CHAPTER 14

Simple Linear Regression

CONTENTS

STATISTICS *in* PRACTICE

ALLIANCE DATA SYSTEMS*

DALLAS, TEXAS

Alliance Data Systems (ADS) provides transaction processing, credit services, and marketing services for clients in the rapidly growing customer relationship management (CRM) industry. ADS clients are concentrated in four industries: retail, petroleum/convenience stores, utilities, and transportation. In 1983, Alliance began offering end-to-end credit processing services to the retail, petroleum, and casual dining industries; today they employ more than 6500 employees who provide services to clients around the world. Operating more than 140,000 point-of-sale terminals in the United States alone, ADS processes in excess of 2.5 billion transactions annually. The company ranks second in the United States in private label credit services by representing 49 private label programs with nearly 72 million cardholders. In 2001, ADS made an initial public offering and is now listed on the New York Stock Exchange.

As one of its marketing services, ADS designs direct mail campaigns and promotions. With its database containing information on the spending habits of more than 100 million consumers, ADS can target those consumers most likely to benefit from a direct mail promotion. The Analytical Development Group uses regression analysis to build models that measure and predict the responsiveness of consumers to direct market campaigns. Some regression models predict the probability of purchase for individuals receiving a promotion, and others predict the amount spent by those consumers making a purchase.

For one particular campaign, a retail store chain wanted to attract new customers. To predict the effect of the campaign, ADS analysts selected a sample from the consumer database, sent the sampled individuals promotional materials, and then collected transaction data on the consumers' response. Sample data were collected on the amount of purchase made by the consumers responding to the campaign, as well as a variety of consumer-specific variables thought to be useful in predicting sales. The consumer-specific variable that contributed most to predicting the amount purchased was the total amount of credit purchases at related stores over the past 39 months. ADS analysts developed an estimated regression equation relating the amount of purchase to the amount spent at related stores:

Alliance Data Systems analysts discuss use of a regression model to predict sales for a direct marketing campaign. © Courtesy of Alliance Data Systems.

$$\hat{y} = 26.7 + 0.00205x$$

where

$\hat{y}$ = amount of purchase
x = amount spent at related stores

Using this equation, we could predict that someone spending $10,000 over the past 39 months at related stores would spend $47.20 when responding to the direct mail promotion. In this chapter, you will learn how to develop this type of estimated regression equation.

The final model developed by ADS analysts also included several other variables that increased the predictive power of the preceding equation. Some of these variables included the absence/presence of a bank credit card, estimated income, and the average amount spent per trip at a selected store. In the following chapter, we will learn how such additional variables can be incorporated into a multiple regression model.

*The authors are indebted to Philip Clemance, Director of Analytical Development at Alliance Data Systems, for providing this Statistics in Practice.

Managerial decisions often are based on the relationship between two or more variables. For example, after considering the relationship between advertising expenditures and sales, a marketing manager might attempt to predict sales for a given level of advertising expenditures. In another case, a public utility might use the relationship between the daily high temperature and the demand for electricity to predict electricity usage on the basis of next month's anticipated daily high temperatures. Sometimes a manager will rely on intuition to judge how two variables are related. However, if data can be obtained, a statistical procedure called *regression analysis* can be used to develop an equation showing how the variables are related.

The statistical methods used in studying the relationship between two variables were first employed by Sir Francis Galton (1822–1911). Galton was interested in studying the relationship between a father's height and the son's height. Galton's disciple, Karl Pearson (1857–1936), analyzed the relationship between the father's height and the son's height for 1078 pairs of subjects.

In regression terminology, the variable being predicted is called the **dependent variable**. The variable or variables being used to predict the value of the dependent variable are called the **independent variables**. For example, in analyzing the effect of advertising expenditures on sales, a marketing manager's desire to predict sales would suggest making sales the dependent variable. Advertising expenditure would be the independent variable used to help predict sales. In statistical notation, y denotes the dependent variable and x denotes the independent variable.

In this chapter we consider the simplest type of regression analysis involving one independent variable and one dependent variable in which the relationship between the variables is approximated by a straight line. It is called **simple linear regression**. Regression analysis involving two or more independent variables is called multiple regression analysis; multiple regression and cases involving curvilinear relationships are covered in Chapters 15 and 16.

14.1 Simple Linear Regression Model

Armand's Pizza Parlors is a chain of Italian-food restaurants located in a five-state area. Armand's most successful locations are near college campuses. The managers believe that quarterly sales for these restaurants (denoted by y) are related positively to the size of the student population (denoted by x); that is, restaurants near campuses with a large student population tend to generate more sales than those located near campuses with a small student population. Using regression analysis, we can develop an equation showing how the dependent variable y is related to the independent variable x.

Regression Model and Regression Equation

In the Armand's Pizza Parlors example, the population consists of all the Armand's restaurants. For every restaurant in the population, there is a value of x (student population) and a corresponding value of y (quarterly sales). The equation that describes how y is related to x and an error term is called the **regression model**. The regression model used in simple linear regression follows.

SIMPLE LINEAR REGRESSION MODEL

$$y = \beta_0 + \beta_1 x + \epsilon \qquad (14.1)$$

β_0 and β_1 are referred to as the parameters of the model, and ϵ (the Greek letter epsilon) is a random variable referred to as the error term. The error term accounts for the variability in y that cannot be explained by the linear relationship between x and y.

The population of all Armand's restaurants can also be viewed as a collection of subpopulations, one for each distinct value of x. For example, one subpopulation consists of all Armand's restaurants located near college campuses with 8000 students; another subpopulation consists of all Armand's restaurants located near college campuses with 9000 students; and so on. Each subpopulation has a corresponding distribution of y values. Thus, a distribution of y values is associated with restaurants located near campuses with 8000 students; a distribution of y values is associated with restaurants located near campuses with 9000 students; and so on. Each distribution of y values has its own mean or expected value. The equation that describes how the expected value of y, denoted $E(y)$, is related to x is called the **regression equation**. The regression equation for simple linear regression follows.

SIMPLE LINEAR REGRESSION EQUATION

$$E(y) = \beta_0 + \beta_1 x \tag{14.2}$$

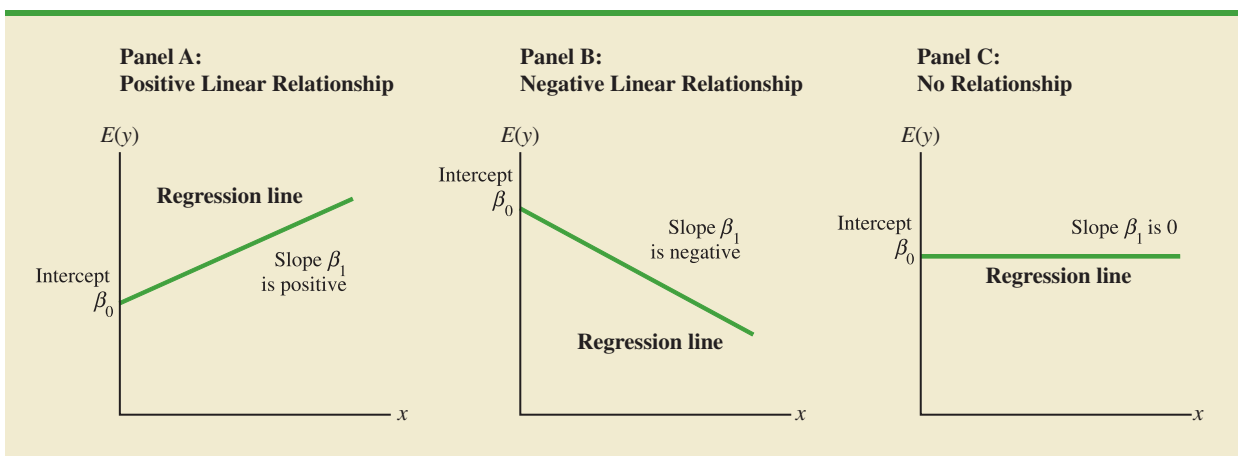
The graph of the simple linear regression equation is a straight line; β_0 is the y-intercept of the regression line, β_1 is the slope, and $E(y)$ is the mean or expected value of y for a given value of x.

Examples of possible regression lines are shown in Figure 14.1. The regression line in Panel A shows that the mean value of y is related positively to x, with larger values of $E(y)$ associated with larger values of x. The regression line in Panel B shows the mean value of y is related negatively to x, with smaller values of $E(y)$ associated with larger values of x. The regression line in Panel C shows the case in which the mean value of y is not related to x; that is, the mean value of y is the same for every value of x.

Estimated Regression Equation

If the values of the population parameters β_0 and β_1 were known, we could use equation (14.2) to compute the mean value of y for a given value of x. In practice, the parameter values are not known and must be estimated using sample data. Sample statistics (denoted b_0 and b_1) are computed as estimates of the population parameters β_0 and β_1. Substituting the values of the sample statistics b_0 and b_1 for β_0 and β_1 in the regression equation, we obtain the

FIGURE 14.1 POSSIBLE REGRESSION LINES IN SIMPLE LINEAR REGRESSION



estimated regression equation. The estimated regression equation for simple linear regression follows.

ESTIMATED SIMPLE LINEAR REGRESSION EQUATION

$$\hat{y} = b_0 + b_1 x \qquad (14.3)$$

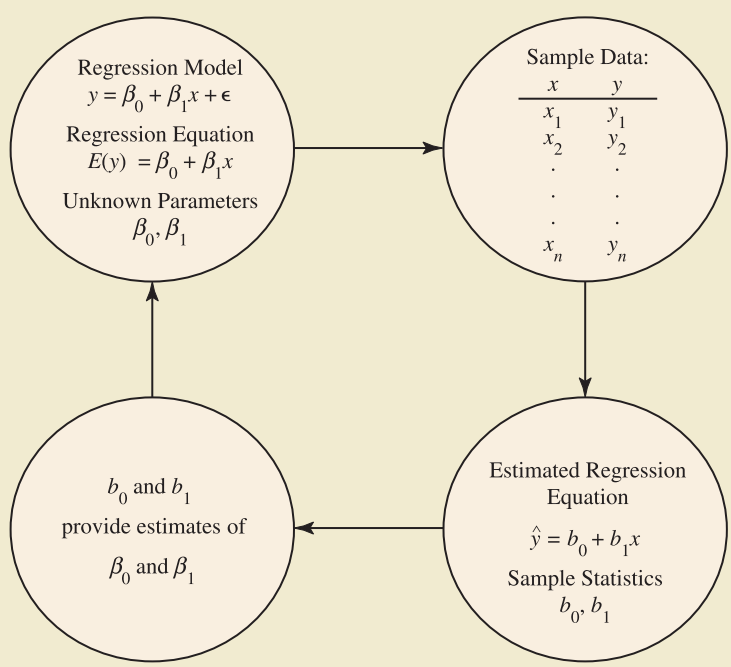
The graph of the estimated simple linear regression equation is called the *estimated regression line*; b_0 is the y intercept and b_1 is the slope. In the next section, we show how the least squares method can be used to compute the values of b_0 and b_1 in the estimated regression equation.

In general, $\hat{y}$ is the point estimator of $E(y)$, the mean value of y for a given value of x. Thus, to estimate the mean or expected value of quarterly sales for all restaurants located near campuses with 10,000 students, Armand's would substitute the value of 10,000 for x in equation (14.3). In some cases, however, Armand's may be more interested in predicting sales for one particular restaurant. For example, suppose Armand's would like to predict quarterly sales for the restaurant located near Talbot College, a school with 10,000 students. As it turns out, the best estimate of y for a given value of x is also provided by $\hat{y}$. Thus, to predict quarterly sales for the restaurant located near Talbot College, Armand's would also substitute the value of 10,000 for x in equation (14.3).

Because the value of $\hat{y}$ provides both a point estimate of $E(y)$ for a given value of x and a point estimate of an individual value of y for a given value of x, we will refer to $\hat{y}$ simply as the *estimated value of y*. Figure 14.2 provides a summary of the estimation process for simple linear regression.

The estimation of β_0 and β_1 is a statistical process much like the estimation of μ discussed in Chapter 7. β_0 and β_1 are the unknown parameters of interest, and b_0 and b_1 are the sample statistics used to estimate the parameters.

FIGURE 14.2 THE ESTIMATION PROCESS IN SIMPLE LINEAR REGRESSION



NOTES AND COMMENTS

1. Regression analysis cannot be interpreted as a procedure for establishing a cause-and-effect relationship between variables. It can only indicate how or to what extent variables are associated with each other. Any conclusions about cause and effect must be based upon the judgment of those individuals most knowledgeable about the application.
2. The regression equation in simple linear regression is $E(y) = \beta_0 + \beta_1 x$. More advanced texts in regression analysis often write the regression equation as $E(y|x) = \beta_0 + \beta_1 x$ to emphasize that the regression equation provides the mean value of y for a given value of x.

14.2 Least Squares Method

In simple linear regression, each observation consists of two values: one for the independent variable and one for the dependent variable.

The **least squares method** is a procedure for using sample data to find the estimated regression equation. To illustrate the least squares method, suppose data were collected from a sample of 10 Armand's Pizza Parlor restaurants located near college campuses. For the ith observation or restaurant in the sample, x_i is the size of the student population (in thousands) and y_i is the quarterly sales (in thousands of dollars). The values of x_i and y_i for the 10 restaurants in the sample are summarized in Table 14.1. We see that restaurant 1, with $x_1 = 2$ and $y_1 = 58$, is near a campus with 2000 students and has quarterly sales of $58,000. Restaurant 2, with $x_2 = 6$ and $y_2 = 105$, is near a campus with 6000 students and has quarterly sales of $105,000. The largest sales value is for restaurant 10, which is near a campus with 26,000 students and has quarterly sales of $202,000.

Figure 14.3 is a scatter diagram of the data in Table 14.1. Student population is shown on the horizontal axis and quarterly sales is shown on the vertical axis. **Scatter diagrams** for regression analysis are constructed with the independent variable x on the horizontal axis and the dependent variable y on the vertical axis. The scatter diagram enables us to observe the data graphically and to draw preliminary conclusions about the possible relationship between the variables.

