression analysis is $n = 150$. The results of the regression estimation procedure are given in Table 11–4.

From the information in Table 11–4, which variables are important, and which are not? Note that the first variable listed is "Constant." This is the Y intercept. As we noted earlier, testing whether the intercept is zero is less important than testing whether the coefficient parameter of any of the k variables is zero. Still, we may do so by dividing the reported coefficient estimate, 53.12, by its standard error, 5.43. The result is the value of the test statistic that has a t distribution with $n - (k + 1) = 150 - 6 = 144$ degrees of freedom when the null hypothesis that the intercept is zero is true. For manual calculation purposes, we shall approximate this t random variable as a standard normal variable Z. The test statistic value is $z = 53.12/5.43 = 9.78$. This value is greater than 1.96, and we may reject the null hypothesis that β_0 is equal to zero at the $\alpha = 0.05$ level of significance. Actually, the p-value is very small. The regression hyperplane, therefore, most probably does not pass through the origin.

TABLE 11–4 Regression Results for Individual Parameters

Variable	Coefficient Estimate	Standard Error
Constant	53.12	5.43
X_1	2.03	0.22
X_2	5.60	1.30
X_3	10.35	6.88
X_4	3.45	2.70
X_5	−4.25	0.38

[8] Each $s(b_i)$ is the product of $s = \sqrt{\text{MSE}}$ and a term denoted by c_i, which is a diagonal element in a matrix obtained in the regression computations. You need not worry about matrices. However, the matrix approach to multiple regression is discussed in a section at the end of this chapter for the benefit of students familiar with matrix theory.

Let us now turn to the tests of significance of the slope parameters of the variables in the regression equation. We start with the test for the significance of variable X_1 as a predictor variable. The hypothesis test is H_0: $\beta_1 = 0$ versus H_1: $\beta_1 \neq 0$. We now compute our test statistic (again, we will use Z for $t_{(144)}$):

$$z = \frac{b_1 - 0}{s(b_1)} = \frac{2.03}{0.22} = 9.227$$

The value of the test statistic, 9.227, lies far in the right-hand rejection region of Z for any conventional level of significance; the p-value is very small. We therefore conclude that there is statistical evidence that the slope of Y with respect to X_1, the population parameter β_1, is not zero. Variable X_1 is shown to have some explanatory power with respect to the dependent variable.

If it is not zero, what is the value of β_1? The parameter, as in the case of all population parameters, is not known to us. An unbiased estimate of the parameter's value is $b_1 = 2.03$. We can also compute a confidence interval for β_1. A 95% confidence interval for β_1 is $b_1 \pm 1.96s(b_1) = 2.03 \pm 1.96(0.22) = [1.599, 2.461]$. Based on our data and the validity of our assumptions, we can be 95% confident that the true slope of Y with respect to X_1 is anywhere from 1.599 to 2.461. Figure 11–10 shows the hypothesis test for the significance of variable X_1.

For the other variables X_2 through X_5, we show the hypothesis tests without figures. The tests are carried out in the same way, with the same distribution. We also do not show the computation of confidence intervals for the slope parameters. These are done exactly as shown for β_1. Note that when the hypothesis test for the significance of a slope parameter leads to nonrejection of the null hypothesis that the slope parameter is zero, the point zero will be included in a confidence interval with the same confidence level as the level of significance of the test.

The hypothesis test for β_2 is H_0: $\beta_2 = 0$ versus H_1: $\beta_2 \neq 0$. The test statistic value is $z = 5.60/1.30 = 4.308$. This value, too, is in the right-hand rejection region for usual levels of significance; the p-value is small. We conclude that X_2 is also an important variable in the regression equation.

The hypothesis test for β_3 is H_0: $\beta_3 = 0$ versus H_1: $\beta_3 \neq 0$. Here the test statistic value is $z = 10.35/6.88 = 1.504$. This value lies in the nonrejection region for levels of α even larger than 0.10. The p-value is greater than 0.133, as you can verify from a normal table. We conclude that variable X_3 is probably not important. Remember our cautionary comments that preceded this discussion—there is a possibility that X_3 is actually an important variable. The variable may *appear* to have a

FIGURE 11–10 **Testing Whether $\beta_1 = 0$**

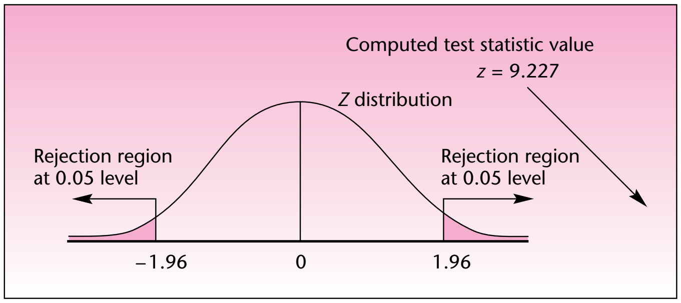
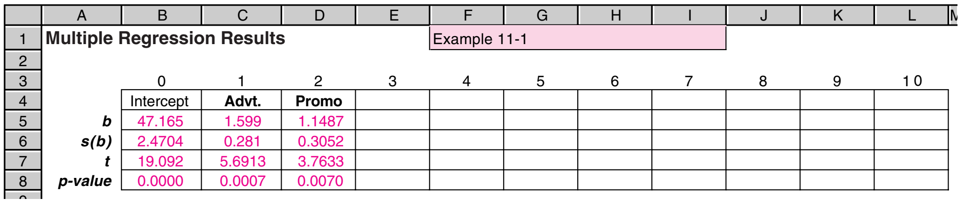


TABLE 11–5 **Multiple Regression Results from the Template [Multiple Regression.xls; Sheet: Results]**



	A	B	C	D	E	F	G	H	I	J	K	L
1	Multiple Regression Results					Example 11-1						
2												
3		0	1	2	3	4	5	6	7	8	9	10
4		Intercept	Advt.	Promo								
5	*b*	47.165	1.599	1.1487								
6	*s(b)*	2.4704	0.281	0.3052								
7	*t*	19.092	5.6913	3.7633								
8	*p-value*	0.0000	0.0007	0.0070								

slope that is not different from zero because its standard error, $s(b_3) = 6.88$, may be unduly inflated; the variable may be correlated with another explanatory variable (the problem of multicollinearity). A way out of this problem is to drop another variable, one that we suspect to be correlated with X_3, and see if X_3 becomes significant in the new regression model. We will come back to this problem in the section on multicollinearity and in the section on selection of variables to be included in a regression model.

The hypothesis test about β_4 is H_0: $\beta_4 = 0$ versus H_1: $\beta_4 \neq 0$. The value of the test statistic for this test is $z = 3.45/2.70 = 1.278$. Again, we cannot reject the null hypothesis that the slope parameter of X_4 is zero and that the variable has no explanatory power. Note, however, the caution in our discussion of the test of β_3. It is possible, for example, that X_3 and X_4 are collinear and that this is the reason for their respective tests resulting in nonsignificance. It would be wise to drop one of these two variables and check whether the other variable then becomes significant. If it does, the reason for our test result is multicollinearity, and not the absence of explanatory power of the variable in question. Another point worth mentioning is the idea of joint inference, discussed earlier. Although the separate tests of β_3 and β_4 both may lead to the nonrejection of the hypothesis that the parameters are zero, it may be that the two parameters are not jointly equal to zero. This would be the situation if, in Figure 11–9, the rectangle contained the point zero while the ellipse—the true joint confidence region for both parameters—did not contain that point. Note that the *t* tests are *conditional.* The significance or nonsignificance of a variable in the equation is conditional on the fact that the regression equation contains the other variables.

Finally, the test for parameter β_5 is H_0: $\beta_5 = 0$ versus H_1: $\beta_5 \neq 0$. The computed value of the test statistic is $z = -4.25/0.38 = -11.184$. This value falls far in the left-hand rejection region, and we conclude that variable X_5 has explanatory power with respect to the dependent variable and therefore should be included in the regression equation. The slope parameter is negative, which means that, everything else staying constant, the dependent variable Y decreases on average as X_5 increases. We note that these tests can be carried out very quickly by just considering the *p*-values.

We now return to Example 11–1 and look at the rest of the results from the template. For easy reference the results from Table 11–4 are repeated here in Table 11–5. As seen in the table, the test statistic *t* is very significant for both advertisement and promotion variables, because the *p*-value is less than 1% in both cases. We therefore declare that both of these variables affect the sales.

EXAMPLE 11–2

In recent years, many U.S. firms have intensified their efforts to market their products in the Pacific Rim. Among the major economic powers in that area are Japan, Hong Kong, and Singapore. A consortium of U.S. firms that produce raw materials used in Singapore is interested in predicting the level of exports from the United

States to Singapore, as well as understanding the relationship between U.S. exports to Singapore and certain variables affecting the economy of that country. Understanding this relationship would allow the consortium members to time their marketing efforts to coincide with favorable conditions in the Singapore economy. Understanding the relationship would also allow the exporters to determine whether expansion of exports to Singapore is feasible. The economist hired to do the analysis obtained from the Monetary Authority of Singapore (MAS) monthly data on five economic variables for the period of January 1989 to August 1995. The variables were U.S. exports to Singapore in billions of Singapore dollars (the dependent variable, Exports), money supply figures in billions of Singapore dollars (variable M1), minimum Singapore bank lending rate in percentages (variable Lend), an index of local prices where the base year is 1974 (variable Price), and the exchange rate of Singapore dollars per U.S. dollar (variable Exchange). The monthly data are given in Table 11–6.

TABLE 11–6 Example 11–2 Data

Row	Exports	M1	Lend	Price	Exchange
1	2.6	5.1	7.8	114	2.16
2	2.6	4.9	8.0	116	2.17
3	2.7	5.1	8.1	117	2.18
4	3.0	5.1	8.1	122	2.20
5	2.9	5.1	8.1	124	2.21
6	3.1	5.2	8.1	128	2.17
7	3.2	5.1	8.3	132	2.14
8	3.7	5.2	8.8	133	2.16
9	3.6	5.3	8.9	133	2.15
10	3.4	5.4	9.1	134	2.16
11	3.7	5.7	9.2	135	2.18
12	3.6	5.7	9.5	136	2.17
13	4.1	5.9	10.3	140	2.15
14	3.5	5.8	10.6	147	2.16
15	4.2	5.7	11.3	150	2.21
16	4.3	5.8	12.1	151	2.24
17	4.2	6.0	12.0	151	2.16
18	4.1	6.0	11.4	151	2.12
19	4.6	6.0	11.1	153	2.11
20	4.4	6.0	11.0	154	2.13
21	4.5	6.1	11.3	154	2.11
22	4.6	6.0	12.6	154	2.09
23	4.6	6.1	13.6	155	2.09
24	4.2	6.7	13.6	155	2.10
25	5.5	6.2	14.3	156	2.08
26	3.7	6.3	14.3	156	2.09
27	4.9	7.0	13.7	159	2.10
28	5.2	7.0	12.7	161	2.11
29	4.9	6.6	12.6	161	2.15
30	4.6	6.4	13.4	161	2.14
31	5.4	6.3	14.3	162	2.16
32	5.0	6.5	13.9	160	2.17
33	4.8	6.6	14.5	159	2.15
34	5.1	6.8	15.0	159	2.10
35	4.4	7.2	13.2	158	2.06
36	5.0	7.6	11.8	155	2.05

(*Continued*)

Row	Exports	M1	Lend	Price	Exchange
37	5.1	7.2	11.2	155	2.06
38	4.8	7.1	10.1	154	2.11
39	5.4	7.0	10.0	154	2.12
40	5.0	7.5	10.2	154	2.13
41	5.2	7.4	11.0	153	2.04
42	4.7	7.4	11.0	152	2.14
43	5.1	7.3	10.7	152	2.15
44	4.9	7.6	10.2	152	2.16
45	4.9	7.8	10.0	151	2.17
46	5.3	7.8	9.8	152	2.20
47	4.8	8.2	9.3	152	2.21
48	4.9	8.2	9.3	152	2.15
49	5.1	8.3	9.5	152	2.08
50	4.3	8.3	9.2	150	2.08
51	4.9	8.0	9.1	147	2.09
52	5.3	8.2	9.0	147	2.10
53	4.8	8.2	9.0	146	2.09
54	5.3	8.0	8.9	145	2.12
55	5.0	8.1	9.0	145	2.13
56	5.1	8.1	9.0	146	2.14
57	4.8	8.1	9.0	147	2.14
58	4.8	8.1	8.9	147	2.13
59	5.2	8.6	8.9	147	2.13
60	4.9	8.8	9.0	146	2.13
61	5.5	8.4	9.1	147	2.13
62	4.3	8.2	9.0	146	2.13
63	5.2	8.3	9.2	146	2.09
64	4.7	8.3	9.6	146	2.09
65	5.4	8.4	10.0	146	2.10
66	5.2	8.3	10.0	147	2.11
67	5.6	8.2	10.1	146	2.15

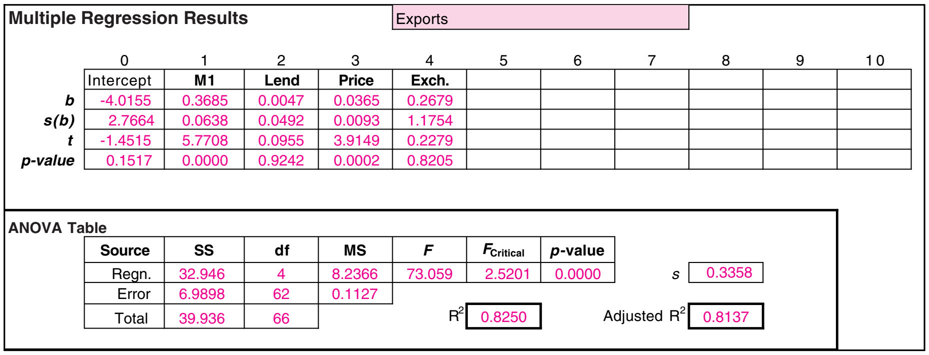
Solution

Use the template to perform a multiple regression analysis with Exports as the dependent variable and the four economic variables M1, Lend, Price, and Exchange as the predictor variables. Table 11–7 shows the results.