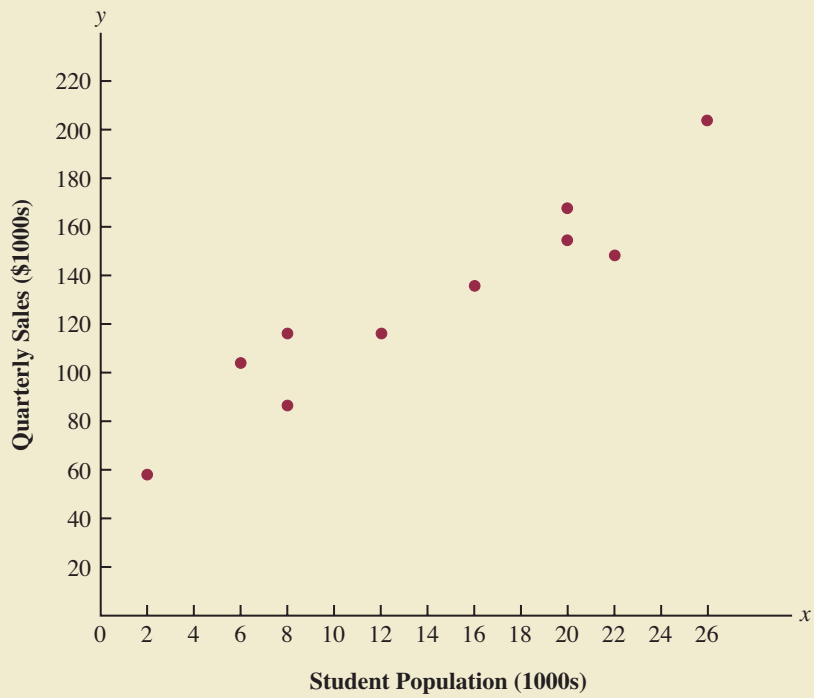
What preliminary conclusions can be drawn from Figure 14.3? Quarterly sales appear to be higher at campuses with larger student populations. In addition, for these data the relationship between the size of the student population and quarterly sales appears to be approximated by a straight line; indeed, a positive linear relationship is indicated between x

TABLE 14.1 STUDENT POPULATION AND QUARTERLY SALES DATA FOR 10 ARMAND'S PIZZA PARLORS

WEB file Armand's

Restaurant i	Student Population (1000s) x_i	Quarterly Sales ($1000s) y_i
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202

FIGURE 14.3 SCATTER DIAGRAM OF STUDENT POPULATION AND QUARTERLY SALES FOR ARMAND'S PIZZA PARLORS



and y. We therefore choose the simple linear regression model to represent the relationship between quarterly sales and student population. Given that choice, our next task is to use the sample data in Table 14.1 to determine the values of b_0 and b_1 in the estimated simple linear regression equation. For the ith restaurant, the estimated regression equation provides

$$\hat{y}_i = b_0 + b_1 x_i \qquad \textbf{(14.4)}$$

where

$\hat{y}_i$ = estimated value of quarterly sales ($1000s) for the ith restaurant
b_0 = the y intercept of the estimated regression line
b_1 = the slope of the estimated regression line
x_i = size of the student population (1000s) for the ith restaurant

With y_i denoting the observed (actual) sales for restaurant i and $\hat{y}_i$ in equation (14.4) representing the estimated value of sales for restaurant i, every restaurant in the sample will have an observed value of sales y_i and an estimated value of sales $\hat{y}_i$. For the estimated regression line to provide a good fit to the data, we want the differences between the observed sales values and the estimated sales values to be small.

The least squares method uses the sample data to provide the values of b_0 and b_1 that minimize the *sum of the squares of the deviations* between the observed values of the dependent variable y_i and the estimated values of the dependent variable $\hat{y}_i$. The criterion for the least squares method is given by expression (14.5).

Carl Friedrich Gauss (1777–1855) proposed the least squares method.

LEAST SQUARES CRITERION

$$\min \Sigma(y_i - \hat{y}_i)^2 \qquad \textbf{(14.5)}$$

where

y_i = observed value of the dependent variable for the ith observation
$\hat{y}_i$ = estimated value of the dependent variable for the ith observation

Differential calculus can be used to show (see Appendix 14.1) that the values of b_0 and b_1 that minimize expression (14.5) can be found by using equations (14.6) and (14.7).

In computing b_1 with a calculator, carry as many significant digits as possible in the intermediate calculations. We recommend carrying at least four significant digits.

SLOPE AND y-INTERCEPT FOR THE ESTIMATED REGRESSION EQUATION[1]

$$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} \qquad \textbf{(14.6)}$$

$$b_0 = \bar{y} - b_1\bar{x} \qquad \textbf{(14.7)}$$

where

x_i = value of the independent variable for the ith observation
y_i = value of the dependent variable for the ith observation
$\bar{x}$ = mean value for the independent variable
$\bar{y}$ = mean value for the dependent variable
n = total number of observations

Some of the calculations necessary to develop the least squares estimated regression equation for Armand's Pizza Parlors are shown in Table 14.2. With the sample of 10 restaurants, we have $n = 10$ observations. Because equations (14.6) and (14.7) require $\bar{x}$ and $\bar{y}$ we begin the calculations by computing $\bar{x}$ and $\bar{y}$.

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{140}{10} = 14$$

$$\bar{y} = \frac{\Sigma y_i}{n} = \frac{1300}{10} = 130$$

Using equations (14.6) and (14.7) and the information in Table 14.2, we can compute the slope and intercept of the estimated regression equation for Armand's Pizza Parlors. The calculation of the slope (b_1) proceeds as follows.

[1] An alternate formula for b_1 is

$$b_1 = \frac{\Sigma x_i y_i - (\Sigma x_i \Sigma y_i)/n}{\Sigma x_i^2 - (\Sigma x_i)^2/n}$$

This form of equation (14.6) is often recommended when using a calculator to compute b_1.

TABLE 14.2 CALCULATIONS FOR THE LEAST SQUARES ESTIMATED REGRESSION EQUATION FOR ARMAND PIZZA PARLORS

Restaurant i	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	2	58	−12	−72	864	144
2	6	105	−8	−25	200	64
3	8	88	−6	−42	252	36
4	8	118	−6	−12	72	36
5	12	117	−2	−13	26	4
6	16	137	2	7	14	4
7	20	157	6	27	162	36
8	20	169	6	39	234	36
9	22	149	8	19	152	64
10	26	202	12	72	864	144
Totals	140	1300			2840	568
	Σx_i	Σy_i			$\Sigma(x_i - \bar{x})(y_i - \bar{y})$	$\Sigma(x_i - \bar{x})^2$

$$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} = \frac{2840}{568} = 5$$

The calculation of the y intercept (b_0) follows.

$$b_0 = \bar{y} - b_1\bar{x} = 130 - 5(14) = 60$$

Thus, the estimated regression equation is

$$\hat{y} = 60 + 5x$$

Figure 14.4 shows the graph of this equation on the scatter diagram.

The slope of the estimated regression equation ($b_1 = 5$) is positive, implying that as student population increases, sales increase. In fact, we can conclude (based on sales measured in \$1000s and student population in 1000s) that an increase in the student population of 1000 is associated with an increase of \$5000 in expected sales; that is, quarterly sales are expected to increase by \$5 per student.

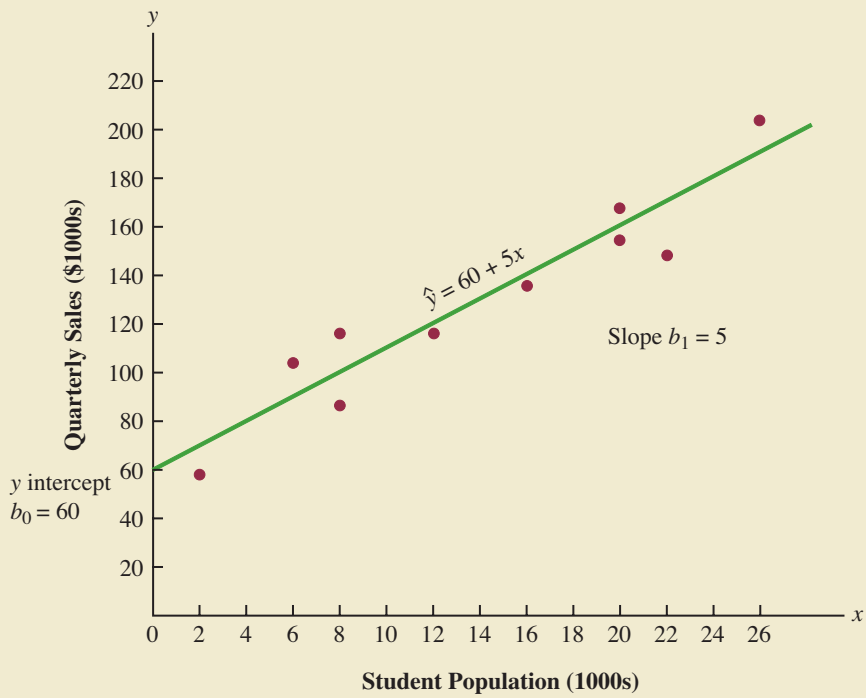
Using the estimated regression equation to make predictions outside the range of the values of the independent variable should be done with caution because outside that range we cannot be sure that the same relationship is valid.

If we believe the least squares estimated regression equation adequately describes the relationship between x and y, it would seem reasonable to use the estimated regression equation to predict the value of y for a given value of x. For example, if we wanted to predict quarterly sales for a restaurant to be located near a campus with 16,000 students, we would compute

$$\hat{y} = 60 + 5(16) = 140$$

Hence, we would predict quarterly sales of \$140,000 for this restaurant. In the following sections we will discuss methods for assessing the appropriateness of using the estimated regression equation for estimation and prediction.

FIGURE 14.4 GRAPH OF THE ESTIMATED REGRESSION EQUATION FOR ARMAND'S PIZZA PARLORS: $\hat{y} = 60 + 5x$



NOTES AND COMMENTS

The least squares method provides an estimated regression equation that minimizes the sum of squared deviations between the observed values of the dependent variable y_i and the estimated values of the dependent variable $\hat{y}_i$. This least squares criterion is used to choose the equation that provides the best fit. If some other criterion were used, such as minimizing the sum of the absolute deviations between y_i and $\hat{y}_i$, a different equation would be obtained. In practice, the least squares method is the most widely used.

Exercises

Methods

SELF test

1. Given are five observations for two variables, x and y.

x_i	1	2	3	4	5
y_i	3	7	5	11	14

   a. Develop a scatter diagram for these data.
   b. What does the scatter diagram developed in part (a) indicate about the relationship between the two variables?

c. Try to approximate the relationship between x and y by drawing a straight line through the data.
d. Develop the estimated regression equation by computing the values of b_0 and b_1 using equations (14.6) and (14.7).
e. Use the estimated regression equation to predict the value of y when $x = 4$.

2. Given are five observations for two variables, x and y.

x_i	3	12	6	20	14
y_i	55	40	55	10	15

a. Develop a scatter diagram for these data.
b. What does the scatter diagram developed in part (a) indicate about the relationship between the two variables?
c. Try to approximate the relationship between x and y by drawing a straight line through the data.
d. Develop the estimated regression equation by computing the values of b_0 and b_1 using equations (14.6) and (14.7).
e. Use the estimated regression equation to predict the value of y when $x = 10$.

3. Given are five observations collected in a regression study on two variables.

x_i	2	6	9	13	20
y_i	7	18	9	26	23

a. Develop a scatter diagram for these data.
b. Develop the estimated regression equation for these data.
c. Use the estimated regression equation to predict the value of y when $x = 6$.

Applications

SELF test

4. The following data were collected on the height (inches) and weight (pounds) of women swimmers.

Height	68	64	62	65	66
Weight	132	108	102	115	128

a. Develop a scatter diagram for these data with height as the independent variable.
b. What does the scatter diagram developed in part (a) indicate about the relationship between the two variables?
c. Try to approximate the relationship between height and weight by drawing a straight line through the data.
d. Develop the estimated regression equation by computing the values of b_0 and b_1.
e. If a swimmer's height is 63 inches, what would you estimate her weight to be?

5. Elliptical trainers are becoming one of the more popular exercise machines. Their smooth and steady low-impact motion makes them a preferred choice for individuals with knee and ankle problems. But selecting the right trainer can be a difficult process. Price and quality are two important factors in any purchase decision. Are higher prices generally associated with higher quality elliptical trainers? *Consumer Reports* conducted extensive tests to develop an overall rating based on ease of use, ergonomics, construction, and

exercise range. The following data show the price and rating for eight elliptical trainers tested (*Consumer Reports,* February 2008).

WEB file
Ellipticals

Brand and Model	Price ($)	Rating
Precor 5.31	3700	87
Keys Fitness CG2	2500	84
Octane Fitness Q37e	2800	82
LifeFitness X1 Basic	1900	74
NordicTrack AudioStrider 990	1000	73
Schwinn 430	800	69
Vision Fitness X6100	1700	68
ProForm XP 520 Razor	600	55

a. Develop a scatter diagram with price as the independent variable.
b. An exercise equipment store that sells primarily higher priced equipment has a sign over the display area that says "Quality: You Get What You Pay For." Based upon your analysis of the data for ellipical trainers, do you think this sign fairly reflects the price-quality relationship for elliptical trainers?
c. Use the least squares method to develop the estimated regression equation.
d. Use the estimated regression equation to predict the rating for an ellipitical trainer with a price of $1500.

6. The cost of a previously owned car depends upon factors such as make and model, model year, mileage, condition, and whether the car is purchased from a dealer or from a private seller. To investigate the relationship between the car's mileage and the sales price, data were collected on the mileage and the sale price for 10 private sales of model year 2000 Honda Accords (PriceHub website, October 2008).

WEB file
HondaAccord

Miles (1000s)	Price ($1000s)
90	7.0
59	7.5
66	6.6
87	7.2
90	7.0
106	5.4
94	6.4
57	7.0
138	5.1
87	7.2

a. Develop a scatter diagram with miles as the independent variable.
b. What does the scatter diagram developed in part (a) indicate about the relationship between the two variables?
c. Use the least squares method to develop the estimated regression equation.
d. Provide an interpretation for the slope of the estimated regression equation.
e. Predict the sales price for a 2000 Honda Accord with 100,000 miles.

7. A sales manager collected the following data on annual sales and years of experience.

WEB file
Sales

Salesperson	Years of Experience	Annual Sales ($1000s)
1	1	80
2	3	97
3	4	92
4	4	102
5	6	103
6	8	111
7	10	119
8	10	123
9	11	117
10	13	136

   a. Develop a scatter diagram for these data with years of experience as the independent variable.
   b. Develop an estimated regression equation that can be used to predict annual sales given the years of experience.
   c. Use the estimated regression equation to predict annual sales for a salesperson with 9 years of experience.

8. Bergans of Norway has been making outdoor gear since 1908. The following data show the temperature rating (F°) and the price ($) for 11 models of sleeping bags produced by Bergans (*Backpacker* 2006 Gear Guide).

WEB file
SleepingBags

Model	Temperature Rating (F°)	Price ($)
Ranger 3-Seasons	12	319
Ranger Spring	24	289
Ranger Winter	3	389
Rondane 3-Seasons	13	239
Rondane Summer	38	149
Rondane Winter	4	289
Senja Ice	5	359
Senja Snow	15	259
Senja Zero	25	229
Super Light	45	129
Tight & Light	25	199

   a. Develop a scatter diagram for these data with temperature rating (F°) as the independent variable.
   b. What does the scatter diagram developed in part (a) indicate about the relationship between temperature rating (F°) and price?
   c. Use the least squares method to develop the estimated regression equation.
   d. Predict the price for a sleeping bag with a temperature rating (F°) of 20.