Let us analyze the regression results. We start with the ANOVA table and the F test for the existence of linear relationships between the independent variables and exports from the United States to Singapore. We have $F_{(4,\,62)} = 73.059$ with a p-value of "0.000." We conclude that there is strong evidence of a linear regression relationship here. This is further confirmed by noting that the coefficient of determination is high: $R^2 = 0.825$. Thus, the combination of the four economic variables explains 82.5% of the variation in exports to Singapore. The adjusted coefficient of determination $\overline{R}^2$ is a little smaller: 0.8137. Now the question is, Which of the four variables are important as predictors of export volume to Singapore and which are not? Looking at the reported p-values, we see that the Singapore money supply M1 is an important variable; the level of prices in Singapore is also an important variable. The remaining two variables, minimum lending rate and exchange rate, have very large p-values. Surprisingly, the lending rate and the exchange rate of Singapore dollars to U.S. dollars seem to have no effect on the volume of Singapore's imports from the United States. Remember, however, that we may have a problem of multicollinearity.

CHAPTER 17

TABLE 11–7 Regression Results from the Template for Exports to Singapore [Multiple Regression.xls]



Multiple Regression Results — Exports

	0	1	2	3	4	5	6	7	8	9	10
	Intercept	M1	Lend	Price	Exch.						
b	-4.0155	0.3685	0.0047	0.0365	0.2679						
s(b)	2.7664	0.0638	0.0492	0.0093	1.1754						
t	-1.4515	5.7708	0.0955	3.9149	0.2279						
p-value	0.1517	0.0000	0.9242	0.0002	0.8205						

ANOVA Table

Source	SS	df	MS	F	F_{Critical}	p-value
Regn.	32.946	4	8.2366	73.059	2.5201	0.0000
Error	6.9898	62	0.1127			
Total	39.936	66				

s 0.3358

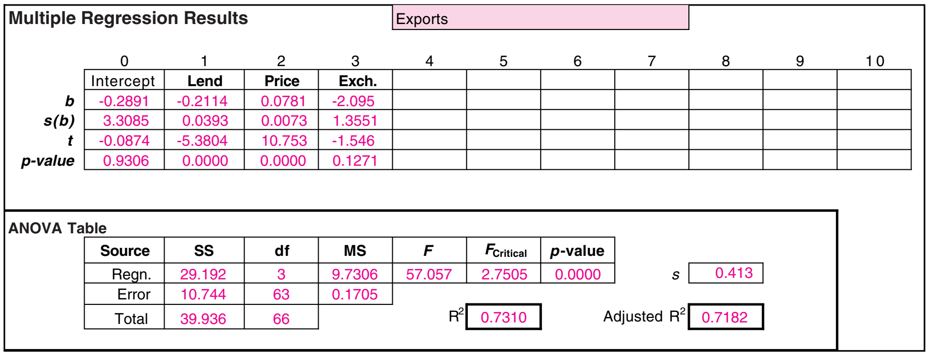
R^2 0.8250 Adjusted R^2 0.8137

This is especially true when we are dealing with economic variables, which tend to be correlated with one another.[9]

When M1 is dropped from the equation and the new regression analysis considers the independent variables Lend, Price, and Exchange, we see that the lending rate, which was not significant in the full regression equation, now becomes significant! This is seen in Table 11–8. Note that R^2 has dropped greatly with the removal of M1. The fact that the lending rate is significant in the new equation is an indication of *multicollinearity;* variables M1 and Lend are correlated with each other. Therefore, Lend is not significant when M1 is in the equation, but in the absence of M1, Lend does have explanatory power.

Note that the exchange rate is still not significant. Since R^2 and the adjusted R^2 both decrease significantly when the money supply M1 is dropped, let us put that variable back into the equation and run U.S. exports to Singapore versus the independent variables M1 and Price only. The results are shown in Table 11–9. In this regression equation, both independent variables are significant. Note that R^2 in this regression is virtually the same as R^2 with all four variables in the equation (see

TABLE 11–8 Regression Results for Singapore Exports without M1



Multiple Regression Results — Exports

	0	1	2	3	4	5	6	7	8	9	10
	Intercept	Lend	Price	Exch.							
b	-0.2891	-0.2114	0.0781	-2.095							
s(b)	3.3085	0.0393	0.0073	1.3551							
t	-0.0874	-5.3804	10.753	-1.546							
p-value	0.9306	0.0000	0.0000	0.1271							

ANOVA Table

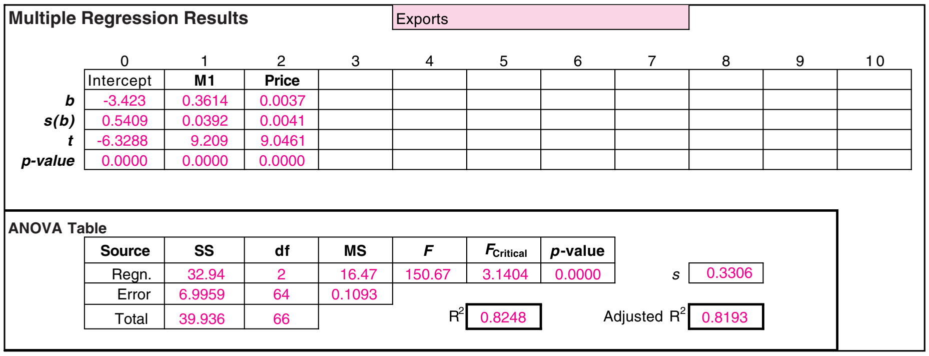
Source	SS	df	MS	F	F_{Critical}	p-value
Regn.	29.192	3	9.7306	57.057	2.7505	0.0000
Error	10.744	63	0.1705			
Total	39.936	66				

s 0.413

R^2 0.7310 Adjusted R^2 0.7182

[9] The analysis of economic variables presents special problems. Economists have developed methods that account for the intricate interrelations among economic variables. These methods, based on multiple regression and time series analysis, are usually referred to as *econometric methods.*

TABLE 11–9 Regressing Exports against M1 and Price



Multiple Regression Results Exports

	0	1	2	3	4	5	6	7	8	9	10
	Intercept	M1	Price								
b	-3.423	0.3614	0.0037								
s(b)	0.5409	0.0392	0.0041								
t	-6.3288	9.209	9.0461								
p-value	0.0000	0.0000	0.0000								

ANOVA Table

Source	SS	df	MS	*F*	$F_{Critical}$	*p*-value
Regn.	32.94	2	16.47	150.67	3.1404	0.0000
Error	6.9959	64	0.1093			
Total	39.936	66				

s 0.3306

R^2 0.8248

Adjusted R^2 0.8193

Table 11–7). However, the adjusted coefficient of determination $\bar{R}^2$ is different. The adjusted R^2 actually *increases* as we drop the variables Lend and Exchange. In the full model with the four variables (Table 11–7), $\bar{R}^2 = 0.8137$, while in the reduced model, with variables M1 and Price only (Table 11–9), $\bar{R}^2 = 0.8193$. This demonstrates the usefulness of the adjusted R^2. When unimportant variables are added to the equation (unimportant in the presence of other variables), $\bar{R}^2$ decreases even if R^2 increases. The best model, in terms of explanatory power gauged against the loss of degrees of freedom, is the reduced model in Table 11–9, which relates exports to Singapore with only the money supply and price level. This is also seen by the fact that the other two variables are not significant once M1 and Price are in the equation. Later, when we discuss stepwise regression–a method of letting the computer choose the best variables to be included in the model–we will see that this automatic procedure also chooses the variables M1 and Price as the best combination for predicting U.S. exports to Singapore.

PROBLEMS

11–28. A regression analysis is carried out, and a confidence interval for β_1 is computed to be [1.25, 1.55]; a confidence interval for β_2 is [2.01, 2.12]. Both are 95% confidence intervals. Explain the possibility that the point (1.26, 2.02) may not lie inside a joint confidence region for (β_1, β_2) at a confidence level of 95%.

11–29. A multiple regression model was developed for predicting firms' governance level, measured on a scale, based on firm size, firm profitability, fixed-asset ratio, growth opportunities, and nondebt tax shield size. For firm size, the coefficient estimate was 0.06 and the standard error was 0.005. For firm profitability, the estimate was −0.166 and the standard error was 0.03. For fixed-asset ratio the estimate was −0.004 and standard error 0.05. For growth opportunities the estimate was –0.018 and standard error 0.025. And for nondebt tax shield the estimate was 0.649 and standard error 0.151. The F statistic was 44.11 and the adjusted R^2 was 16.5%.[10] Explain these results completely and offer a next step in this analysis. Assume a very large sample size.

[10] Pornsit Jiraporn and Kimberly C. Gleason, "Capital Structure, Shareholder Rights, and Corporate Governance," *Journal of Financial Research* 30, no. 1 (2007), pp. 21–33.

11–30. Give three reasons why caution must be exercised in interpreting the significance of single regression slope parameters.

11–31. Give 95% confidence intervals for the slope parameters β_2 through β_5, using the information in Table 11–4. Which confidence intervals contain the point (0, 0)? Explain the interpretation of such outcomes.

11–32. A regression analysis was carried out to predict a firm's reputation (defined on a scale called the Carter-Manaster reputation ranking) on the basis of unexpected accruals, auditor quality, return on investment, and expenditure on research and development. The parameter estimates (and standard errors, in parentheses), in the order these predictor variables are listed, are $-2.0775(0.4111)$, $-0.1116(0.2156)$, $0.4192(0.2357)$, and $0.0328(0.0155)$. The number of observations was 487, and the R^2 was 36.51%.[11] Interpret these findings.

11–33. A computer program for regression analysis produces a joint confidence region for the two slope parameters considered in the regression equation, β_1 and β_2. The elliptical region of confidence level 95% does not contain the point (0, 0). Not knowing the value of the F statistic, or R^2, do you believe there is a linear regression relationship between Y and at least one of the two explanatory variables? Explain.

11–34. In the Nissan Motor Company situation of problems 11–14 and 11–21, the regression results, using MINITAB, are as follows. Give a complete interpretation of these results.

```
The regression equation is
RATING = 24.1 − 0.166 PRESTIGE + 0.324 COMFORT + 0.514 ECONOMY

Predictor       Coef        Stdev
Constant        24.14       18.22
PRESTIGE     −0.1658       0.1215
COMFORT       0.3236       0.1228
ECONOMY       0.5139       0.1143
```

11–35. Refer to Example 11–2, where exports to Singapore were regressed on several economic variables. Interpret the results of the following MINITAB regression analysis, and compare them with the results reported in the text. How does the present model fit with the rest of the analysis? Explain.

```
The regression equation is
EXPORTS = − 3.40 + 0.363 M1 + 0.0021 LEND + 0.0367 PRICE

 Predictor        Coef         Stdev    t-ratio       P
 CONSTANT      −3.4047        0.6821     −4.99     0.000
 M1            0.36339       0.05940      6.12     0.000
 LEND          0.00211       0.04753      0.04     0.965
 PRICE        0.036666      0.009231      3.97     0.000

s = 0.3332      R-sq = 82.5%     R-sq (adj) = 81.6%
```

11–36. After the model of problem 11–35, the next model was run:

```
The regression equation is
EXPORTS = − 1.09 + 0.552 M1 + 0.171 LEND
```

[11] Hoje Jo, Yongtae Kim, and Myung Seok Park, "Underwriter Choice and Earnings Management: Evidence from Seasoned Equity Offerings," *Review of Accounting Studies* 12, no. 1 (2007), pp. 23–59.

```
Predictor      Coef       Stdev     t-ratio      P
Constant     −1.0859      0.3914     −2.77     0.007
M1           0.55222     0.03950     13.98     0.000
LEND         0.17100     0.02357      7.25     0.000

s = 0.3697     R-sq = 78.1%     R-sq (adj) = 77.4%

Analysis of Variance

  SOURCE      DF       SS         MS          F         P
Regression     2     31.189     15.594     114.09     0.000
Error         64      8.748      0.137
Total         66     39.936
```

a. What happened when Price was dropped from the regression equation? Why?

b. Compare this model with all previous models of exports versus the economic variables, and draw conclusions.

c. Which model is best overall? Why?

d. Conduct the *F* test for this particular model.

e. Compare the reported value of *s* in this model with the reported *s* value in the model of problem 11–35. Why is *s* higher in this model?

f. For the model in problem 11–35, what is the mean square error?

11–37. A regression analysis of monthly sales versus four independent variables is carried out. One of the variables is known not to have any effect on sales, yet its slope parameter in the regression is significant. In your opinion, what may have caused this to happen?

11–38. A study of 14,537 French firms was carried out to assess employment growth based on levels of new technological process, organizational innovation, commercial innovation, and research and development. The R^2 was 74.3%. The coefficient estimates for these variables (and standard errors) were reported, in order, as follows: −0.014(0.004), 0.001(0.004), 0.016(0.005), and 0.027(0.006).[12] Which of these variables have explanatory power over a firm's employment growth? Explain.

11–39. Run a regression of profits against revenues and number of employees for the airline industry using the data in the following table. Interpret all your findings.

Profit ($ billion)	Revenue ($ billion)	Employees (thousands)
−1.2	17	96
−2.8	13	68
−0.2	13	70
0.2	9.5	39
0.03	8.8	38
1.4	6.8	32
0.4	5.9	33
0.01	2.4	13
0.06	2.3	11
0.1	1.3	6

[12] Pierre Biscourp and Francis Kramarz, "Employment, Skill Structure and Internal Trade: Firm-Level Evidence for France," *Journal of International Economics* 72 (May 2007), pp. 22–51.

11–6 Testing the Validity of the Regression Model

In Chapter 10, we stressed the importance of the three stages of statistical model building: model specification, estimation of parameters, and testing the validity of the model assumptions. We will now discuss the third and very important stage of checking the validity of the model assumptions in multiple regression analysis.

Residual Plots

CHAPTER 17

As with simple linear regression, the analysis of regression residuals is an important tool for determining whether the assumptions of the multiple regression model are met. Residual plots are easy to use, and they convey much information quickly. The saying "A picture is worth a thousand words" is a good description of the technique of examining plots of regression residuals. As with simple linear regression, we may plot the residuals against the predicted values of the dependent variable, against each independent variable, against time (or the order of selection of the data points), and on a probability scale, to check the normality assumption. Since we have already discussed the use of residual plots in Chapter 10, we will demonstrate only some of the residual plots, using Example 11–2. Figure 11–11 is a plot of the residuals produced from the model with the two independent variables M1 and Price (Table 11–9) against variable M1. It appears that the residuals are randomly distributed with no pattern and with equal variance as M1 increases.

Figure 11–12 is a plot of the regression residuals against the variable Price. Here the picture is quite different. As we examine this figure carefully, we see that the spread of the residuals increases as Price increases. Thus, the variance of the residuals is not constant. We have the situation called *heteroscedasticity*—a violation of the assumption of equal error variance. In such cases, the ordinary least-squares (OLS) estimation method is not efficient, and an alternative method, called *weighted least squares* (*WLS*), should be used instead. The WLS procedure is discussed in advanced texts on regression analysis.

Figure 11–13 is a plot of the regression residuals against the variable Time, that is, the order of the observations. (The observations are a time sequence of monthly data.) This variable was not included in the model, and the plot could reveal whether time should have been included as a variable in our regression model. The plot of the residuals against time reveals no pattern in the residuals as time increases. The residuals seem to be more or less randomly distributed about their mean of zero.

Figure 11–14 is a plot of the regression residuals against the predicted export values $\hat{Y}$. We leave it as an exercise to the reader to interpret the information in this plot.

Standardized Residuals

Remember that under the assumptions of the regression model, the population errors ϵ_j are normally distributed with mean zero and standard deviation σ. As a result,

FIGURE 11–11 Residuals versus M1

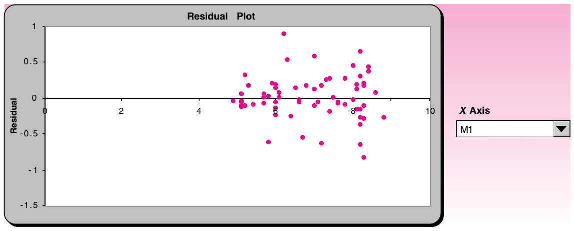


FIGURE 11–12 Residuals versus Price

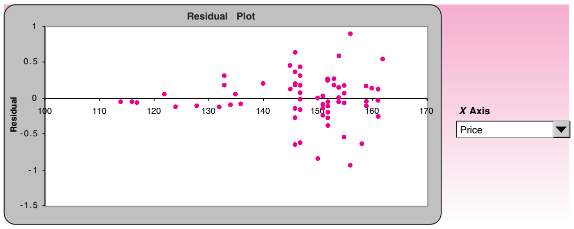


FIGURE 11–13 Residuals versus Time

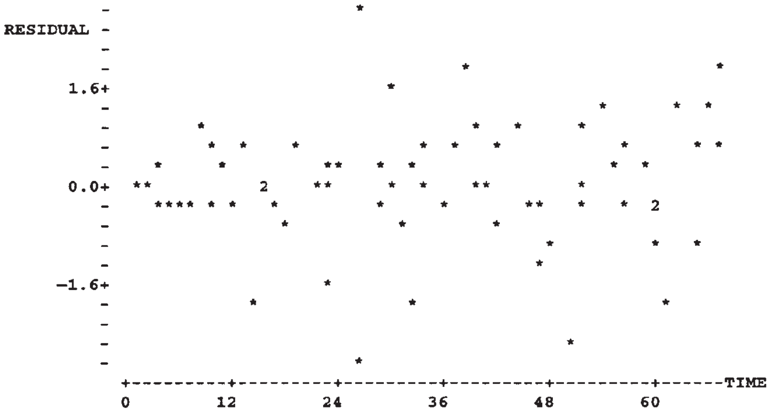


FIGURE 11–14 Residuals versus Predicted *Y* Values

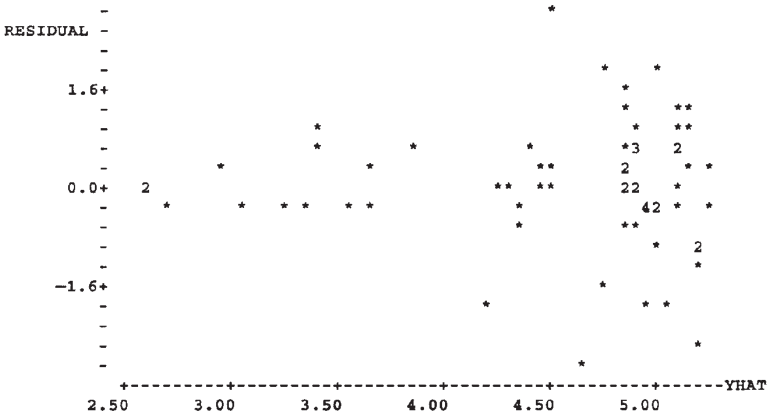
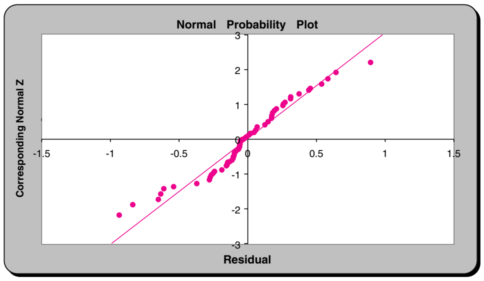


FIGURE 11–15 The Normal Probability Plot of the Residuals [Multiple Regression.xls; Sheet: Residuals]



CHAPTER 17

the errors divided by their standard deviation should follow the standard normal distribution:

$$\frac{\epsilon_j}{\sigma} \sim \mathrm{N}(0, 1) \ \circ \ \text{for all } j$$

Therefore, dividing the observed regression errors e_j by their estimated standard deviation s will give us standardized residuals. Examination of a histogram of these residuals may give us an idea as to whether the normal assumption is valid.[13]

CHAPTER 17

The Normal Probability Plot

Just as we saw in the simple regression template, the multiple regression template also produces a normal probability plot of the residuals. If the residuals are perfectly normally distributed, they will lie along the diagonal straight line in the plot. The more they deviate from the diagonal line, the more they deviate from the normal distribution. In Figure 11–15, the deviations do not appear to be significant. Consequently, we assume that the residuals are normally distributed.

Outliers and Influential Observations

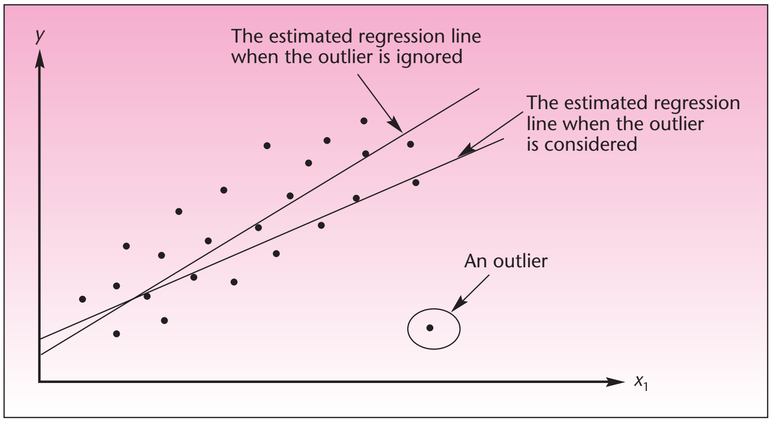
An **outlier** is an extreme observation. It is a point that lies away from the rest of the data set. Because of this, outliers may exert greater influence on the least-squares estimates of the regression parameters than do other observations. To see why, consider the data in Figure 11–16. The graph shows the estimated least-squares regression line without the outlier and the line obtained when the outlier is considered.

As can be seen from Figure 11–16, the outlier has a strong effect on the estimation of model parameters. (We used a line showing Y versus variable X_1. The same is true for a regression plane or hyperplane: The outlier "tilts" the regression surface away from the other points.) The reason for this effect is the nature of least squares: The procedure minimizes the squared deviations of the data points from the regression surface. A point with an unusually large deviation "attracts" the surface toward itself so as to make its squared deviation smaller.

We must, therefore, pay special attention to outliers. If an outlier can be traced to an error in recording the data or to another type of error, it should, of course, be removed. On the other hand, if an outlier is not due to error, it may have been caused by special circumstances, and the information it provides may be important. For example, an outlier may be an indication of a missing variable in the regression equation.

[13] Actually, the residuals are not independent and do not have equal variance; therefore, we really should divide the residuals e_j by something a little more complicated than s. However, the simpler procedure outlined here and implemented in some computer packages is usually sufficiently accurate.

FIGURE 11–16 A Least-Squares Regression Line Estimated with and without the Outlier



The data shown in Figure 11–16 may be maximum speed for an automobile as a function of engine displacement. The outlier may be an automobile with four cylinders, while all others are six-cylinder cars. Thus, the fact that the point lies away from the rest may be explained. Because of the possible information content in outliers, they should be carefully scrutinized before being discarded. Some alternative regression methods do not use a squared-distance approach and are therefore more robust—less sensitive to the influence of outliers.

Sometimes an outlier is actually a point that is distant from the rest because the value of one of its independent variables is larger than the rest of the data. For example, suppose we measure chemical yield Y as a function of temperature X_1. There may be other variables, but we will consider only these two. Suppose that most of our data are obtained at low temperatures within a certain range, but one observation is taken at a high temperature. This outlying point, far in the X_1 direction, exerts strong influence on the estimation of the model parameters. This is shown in Figure 11–17. Without the point at high temperature, the regression line may have slope zero, and no relationship may be detected, as can be seen from the figure. We must also be

FIGURE 11–17 Influence of an Observation Far in the X_1 Direction

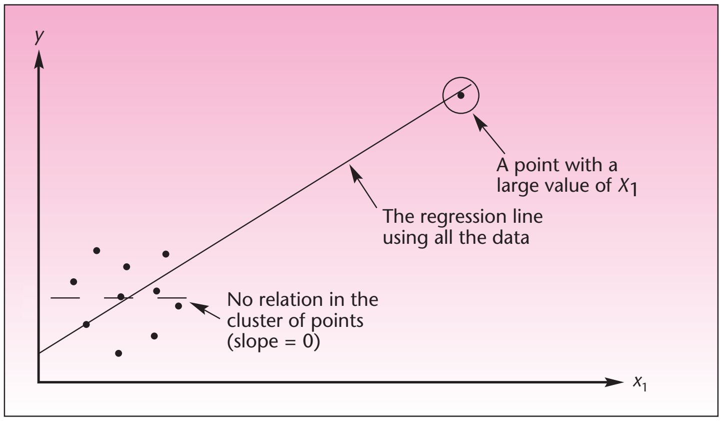
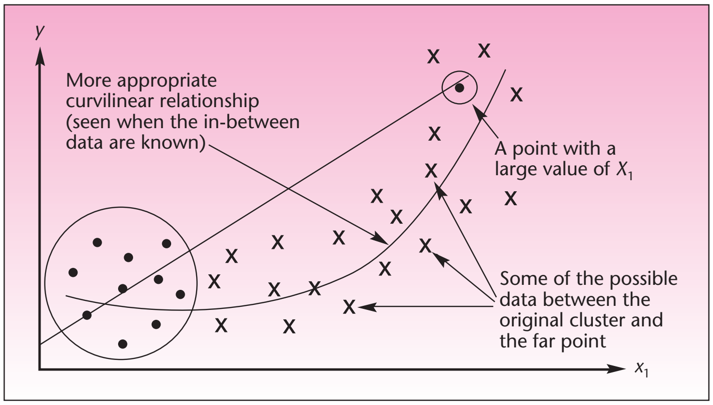


FIGURE 11–18 **Possible Relation in the Region between the Available Cluster of Data and the Far Point**


y
More appropriate curvilinear relationship (seen when the in-between data are known)
A point with a large value of X_1
Some of the possible data between the original cluster and the far point
x_1

careful in such cases to guard against estimating a straight-line relation where a curvilinear one may be more appropriate. This could become evident if we had more data points in the region between the far point and the rest of the data. This is shown in Figure 11–18.