9. To avoid extra checked-bag fees, airline travelers often pack as much as they can into their suitcase. Finding a rolling suitcase that is durable, has good capacity, and is easy to pull can be difficult. The following table shows the results of tests conducted by *Consumer Reports* for 10 rolling suitcases; higher scores indicate better overall test results (Consumer Reports website, October 2008).

WEB file
Suitcases

Brand	Price ($)	Score
Briggs & Riley	325	72
Hartman	350	74
Heys	67	54
Kenneth Cole Reaction	120	54
Liz Claiborne	85	64
Samsonite	180	57
Titan	360	66
TravelPro	156	67
Tumi	595	87
Victorinox	400	77

a. Develop a scatter diagram with price as the independent variable.
b. What does the scatter diagram developed in part (a) indicate about the relationship between the two variables?
c. Use the least squares method to develop the estimated regression equation.
d. Provide an interpretation for the slope of the estimated regression equation.
e. The Eagle Creek Hovercraft suitcase has a price of $225. Predict the score for this suitcase using the estimated regression equation developed in part (c).

10. According to *Advertising Age*'s annual salary review, Mark Hurd, the 49-year-old chairman, president, and CEO of Hewlett-Packard Co., received an annual salary of $817,000, a bonus of more than $5 million, and other compensation exceeding $17 million. His total compensation was slightly better than the average CEO total pay of $12.4 million. The following table shows the age and annual salary (in thousands of dollars) for Mark Hurd and 14 other executives who led publicly held companies (*Advertising Age,* December 5, 2006).

WEB file
ExecSalary

Executive	Title	Company	Age	Salary ($1000s)
Charles Prince	Chmn/CEO	Citigroup	56	1000
Harold McGraw III	Chmn/Pres/CEO	McGraw-Hill Cos.	57	1172
James Dimon	Pres/CEO	JP Morgan Chase & Co.	50	1000
K. Rupert Murdoch	Chmn/CEO	News Corp.	75	4509
Kenneth D. Lewis	Chmn/Pres/CEO	Bank of America	58	1500
Kenneth I. Chenault	Chmn/CEO	American Express Co.	54	1092
Louis C. Camilleri	Chmn/CEO	Altria Group	51	1663
Mark V. Hurd	Chmn/Pres/CEO	Hewlett-Packard Co.	49	817
Martin S. Sorrell	CEO	WPP Group	61	1562
Robert L. Nardelli	Chmn/Pres/CEO	Home Depot	57	2164
Samuel J. Palmisano	Chmn/Pres/CEO	IBM Corp.	55	1680
David C. Novak	Chmn/Pres/CEO	Yum Brands	53	1173
Henry R. Silverman	Chmn/CEO	Cendant Corp.	65	3300
Robert C. Wright	Chmn/CEO	NBC Universal	62	2500
Sumner Redstone	Exec Chmn/Founder	Viacom	82	5807

a. Develop a scatter diagram for these data with the age of the executive as the independent variable.
b. What does the scatter diagram developed in part (a) indicate about the relationship between the two variables?
c. Develop the least squares estimated regression equation.
d. Suppose Bill Gustin is the 72-year-old chairman, president, and CEO of a major electronics company. Predict the annual salary for Bill Gustin.

11. Sporty cars are designed to provide better handling, acceleration, and a more responsive driving experience than a typical sedan. But, even within this select group of cars, performance as well as price can vary. *Consumer Reports* provided road-test scores and prices for the following 12 sporty cars (Consumer Reports website, October 2008). Prices are in thousands of dollars and road-test scores are based on a 0–100 rating scale, with higher values indicating better performance.

WEB file
SportyCars

Car	Price ($1000s)	Road-Test Score
Chevrolet Cobalt SS	24.5	78
Dodge Caliber SRT4	24.9	56
Ford Mustang GT (V8)	29.0	73
Honda Civic Si	21.7	78
Mazda RX-8	31.3	86
Mini Cooper S	26.4	74
Mitsubishi Lancer Evolution GSR	38.1	83
Nissan Sentra SE-R Spec V	23.3	66
Suburu Impreza WRX	25.2	81
Suburu Impreza WRX Sti	37.6	89
Volkswagen GTI	24.0	83
Volkswagen R32	33.6	83

a. Develop a scatter diagram with price as the independent variable.
b. What does the scatter diagram developed in part (a) indicate about the relationship between the two variables?
c. Use the least squares method to develop the estimated regression equation.
d. Provide an interpretation for the slope of the estimated regression equation.
e. Another sporty car that *Consumer Reports* tested is the BMW 135i; the price for this car was $36,700. Predict the road-test score for the BMW 135i using the estimated regression equation developed in part (c).

12. A personal watercraft (PWC) is a vessel propelled by water jets, designed to be operated by a person sitting, standing, or kneeling on the vessel. In the early 1970s, Kawasaki Motors Corp. U.S.A. introduced the JET SKI® watercraft, the first commercially successful PWC. Today, *jet ski* is commonly used as a generic term for personal watercraft. The following data show the weight (rounded to the nearest 10 lbs.) and the price (rounded to the nearest $50) for 10 three-seater personal watercraft (Jetski News website, 2006).

WEB file
JetSki

Make and Model	Weight (lbs.)	Price ($)
Honda AquaTrax F-12	750	9500
Honda AquaTrax F-12X	790	10500
Honda AquaTrax F-12X GPScape	800	11200
Kawasaki STX-12F Jetski	740	8500
Yamaha FX Cruiser Waverunner	830	10000
Yamaha FX High Output Waverunner	770	10000
Yamaha FX Waverunner	830	9300
Yamaha VX110 Deluxe Waverunner	720	7700
Yamaha VX110 Sport Waverunner	720	7000
Yamaha XLT1200 Waverunner	780	8500

a. Develop a scatter diagram for these data with weight as the independent variable.
b. What does the scatter diagram developed in part (a) indicate about the relationship between weight and price?
c. Use the least squares method to develop the estimated regression equation.
d. Predict the price for a three-seater PWC with a weight of 750 pounds.

e. The Honda AquaTrax F-12 weighs 750 pounds and has a price of $9500. Shouldn't the predicted price you developed in part (d) for a PWC with a weight of 750 pounds also be $9500?

f. The Kawasaki SX-R 800 Jetski has a seating capacity of one and weighs 350 pounds. Do you think the estimated regression equation developed in part (c) should be used to predict the price for this model?

13. To the Internal Revenue Service, the reasonableness of total itemized deductions depends on the taxpayer's adjusted gross income. Large deductions, which include charity and medical deductions, are more reasonable for taxpayers with large adjusted gross incomes. If a taxpayer claims larger than average itemized deductions for a given level of income, the chances of an IRS audit are increased. Data (in thousands of dollars) on adjusted gross income and the average or reasonable amount of itemized deductions follow.

Adjusted Gross Income ($1000s)	Reasonable Amount of Itemized Deductions ($1000s)
22	9.6
27	9.6
32	10.1
48	11.1
65	13.5
85	17.7
120	25.5

a. Develop a scatter diagram for these data with adjusted gross income as the independent variable.

b. Use the least squares method to develop the estimated regression equation.

c. Estimate a reasonable level of total itemized deductions for a taxpayer with an adjusted gross income of $52,500. If this taxpayer claimed itemized deductions of $20,400, would the IRS agent's request for an audit appear justified? Explain.

14. *PCWorld* rated four component characteristics for 10 ultraportable laptop computers: features, performance, design, and price. Each characteristic was rated using a 0–100 point scale. An overall rating, referred to as the *PCW World* Rating, was then developed for each laptop. The following table shows the features rating and the *PCW World* Rating for the 10 laptop computers (PC World website, February 5, 2009).

WEB file
Laptop

Model	Features Rating	*PCW World* Rating
Thinkpad X200	87	83
VGN-Z598U	85	82
U6V	80	81
Elitebook 2530P	75	78
X360	80	78
Thinkpad X300	76	78
Ideapad U110	81	77
Micro Express JFT2500	73	75
Toughbook W7	79	73
HP Voodoo Envy133	68	72

a. Develop a scatter diagram with the features rating as the independent variable.

b. What does the scatter diagram developed in part (a) indicate about the relationship between the two variables?

c. Use the least squares method to develop the estimated regression equation.

d. Estimate the *PCW World* Rating for a new laptop computer that has a features rating of 70.

14.3 Coefficient of Determination

For the Armand's Pizza Parlors example, we developed the estimated regression equation $\hat{y} = 60 + 5x$ to approximate the linear relationship between the size of the student population x and quarterly sales y. A question now is: How well does the estimated regression equation fit the data? In this section, we show that the **coefficient of determination** provides a measure of the goodness of fit for the estimated regression equation.

For the ith observation, the difference between the observed value of the dependent variable, y_i, and the estimated value of the dependent variable, $\hat{y}_i$, is called the ***i*th residual**. The ith residual represents the error in using $\hat{y}_i$ to estimate y_i. Thus, for the ith observation, the residual is $y_i - \hat{y}_i$. The sum of squares of these residuals or errors is the quantity that is minimized by the least squares method. This quantity, also known as the *sum of squares due to error,* is denoted by SSE.

SUM OF SQUARES DUE TO ERROR

$$\text{SSE} = \Sigma(y_i - \hat{y}_i)^2 \qquad (14.8)$$

The value of SSE is a measure of the error in using the estimated regression equation to estimate the values of the dependent variable in the sample.

In Table 14.3 we show the calculations required to compute the sum of squares due to error for the Armand's Pizza Parlors example. For instance, for restaurant 1 the values of the independent and dependent variables are $x_1 = 2$ and $y_1 = 58$. Using the estimated regression equation, we find that the estimated value of quarterly sales for restaurant 1 is $\hat{y}_1 = 60 + 5(2) = 70$. Thus, the error in using $\hat{y}_1$ to estimate y_1 for restaurant 1 is $y_1 - \hat{y}_1 = 58 - 70 = -12$. The squared error, $(-12)^2 = 144$, is shown in the last column of Table 14.3. After computing and squaring the residuals for each restaurant in the sample, we sum them to obtain SSE = 1530. Thus, SSE = 1530 measures the error in using the estimated regression equation $\hat{y} = 60 + 5x$ to predict sales.

Now suppose we are asked to develop an estimate of quarterly sales without knowledge of the size of the student population. Without knowledge of any related variables, we would

TABLE 14.3 CALCULATION OF SSE FOR ARMAND'S PIZZA PARLORS

Restaurant i	x_i = Student Population (1000s)	y_i = Quarterly Sales ($1000s)	Predicted Sales $\hat{y}_i = 60 + 5x_i$	Error $y_i - \hat{y}_i$	Squared Error $(y_i - \hat{y}_i)^2$
1	2	58	70	−12	144
2	6	105	90	15	225
3	8	88	100	−12	144
4	8	118	100	18	324
5	12	117	120	−3	9
6	16	137	140	−3	9
7	20	157	160	−3	9
8	20	169	160	9	81
9	22	149	170	−21	441
10	26	202	190	12	144
					SSE = 1530

TABLE 14.4 COMPUTATION OF THE TOTAL SUM OF SQUARES FOR ARMAND'S PIZZA PARLORS

Restaurant i	x_i = Student Population (1000s)	y_i = Quarterly Sales ($1000s)	Deviation $y_i - \bar{y}$	Squared Deviation $(y_i - \bar{y})^2$
1	2	58	−72	5,184
2	6	105	−25	625
3	8	88	−42	1,764
4	8	118	−12	144
5	12	117	−13	169
6	16	137	7	49
7	20	157	27	729
8	20	169	39	1,521
9	22	149	19	361
10	26	202	72	5,184
				SST = 15,730

use the sample mean as an estimate of quarterly sales at any given restaurant. Table 14.2 showed that for the sales data, $\Sigma y_i = 1300$. Hence, the mean value of quarterly sales for the sample of 10 Armand's restaurants is $\bar{y} = \Sigma y_i/n = 1300/10 = 130$. In Table 14.4 we show the sum of squared deviations obtained by using the sample mean $\bar{y} = 130$ to estimate the value of quarterly sales for each restaurant in the sample. For the ith restaurant in the sample, the difference $y_i - \bar{y}$ provides a measure of the error involved in using $\bar{y}$ to estimate sales. The corresponding sum of squares, called the *total sum of squares,* is denoted SST.

TOTAL SUM OF SQUARES

$$\text{SST} = \Sigma(y_i - \bar{y})^2 \qquad \textbf{(14.9)}$$

The sum at the bottom of the last column in Table 14.4 is the total sum of squares for Armand's Pizza Parlors; it is SST = 15,730.

With SST = *15,730 and* SSE = *1530, the estimated regression line provides a much better fit to the data than the line* $y = \bar{y}$.

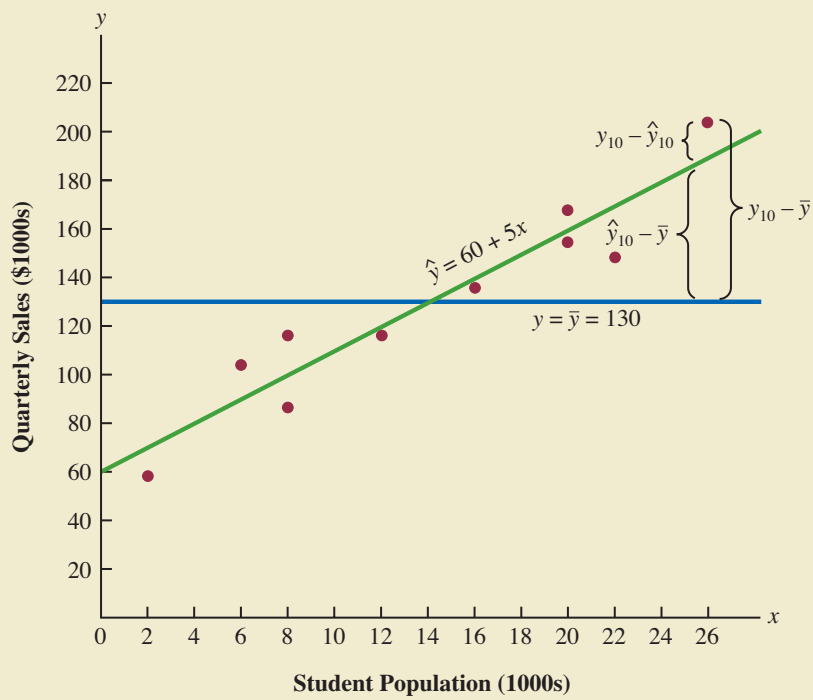
In Figure 14.5 we show the estimated regression line $\hat{y} = 60 + 5x$ and the line corresponding to $\bar{y} = 130$. Note that the points cluster more closely around the estimated regression line than they do about the line $\bar{y} = 130$. For example, for the 10th restaurant in the sample we see that the error is much larger when $\bar{y} = 130$ is used as an estimate of y_{10} than when $\hat{y}_{10} = 60 + 5(26) = 190$ is used. We can think of SST as a measure of how well the observations cluster about the $\bar{y}$ line and SSE as a measure of how well the observations cluster about the $\hat{y}$ line.

To measure how much the $\hat{y}$ values on the estimated regression line deviate from $\bar{y}$, another sum of squares is computed. This sum of squares, called the *sum of squares due to regression,* is denoted SSR.

SUM OF SQUARES DUE TO REGRESSION

$$\text{SSR} = \Sigma(\hat{y}_i - \bar{y})^2 \qquad \textbf{(14.10)}$$

FIGURE 14.5 DEVIATIONS ABOUT THE ESTIMATED REGRESSION LINE AND THE LINE $y = \bar{y}$ FOR ARMAND'S PIZZA PARLORS



From the preceding discussion, we should expect that SST, SSR, and SSE are related. Indeed, the relationship among these three sums of squares provides one of the most important results in statistics.

RELATIONSHIP AMONG SST, SSR, AND SSE

$$\text{SST} = \text{SSR} + \text{SSE} \qquad (14.11)$$

where

SST = total sum of squares
SSR = sum of squares due to regression
SSE = sum of squares due to error

SSR *can be thought of as the explained portion of* SST, *and* SSE *can be thought of as the unexplained portion of* SST.

Equation (14.11) shows that the total sum of squares can be partitioned into two components, the sum of squares due to regression and the sum of squares due to error. Hence, if the values of any two of these sum of squares are known, the third sum of squares can be computed easily. For instance, in the Armand's Pizza Parlors example, we already know that SSE = 1530 and SST = 15,730; therefore, solving for SSR in equation (14.11), we find that the sum of squares due to regression is

$$\text{SSR} = \text{SST} - \text{SSE} = 15{,}730 - 1530 = 14{,}200$$

Now let us see how the three sums of squares, SST, SSR, and SSE, can be used to provide a measure of the goodness of fit for the estimated regression equation. The estimated regression equation would provide a perfect fit if every value of the dependent variable y_i happened to lie on the estimated regression line. In this case, $y_i - \hat{y}_i$ would be zero for each observation, resulting in SSE = 0. Because SST = SSR + SSE, we see that for a perfect fit SSR must equal SST, and the ratio (SSR/SST) must equal one. Poorer fits will result in larger values for SSE. Solving for SSE in equation (14.11), we see that SSE = SST − SSR. Hence, the largest value for SSE (and hence the poorest fit) occurs when SSR = 0 and SSE = SST.

The ratio SSR/SST, which will take values between zero and one, is used to evaluate the goodness of fit for the estimated regression equation. This ratio is called the *coefficient of determination* and is denoted by r^2.

COEFFICIENT OF DETERMINATION

$$r^2 = \frac{\text{SSR}}{\text{SST}} \tag{14.12}$$

For the Armand's Pizza Parlors example, the value of the coefficient of determination is

$$r^2 = \frac{\text{SSR}}{\text{SST}} = \frac{14{,}200}{15{,}730} = .9027$$

When we express the coefficient of determination as a percentage, r^2 can be interpreted as the percentage of the total sum of squares that can be explained by using the estimated regression equation. For Armand's Pizza Parlors, we can conclude that 90.27% of the total sum of squares can be explained by using the estimated regression equation $\hat{y} = 60 + 5x$ to predict quarterly sales. In other words, 90.27% of the variability in sales can be explained by the linear relationship between the size of the student population and sales. We should be pleased to find such a good fit for the estimated regression equation.

Correlation Coefficient

In Chapter 3 we introduced the **correlation coefficient** as a descriptive measure of the strength of linear association between two variables, x and y. Values of the correlation coefficient are always between −1 and +1. A value of +1 indicates that the two variables x and y are perfectly related in a positive linear sense. That is, all data points are on a straight line that has a positive slope. A value of −1 indicates that x and y are perfectly related in a negative linear sense, with all data points on a straight line that has a negative slope. Values of the correlation coefficient close to zero indicate that x and y are not linearly related.

In Section 3.5 we presented the equation for computing the sample correlation coefficient. If a regression analysis has already been performed and the coefficient of determination r^2 computed, the sample correlation coefficient can be computed as follows.

SAMPLE CORRELATION COEFFICIENT

$$\begin{aligned} r_{xy} &= (\text{sign of } b_1)\sqrt{\text{Coefficient of determination}} \\ &= (\text{sign of } b_1)\sqrt{r^2} \end{aligned} \tag{14.13}$$

where

b_1 = the slope of the estimated regression equation $\hat{y} = b_0 + b_1 x$

The sign for the sample correlation coefficient is positive if the estimated regression equation has a positive slope ($b_1 > 0$) and negative if the estimated regression equation has a negative slope ($b_1 < 0$).

For the Armand's Pizza Parlor example, the value of the coefficient of determination corresponding to the estimated regression equation $\hat{y} = 60 + 5x$ is .9027. Because the slope of the estimated regression equation is positive, equation (14.13) shows that the sample correlation coefficient is $+\sqrt{.9027} = +.9501$. With a sample correlation coefficient of $r_{xy} = +.9501$, we would conclude that a strong positive linear association exists between x and y.

In the case of a linear relationship between two variables, both the coefficient of determination and the sample correlation coefficient provide measures of the strength of the relationship. The coefficient of determination provides a measure between zero and one, whereas the sample correlation coefficient provides a measure between -1 and $+1$. Although the sample correlation coefficient is restricted to a linear relationship between two variables, the coefficient of determination can be used for nonlinear relationships and for relationships that have two or more independent variables. Thus, the coefficient of determination provides a wider range of applicability.

NOTES AND COMMENTS

1. In developing the least squares estimated regression equation and computing the coefficient of determination, we made no probabilistic assumptions about the error term ϵ, and no statistical tests for significance of the relationship between x and y were conducted. Larger values of r^2 imply that the least squares line provides a better fit to the data; that is, the observations are more closely grouped about the least squares line. But, using only r^2, we can draw no conclusion about whether the relationship between x and y is statistically significant. Such a conclusion must be based on considerations that involve the sample size and the properties of the appropriate sampling distributions of the least squares estimators.
2. As a practical matter, for typical data found in the social sciences, values of r^2 as low as .25 are often considered useful. For data in the physical and life sciences, r^2 values of .60 or greater are often found; in fact, in some cases, r^2 values greater than .90 can be found. In business applications, r^2 values vary greatly, depending on the unique characteristics of each application.

Exercises

Methods

SELF test

15. The data from exercise 1 follow.

x_i	1	2	3	4	5
y_i	3	7	5	11	14

The estimated regression equation for these data is $\hat{y} = .20 + 2.60x$.

a. Compute SSE, SST, and SSR using equations (14.8), (14.9), and (14.10).
b. Compute the coefficient of determination r^2. Comment on the goodness of fit.
c. Compute the sample correlation coefficient.

16. The data from exercise 2 follow.

x_i	3	12	6	20	14
y_i	55	40	55	10	15

The estimated regression equation for these data is $\hat{y} = 68 - 3x$.
a. Compute SSE, SST, and SSR.
b. Compute the coefficient of determination r^2. Comment on the goodness of fit.
c. Compute the sample correlation coefficient.

17. The data from exercise 3 follow.

x_i	2	6	9	13	20
y_i	7	18	9	26	23

The estimated regression equation for these data is $\hat{y} = 7.6 + .9x$. What percentage of the total sum of squares can be accounted for by the estimated regression equation? What is the value of the sample correlation coefficient?

Applications

SELF test

18. The following data are the monthly salaries y and the grade point averages x for students who obtained a bachelor's degree in business administration with a major in information systems. The estimated regression equation for these data is $\hat{y} = 1790.5 + 581.1x$.

GPA	Monthly Salary ($)
2.6	3300
3.4	3600
3.6	4000
3.2	3500
3.5	3900
2.9	3600

a. Compute SST, SSR, and SSE.
b. Compute the coefficient of determination r^2. Comment on the goodness of fit.
c. What is the value of the sample correlation coefficient?

19. In exercise 7 a sales manager collected the following data on x = annual sales and y = years of experience. The estimated regression equation for these data is $\hat{y} = 80 + 4x$.

WEB file
Sales

Salesperson	Years of Experience	Annual Sales ($1000s)
1	1	80
2	3	97
3	4	92
4	4	102
5	6	103
6	8	111
7	10	119
8	10	123
9	11	117
10	13	136

a. Compute SST, SSR, and SSE.
b. Compute the coefficient of determination r^2. Comment on the goodness of fit.
c. What is the value of the sample correlation coefficient?

20. *Consumer Reports* provided extensive testing and ratings for more than 100 HDTVs. An overall score, based primarily on picture quality, was developed for each model. In general, a higher overall score indicates better performance. The following data show the price and overall score for the ten 42-inch plasma televisions (*Consumer Reports,* March 2006).

WEB file
PlasmaTV

Brand	Price	Score
Dell	2800	62
Hisense	2800	53
Hitachi	2700	44
JVC	3500	50
LG	3300	54
Maxent	2000	39
Panasonic	4000	66
Phillips	3000	55
Proview	2500	34
Samsung	3000	39

a. Use these data to develop an estimated regression equation that could be used to estimate the overall score for a 42-inch plasma television given the price.
b. Compute r^2. Did the estimated regression equation provide a good fit?
c. Estimate the overall score for a 42-inch plasma television with a price of $3200.

21. An important application of regression analysis in accounting is in the estimation of cost. By collecting data on volume and cost and using the least squares method to develop an estimated regression equation relating volume and cost, an accountant can estimate the cost associated with a particular manufacturing volume. Consider the following sample of production volumes and total cost data for a manufacturing operation.

Production Volume (units)	Total Cost ($)
400	4000
450	5000
550	5400
600	5900
700	6400
750	7000

a. Use these data to develop an estimated regression equation that could be used to predict the total cost for a given production volume.
b. What is the variable cost per unit produced?
c. Compute the coefficient of determination. What percentage of the variation in total cost can be explained by production volume?
d. The company's production schedule shows 500 units must be produced next month. What is the estimated total cost for this operation?

22. Refer to exercise 5 where the following data were used to investigate whether higher prices are generally associated with higher ratings for elliptical trainers (*Consumer Reports,* February 2008).

WEB file
Ellipticals

Brand and Model	Price ($)	Rating
Precor 5.31	3700	87
Keys Fitness CG2	2500	84
Octane Fitness Q37e	2800	82
LifeFitness X1 Basic	1900	74
NordicTrack AudioStrider 990	1000	73
Schwinn 430	800	69
Vision Fitness X6100	1700	68
ProForm XP 520 Razor	600	55

With x = price ($) and y = rating, the estimated regression equation is $\hat{y} = 58.158 + .008449x$. For these data, SSE = 173.88.

a. Compute the coefficient of determination r^2.
b. Did the estimated regression equation provide a good fit? Explain.
c. What is the value of the sample correlation coefficient? Does it reflect a strong or weak relationship between price and rating?

14.4 Model Assumptions

In conducting a regression analysis, we begin by making an assumption about the appropriate model for the relationship between the dependent and independent variable(s). For the case of simple linear regression, the assumed regression model is

$$y = \beta_0 + \beta_1 x + \epsilon$$

Then the least squares method is used to develop values for b_0 and b_1, the estimates of the model parameters β_0 and β_1, respectively. The resulting estimated regression equation is

$$\hat{y} = b_0 + b_1 x$$

We saw that the value of the coefficient of determination (r^2) is a measure of the goodness of fit of the estimated regression equation. However, even with a large value of r^2, the estimated regression equation should not be used until further analysis of the appropriateness of the assumed model has been conducted. An important step in determining whether the assumed model is appropriate involves testing for the significance of the relationship. The tests of significance in regression analysis are based on the following assumptions about the error term ϵ.

ASSUMPTIONS ABOUT THE ERROR TERM ϵ IN THE REGRESSION MODEL

$$y = \beta_0 + \beta_1 x + \epsilon$$

1. The error term ϵ is a random variable with a mean or expected value of zero; that is, $E(\epsilon) = 0$.
Implication: β_0 and β_1 are constants, therefore $E(\beta_0) = \beta_0$ and $E(\beta_1) = \beta_1$; thus, for a given value of x, the expected value of y is

$$E(y) = \beta_0 + \beta_1 x \qquad \textbf{(14.14)}$$

(*continued*)

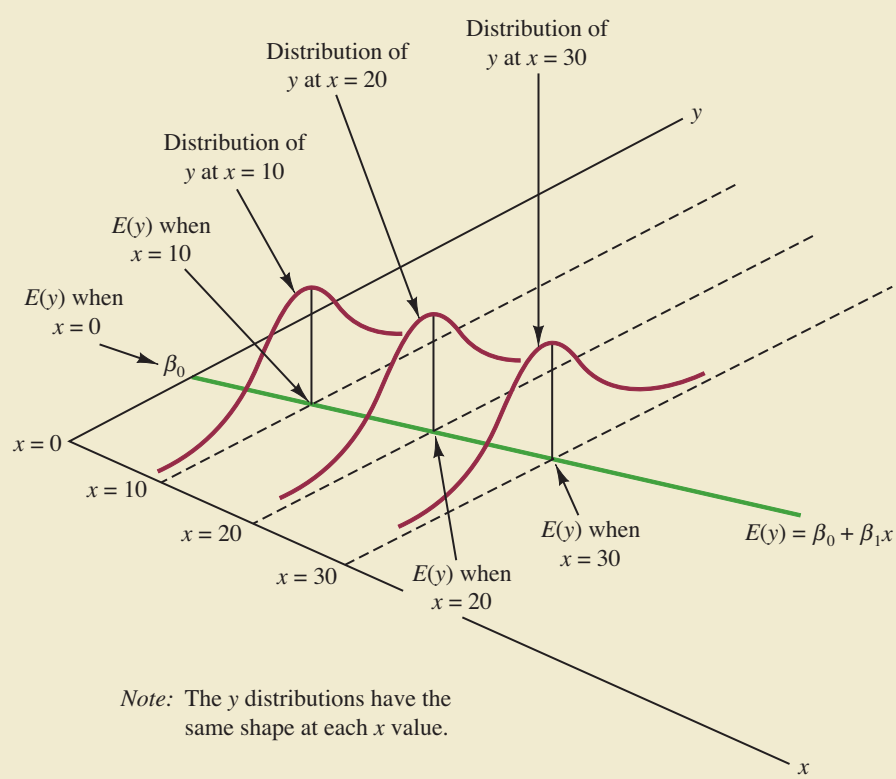
As we indicated previously, equation (14.14) is referred to as the regression equation.

2. The variance of ϵ, denoted by σ^2, is the same for all values of x.
   Implication: The variance of y about the regression line equals σ^2 and is the same for all values of x.
3. The values of ϵ are independent.
   Implication: The value of ϵ for a particular value of x is not related to the value of ϵ for any other value of x; thus, the value of y for a particular value of x is not related to the value of y for any other value of x.
4. The error term ϵ is a normally distributed random variable.
   Implication: Because y is a linear function of ϵ, y is also a normally distributed random variable.

Figure 14.6 illustrates the model assumptions and their implications; note that in this graphical interpretation, the value of $E(y)$ changes according to the specific value of x considered. However, regardless of the x value, the probability distribution of ϵ and hence the probability distributions of y are normally distributed, each with the same variance. The specific value of the error ϵ at any particular point depends on whether the actual value of y is greater than or less than $E(y)$.