Figure 11–18 serves as a good reminder that regression analysis should not be used for extrapolation. We do not know what happens in the region in which we have no data. This region may be between two regions where we have data, or it may lie beyond the last observation in a given direction. The relationship may be quite different from what we estimate from the data. This is also a reason why forcing the regression surface to go through the origin (that is, carrying out a regression with no constant term $\beta_0 = 0$), as is done in some applications, is not a good idea. The reasoning in such cases follows the idea expressed in the statement "In this particular case, when there is zero input, there must be zero output," which may very well be true. Forcing the regression to go through the origin, however, may make the estimation procedure biased. This is because in the region where the data points are located—assuming they are not near the origin—the best straight line to describe the data may not have an intercept of zero. This happens when the relationship is not a straight-line relationship. We mentioned this problem in Chapter 10.

A data point far from the other point in some X_i direction is called an *influential observation* if it strongly affects the regression fit. Statistical techniques can be used to test whether the regression fit is strongly affected by a given observation. Computer routines such as MINITAB automatically search for outliers and influential observations, reporting them in the regression output so that the user is alerted to the possible effects of these observations. Table 11–10 shows part of the MINITAB output for the analysis of Example 11–2. The table reports "unusual observations": large residuals and influential observations that affect the estimation of the regression relationship.

Lack of Fit and Other Problems

Model lack of fit occurs if, for example, we try to fit a straight line to curved data. The statistical method of determining the existence of lack of fit consists of breaking down the sum of squares for error to a sum of squares due to pure error and a sum of squares due to lack of fit. The method requires that we have observations at equal values of the independent variables or near-neighbor points. This method is described in advanced texts on regression.

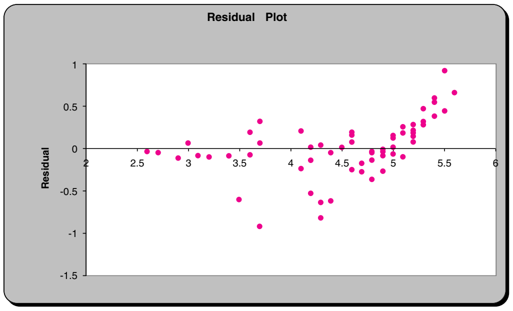
TABLE 11–10 Part of the MINITAB Output for Example 11–2

```
Unusual Observations
Obs.    M1     EXPORTS      Fit      Stdev.Fit    Residual     St.Resid

  1    5.10    2.6000     2.6420     0.1288      -0.0420      -0.14 X
  2    4.90    2.6000     2.6438     0.1234      -0.0438      -0.14 X
 25    6.20    5.5000     4.5949     0.0676       0.9051       2.80R
 26    6.30    3.7000     4.6311     0.0651      -0.9311      -2.87R
 50    8.30    4.3000     5.1317     0.0648      -0.8317      -2.57R
 67    8.20    5.6000     4.9474     0.0668       0.6526       2.02R

R denotes an obs. with a large st.resid.
X denotes an obs. whose X value gives it large influence.
```

A statistical method for determining whether the errors in a regression model are correlated through time (thus violating the regression model assumptions) is the Durbin-Watson test. This test is discussed in a later section of this chapter. Once we determine that our regression model is valid and that there are no serious violations of assumptions, we can use the model for its intended purpose.

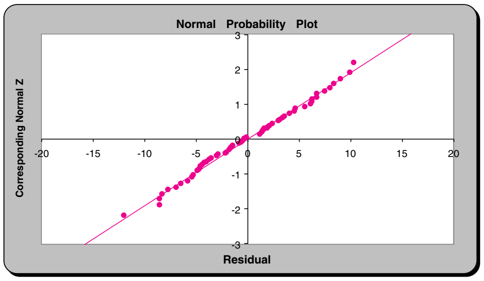
PROBLEMS

11–40. Analyze the following plot of the residuals versus $\hat{Y}$.

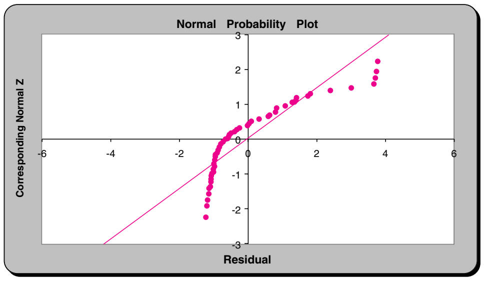


11–41. The normal probability plots of two regression experiments are given below. For each case, give your comments.

a.



b.



11–42. Explain what an outlier is.

11–43. How can you detect outliers? Discuss two ways of doing so.

11–44. Why should outliers not be discarded and the regression run without them?

11–45. Discuss the possible effects of an outlier on the regression analysis.

11–46. What is an influential observation? Give a few examples.

11–47. What are the limitations of forcing the regression surface to go through the origin?

11–48. Analyze the residual plot of Figure 11–14.

11–7 Using the Multiple Regression Model for Prediction

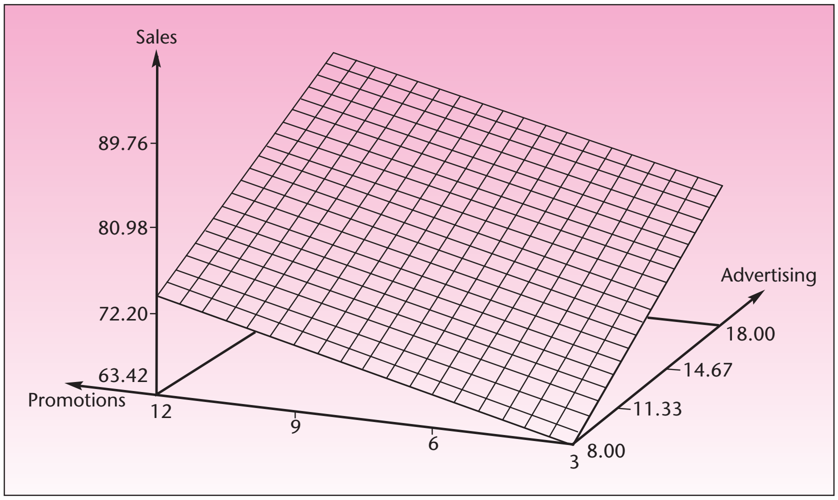
The use of the multiple regression model for prediction follows the same lines as in the case of simple linear regression, discussed in Chapter 10. We obtain a regression model prediction of a value of the dependent variable Y, based on given values of the independent variables, by substituting the values of the independent variables into the prediction equation. That is, we substitute the values of X_i variables into the equation for $\hat{Y}$. We demonstrate this in Example 11–1.

The predicted value of Y is given by substituting the given values of advertising X_1 and in-store promotions X_2 for which we want to predict sales Y into equation 11–6, using the parameter estimates obtained in Section 11–2. Let us predict sales when advertising is at a level of \$10,000 and in-store promotions are at a level of \$5,000.

$$\begin{aligned}\hat{Y} &= 47.165 + 1.599X_1 + 1.149X_2\\ &= 47.165 + (1.599)(10) + (1.149)(5) = 68.9 \text{ (thousand dollars)}\end{aligned}$$

This prediction is not bad, since the value of Y actually occurring for these values of X_1 and X_2 is known from Table 11–1 to be $Y = 70$ (thousand dollars). Our point estimate of the expected value of Y, denoted $E(Y)$, given these values of X_1 and X_2, is also 68.9 (thousand dollars). Note that our predictions lie *on* the estimated regression surface. The estimated regression surface for Example 11–1 is the plane shown in Figure 11–19.

FIGURE 11–19 **Estimated Regression Plane for Example 11–1**



We may also compute prediction intervals as well as confidence intervals for $E(Y)$, given values of the independent variables. As you recall, while the predicted value and the estimate of the mean value of Y are equal, the prediction interval is wider than a confidence interval for $E(Y)$ using the same confidence level. There is more uncertainty about the predicted value than there is about the average value of Y given the values X_i. The equation for a $(1 - \alpha)$ 100% prediction interval is an extension of equation 10–32 for simple linear regression. The only difference is that the degrees of freedom of the t distribution are $n - (k + 1)$ rather than just $n - 2$, as is the case for $k = 1$. The standard error, when there are several explanatory variables, is a complicated expression, and we will not give it here; we will denote it by $s(\hat{Y})$. The prediction interval is given in equation 11–17.

A $(1 - \alpha)$ 100% prediction interval for a value of Y given values of X_i is

$$\hat{y} \pm t_{[\alpha/2,\, n-(k+1)]} \sqrt{s^2(\hat{Y}) + \text{MSE}} \qquad (11\text{–}17)$$

While the expression in the square root is complex, it is computed by most computer packages for regression. The prediction intervals for any values of the independent variables and a given level of confidence are produced as output.

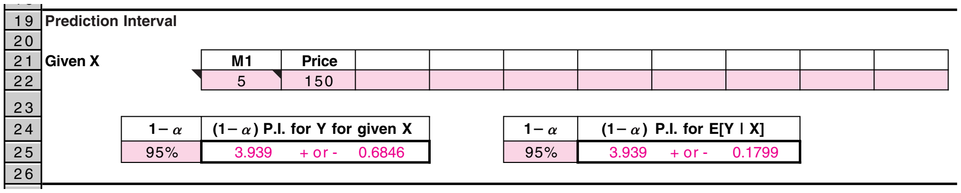
Similarly, the equation for a $(1 - \alpha)$ 100% confidence interval for the conditional mean of Y is an extension of equation 10–33 for the simple linear regression. It is given as equation 11–18. Again, the degrees of freedom are $n - (k + 1)$. The formula for the standard error is complex and will not be given here. We will call the standard error $s[E(\hat{Y})]$. The confidence interval for the conditional mean of Y is computable and may be reported, upon request, in the output of most computer packages that include regression analysis.

A $(1 - \alpha)$ 100% confidence interval for the conditional mean of Y is

$$\hat{y} \pm t_{[\alpha/2,\, n-(k+1)]} s\,[E(\hat{y})] \qquad (11\text{–}18)$$

Equations 11–17 and 11–18 are implemented in the template on the Results sheet. These equations are also produced by other computer packages for regression, and are presented here–as many other formulas–for information only.[14] To make a prediction, we enter the values of the independent variables in row 22 and the confidence level desired in row 25. Table 11–11 shows the case of Example 11–2 with

TABLE 11–11 Prediction Using Multiple Regression [Multiple Regression.xls; Sheet: Results]



19	Prediction Interval						
20							
21	Given X		M1	Price			
22			5	150			
23							
24		$1-\alpha$	$(1-\alpha)$ P.I. for Y for given X			$1-\alpha$	$(1-\alpha)$ P.I. for $E[Y \mid X]$
25		95%	3.939 + or - 0.6846			95%	3.939 + or - 0.1799
26							

[14] Note also that equations 11–17 and 11–18 are extensions to multiple regression of the analogous equations, 10–32 and 10–33, of simple linear regression–which is a special case of multiple regression with one explanatory variable.

independent variables M1 and Price. The 95% prediction interval has been computed for the exports when M1 = 5 and Price = 150. A similar interval for the expected value of the exports for the given M1 and Price values has also been computed. The two prediction intervals appear in row 24.

The predictions are not very reliable because of the heteroscedasticity we discovered in the last section, but they are useful as a demonstration of the procedure. Remember that it is never a good idea to try to predict values outside the region of the data used in the estimation of the regression parameters, because the regression relationship may be different outside that range. In this example, all predictions use values of the independent variables within the range of the estimation data.

When using regression models, remember that a regression relationship between the dependent variable and some independent variables does not imply causality. Thus, if we find a linear relationship between Y and X, it does not necessarily mean that X causes Y. Causality is very different to determine and to prove. There is also the issue of spurious correlations between variables—correlations that are not real. Montgomery and Peck give an example of a regression analysis of the number of mentally disturbed people in the United Kingdom versus the number of radio receiver licenses issued in that country.[15] The regression relationship is close to a perfect straight line, with $r^2 = 0.9842$. Can the conclusion be drawn that there is a relationship between the number of radio receiver licenses and the incidence of mental illness? Probably not. Both variables—the number of licenses and the incidence of mental illness—are related to a third variable: population size. The increase in both of these variables reflects the growth of the population in general, and there is probably no *direct* connection between the two variables. We must be very careful in our interpretation of regression results.

The Template

The multiple regression template [Multiple Regression.xls] consists of a total of five sheets. The sheet titled "Data" is used to enter the data (see Figure 11–28 for an example). The sheet titled "Results" contains the regression coefficients, their standard errors, the corresponding t tests, the ANOVA table, and a panel for prediction intervals. The sheet titled "Residuals" contains a plot of the residuals, the Durbin-Watson statistic (described later), and a normal probability plot for testing the normality assumption of the error term. The sheet titled "Correl" displays the correlation coefficient between every pair of variables. The use of the correlation matrix is described later in this chapter. The sheet titled "Partial F" can be used to find partial F, which is also described later in this chapter.

Setting Recalculation to "Manual" on the Template

Since the calculations performed in the multiple regression template are voluminous, a recalculation can take a little longer than in other templates. Therefore, entering data in the Data sheet may be difficult, especially on slower PCs (Pentium II or earlier), because the computer will recalculate every result before taking in the next data entry. If this problem occurs, set the Recalculation feature to manual. This can be done by clicking the Microsoft Office button and then Formulas. Choose Manual under the Calculation options. When this is done, a change made in the data or in any cell *will not cause the spreadsheet to automatically update itself.* Only when recalculation is manually initiated will the spreadsheet update itself. To initiate recalculation, *press the F9 key* on the keyboard. A warning message about pressing the F9 key is displayed at a few places in the template. If the recalculation has not been set to manual, this message can be ignored.

[15] D. Montgomery, E. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis,* 4th ed. (New York: Wiley, 2006).

Note also that when the recalculation is set to manual, *none of the open spreadsheets* will update itself. That is, if other spreadsheets were open, they will not update themselves either. The F9 key needs to be pressed on every open spreadsheet to initiate recalculation. This state of manual recalculation will continue until the Excel program is closed and reopened. For this reason, set the recalculation to manual only after careful consideration.

PROBLEMS

11–49. Explain why it is not a good idea to use the regression equation for predicting values outside the range of the estimation data set.

11–50. Use equation 11–6 to predict sales in Example 11–1 when the level of advertising is $8,000 and in-store promotions are at a level of $12,000.

11–51. Using the regression relationship you estimated in problem 11–8, predict the value of a home 1,800 square feet located 2.0 miles from the center of the town.

11–52. Using the regression equation from problem 11–25, predict excess stock return when SIZRNK = 5 and PRCRNK = 6.

11–53. Using the information in Table 11–11, what is the standard error of $\hat{Y}$? What is the standard error of $E(\hat{Y})$?

11–54. Use a computer to produce a prediction interval and a confidence interval for the conditional mean of Y for the prediction in problem 11–50. Use the data in Table 11–1.

11–55. What is the difference between a predicted value of the dependent variable and the conditional mean of the dependent variable?

11–56. Why is the prediction interval of 95% wider than the 95% confidence interval for the conditional mean, using the same values of the independent variables?

11–8 Qualitative Independent Variables

The variables we have encountered so far in this chapter have all been *quantitative* variables: variables that can take on values on a scale. Sales volume, advertising expenditure, exports, the money supply, and people's ratings of an automobile are all examples of quantitative variables. In this section, we will discuss the use of *qualitative* variables as explanatory variables in a regression model. Qualitative variables are variables that describe a quality rather than a quantity. This should remind you of analysis of variance in Chapter 9. There we had qualitative variables: the kind of resort in the Club Med example, type of airplane, type of coffee, and so on.

CHAPTER 19

In some cases, including information on one or more qualitative variables in our multiple regression model is very useful. For example, a hotel chain may be interested in predicting the number of occupied rooms as a function of the economy of the area in which the hotel is located, as well as advertising level and some other quantitative variables. The hotel may also want to know whether the peak season is in progress—a qualitative variable that may have a lot to do with the level of occupancy at the hotel. A property appraiser may be interested in predicting the value of different residential units on the basis of several quantitative variables, such as age of the unit and area in square feet, as well as the qualitative variable of whether the unit is owned or rented.

Each of these qualitative variables has only two *levels:* peak season versus nonpeak season, rental unit versus nonrental unit. An easy way to quantify such a qualitative variable is by way of a single **indicator variable,** also called a **dummy variable.** An indicator variable is a variable that indicates whether some condition holds. It has the value 1 when the condition holds and the value 0 when the condition does not hold. If you are familiar with computer science, you probably know the indicator variable by another name: *binary variable,* because it takes on only two possible values, 0 and 1.

When included in the model of hotel occupancy, the indicator variable will equal 0 if it is not peak season and 1 if it is (or vice versa; it makes no difference). Similarly, in the property value analysis, the dummy variable will have the value 0 when the unit is rented and the value 1 when the unit is owned, or vice versa. We define the general form of an indicator variable in equation 11–19.

An indicator variable of qualitative level A is

$$X_h = \begin{cases} 1 & \text{if level A is obtained} \\ 0 & \text{if level A is not obtained} \end{cases} \qquad (11\text{–}19)$$

The use of indicator variables in regression analysis is very simple. No special computational routines are required. All we do is code the indicator variable as 1 whenever the quality of interest is obtained for a particular data point and as 0 when it is not obtained. The rest of the variables in the regression equation are left the same. We demonstrate the use of an indicator variable in modeling a qualitative variable with two levels in the following example.

EXAMPLE 11–3

A motion picture industry analyst wants to estimate the gross earnings generated by a movie. The estimate will be based on different variables involved in the film's production. The independent variables considered are X_1 = production cost of the movie and X_2 = total cost of all promotional activities. A third variable that the analyst wants to consider is the qualitative variable of whether the movie is based on a book published before the release of the movie. This third, qualitative variable is handled by the use of an indicator variable: $X_3 = 0$ if the movie is not based on a book, and $X_3 = 1$ if it is. The analyst obtains information on a random sample of 20 Hollywood movies made within the last 5 years (the inference is to be made only about the population of movies in this particular category). The data are given in Table 11–12. The variable Y is gross earnings, in millions of dollars. The two quantitative independent variables are also in millions of dollars.

Solution

The data are entered into the template. The resulting output is presented in Figure 11–20. The coefficient of determination of this regression is very high; the F statistic value is very significant, and we have a good regression relationship. From the individual t ratios and their p-values, we find that all three independent variables are important in the equation.

From the intercept of 7.84, we could (erroneously, of course) deduce that a movie costing nothing to produce or promote, and that is not based on a book, would still gross $7.84 million! The point 0 ($X_1 = 0$, $X_2 = 0$, $X_3 = 0$) is outside the estimation region, and the regression relationship may not hold for that region. In our case, it evidently does not. The intercept is merely a reference point used to move the regression surface upward to where it should be in the estimation region.

The estimated slope for the cost variable, 2.85, means that—within the estimation region—an increase of $1 million in a movie's production cost (the other variables held constant) increases the movie's gross earnings by an average of $2.85 million. Similarly, the estimated slope coefficient for the promotion variable means that, in the estimation region of the variables, an increase of $1 million in promotional

TABLE 11–12 Data for Example 11–3

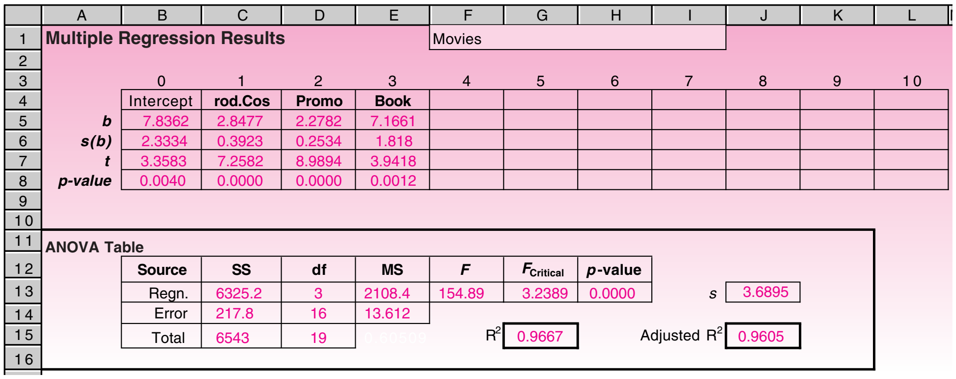
Movie	Gross Earnings Y, Million $	Production Cost X_1, Million $	Promotion Cost X_2, Million $	Book X_3
1	28	4.2	1	0
2	35	6.0	3	1
3	50	5.5	6	1
4	20	3.3	1	0
5	75	12.5	11	1
6	60	9.6	8	1
7	15	2.5	0.5	0
8	45	10.8	5	0
9	50	8.4	3	1
10	34	6.6	2	0
11	48	10.7	1	1
12	82	11.0	15	1
13	24	3.5	4	0
14	50	6.9	10	0
15	58	7.8	9	1
16	63	10.1	10	0
17	30	5.0	1	1
18	37	7.5	5	0
19	45	6.4	8	1
20	72	10.0	12	1

activities (with the other variables constant) increases the movie's gross earnings by an average of $2.28 million.

How do we interpret the estimated coefficient of variable X_3? The estimated coefficient of 7.17 means that having the movie based on a published book ($X_3 = 1$) increases the movie's gross earnings by an average of $7.17 million. Again, the inference is valid only for the region of the data used in the estimation. When $X_3 = 0$, that is, when the movie is not based on a book, the last term in the estimated equation for $\hat{Y}$ drops out—there is no added $7.17 million.

What do we learn from this example about the function of the indicator variable? Note that the predicted value of Y, given the values of the quantitative independent

FIGURE 11–20 Multiple Regression Results for Example 11–3. [Multiple Regression.xls; Sheet: Results]



Multiple Regression Results

Movies

	0	1	2	3	4	5	6	7	8	9	10
	Intercept	rod.Cos	Promo	Book							
b	7.8362	2.8477	2.2782	7.1661							
s(b)	2.3334	0.3923	0.2534	1.818							
t	3.3583	7.2582	8.9894	3.9418							
p-value	0.0040	0.0000	0.0000	0.0012							

ANOVA Table

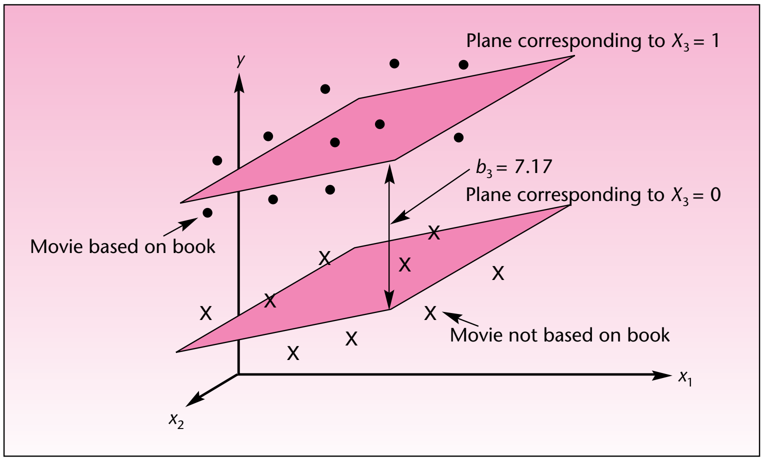
Source	SS	df	MS	*F*	F_{Critical}	*p*-value
Regn.	6325.2	3	2108.4	154.89	3.2389	0.0000
Error	217.8	16	13.612			
Total	6543	19				

s 3.6895

R^2 0.9667

Adjusted R^2 0.9605

FIGURE 11–21 Two Regression Planes of Example 11–3



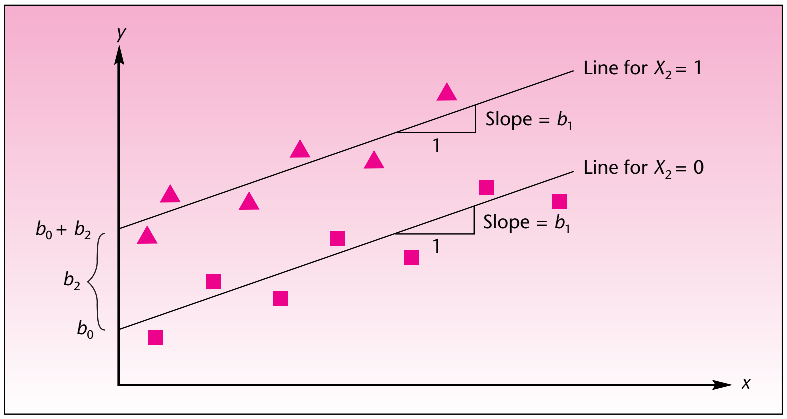
variables, shifts upward (or downward, depending on the sign of the estimated coefficient) by an amount equal to the coefficient of the indicator variable whenever the variable is equal to 1. In this particular case, the surface of the regression—the plane formed by the variables Y, X_1, and X_2—is split into two surfaces: one corresponding to movies based on books and the other corresponding to movies not based on books. The appropriate surface depends on whether $X_3 = 0$ or $X_3 = 1$; the two estimated surfaces are separated by a distance equal to $b_3 = 7.17$. This is demonstrated in Figure 11–21. The regression surface in this example is a plane, so we can draw its image (for a higher-dimensional surface, the same idea holds).