At this point, we must keep in mind that we are also making an assumption or hypothesis about the form of the relationship between x and y. That is, we assume that a straight

FIGURE 14.6 ASSUMPTIONS FOR THE REGRESSION MODEL



Note: The y distributions have the same shape at each x value.

line represented by $\beta_0 + \beta_1 x$ is the basis for the relationship between the variables. We must not lose sight of the fact that some other model, for instance $y = \beta_0 + \beta_1 x^2 + \epsilon$, may turn out to be a better model for the underlying relationship.

14.5 Testing for Significance

In a simple linear regression equation, the mean or expected value of y is a linear function of x: $E(y) = \beta_0 + \beta_1 x$. If the value of β_1 is zero, $E(y) = \beta_0 + (0)x = \beta_0$. In this case, the mean value of y does not depend on the value of x and hence we would conclude that x and y are not linearly related. Alternatively, if the value of β_1 is not equal to zero, we would conclude that the two variables are related. Thus, to test for a significant regression relationship, we must conduct a hypothesis test to determine whether the value of β_1 is zero. Two tests are commonly used. Both require an estimate of σ^2, the variance of ϵ in the regression model.

Estimate of σ^2

From the regression model and its assumptions we can conclude that σ^2, the variance of ϵ, also represents the variance of the y values about the regression line. Recall that the deviations of the y values about the estimated regression line are called residuals. Thus, SSE, the sum of squared residuals, is a measure of the variability of the actual observations about the estimated regression line. The **mean square error** (MSE) provides the estimate of σ^2; it is SSE divided by its degrees of freedom.

With $\hat{y}_i = b_0 + b_1 x_i$, SSE can be written as

$$\text{SSE} = \Sigma(y_i - \hat{y}_i)^2 = \Sigma(y_i - b_0 - b_1 x_i)^2$$

Every sum of squares has associated with it a number called its degrees of freedom. Statisticians have shown that SSE has $n - 2$ degrees of freedom because two parameters (β_0 and β_1) must be estimated to compute SSE. Thus, the mean square error is computed by dividing SSE by $n - 2$. MSE provides an unbiased estimator of σ^2. Because the value of MSE provides an estimate of σ^2, the notation s^2 is also used.

MEAN SQUARE ERROR (ESTIMATE OF σ^2)

$$s^2 = \text{MSE} = \frac{\text{SSE}}{n - 2} \qquad (14.15)$$

In Section 14.3 we showed that for the Armand's Pizza Parlors example, SSE = 1530; hence,

$$s^2 = \text{MSE} = \frac{1530}{8} = 191.25$$

provides an unbiased estimate of σ^2.

To estimate σ we take the square root of s^2. The resulting value, s, is referred to as the **standard error of the estimate**.

STANDARD ERROR OF THE ESTIMATE

$$s = \sqrt{\text{MSE}} = \sqrt{\frac{\text{SSE}}{n - 2}} \qquad (14.16)$$

For the Armand's Pizza Parlors example, $s = \sqrt{\text{MSE}} = \sqrt{191.25} = 13.829$. In the following discussion, we use the standard error of the estimate in the tests for a significant relationship between x and y.

t Test

The simple linear regression model is $y = \beta_0 + \beta_1 x + \epsilon$. If x and y are linearly related, we must have $\beta_1 \neq 0$. The purpose of the t test is to see whether we can conclude that $\beta_1 \neq 0$. We will use the sample data to test the following hypotheses about the parameter β_1.

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

If H_0 is rejected, we will conclude that $\beta_1 \neq 0$ and that a statistically significant relationship exists between the two variables. However, if H_0 cannot be rejected, we will have insufficient evidence to conclude that a significant relationship exists. The properties of the sampling distribution of b_1, the least squares estimator of β_1, provide the basis for the hypothesis test.

First, let us consider what would happen if we used a different random sample for the same regression study. For example, suppose that Armand's Pizza Parlors used the sales records of a different sample of 10 restaurants. A regression analysis of this new sample might result in an estimated regression equation similar to our previous estimated regression equation $\hat{y} = 60 + 5x$. However, it is doubtful that we would obtain exactly the same equation (with an intercept of exactly 60 and a slope of exactly 5). Indeed, b_0 and b_1, the least squares estimators, are sample statistics with their own sampling distributions. The properties of the sampling distribution of b_1 follow.

SAMPLING DISTRIBUTION OF b_1

Expected Value

$$E(b_1) = \beta_1$$

Standard Deviation

$$\sigma_{b_1} = \frac{\sigma}{\sqrt{\Sigma(x_i - \bar{x})^2}} \qquad \textbf{(14.17)}$$

Distribution Form

Normal

Note that the expected value of b_1 is equal to β_1, so b_1 is an unbiased estimator of β_1.

Because we do not know the value of σ, we develop an estimate of σ_{b_1}, denoted s_{b_1}, by estimating σ with s in equation (14.17). Thus, we obtain the following estimate of σ_{b_1}.

The standard deviation of b_1 is also referred to as the standard error of b_1. Thus, s_{b_1} provides an estimate of the standard error of b_1.

ESTIMATED STANDARD DEVIATION OF b_1

$$s_{b_1} = \frac{s}{\sqrt{\Sigma(x_i - \bar{x})^2}} \qquad \textbf{(14.18)}$$

For Armand's Pizza Parlors, $s = 13.829$. Hence, using $\Sigma(x_i - \bar{x})^2 = 568$ as shown in Table 14.2, we have

$$s_{b_1} = \frac{13.829}{\sqrt{568}} = .5803$$

as the estimated standard deviation of b_1.

The t test for a significant relationship is based on the fact that the test statistic

$$\frac{b_1 - \beta_1}{s_{b_1}}$$

follows a t distribution with $n - 2$ degrees of freedom. If the null hypothesis is true, then $\beta_1 = 0$ and $t = b_1/s_{b_1}$.

Let us conduct this test of significance for Armand's Pizza Parlors at the $\alpha = .01$ level of significance. The test statistic is

$$t = \frac{b_1}{s_{b_1}} = \frac{5}{.5803} = 8.62$$

Appendixes 14.3 and 14.4 show how Minitab and Excel can be used to compute the p-value.

The t distribution table shows that with $n - 2 = 10 - 2 = 8$ degrees of freedom, $t = 3.355$ provides an area of .005 in the upper tail. Thus, the area in the upper tail of the t distribution corresponding to the test statistic $t = 8.62$ must be less than .005. Because this test is a two-tailed test, we double this value to conclude that the p-value associated with $t = 8.62$ must be less than $2(.005) = .01$. Excel or Minitab show the p-value $= .000$. Because the p-value is less than $\alpha = .01$, we reject H_0 and conclude that β_1 is not equal to zero. This evidence is sufficient to conclude that a significant relationship exists between student population and quarterly sales. A summary of the t test for significance in simple linear regression follows.

t TEST FOR SIGNIFICANCE IN SIMPLE LINEAR REGRESSION

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

TEST STATISTIC

$$t = \frac{b_1}{s_{b_1}} \qquad \textbf{(14.19)}$$

REJECTION RULE

p-value approach:	Reject H_0 if p-value $\leq \alpha$
Critical value approach:	Reject H_0 if $t \leq -t_{\alpha/2}$ or if $t \geq t_{\alpha/2}$

where $t_{\alpha/2}$ is based on a t distribution with $n - 2$ degrees of freedom.

Confidence Interval for β_1

The form of a confidence interval for β_1 is as follows:

$$b_1 \pm t_{\alpha/2} s_{b_1}$$

The point estimator is b_1 and the margin of error is $t_{\alpha/2}s_{b_1}$. The confidence coefficient associated with this interval is $1 - \alpha$, and $t_{\alpha/2}$ is the t value providing an area of $\alpha/2$ in the upper tail of a t distribution with $n - 2$ degrees of freedom. For example, suppose that we wanted to develop a 99% confidence interval estimate of β_1 for Armand's Pizza Parlors. From Table 2 of Appendix B we find that the t value corresponding to $\alpha = .01$ and $n - 2 = 10 - 2 = 8$ degrees of freedom is $t_{.005} = 3.355$. Thus, the 99% confidence interval estimate of β_1 is

$$b_1 \pm t_{\alpha/2}s_{b_1} = 5 \pm 3.355(.5803) = 5 \pm 1.95$$

or 3.05 to 6.95.

In using the t test for significance, the hypotheses tested were

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

At the $\alpha = .01$ level of significance, we can use the 99% confidence interval as an alternative for drawing the hypothesis testing conclusion for the Armand's data. Because 0, the hypothesized value of β_1, is not included in the confidence interval (3.05 to 6.95), we can reject H_0 and conclude that a significant statistical relationship exists between the size of the student population and quarterly sales. In general, a confidence interval can be used to test any two-sided hypothesis about β_1. If the hypothesized value of β_1 is contained in the confidence interval, do not reject H_0. Otherwise, reject H_0.

F Test

An F test, based on the F probability distribution, can also be used to test for significance in regression. With only one independent variable, the F test will provide the same conclusion as the t test; that is, if the t test indicates $\beta_1 \neq 0$ and hence a significant relationship, the F test will also indicate a significant relationship. But with more than one independent variable, only the F test can be used to test for an overall significant relationship.

The logic behind the use of the F test for determining whether the regression relationship is statistically significant is based on the development of two independent estimates of σ^2. We explained how MSE provides an estimate of σ^2. If the null hypothesis $H_0: \beta_1 = 0$ is true, the sum of squares due to regression, SSR, divided by its degrees of freedom provides another independent estimate of σ^2. This estimate is called the *mean square due to regression,* or simply the *mean square regression,* and is denoted MSR. In general,

$$\text{MSR} = \frac{\text{SSR}}{\text{Regression degrees of freedom}}$$

For the models we consider in this text, the regression degrees of freedom is always equal to the number of independent variables in the model:

$$\text{MSR} = \frac{\text{SSR}}{\text{Number of independent variables}} \qquad \textbf{(14.20)}$$

Because we consider only regression models with one independent variable in this chapter, we have MSR = SSR/1 = SSR. Hence, for Armand's Pizza Parlors, MSR = SSR = 14,200.

If the null hypothesis ($H_0: \beta_1 = 0$) is true, MSR and MSE are two independent estimates of σ^2 and the sampling distribution of MSR/MSE follows an F distribution with numerator

degrees of freedom equal to one and denominator degrees of freedom equal to $n - 2$. Therefore, when $\beta_1 = 0$, the value of MSR/MSE should be close to one. However, if the null hypothesis is false ($\beta_1 \neq 0$), MSR will overestimate σ^2 and the value of MSR/MSE will be inflated; thus, large values of MSR/MSE lead to the rejection of H_0 and the conclusion that the relationship between x and y is statistically significant.

Let us conduct the F test for the Armand's Pizza Parlors example. The test statistic is

$$F = \frac{\text{MSR}}{\text{MSE}} = \frac{14{,}200}{191.25} = 74.25$$

The F test and the t test provide identical results for simple linear regression.

The F distribution table (Table 4 of Appendix B) shows that with one degree of freedom in the numerator and $n - 2 = 10 - 2 = 8$ degrees of freedom in the denominator, $F = 11.26$ provides an area of .01 in the upper tail. Thus, the area in the upper tail of the F distribution corresponding to the test statistic $F = 74.25$ must be less than .01. Thus, we conclude that the p-value must be less than .01. Excel or Minitab show the p-value = .000. Because the p-value is less than $\alpha = .01$, we reject H_0 and conclude that a significant relationship exists between the size of the student population and quarterly sales. A summary of the F test for significance in simple linear regression follows.

If H_0 is false, MSE *still provides an unbiased estimate of σ^2 and* MSR *overestimates σ^2. If H_0 is true, both* MSE *and* MSR *provide unbiased estimates of σ^2; in this case the value of* MSR/MSE *should be close to 1.*

F TEST FOR SIGNIFICANCE IN SIMPLE LINEAR REGRESSION

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

TEST STATISTIC

$$F = \frac{\text{MSR}}{\text{MSE}} \qquad \textbf{(14.21)}$$

REJECTION RULE

p-value approach:	Reject H_0 if p-value $\leq \alpha$
Critical value approach:	Reject H_0 if $F \geq F_\alpha$

where F_α is based on an F distribution with 1 degree of freedom in the numerator and $n - 2$ degrees of freedom in the denominator.

In Chapter 13 we covered analysis of variance (ANOVA) and showed how an **ANOVA table** could be used to provide a convenient summary of the computational aspects of analysis of variance. A similar ANOVA table can be used to summarize the results of the F test for significance in regression. Table 14.5 is the general form of the ANOVA table for simple linear regression. Table 14.6 is the ANOVA table with the F test computations performed for Armand's Pizza Parlors. Regression, Error, and Total are the labels for the three sources of variation, with SSR, SSE, and SST appearing as the corresponding sum of squares in column 2. The degrees of freedom, 1 for SSR, $n - 2$ for SSE, and $n - 1$ for SST, are shown in column 3. Column 4 contains the values of MSR and MSE, column 5 contains the value of F = MSR/MSE, and column 6 contains the p-value corresponding to the F value in column 5. Almost all computer printouts of regression analysis include an ANOVA table summary of the F test for significance.

TABLE 14.5 GENERAL FORM OF THE ANOVA TABLE FOR SIMPLE LINEAR REGRESSION

In every analysis of variance table the total sum of squares is the sum of the regression sum of squares and the error sum of squares; in addition, the total degrees of freedom is the sum of the regression degrees of freedom and the error degrees of freedom.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	*F*	*p*-value
Regression	SSR	1	$\text{MSR} = \frac{\text{SSR}}{1}$	$F = \frac{\text{MSR}}{\text{MSE}}$	
Error	SSE	$n - 2$	$\text{MSE} = \frac{\text{SSE}}{n-2}$		
Total	SST	$n - 1$			

Some Cautions About the Interpretation of Significance Tests

Regression analysis, which can be used to identify how variables are associated with one another, cannot be used as evidence of a cause-and-effect relationship.

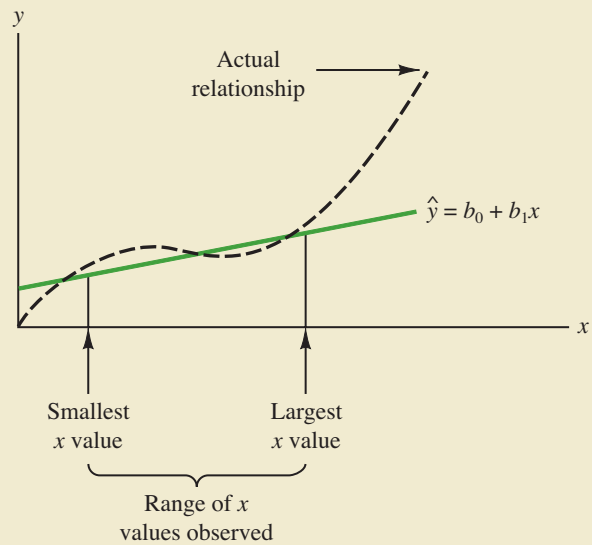
Rejecting the null hypothesis H_0: $\beta_1 = 0$ and concluding that the relationship between x and y is significant does not enable us to conclude that a cause-and-effect relationship is present between x and y. Concluding a cause-and-effect relationship is warranted only if the analyst can provide some type of theoretical justification that the relationship is in fact causal. In the Armand's Pizza Parlors example, we can conclude that there is a significant relationship between the size of the student population x and quarterly sales y; moreover, the estimated regression equation $\hat{y} = 60 + 5x$ provides the least squares estimate of the relationship. We cannot, however, conclude that changes in student population x *cause* changes in quarterly sales y just because we identified a statistically significant relationship. The appropriateness of such a cause-and-effect conclusion is left to supporting theoretical justification and to good judgment on the part of the analyst. Armand's managers felt that increases in the student population were a likely cause of increased quarterly sales. Thus, the result of the significance test enabled them to conclude that a cause-and-effect relationship was present.