We will now look at the simpler case, with one independent quantitative variable and one indicator variable. Here we assume an estimated regression relationship of the form $\hat{Y} = b_0 + b_1X_1 + b_2X_2$, where X_1 is a quantitative variable and X_2 is an indicator variable. The regression relationship is a straight line, and the indicator variable splits the line into two parallel straight lines, one for each level (0 or 1) of the qualitative variable. The points belonging to one level (a level could be Book, as in Example 11–3) are shown as triangles, and the points belonging to the other level are shown as squares. The distance between the two parallel lines (measured as the difference between the two intercepts) is equal to the estimated coefficient of the dummy variable X_2. The situation is demonstrated in Figure 11–22.

We have been dealing with qualitative variables that have only two levels. Therefore, it has sufficed to use an indicator variable with two possible values, 0 and 1. What about situations where we have a qualitative variable with more than two levels? Should we use an "indicator" variable with more than two values? The answer is no. Were we to do this and give our variable values such as 0, 1, 2, 3, . . . , to indicate qualitative levels, we would be using a quantitative variable that has several discrete values but no values in between. Also, the assignment of the qualities to the values would be arbitrary. Since there may be no justification for using the values 1, 2, 3, etc., we would be imposing a very special measuring scale on the regression problem—a scale that may not be appropriate. Instead, we will use several indicator variables.

We account for a qualitative variable with r levels by the use of $r - 1$ indicator (0/1) variables.

FIGURE 11–22 A Regression with One Quantitative Variable and One Dummy Variable



We will now demonstrate the use of this rule by changing Example 11–3 somewhat. Suppose that the analyst is interested not in whether a movie is based on a book, but rather in using an explanatory variable that represents the category to which each movie belongs: adventure, drama, or romance. Since this qualitative variable has $r = 3$ levels, the rule tells us that we need to model this variable by using $r - 1 = 2$ indicator variables. Each of the two indicator variables will have one of two possible values, as before: 0 or 1. The setup of the two dummy variables indicating the level of the qualitative variable, movie category, is shown in the following table. For simplicity, let us also assume that the only quantitative variable in the equation is production cost (we leave out the promotion variable). This will allow us to have lines rather than planes. We let X_1 = production cost, as before. We now define the two dummy variables X_2 and X_3.

Category	X_2	X_3
Adventure	0	0
Drama	0	1
Romance	1	0

The definition of the values of X_2 and X_3 for representing the different categories is arbitrary; we could just as well have assigned the values $X_2 = 0$, $X_3 = 0$ to drama or to romance as to adventure. The important thing to remember is that the number of dummy variables is 1 less than the number of categories they represent. Otherwise our model will be overspecified, and problems will occur. In this example, variable X_2 is the indicator variable for romance; when a movie is in the romance category, this variable has the value 1. Similarly, X_3 is the indicator for drama and has the value 1 in cases where a movie is in the drama category. Only three categories are under consideration, so when both X_2 and X_3 are zero, the movie is neither a drama nor a romance; therefore, it must be an adventure movie.

If we use the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon \qquad (11\text{–}20)$$

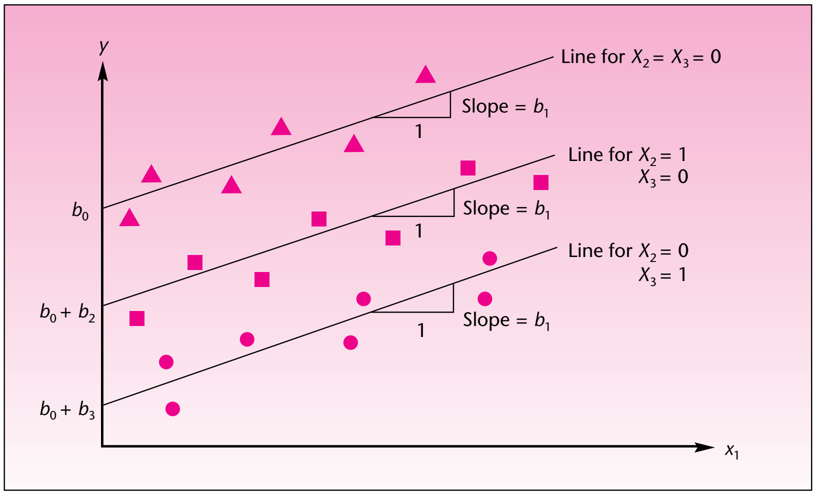
with X_2 and X_3 as defined, we will be estimating three regression lines, one line per category. The line for adventure movies will be $\hat{Y} = b_0 + b_1 X_1$ because here both X_2

and X_3 are zero. The drama line will be $\hat{Y} = b_0 + b_3 + b_1X_1$ because here $X_3 = 1$ and $X_2 = 0$. In the case of romance movies, our line will be $\hat{Y} = b_0 + b_2 + b_1X_1$ because in this case $X_2 = 1$ and $X_3 = 0$. Since the estimated coefficients b_i may be negative as well as positive, the different parallel lines may position themselves above or below one another, as determined by the data. Of course, the b_i may be estimates of zero. If we did not reject the null hypothesis H_0: $\beta_3 = 0$, using the usual t test, it would mean that there was no evidence that the adventure and the drama lines were different. That is, it would mean that, on average, adventure movies and drama movies have the same gross earnings as determined by the production costs. If we determine that β_2 is not different from zero, the adventure and romance lines will be the same and the drama line may be different. In case the adventure line is different from drama and romance, these two being the same, we would determine statistically that both β_2 and β_3 are different from zero, but not different from each other.

If we have three regression lines, why bother with indicator variables at all? Why not just run three separate regressions, each for a different movie category? One answer to this question has already been given: The use of indicator variables and their estimated regression coefficients with their standard errors allows us to *test statistically* whether the qualitative variable of interest has any effect on the dependent variable. We are able to test whether we have one distinct line, two lines, three lines, or as many lines as there are levels of the qualitative variable. Another reason is that even if we know that there are, say, three distinct lines, estimating them together via a regression analysis with dummy variables allows us to pool the degrees of freedom for the three regressions, leading to better estimation and a more efficient analysis.

Figure 11–23 shows the three regression lines of our new version of Example 11–3; each line shows the regression relationship between a movie's production cost and the resulting movie's gross earnings in its category. In case there are two independent quantitative variables, say, if we add promotions as a second quantitative variable, we will have three regression *planes* like the two planes shown in Figure 11–21. In Figure 11–23, we show adventure movies as triangles, romance movies as squares, and drama movies as circles. Assuming that adventure movies have the highest average

FIGURE 11–23 **The Three Possible Regression Lines, Depending on Movie Category (modified Example 11–3)**



gross earnings, followed by romance and drama, the estimated coefficients b_2 and b_3 have to be negative, as can be seen from the figure.

Can we run a regression on a qualitative variable (by use of dummy variables) only? Yes. You have already seen this model, essentially. Running a regression on a qualitative variable only means modeling some quantitative response by levels of a qualitative factor: it is the *analysis of variance,* discussed in Chapter 9. Doing the analysis by regression means using a different computational procedure than was done in Chapter 9, but it is still the analysis of variance. Two qualitative variables make the analysis a two-way ANOVA, and interaction terms are cross-products of the appropriate dummy variables, such as $X_2 X_3$. We will say more about cross-products a little later. For now, we note that the regression approach to ANOVA allows us more freedom. Remember that a two-way ANOVA, using the method in Chapter 9, required a balanced design (equal sample size in each cell). If we use the regression approach, we are no longer restricted to the balanced design and may use any sample size.

Let us go back to regressions using quantitative independent variables with some qualitative variables. In some situations, we are not interested in using a regression equation for prediction or for any of the other common uses of regression analysis. Instead, we are intrinsically interested in a qualitative variable used in the regression. Let us be more specific. Recall our original Example 11–3. Suppose we are not interested in predicting a movie's gross earnings based on the production cost, promotions, and whether the movie is based on a book. Suppose instead that we are interested in answering the question: Is there a difference in average gross earnings between movies based on books and movies not based on books?

To answer this question, we use the estimated regression relationship. We use the estimate b_3 and its standard error in testing the null hypothesis H_0: $\beta_3 = 0$ versus the alternative H_1: $\beta_3 \neq 0$. The question is really an ANOVA question. We want to know whether a difference exists in the population means of the two groups of movies based on books and movies not based on books. However, we have some quantitative variables that affect the variable we are measuring (gross earnings). We therefore incorporate information on these variables (production cost and promotions) in a regression model aimed at answering our ANOVA question. When we do this, that is, when we attempt to answer the question of whether differences in population means exist, using a regression equation to account for other sources of variation in our data (the quantitative independent variables), we are conducting an **analysis of covariance.** The independent variables used in the analysis of covariance are called **concomitant variables,** and their purpose in the analysis is not to explain or predict the independent variable, but rather to reduce the errors in the test of significance of the indicator variable or variables.

One of the interesting applications of analysis of covariance is in providing statistical evidence in cases of sex or race discrimination. We demonstrate this particular use in the following example.

EXAMPLE 11–4

A large service company was sued by its female employees in a class action suit alleging sex discrimination in salary levels. The claim was that, on average, a man and a woman of the same education and experience received different salaries: the man's salary was believed to be higher than the woman's salary. The attorney representing the women employees hired a statistician to provide statistical evidence supporting the women's side of the case. The statistician was allowed access to the company's payroll files and obtained a random sample of 100 employees, 40 of whom were women. In addition to salary, the files contained information on education and experience. The statistician then ran a regression analysis of salary Y versus three variables: education level X_1 (on a scale based on the total number of years in school, with an additional value added to the score for each college degree earned, by type),

TABLE 11–13 Regression Results for Example 11–4

Variable	Coefficient Estimate	Standard Error
Constant	8,547	32.6
Education	949	45.1
Experience	1,258	78.5
Sex	−3,256	212.4

years of experience X_2 (on a scale that combined the number of years of experience directly related to the job assignment with the number of years of similar job experience), and gender X_3 (0 if the employee was a man and 1 if the employee was a woman). The computer output for the regression included the results F ratio $=$ 1,237.56 and $R^2 = 0.67$, as well as the coefficient estimates and standard errors given in Table 11–13. Based on this information, does the attorney for the women employees have a case against the company?

Solution Let us analyze the regression results. Remember that we are using a regression with a dummy variable to perform an analysis of covariance. There is certainly a regression relationship between salary and at least some of the variables, as evidenced by the very large F value, which is beyond any critical point we can find in a table. The p-value is very small. The coefficient of determination is not extremely high, but then we are using very few variables to explain variation in salary levels. This being the case, 67% explained variation, based on these variables only, is quite respectable. Now we consider the information in Table 11–13.

Dividing the four coefficient estimates by their standard errors, we find that all three variables are important, and the intercept is different from zero. However, we are particularly interested in the hypothesis test:

$$
\begin{aligned}
&H_0\colon \beta_3 = 0\\
&H_1\colon \beta_3 \neq 0
\end{aligned}
$$

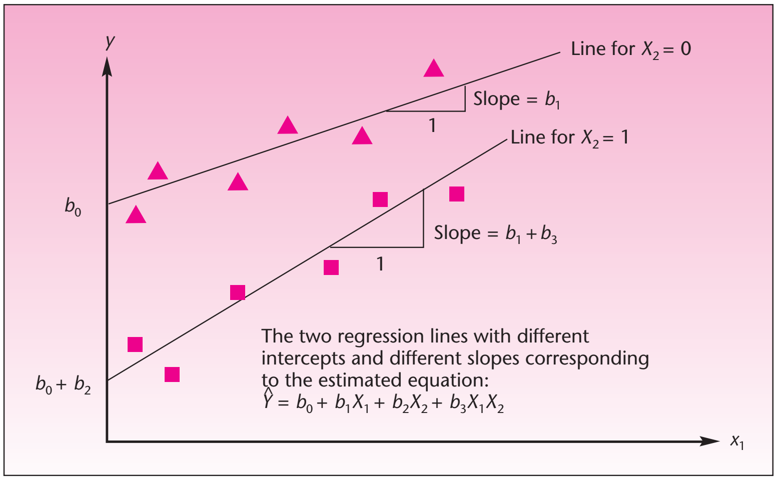
Our test statistic is $t_{(96)} = b_3/s(b_3) = -3{,}256/212.4 = -15.33$. Since t with 96 degrees of freedom [df $= n - (k+1) = 100 - 4 = 96$] is virtually a standard normal random variable, we conduct this as a Z test. The computed test statistic value of -15.33 lies very far in the left-hand rejection region. This means that there are two regressions: one for men and one for women. Since we coded X_3 as 0 for a man and 1 for a woman, the women's estimated regression plane lies $3,256 below the regression plane for men. Since the parameter of the sex variable is significantly different from zero (with an extremely small p-value) and is negative, there is statistical evidence of sex discrimination in this case. The situation here is as seen in Figure 11–21 for the previous example: We have two regression planes, one below the other. The only difference is that in this example, we were not interested in using the regression for prediction, but rather for an ANOVA-type statistical test.