In addition, just because we are able to reject H_0: $\beta_1 = 0$ and demonstrate statistical significance does not enable us to conclude that the relationship between x and y is linear. We can state only that x and y are related and that a linear relationship explains a significant portion of the variability in y over the range of values for x observed in the sample. Figure 14.7 illustrates this situation. The test for significance calls for the rejection of the null hypothesis H_0: $\beta_1 = 0$ and leads to the conclusion that x and y are significantly related, but the figure shows that the actual relationship between x and y is not linear. Although the

TABLE 14.6 ANOVA TABLE FOR THE ARMAND'S PIZZA PARLORS PROBLEM

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	*F*	*p*-value
Regression	14,200	1	$\frac{14{,}200}{1} = 14{,}200$	$\frac{14{,}200}{191.25} = 74.25$	.000
Error	1,530	8	$\frac{1530}{8} = 191.25$		
Total	15,730	9			

FIGURE 14.7 EXAMPLE OF A LINEAR APPROXIMATION OF A NONLINEAR RELATIONSHIP



linear approximation provided by $\hat{y} = b_0 + b_1x$ is good over the range of x values observed in the sample, it becomes poor for x values outside that range.

Given a significant relationship, we should feel confident in using the estimated regression equation for predictions corresponding to x values within the range of the x values observed in the sample. For Armand's Pizza Parlors, this range corresponds to values of x between 2 and 26. Unless other reasons indicate that the model is valid beyond this range, predictions outside the range of the independent variable should be made with caution. For Armand's Pizza Parlors, because the regression relationship has been found significant at the .01 level, we should feel confident using it to predict sales for restaurants where the associated student population is between 2000 and 26,000.

NOTES AND COMMENTS

1. The assumptions made about the error term (Section 14.4) are what allow the tests of statistical significance in this section. The properties of the sampling distribution of b_1 and the subsequent t and F tests follow directly from these assumptions.
2. Do not confuse statistical significance with practical significance. With very large sample sizes, statistically significant results can be obtained for small values of b_1; in such cases, one must exercise care in concluding that the relationship has practical significance.
3. A test of significance for a linear relationship between x and y can also be performed by using the sample correlation coefficient r_{xy}. With ρ_{xy} denoting the population correlation coefficient, the hypotheses are as follows.

$$H_0: \rho_{xy} = 0$$
$$H_a: \rho_{xy} \neq 0$$

A significant relationship can be concluded if H_0 is rejected. The details of this test are provided in Appendix 14.2. However, the t and F tests presented previously in this section give the same result as the test for significance using the correlation coefficient. Conducting a test for significance using the correlation coefficient therefore is not necessary if a t or F test has already been conducted.

Exercises

Methods

SELF test

23. The data from exercise 1 follow.

x_i	1	2	3	4	5
y_i	3	7	5	11	14

a. Compute the mean square error using equation (14.15).
b. Compute the standard error of the estimate using equation (14.16).
c. Compute the estimated standard deviation of b_1 using equation (14.18).
d. Use the t test to test the following hypotheses ($\alpha = .05$):

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

e. Use the F test to test the hypotheses in part (d) at a .05 level of significance. Present the results in the analysis of variance table format.

24. The data from exercise 2 follow.

x_i	3	12	6	20	14
y_i	55	40	55	10	15

a. Compute the mean square error using equation (14.15).
b. Compute the standard error of the estimate using equation (14.16).
c. Compute the estimated standard deviation of b_1 using equation (14.18).
d. Use the t test to test the following hypotheses ($\alpha = .05$):

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

e. Use the F test to test the hypotheses in part (d) at a .05 level of significance. Present the results in the analysis of variance table format.

25. The data from exercise 3 follow.

x_i	2	6	9	13	20
y_i	7	18	9	26	23

a. What is the value of the standard error of the estimate?
b. Test for a significant relationship by using the t test. Use $\alpha = .05$.
c. Use the F test to test for a significant relationship. Use $\alpha = .05$. What is your conclusion?

Applications

SELF test

26. In exercise 18 the data on grade point average and monthly salary were as follows.

GPA	Monthly Salary ($)	GPA	Monthly Salary ($)
2.6	3300	3.2	3500
3.4	3600	3.5	3900
3.6	4000	2.9	3600

a. Does the t test indicate a significant relationship between grade point average and monthly salary? What is your conclusion? Use $\alpha = .05$.
b. Test for a significant relationship using the F test. What is your conclusion? Use $\alpha = .05$.
c. Show the ANOVA table.

27. *Outside Magazine* tested 10 different models of day hikers and backpacking boots. The following data show the upper support and price for each model tested. Upper support was measured using a rating from 1 to 5, with a rating of 1 denoting average upper support and a rating of 5 denoting excellent upper support (*Outside Magazine Buyer's Guide,* 2001).

WEB file
Boots

Manufacturer and Model	Upper Support	Price ($)
Salomon Super Raid	2	120
Merrell Chameleon Prime	3	125
Teva Challenger	3	130
Vasque Fusion GTX	3	135
Boreal Maigmo	3	150
L.L. Bean GTX Super Guide	5	189
Lowa Kibo	5	190
Asolo AFX 520 GTX	4	195
Raichle Mt. Trail GTX	4	200
Scarpa Delta SL M3	5	220

a. Use these data to develop an estimated regression equation to estimate the price of a day hiker and backpacking boot given the upper support rating.
b. At the .05 level of significance, determine whether upper support and price are related.
c. Would you feel comfortable using the estimated regression equation developed in part (a) to estimate the price for a day hiker or backpacking boot given the upper support rating?
d. Estimate the price for a day hiker with an upper support rating of 4.

WEB file
SleepingBags

28. In exercise 8, data on x = temperature rating (F°) and y = price ($) for 11 sleeping bags manufactured by Bergans of Norway provided the estimated regression equation $\hat{y} = 359.2668 - 5.2772x$. At the .05 level of significance, test whether temperature rating and price are related. Show the ANOVA table. What is your conclusion?

29. Refer to exercise 21, where data on production volume and cost were used to develop an estimated regression equation relating production volume and cost for a particular manufacturing operation. Use $\alpha = .05$ to test whether the production volume is significantly related to the total cost. Show the ANOVA table. What is your conclusion?

30. Refer to excercise 5 where the following data were used to investigate whether higher prices are generally associated with higher ratings for elliptical trainers (*Consumer Reports*, February 2008).

WEB file
Ellipticals

Brand and Model	Price ($)	Rating
Precor 5.31	3700	87
Keys Fitness CG2	2500	84
Octane Fitness Q37e	2800	82
LifeFitness X1 Basic	1900	74
NordicTrack AudioStrider 990	1000	73
Schwinn 430	800	69
Vision Fitness X6100	1700	68
ProForm XP 520 Razor	600	55

With x = price ($) and y = rating, the estimated regression equation is $\hat{y} = 58.158 + .008449x$. For these data, SSE = 173.88 and SST = 756. Does the evidence indicate a significant relationship between price and rating?

31. In exercise 20, data on x = price ($) and y = overall score for ten 42-inch plasma televisions tested by *Consumer Reports* provided the estimated regression equation $\hat{y} = 12.0169 + .0127x$. For these data SSE = 540.04 and SST = 982.40. Use the F test to determine whether the price for a 42-inch plasma television and the overall score are related at the .05 level of significance.

14.6 Using the Estimated Regression Equation for Estimation and Prediction

When using the simple linear regression model we are making an assumption about the relationship between x and y. We then use the least squares method to obtain the estimated simple linear regression equation. If a significant relationship exists between x and y, and the coefficient of determination shows that the fit is good, the estimated regression equation should be useful for estimation and prediction.

Point Estimation

In the Armand's Pizza Parlors example, the estimated regression equation $\hat{y} = 60 + 5x$ provides an estimate of the relationship between the size of the student population x and quarterly sales y. We can use the estimated regression equation to develop a point estimate of the mean value of y for a particular value of x or to predict an individual value of y corresponding to a given value of x. For instance, suppose Armand's managers want a point estimate of the mean quarterly sales for all restaurants located near college campuses with 10,000 students. Using the estimated regression equation $\hat{y} = 60 + 5x$, we see that for $x = 10$ (or 10,000 students), $\hat{y} = 60 + 5(10) = 110$. Thus, a point estimate of the mean quarterly sales for all restaurants located near campuses with 10,000 students is $110,000.

Now suppose Armand's managers want to predict sales for an individual restaurant located near Talbot College, a school with 10,000 students. In this case we are not interested in the mean value for all restaurants located near campuses with 10,000 students; we are just interested in predicting quarterly sales for one individual restaurant. As it turns out, the point estimate for an individual value of y is the same as the point estimate for the mean value of y. Hence, we would predict quarterly sales of $\hat{y} = 60 + 5(10) = 110$ or $110,000 for this one restaurant.

Interval Estimation

Confidence intervals and prediction intervals show the precision of the regression results. Narrower intervals provide a higher degree of precision.

Point estimates do not provide any information about the precision associated with an estimate. For that we must develop interval estimates much like those in Chapters 8, 10, and 11. The first type of interval estimate, a **confidence interval**, is an interval estimate of the *mean value of y* for a given value of x. The second type of interval estimate, a **prediction interval**, is used whenever we want an interval estimate of an *individual value of y* for a given value of x. The point estimate of the mean value of y is the same as the point estimate of an individual value of y. But the interval estimates we obtain for the two cases are different. The margin of error is larger for a prediction interval.

Confidence Interval for the Mean Value of y

The estimated regression equation provides a point estimate of the mean value of y for a given value of x. In developing the confidence interval, we will use the following notation.

x_p = the particular or given value of the independent variable x
y_p = the value of the dependent variable y corresponding to the given x_p
$E(y_p)$ = the mean or expected value of the dependent variable y corresponding to the given x_p
$\hat{y}_p = b_0 + b_1 x_p$ = the point estimate of $E(y_p)$ when $x = x_p$

Using this notation to estimate the mean sales for all Armand's restaurants located near a campus with 10,000 students, we have $x_p = 10$, and $E(y_p)$ denotes the unknown mean value of sales for all restaurants where $x_p = 10$. The point estimate of $E(y_p)$ is provided by $\hat{y}_p = 60 + 5(10) = 110$.

In general, we cannot expect $\hat{y}_p$ to equal $E(y_p)$ exactly. If we want to make an inference about how close $\hat{y}_p$ is to the true mean value $E(y_p)$, we will have to estimate the variance of $\hat{y}_p$. The formula for estimating the variance of $\hat{y}_p$ given x_p, denoted by $s^2_{\hat{y}_p}$, is

$$s^2_{\hat{y}_p} = s^2\left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}\right] \tag{14.22}$$

The estimate of the standard deviation of $\hat{y}_p$ is given by the square root of equation (14.22).

$$s_{\hat{y}_p} = s\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}} \tag{14.23}$$

The computational results for Armand's Pizza Parlors in Section 14.5 provided $s = 13.829$. With $x_p = 10$, $\bar{x} = 14$, and $\Sigma(x_i - \bar{x})^2 = 568$, we can use equation (14.23) to obtain

$$s_{\hat{y}_p} = 13.829\sqrt{\frac{1}{10} + \frac{(10 - 14)^2}{568}}$$
$$= 13.829\sqrt{.1282} = 4.95$$

The general expression for a confidence interval follows.

CONFIDENCE INTERVAL FOR $E(y_p)$

$$\hat{y}_p \pm t_{\alpha/2} s_{\hat{y}_p} \tag{14.24}$$

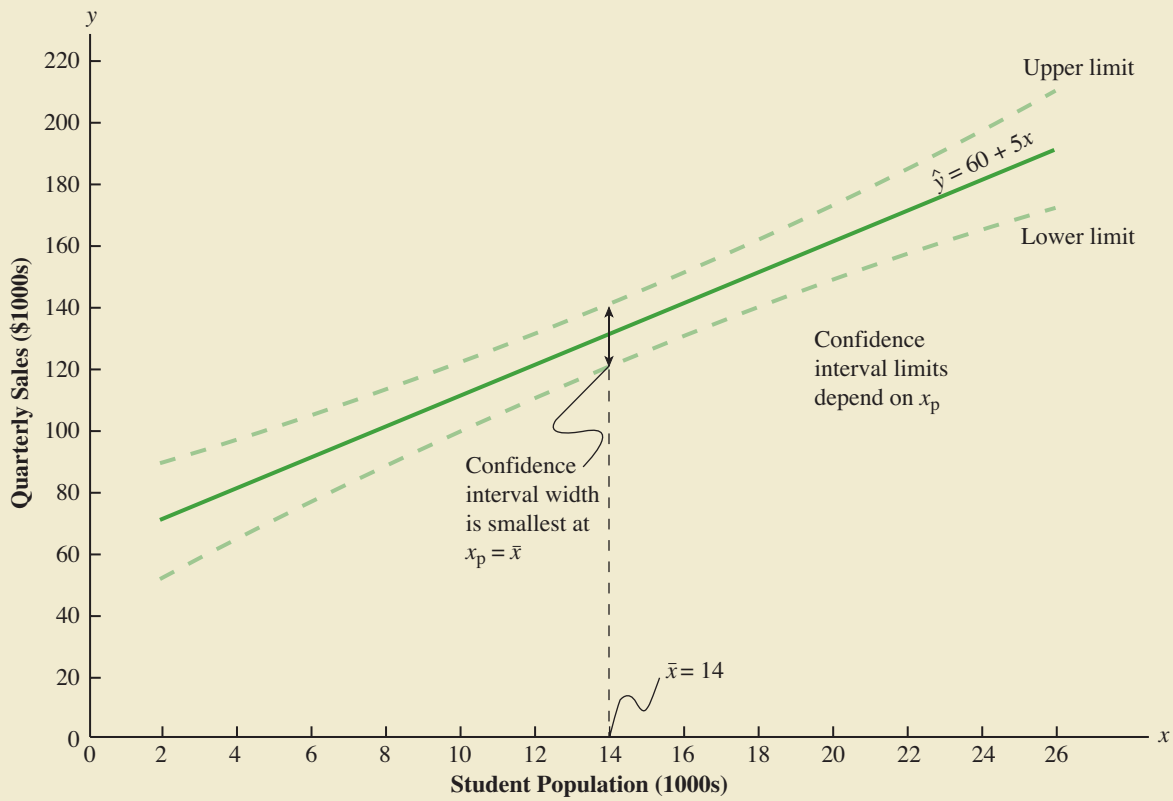
where the confidence coefficient is $1 - \alpha$ and $t_{\alpha/2}$ is based on a t distribution with $n - 2$ degrees of freedom.

The margin of error associated with this internal estimate is $t_{\alpha/2} s_{\hat{y}_p}$.

Using expression (14.24) to develop a 95% confidence interval of the mean quarterly sales for all Armand's restaurants located near campuses with 10,000 students, we need the value of t for $\alpha/2 = .025$ and $n - 2 = 10 - 2 = 8$ degrees of freedom. Using Table 2 of Appendix B, we have $t_{.025} = 2.306$. Thus, with $\hat{y}_p = 110$ and a margin of error of $t_{\alpha/2} s_{\hat{y}_p} = 2.306(4.95) = 11.415$, the 95% confidence interval estimate is

$$110 \pm 11.415$$

FIGURE 14.8 CONFIDENCE INTERVALS FOR THE MEAN SALES y AT GIVEN VALUES OF STUDENT POPULATION x



In dollars, the 95% confidence interval for the mean quarterly sales of all restaurants near campuses with 10,000 students is \$110,000 ± \$11,415. Therefore, the 95% confidence interval for the mean quarterly sales when the student population is 10,000 is \$98,585 to \$121,415.

Note that the estimated standard deviation of $\hat{y}_p$ given by equation (14.23) is smallest when $x_p = \bar{x}$ and the quantity $x_p - \bar{x} = 0$. In this case, the estimated standard deviation of $\hat{y}_p$ becomes

$$s_{\hat{y}_p} = s\sqrt{\frac{1}{n} + \frac{(\bar{x} - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}} = s\sqrt{\frac{1}{n}}$$

This result implies that we can make the best or most precise estimate of the mean value of y whenever $x_p = \bar{x}$. In fact, the further x_p is from $\bar{x}$ the larger $x_p - \bar{x}$ becomes. As a result, confidence intervals for the mean value of y will become wider as x_p deviates more from $\bar{x}$. This pattern is shown graphically in Figure 14.8.

Prediction Interval for an Individual Value of y

Suppose that instead of estimating the mean value of sales for all Armand's restaurants located near campuses with 10,000 students, we want to estimate the sales for an individual restaurant located near Talbot College, a school with 10,000 students. As noted previously,

the point estimate of y_p, the value of y corresponding to the given x_p, is provided by the estimated regression equation $\hat{y}_p = b_0 + b_1 x_p$. For the restaurant at Talbot College, we have $x_p = 10$ and a corresponding predicted quarterly sales of $\hat{y}_p = 60 + 5(10) = 110$, or \$110,000. Note that this value is the same as the point estimate of the mean sales for all restaurants located near campuses with 10,000 students.

To develop a prediction interval, we must first determine the variance associated with using $\hat{y}_p$ as an estimate of an individual value of y when $x = x_p$. This variance is made up of the sum of the following two components.

1. The variance of individual y values about the mean $E(y_p)$, an estimate of which is given by s^2
2. The variance associated with using $\hat{y}_p$ to estimate $E(y_p)$, an estimate of which is given by $s^2_{\hat{y}_p}$

The formula for estimating the variance of an individual value of y_p, denoted by s^2_{ind}, is

$$\begin{aligned} s^2_{\text{ind}} &= s^2 + s^2_{\hat{y}_p} \\ &= s^2 + s^2\left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}\right] \\ &= s^2\left[1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}\right] \end{aligned} \qquad \textbf{(14.25)}$$

Hence, an estimate of the standard deviation of an individual value of y_p is given by

$$s_{\text{ind}} = s\sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}} \qquad \textbf{(14.26)}$$

For Armand's Pizza Parlors, the estimated standard deviation corresponding to the prediction of sales for one specific restaurant located near a campus with 10,000 students is computed as follows.

$$\begin{aligned} s_{\text{ind}} &= 13.829\sqrt{1 + \frac{1}{10} + \frac{(10 - 14)^2}{568}} \\ &= 13.829\sqrt{1.1282} \\ &= 14.69 \end{aligned}$$

The general expression for a prediction interval follows.

PREDICTION INTERVAL FOR y_p

$$\hat{y}_p \pm t_{\alpha/2} s_{\text{ind}} \qquad \textbf{(14.27)}$$

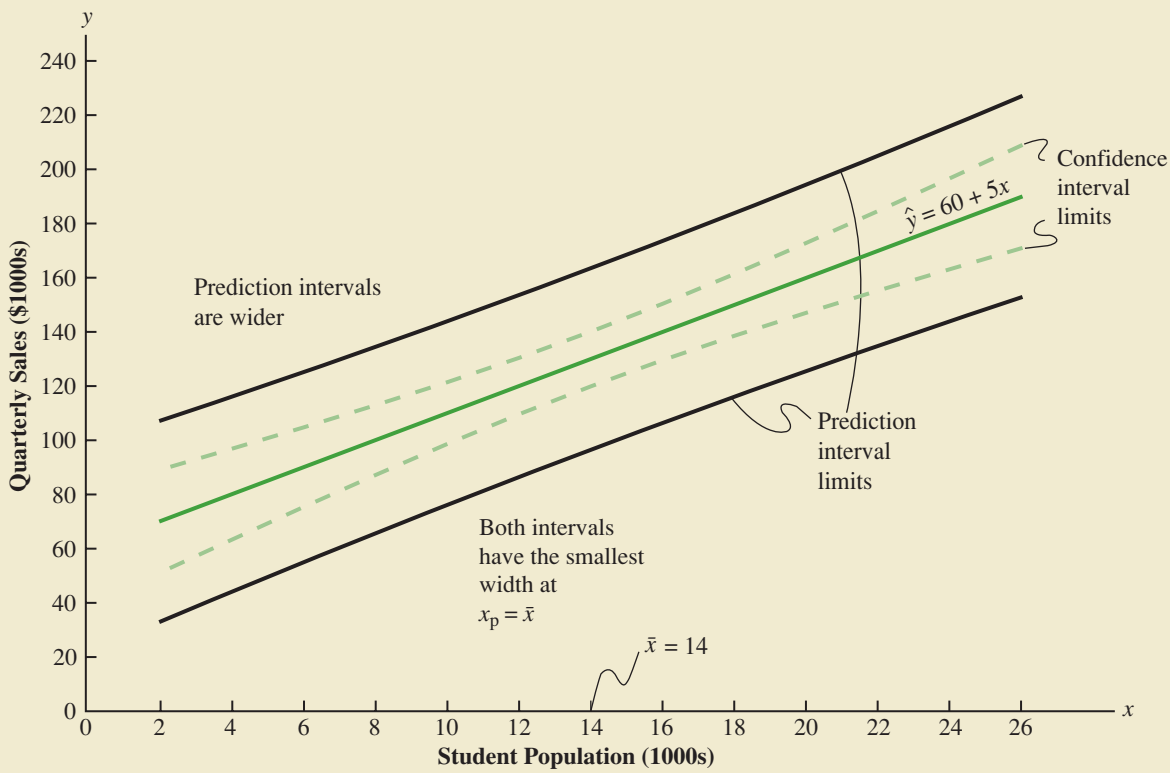
where the confidence coefficient is $1 - \alpha$ and $t_{\alpha/2}$ is based on a t distribution with $n - 2$ degrees of freedom.

The margin of error associated with this interval estimate is $t_{\alpha/2} s_{\text{ind}}$.

The 95% prediction interval for quarterly sales at Armand's Talbot College restaurant can be found by using $t_{.025} = 2.306$ and $s_{\text{ind}} = 14.69$. Thus, with $\hat{y}_p = 110$ and a margin of error of $t_{\alpha/2} s_{\text{ind}} = 2.306(14.69) = 33.875$, the 95% prediction interval is

$$110 \pm 33.875$$

FIGURE 14.9 CONFIDENCE AND PREDICTION INTERVALS FOR SALES y AT GIVEN VALUES OF STUDENT POPULATION x



In dollars, this prediction interval is \$110,000 ± \$33,875 or \$76,125 to \$143,875. Note that the prediction interval for an individual restaurant located near a campus with 10,000 students is wider than the confidence interval for the mean sales of all restaurants located near campuses with 10,000 students. The difference reflects the fact that we are able to estimate the mean value of y more precisely than we can an individual value of y.

In general, the lines for the confidence interval limits and the prediction interval limits both have curvature.

Both confidence interval estimates and prediction interval estimates are most precise when the value of the independent variable is $x_p = \bar{x}$. The general shapes of confidence intervals and the wider prediction intervals are shown together in Figure 14.9.

Exercises

Methods

SELF test

32. The data from exercise 1 follow.

x_i	1	2	3	4	5
y_i	3	7	5	11	14

a. Use equation (14.23) to estimate the standard deviation of $\hat{y}_p$ when $x = 4$.
b. Use expression (14.24) to develop a 95% confidence interval for the expected value of y when $x = 4$.

c. Use equation (14.26) to estimate the standard deviation of an individual value of y when $x = 4$.
d. Use expression (14.27) to develop a 95% prediction interval for y when $x = 4$.

33. The data from exercise 2 follow.

x_i	3	12	6	20	14
y_i	55	40	55	10	15

a. Estimate the standard deviation of $\hat{y}_p$ when $x = 8$.
b. Develop a 95% confidence interval for the expected value of y when $x = 8$.
c. Estimate the standard deviation of an individual value of y when $x = 8$.
d. Develop a 95% prediction interval for y when $x = 8$.

34. The data from exercise 3 follow.

x_i	2	6	9	13	20
y_i	7	18	9	26	23

Develop the 95% confidence and prediction intervals when $x = 12$. Explain why these two intervals are different.

Applications

SELF test

35. In exercise 18, the data on grade point average x and monthly salary y provided the estimated regression equation $\hat{y} = 1790.5 + 581.1x$.
a. Develop a 95% confidence interval for the mean starting salary for all students with a 3.0 GPA.
b. Develop a 95% prediction interval for the starting salary for Joe Heller, a student with a GPA of 3.0.

WEB file
SleepingBags

36. In exercise 8, data on x = temperature rating (F°) and y = price ($) for 11 sleeping bags manufactured by Bergans of Norway provided the estimated regression equation $\hat{y} = 359.2668 - 5.2772x$. For these data $s = 37.9372$.
a. Develop a point estimate of the price for a sleeping bag with a temperature rating of 30.
b. Develop a 95% confidence interval for the mean overall temperature rating for all sleeping bags with a temperature rating of 30.
c. Suppose that Bergans developed a new model with a temperature rating of 30. Develop a 95% prediction interval for the price of this new model.
d. Discuss the differences in your answers to parts (b) and (c).

37. In exercise 13, data were given on the adjusted gross income x and the amount of itemized deductions taken by taxpayers. Data were reported in thousands of dollars. With the estimated regression equation $\hat{y} = 4.68 + .16x$, the point estimate of a reasonable level of total itemized deductions for a taxpayer with an adjusted gross income of $52,500 is $13,080.
a. Develop a 95% confidence interval for the mean amount of total itemized deductions for all taxpayers with an adjusted gross income of $52,500.
b. Develop a 95% prediction interval estimate for the amount of total itemized deductions for a particular taxpayer with an adjusted gross income of $52,500.
c. If the particular taxpayer referred to in part (b) claimed total itemized deductions of $20,400, would the IRS agent's request for an audit appear to be justified?
d. Use your answer to part (b) to give the IRS agent a guideline as to the amount of total itemized deductions a taxpayer with an adjusted gross income of $52,500 should claim before an audit is recommended.

38. Refer to Exercise 21, where data on the production volume x and total cost y for a particular manufacturing operation were used to develop the estimated regression equation $\hat{y} = 1246.67 + 7.6x$.
a. The company's production schedule shows that 500 units must be produced next month. What is the point estimate of the total cost for next month?

b. Develop a 99% prediction interval for the total cost for next month.
c. If an accounting cost report at the end of next month shows that the actual production cost during the month was $6000, should managers be concerned about incurring such a high total cost for the month? Discuss.

39. Almost all U.S. light-rail systems use electric cars that run on tracks built at street level. The Federal Transit Administration claims light-rail is one of the safest modes of travel, with an accident rate of .99 accidents per million passenger miles as compared to 2.29 for buses. The following data show the miles of track and the weekday ridership in thousands of passengers for six light-rail systems (*USA Today,* January 7, 2003).

City	Miles of Track	Ridership (1000s)
Cleveland	15	15
Denver	17	35
Portland	38	81
Sacramento	21	31
San Diego	47	75
San Jose	31	30
St. Louis	34	42

a. Use these data to develop an estimated regression equation that could be used to predict the ridership given the miles of track.
b. Did the estimated regression equation provide a good fit? Explain.
c. Develop a 95% confidence interval for the mean weekday ridership for all light-rail systems with 30 miles of track.
d. Suppose that Charlotte is considering construction of a light-rail system with 30 miles of track. Develop a 95% prediction interval for the weekday ridership for the Charlotte system. Do you think that the prediction interval you developed would be of value to Charlotte planners in anticipating the number of weekday riders for their new light-rail system? Explain.

14.7 Computer Solution

Performing the regression analysis computations without the help of a computer can be quite time consuming. In this section we discuss how the computational burden can be minimized by using a computer software package such as Minitab.

We entered Armand's student population and sales data into a Minitab worksheet. The independent variable was named Pop and the dependent variable was named Sales to assist with interpretation of the computer output. Using Minitab, we obtained the printout for Armand's Pizza Parlors shown in Figure 14.10.[2] The interpretation of this printout follows.

1. Minitab prints the estimated regression equation as Sales = 60.0 + 5.00 Pop.
2. A table is printed that shows the values of the coefficients b_0 and b_1, the standard deviation of each coefficient, the t value obtained by dividing each coefficient value by its standard deviation, and the p-value associated with the t test. Because the p-value is zero (to three decimal places), the sample results indicate that the null hypothesis (H_0: $\beta_1 = 0$) should be rejected. Alternatively, we could compare 8.62 (located in the t-ratio column) to the appropriate critical value. This procedure for the t test was described in Section 14.5.

[2]The Minitab steps necessary to generate the output are given in Appendix 14.3.