Interactions between Qualitative and Quantitative Variables

CHAPTER 19

Do the different regression lines or higher-dimensional surfaces have to be parallel? The answer is no. Sometimes, there are *interactions* between a qualitative variable and one or more quantitative variables. The idea of an interaction in regression analysis is the same as the idea of interaction between factors in a two-way ANOVA model (as well as higher-order ANOVAs). In regression analysis with qualitative variables,

FIGURE 11–24 Effects of an Interaction between a Qualitative Variable and a Quantitative Variable



the interaction between a qualitative variable and a quantitative variable makes the regression lines or planes at different levels of the dummy variables have *different slopes*. Let us look at the simple case where we have one independent quantitative variable X_1 and one qualitative variable with two levels, modeled by the dummy variable X_2. When an interaction exists between the qualitative and the quantitative variables, the slope of the regression line for $X_2 = 0$ is different from the slope of the regression line for $X_2 = 1$. This is shown in Figure 11–24.

We model the interactions between variables by the cross-product of the variables. The interaction of X_1 with X_2 in this case is modeled by adding the term X_1X_2 to the regression equation. We are thus interested in the model

$$Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_1X_2 + \epsilon \qquad (11\text{–}21)$$

We can use the results of the estimation procedure to test for the existence of an interaction. We do so by testing the significance of parameter β_3.

When regression parameters β_1, β_2, and β_3 are all nonzero, we have two distinct lines with different intercepts and different slopes. When β_2 is zero, we have two lines with the same intercept and different slopes (this is unlikely to happen, except when both intercepts are zero). When β_3 is zero, we have two parallel lines, as in the case of equation 11–20. If β_1 is zero, of course, we have no regression—just an ANOVA model; we then assume that β_3 is also zero. Assuming the full model of equation 11–21, representing two distinct lines with different slopes and different intercepts, the intercept and the slope of each line will be as shown in Figure 11–24. By substituting $X_2 = 0$ or $X_2 = 1$ into equation 11–21, verify the definition of each slope and each intercept.

Again, estimating a single model for the different levels of the indicator variable offers two advantages. These are the pooling of degrees of freedom (we assume that the spread of the data about the two or more lines is equal) and an understanding of the joint process generating the data. More important, we may use the model to statistically test for the equality of intercepts and slopes. Note that when several indicator variables are used in modeling one or more qualitative variables, the model has several possible interaction terms. We will learn more about interactions in general in the next section.

PROBLEMS

11–57. Echlin, Inc., makes parts for automobiles. The company is engaged in strong competition with Japanese, Taiwanese, and Korean manufacturers of the same automobile parts. Recently, the company hired a statistician to study the relationship between monthly sales and the independent variable, number of cars on the road. Data on the explanatory variable are published in national statistical reports. Because of the keen competition with Asian firms, an indicator variable was also used. This variable was given the value 1 during months when restrictions on imports from Asia were in effect and 0 when such restrictions were not in effect. Denoting sales by Y, total number of cars on the road by X_1, and the import restriction dummy variable by X_2, the following regression equation was estimated:

$$\hat{Y} = -567.3 + 0.006X_1 + 26{,}540X_2$$

The standard error of the intercept estimate was 38.5, that of the coefficient of X_1 was 0.0002, and the standard error of the coefficient of X_2 was 1,534.67. The multiple coefficient of determination was $R^2 = 0.783$. The sample size used was $n = 60$ months (5 years of data). Analyze the results presented. What kind of regression model was used? Comment on the significance of the model parameters and the value of R^2. How many distinct regression lines are there? What likely happens during times of restricted trade with Asia?

11–58. A regression analysis was carried out based on 7,016 observations of firms, aimed at assessing the factors that determine the level of a firm's leverage. The independent variables included amount of fixed assets, profitability, firm size, volatility, and abnormal earnings level, as well as a dummy variable that indicated whether the firm was regulated (1) or unregulated (0). The coefficient estimate for this dummy variable was -0.003 and its standard error was -0.29.[16] Does a firm's being regulated affect its leverage level? Explain.

11–59. If we have a regression model with no quantitative variables and only two qualitative variables, represented by some indicator variables and cross-products, what kind of analysis is carried out?

11–60. Recall our Club Med example of Chapter 9. Suppose that not all vacationers at Club Med resorts stay an equal length of time at the resort—different people stay different numbers of days. The club's research director knows that people's ratings of the resorts tend to differ depending on the number of days spent at the resort. Design a new method for studying whether there are differences among the average population ratings of the five Caribbean resorts. What is the name of your method of analysis, and how is the analysis carried out? Explain.

11–61. A financial institution specializing in venture capital is interested in predicting the success of business operations that the institution helps to finance. Success is defined by the institution as return on its investment, as a percentage, after 3 years of operation. The explanatory variables used are Investment (in thousands of dollars), Early investment (in thousands of dollars), and two dummy variables denoting the category of business. The values of these variables are (0, 0) for high-technology industry, (0, 1) for biotechnology companies, and (1, 0) for aerospace firms. Following is part of the computer output for this analysis. Interpret the output, and give a complete analysis of the results of this study based on the provided information.

[16] Matthew T. Billett, Tao-Hsien Dolly King, and David Mauer, "Growth Opportunities and the Choice of Leverage, Debt Maturity, and Covenants," *Journal of Finance* 42, no. 2 (2007), pp. 697–730.

```
The regression equation is
Return = 6.16 + 0.617 INVEST + 0.151 EARLY + 11.1 DUM1 + 4.15 DUM2
 Predictor      Coef         Stdev
  Constant     6.162        1.642
  INVEST       0.6168       0.1581
  EARLY        0.1509       0.1465
  DUM1        11.051        1.355
  DUM2         4.150        1.315
s = 2.148    R-sq = 91.6%     R-sq (adj) = 89.4%

Analysis of Variance

  SOURCE      DF       SS
Regression     4     755.99
Error         15      69.21
Total         19     825.20
```

11–9 Polynomial Regression

Often, the relationship between the dependent variable Y and one or more of the independent X variables is not a straight-line relationship but, rather, has some curvature to it. Several such situations are shown in Figure 11–25 (we show the curved relationship between Y and a *single* explanatory variable X). In each of the situations shown, a straight line provides a poor fit to the data. Instead, polynomials of order higher than 1, that is, functions of higher powers of X, such as X^2 and X^3, provide much better fit to our data. Such polynomials in the X variable or in several X_i variables are still considered linear regression models. Only models where the parameters β_i are not all of the first power are called *nonlinear models.* The multiple linear regression model thus covers situations of fitting data to polynomial functions. The general form of a polynomial regression model in one variable X is given in equation 11–22.

CHAPTER 18

FIGURE 11–25 **Situations Where the Relationship between X and Y Is Curved**

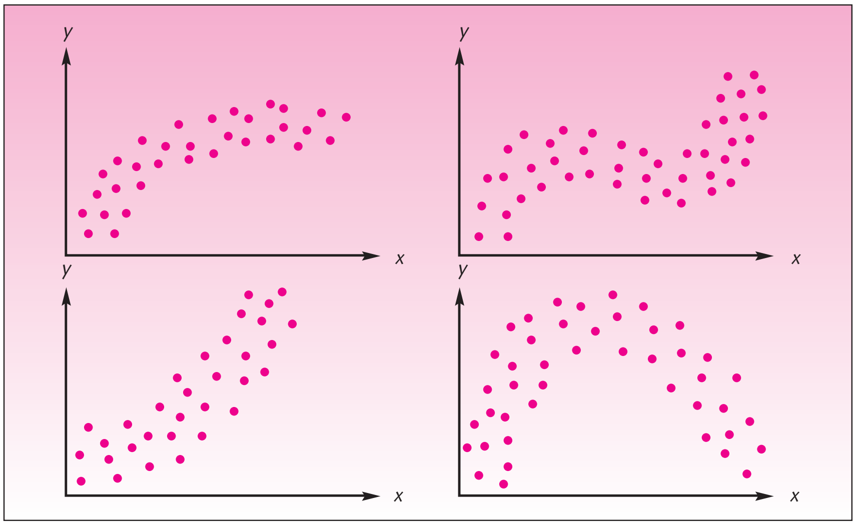
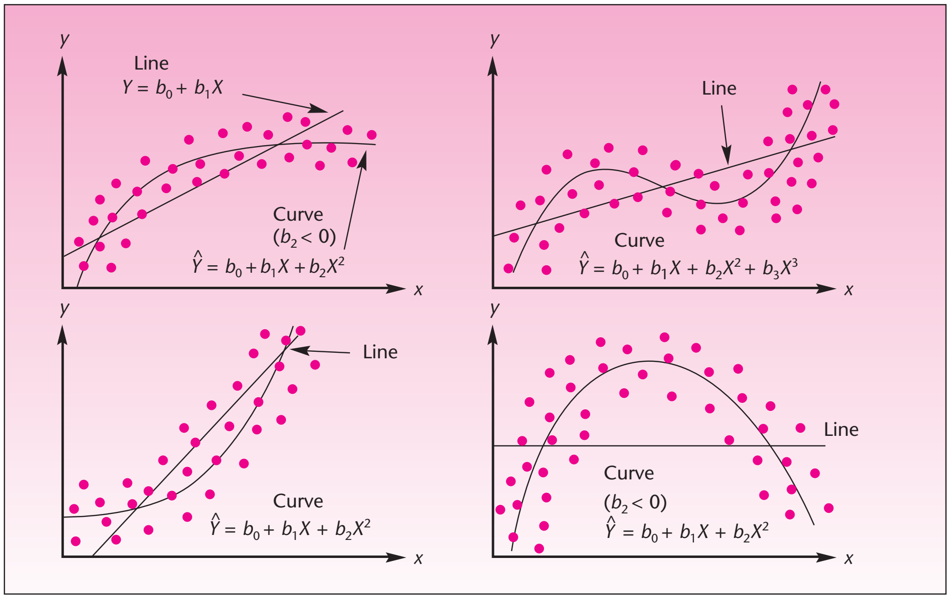


FIGURE 11–26 The Fits Provided for the Data Sets in Figure 11–25 by Polynomial Models



A one-variable polynomial regression model is

$$Y = \beta_0 + \beta_1X + \beta_2X^2 + \beta_3X^3 + \cdots + \beta_mX^m + \epsilon \qquad (11\text{–}22)$$

where m is the *degree* of the polynomial—the highest power of X appearing in the equation. The degree of the polynomial is the *order* of the model.

Figure 11–26 shows how second- and third-degree polynomial models provide good fits for the data sets in Figure 11–25. A straight line is also shown in each case, for comparison. Compare the fit provided in each case by a polynomial with the poor fit provided by a straight line. Some authors, for example, Cook and Weisberg, recommend using polynomials of order no greater than 2 (the third-order example in Figure 11–26 would be an exception) because of the overfitting problem.[17] At any rate, models should never be of order 6 or higher (unless the powers of X have been transformed in a special way). Seber shows that when a polynomial of degree 6 or greater is fit to a data set, a matrix involved in regression computations becomes *ill-conditioned,* which means that very small errors in the data cause relatively large errors in the estimated model parameters.[18] In short, we must be very careful with polynomial regression models and try to obtain the most parsimonious polynomial model that will fit our data. In the next section, we will discuss *transformations* of data that often can change curved data sets into a straight-line form. If we can find such a transformation for a data set, it is always better to use a first-order model on the transformed data set than to use a higher-order polynomial model on the original data. It should be intuitively clear that problems may arise in polynomial regression. The variables X and X^2,

[17] R. Dennis Cook and Sanford Weisberg, *Applied Regression Including Computing and Graphics* (New York: Wiley, 1999).

[18] George A. F. Seber and Alan J. Lee, *Linear Regression Analysis*, 2nd ed. (New York: Wiley, 2003).

for example, are clearly not independent of each other. This may cause the problem of multicollinearity in cases where the data are confined to a narrow range of values.

Having seen what to beware of in using polynomial regression, now we see how these models are used. Since powers of X can be obtained directly from the value of variable X, it is relatively easy to run polynomial models. We enter the data into the computer and add a command that uses X to form a new variable. In a second-order model, we create an X^2 column using spreadsheet commands. Then we run a multiple regression model with two "independent" variables: X and X^2. We demonstrate this with a new example.

EXAMPLE 11–5

Sales response to advertising usually follows a curve reflecting the diminishing returns to advertising expenditure. As a firm increases its advertising expenditure, sales increase, but the rate of increase drops continually after a certain point. If we consider company sales profits as a function of advertising expenditure, we find that the response function can be very well approximated by a second-order (quadratic) model of the form

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$$

A quadratic response function such as this one is shown in Figure 11–27.

It is very important for a firm to identify its own point X_m, shown in the figure. At this point, a maximum benefit is achieved from advertising in terms of the resulting sales profits. Figure 11–27 shows a general form of the sales response to advertising. To find its own maximum point X_m, a firm needs to estimate its response-to-advertising function from its own operation data, obtained by using different levels of advertising at different time periods and observing the resulting sales profits. For a particular firm, the data on monthly sales Y and monthly advertising expenditure X, both in hundred thousand dollars, are given in Table 11–14. The table also shows the values of X^2 used in the regression analysis.

FIGURE 11–27 A Quadratic Response Function of Sales Profits to Advertising Expenditure

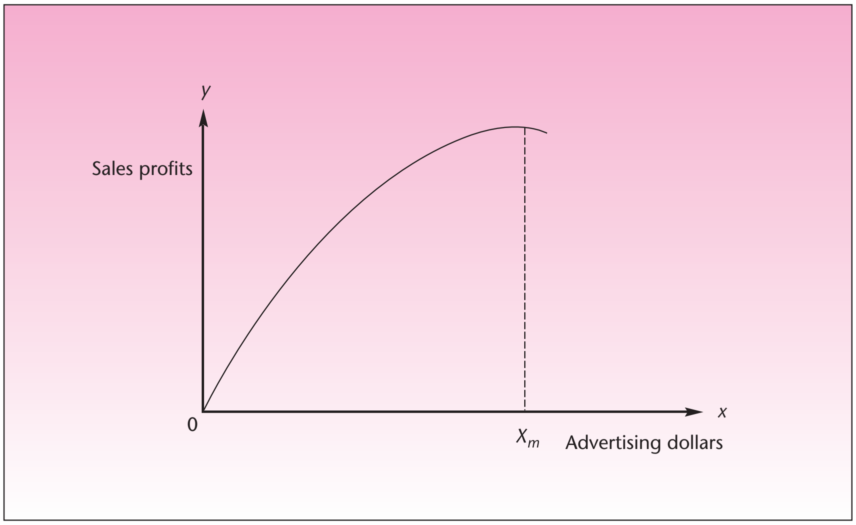
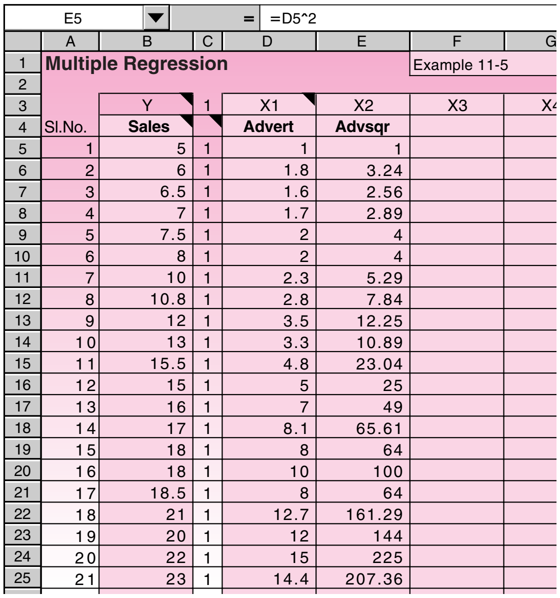


TABLE 11–14 Data for Example 11–5

Row	Sales	Advert	Advsqr
1	5.0	1.0	1.00
2	6.0	1.8	3.24
3	6.5	1.6	2.56
4	7.0	1.7	2.89
5	7.5	2.0	4.00
6	8.0	2.0	4.00
7	10.0	2.3	5.29
8	10.8	2.8	7.84
9	12.0	3.5	12.25
10	13.0	3.3	10.89
11	15.5	4.8	23.04
12	15.0	5.0	25.00
13	16.0	7.0	49.00
14	17.0	8.1	65.61
15	18.0	8.0	64.00
16	18.0	10.0	100.00
17	18.5	8.0	64.00
18	21.0	12.7	161.29
19	20.0	12.0	144.00
20	22.0	15.0	225.00
21	23.0	14.4	207.36

Solution Figure 11–28 shows the data entered in the template. In cell E5, the formula "=D5^2" has been entered. This calculates X^2. The formula has been copied down through cell E25. The regression results from the Results sheet of the template are shown in

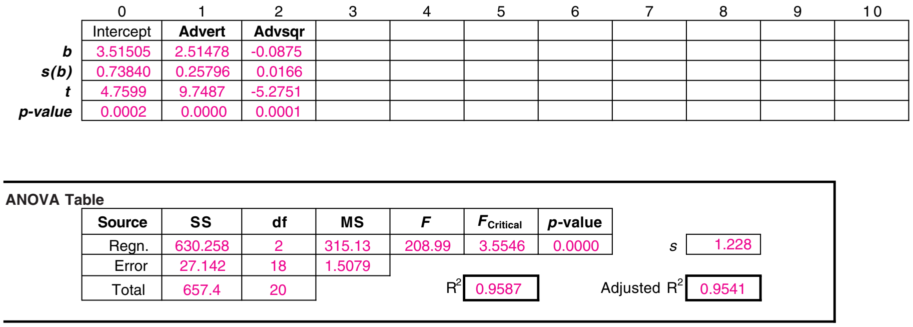
FIGURE 11–28 Data for the Regression [Multiple Regression.xls; Sheet: Data]



E5 = =D5^2

	A	B	C	D	E	F	G
1	Multiple Regression					Example 11-5	
2							
3		Y	1	X1	X2	X3	X4
4	Sl.No.	Sales		Advert	Advsqr		
5	1	5	1	1	1		
6	2	6	1	1.8	3.24		
7	3	6.5	1	1.6	2.56		
8	4	7	1	1.7	2.89		
9	5	7.5	1	2	4		
10	6	8	1	2	4		
11	7	10	1	2.3	5.29		
12	8	10.8	1	2.8	7.84		
13	9	12	1	3.5	12.25		
14	10	13	1	3.3	10.89		
15	11	15.5	1	4.8	23.04		
16	12	15	1	5	25		
17	13	16	1	7	49		
18	14	17	1	8.1	65.61		
19	15	18	1	8	64		
20	16	18	1	10	100		
21	17	18.5	1	8	64		
22	18	21	1	12.7	161.29		
23	19	20	1	12	144		
24	20	22	1	15	225		
25	21	23	1	14.4	207.36		

TABLE 11–15 Results of the Regression [Multiple Regression; Sheet: Results]



	0	1	2	3	4	5	6	7	8	9	10
	Intercept	**Advert**	**Advsqr**								
b	3.51505	2.51478	-0.0875								
s(b)	0.73840	0.25796	0.0166								
t	4.7599	9.7487	-5.2751								
p-value	0.0002	0.0000	0.0001								

ANOVA Table

Source	SS	df	MS	*F*	F_{Critical}	*p*-value
Regn.	630.258	2	315.13	208.99	3.5546	0.0000
Error	27.142	18	1.5079			
Total	657.4	20				

s 1.228

R² 0.9587 Adjusted R² 0.9541

Table 11–15. The coefficient of determination is $R^2 = 0.9587$, the F ratio is significant, and both Advert and Advsqr are very significant. The minus sign of the squared variable, Advsqr, is logical because a quadratic function with a maximum point has a negative leading coefficient (the coefficient of X^2). We may write the estimated quadratic regression model of Y in terms of X and X^2 as follows:

$$Y = 3.52 + 2.51X - 0.0875X^2 + e \qquad (11\text{–}23)$$

The equation of the estimated regression curve itself is given by dropping the error term e, giving an equation for the predicted values $\hat{Y}$ that lie on the quadratic curve

$$\hat{Y} = 3.52 + 2.51X - 0.0875X^2 \qquad (11\text{–}24)$$

In our particular example, the equation of the curve (equation 11–24) is of importance, as it can be differentiated with respect to X, with the derivative then set to zero and the result solved for the maximizing value X_m shown in Figure 11–27. (If you have not studied calculus, you may ignore the preceding statement.) The result here is $x_m = 14.34$ (hundred thousand dollars). This value maximizes sales profits with respect to advertising (within estimation error of the regression). Thus, the firm should set its advertising level at $1.434 million. The fact that polynomials can always be differentiated gives these models an advantage over alternative models. Remember, however, to keep the order of the model low.

Other Variables and Cross-Product Terms

The polynomial regression model in one variable X, given in equation 11–22, can easily be extended to include more than one independent explanatory variable. The new model, which includes several variables at different powers, is a mixture of the usual multiple regression model in k variables (equation 11–1) and the polynomial regression model (equation 11–22). When several variables are in a regression equation, we may also consider interactions among variables. We have already encountered interactions in the previous section, where we discussed interactions between an indicator variable and a quantitative variable. We saw that an interaction term is just the cross-product of the two variables involved. In this section, we discuss the general concept of interactions between variables, quantitative or not.

The interaction term $X_i X_j$ is a second-order term (the product of two variables is classified the same way as an X^2 term). Similarly, $X_i X_j^2$, for example, is a third-order term. Thus, models that incorporate interaction terms find their natural place within the class of polynomial models. Equation 11–25 is a second-order regression model in two variables X_1 and X_2. This model includes both first and second powers of both variables and an interaction term.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1 X_2 + \epsilon \qquad (11\text{–}25)$$

A regression surface of a model like that of equation 11–25 is shown in Figure 11–29. Of course, many surfaces are possible, depending on the values of the coefficients of all terms in the equation. Equation 11–25 may be generalized to more than two explanatory variables, to higher powers of each variable, and to more interaction terms.

When we are considering polynomial regression models in several variables, it is very important not to get carried away by the number of possible terms we can include in the model. The number of variables, as well as the powers of these variables and the number of interaction terms, should be kept to a minimum.

How do we choose the terms to include in a model? This question will be answered in Section 11–13, where we discuss methods of variable selection. You already know several criteria for the inclusion of variables, powers of variables, and interaction terms in a model. One thing to consider is the adjusted coefficient of determination. If this measure decreases when a term is included in the model, then the term should be dropped. Also, the significance of any particular term in a model depends on which other variables, powers, or interaction terms are in the model. We must consider the significance of each term by its t statistic, and we must consider what happens to the significance of regression terms once other terms are added to the model or removed from it. For example, let us consider the regression output in Table 11–16.

The results in the table clearly show that only X_1, X_2, and X_1^2 are significant. The apparent nonsignificance of X_2^2 and $X_1 X_2$ may be due to multicollinearity. At any

FIGURE 11–29 **An Example of the Regression Surface of a Second-Order Model in Two Variables**

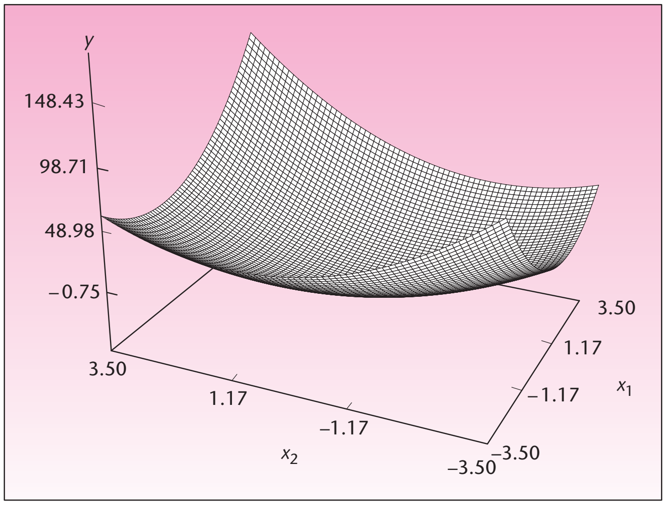


TABLE 11–16 Example of Regression Output for a Second-Order Model in Two Variables

Variable	Estimate	Standard Error	*t* Ratio
X_1	2.34	0.92	2.54
X_2	3.11	1.05	2.96
X_1^2	4.22	1.00	4.22
X_2^2	3.57	2.12	1.68
X_1X_2	2.77	2.30	1.20

rate, a regression without these last two variables should be carried out. We must also look at R^2 and the adjusted R^2 of the different regressions, and find the most parsimonious model with statistically significant parameters that explain as much as possible of the variation in the values of the dependent variable. Incidentally, the surface in Figure 11–29 was generated by computer, using all the coefficient estimates given in Table 11–16 (regardless of their significance) and an intercept of zero.

PROBLEMS

11–62. The following results pertain to a regression analysis of the difference between the mortgage rate and the Treasury bill rate (SPREAD) on the shape of the yield curve (S) and the corporate bond yields spread (R). What kind of regression model is used? Explain.

$$\text{SPREAD} = b_0 + b_1S + b_2R + b_3S^2 + b_4S * R$$

11–63. Use the data in Table 11–6 to run a polynomial regression model of exports to Singapore versus M1 and M1 squared, as well as Price and Price squared, and an interaction term. Also try to add a squared exchange rate variable into the model. Find the best, most parsimonious regression model for the data.

11–64. Use the data of Example 11–3, presented in Table 11–12, to try to fit a polynomial regression model of movie gross earnings on production cost and production cost squared. Also try promotion and promotion squared. What is the best, most parsimonious model?

11–65. An ingenious regression analysis was reported in which the effects of the 1985 French banking deregulation were assessed. Bank equity was the dependent variable, and each data point was a tax return for a particular quarter and bank in France from 1978 to the time the research was done. This resulted in 325,928 data points, assumed a random sample. The independent variables were Bankdep—average debt in the industry during this period; ROA—the given firm's average return on assets for the entire period, and After—0 before 1985, and 1 after 1985. The variables used in this regression were all cross-products. These variables and their coefficient estimates (with their standard errors) are given below.

After * Bankdep	−0.398 (0.035)
After * Bankdep * ROA	0.155 (0.057)
After * ROA	−0.072 (0.024)
Bankdep * ROA	−0.286 (0.073)

The adjusted R^2 was 53%.[19] Carefully analyze these results and try to draw a conclusion about the effects of the 1985 French Banking Deregulation Act.

[19] Marianne Bertrand, Antoinette Schoar, and David Thesmar, "Banking Deregulation and Industry Structure: Evidence from the French Banking Reforms of 1985," *Journal of Finance* 42, no. 2 (2007), pp. 597–628.