FIGURE 14.10 MINITAB OUTPUT FOR THE ARMAND'S PIZZA PARLORS PROBLEM

```
The regression equation is
Sales = 60.0 + 5.00 Pop

Predictor      Coef   SE Coef      T       p
Constant     60.000     9.226   6.50   0.000
Pop          5.0000    0.5803   8.62   0.000

S = 13.8293     R-sq = 90.3%     R-sq(adj) = 89.1%

Analysis of Variance

SOURCE           DF      SS      MS       F       p
Regression        1   14200   14200   74.25   0.000
Residual Error    8    1530     191
Total             9   15730

Predicted Values for New Observations

New
Obs      Fit   SE Fit        95% C.I.          95% P.I.
  1   110.00     4.95   (98.58, 121.42)   (76.13, 143.87)
```

Estimated regression equation

ANOVA table

Interval estimates

3. Minitab prints the standard error of the estimate, $s = 13.8293$, as well as information about the goodness of fit. Note that "R-sq = 90.3%" is the coefficient of determination expressed as a percentage. The value "R-Sq(adj) = 89.1%" is discussed in Chapter 15.
4. The ANOVA table is printed below the heading Analysis of Variance. Minitab uses the label Residual Error for the error source of variation. Note that DF is an abbreviation for degrees of freedom and that MSR is given as 14,200 and MSE as 191. The ratio of these two values provides the F value of 74.25 and the corresponding p-value of 0.000. Because the p-value is zero (to three decimal places), the relationship between Sales and Pop is judged statistically significant.
5. The 95% confidence interval estimate of the expected sales and the 95% prediction interval estimate of sales for an individual restaurant located near a campus with 10,000 students are printed below the ANOVA table. The confidence interval is (98.58, 121.42) and the prediction interval is (76.13, 143.87) as we showed in Section 14.6.

Exercises

Applications

SELF test

40. The commercial division of a real estate firm is conducting a regression analysis of the relationship between x, annual gross rents (in thousands of dollars), and y, selling price (in thousands of dollars) for apartment buildings. Data were collected on several properties recently sold and the following computer output was obtained.

```
The regression equation is
Y = 20.0 + 7.21 X

Predictor      Coef    SE Coef       T
Constant     20.000     3.2213    6.21
X             7.210     1.3626    5.29

Analysis of Variance

SOURCE            DF         SS
Regression         1    41587.3
Residual Error     7
Total              8    51984.1
```

a. How many apartment buildings were in the sample?
b. Write the estimated regression equation.
c. What is the value of s_{b_1}?
d. Use the F statistic to test the significance of the relationship at a .05 level of significance.
e. Estimate the selling price of an apartment building with gross annual rents of $50,000.

41. Following is a portion of the computer output for a regression analysis relating y = maintenance expense (dollars per month) to x = usage (hours per week) of a particular brand of computer terminal.

```
The regression equation is
Y = 6.1092 + .8951 X

Predictor       Coef     SE Coef
Constant      6.1092      0.9361
X             0.8951      0.1490

Analysis of Variance

SOURCE            DF         SS         MS
Regression         1    1575.76    1575.76
Residual Error     8     349.14      43.64
Total              9    1924.90
```

a. Write the estimated regression equation.
b. Use a t test to determine whether monthly maintenance expense is related to usage at the .05 level of significance.
c. Use the estimated regression equation to predict monthly maintenance expense for any terminal that is used 25 hours per week.

42. A regression model relating x, number of salespersons at a branch office, to y, annual sales at the office (in thousands of dollars) provided the following computer output from a regression analysis of the data.

```
The regression equation is
Y = 80.0 + 50.00 X

Predictor       Coef    SE Coef       T
Constant        80.0     11.333    7.06
X               50.0      5.482    9.12

Analysis of Variance

SOURCE             DF         SS         MS
Regression          1     6828.6     6828.6
Residual Error     28     2298.8       82.1
Total              29     9127.4
```

a. Write the estimated regression equation.
b. How many branch offices were involved in the study?
c. Compute the F statistic and test the significance of the relationship at a .05 level of significance.
d. Predict the annual sales at the Memphis branch office. This branch employs 12 salespersons.

43. Health experts recommend that runners drink 4 ounces of water every 15 minutes they run. Although handheld bottles work well for many types of runs, all-day cross-country runs require hip-mounted or over-the-shoulder hydration systems. In addition to carrying more water, hip-mounted or over-the-shoulder hydration systems offer more storage space for food and extra clothing. As the capacity increases, however, the weight and cost of these larger-capacity systems also increase. The following data show the weight (ounces) and the price for 26 hip-mounted or over-the-shoulder hydration systems (*Trail Runner Gear Guide,* 2003).

WEB file
Hydration1

Model	Weight (oz.)	Price ($)
Fastdraw	3	10
Fastdraw Plus	4	12
Fitness	5	12
Access	7	20
Access Plus	8	25
Solo	9	25
Serenade	9	35
Solitaire	11	35
Gemini	21	45
Shadow	15	40
SipStream	18	60
Express	9	30
Lightning	12	40
Elite	14	60
Extender	16	65
Stinger	16	65
GelFlask Belt	3	20
GelDraw	1	7
GelFlask Clip-on Holster	2	10
GelFlask Holster SS	1	10
Strider (W)	8	30

Model	Weight (oz.)	Price ($)
Walkabout (W)	14	40
Solitude I.C.E.	9	35
Getaway I.C.E.	19	55
Profile I.C.E.	14	50
Traverse I.C.E.	13	60

a. Use these data to develop an estimated regression equation that could be used to predict the price of a hydration system given its weight.
b. Test the significance of the relationship at the .05 level of significance.
c. Did the estimated regression equation provide a good fit? Explain.
d. Assume that the estimated regression equation developed in part (a) will also apply to hydration systems produced by other companies. Develop a 95% confidence interval estimate of the price for all hydration systems that weigh 10 ounces.
e. Assume that the estimated regression equation developed in part (a) will also apply to hydration systems produced by other companies. Develop a 95% prediction interval estimate of the price for the Back Draft system produced by Eastern Mountain Sports. The Back Draft system weighs 10 ounces.

44. Automobile racing, high-performance driving schools, and driver education programs run by automobile clubs continue to grow in popularity. All these activities require the participant to wear a helmet that is certified by the Snell Memorial Foundation, a not-for-profit organization dedicated to research, education, testing, and development of helmet safety standards. Snell "SA" (Sports Application) rated professional helmets are designed for auto racing and provide extreme impact resistance and high fire protection. One of the key factors in selecting a helmet is weight, since lower weight helmets tend to place less stress on the neck. The following data show the weight and price for 18 SA helmets (SoloRacer website, April 20, 2008).

WEB file
RaceHelmets

Helmet	Weight (oz)	Price ($)
Pyrotect Pro Airflow	64	248
Pyrotect Pro Airflow Graphics	64	278
RCi Full Face	64	200
RaceQuip RidgeLine	64	200
HJC AR-10	58	300
HJC Si-12	47	700
HJC HX-10	49	900
Impact Racing Super Sport	59	340
Zamp FSA-1	66	199
Zamp RZ-2	58	299
Zamp RZ-2 Ferrari	58	299
Zamp RZ-3 Sport	52	479
Zamp RZ-3 Sport Painted	52	479
Bell M2	63	369
Bell M4	62	369
Bell M4 Pro	54	559
G Force Pro Force 1	63	250
G Force Pro Force 1 Grafx	63	280

a. Develop a scatter diagram with weight as the independent variable.
b. Does there appear to be any relationship between these two variables?

c. Develop the estimated regression equation that could be used to predict the price given the weight.
d. Test for the significance of the relationship at the .05 level of significance.
e. Did the estimated regression equation provide a good fit? Explain.

14.8 Residual Analysis: Validating Model Assumptions

Residual analysis *is the primary tool for determining whether the assumed regression model is appropriate.*

As we noted previously, the *residual* for observation i is the difference between the observed value of the dependent variable (y_i) and the estimated value of the dependent variable ($\hat{y}_i$).

RESIDUAL FOR OBSERVATION i

$$y_i - \hat{y}_i \tag{14.28}$$

where

y_i is the observed value of the dependent variable
$\hat{y}_i$ is the estimated value of the dependent variable

In other words, the ith residual is the error resulting from using the estimated regression equation to predict the value of the dependent variable. The residuals for the Armand's Pizza Parlors example are computed in Table 14.7. The observed values of the dependent variable are in the second column and the estimated values of the dependent variable, obtained using the estimated regression equation $\hat{y} = 60 + 5x$, are in the third column. An analysis of the corresponding residuals in the fourth column will help determine whether the assumptions made about the regression model are appropriate.

Let us now review the regression assumptions for the Armand's Pizza Parlors example. A simple linear regression model was assumed.

$$y = \beta_0 + \beta_1 x + \epsilon \tag{14.29}$$

TABLE 14.7 RESIDUALS FOR ARMAND'S PIZZA PARLORS

Student Population x_i	Sales y_i	Estimated Sales $\hat{y}_i = 60 + 5x_i$	Residuals $y_i - \hat{y}_i$
2	58	70	−12
6	105	90	15
8	88	100	−12
8	118	100	18
12	117	120	−3
16	137	140	−3
20	157	160	−3
20	169	160	9
22	149	170	−21
26	202	190	12

This model indicates that we assumed quarterly sales (y) to be a linear function of the size of the student population (x) plus an error term ϵ. In Section 14.4 we made the following assumptions about the error term ϵ.

1. $E(\epsilon) = 0$.
2. The variance of ϵ, denoted by σ^2, is the same for all values of x.
3. The values of ϵ are independent.
4. The error term ϵ has a normal distribution.

These assumptions provide the theoretical basis for the t test and the F test used to determine whether the relationship between x and y is significant, and for the confidence and prediction interval estimates presented in Section 14.6. If the assumptions about the error term ϵ appear questionable, the hypothesis tests about the significance of the regression relationship and the interval estimation results may not be valid.

The residuals provide the best information about ϵ; hence an analysis of the residuals is an important step in determining whether the assumptions for ϵ are appropriate. Much of residual analysis is based on an examination of graphical plots. In this section, we discuss the following residual plots.

1. A plot of the residuals against values of the independent variable x
2. A plot of residuals against the predicted values of the dependent variable $\hat{y}$
3. A standardized residual plot
4. A normal probability plot

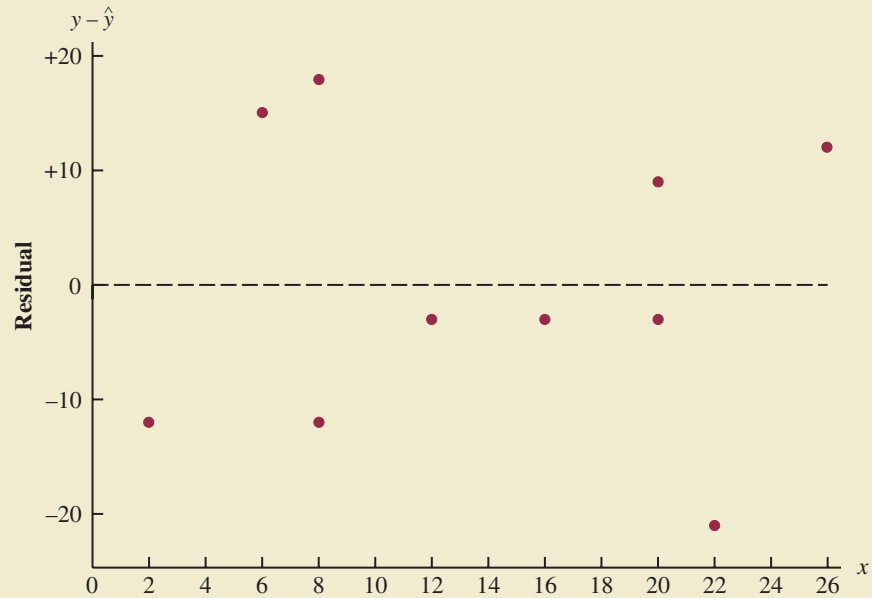
Residual Plot Against x

A **residual plot** against the independent variable x is a graph in which the values of the independent variable are represented by the horizontal axis and the corresponding residual values are represented by the vertical axis. A point is plotted for each residual. The first coordinate for each point is given by the value of x_i and the second coordinate is given by the corresponding value of the residual $y_i - \hat{y}_i$. For a residual plot against x with the Armand's Pizza Parlors data from Table 14.7, the coordinates of the first point are $(2, -12)$, corresponding to $x_1 = 2$ and $y_1 - \hat{y}_1 = -12$; the coordinates of the second point are (6, 15), corresponding to $x_2 = 6$ and $y_2 - \hat{y}_2 = 15$; and so on. Figure 14.11 shows the resulting residual plot.

Before interpreting the results for this residual plot, let us consider some general patterns that might be observed in any residual plot. Three examples appear in Figure 14.12. If the assumption that the variance of ϵ is the same for all values of x and the assumed regression model is an adequate representation of the relationship between the variables, the residual plot should give an overall impression of a horizontal band of points such as the one in Panel A of Figure 14.12. However, if the variance of ϵ is not the same for all values of x—for example, if variability about the regression line is greater for larger values of x—a pattern such as the one in Panel B of Figure 14.12 could be observed. In this case, the assumption of a constant variance of ϵ is violated. Another possible residual plot is shown in Panel C. In this case, we would conclude that the assumed regression model is not an adequate representation of the relationship between the variables. A curvilinear regression model or multiple regression model should be considered.

Now let us return to the residual plot for Armand's Pizza Parlors shown in Figure 14.11. The residuals appear to approximate the horizontal pattern in Panel A of Figure 14.12. Hence, we conclude that the residual plot does not provide evidence that the assumptions made for Armand's regression model should be challenged. At this point, we are confident in the conclusion that Armand's simple linear regression model is valid.

FIGURE 14.11 PLOT OF THE RESIDUALS AGAINST THE INDEPENDENT VARIABLE x FOR ARMAND'S PIZZA PARLORS



Experience and good judgment are always factors in the effective interpretation of residual plots. Seldom does a residual plot conform precisely to one of the patterns in Figure 14.12. Yet analysts who frequently conduct regression studies and frequently review residual plots become adept at understanding the differences between patterns that are reasonable and patterns that indicate the assumptions of the model should be questioned. A residual plot provides one technique to assess the validity of the assumptions for a regression model.

Residual Plot Against $\hat{y}$

Another residual plot represents the predicted value of the dependent variable $\hat{y}$ on the horizontal axis and the residual values on the vertical axis. A point is plotted for each residual. The first coordinate for each point is given by $\hat{y}_i$ and the second coordinate is given by the corresponding value of the ith residual $y_i - \hat{y}_i$. With the Armand's data from Table 14.7, the coordinates of the first point are (70, −12), corresponding to $\hat{y}_1 = 70$ and $y_1 - \hat{y}_1 = -12$; the coordinates of the second point are (90, 15); and so on. Figure 14.13 provides the residual plot. Note that the pattern of this residual plot is the same as the pattern of the residual plot against the independent variable x. It is not a pattern that would lead us to question the model assumptions. For simple linear regression, both the residual plot against x and the residual plot against $\hat{y}$ provide the same pattern. For multiple regression analysis, the residual plot against $\hat{y}$ is more widely used because of the presence of more than one independent variable.

Standardized Residuals

Many of the residual plots provided by computer software packages use a standardized version of the residuals. As demonstrated in preceding chapters, a random variable is standardized by subtracting its mean and dividing the result by its standard deviation. With the

FIGURE 14.12 RESIDUAL PLOTS FROM THREE REGRESSION STUDIES

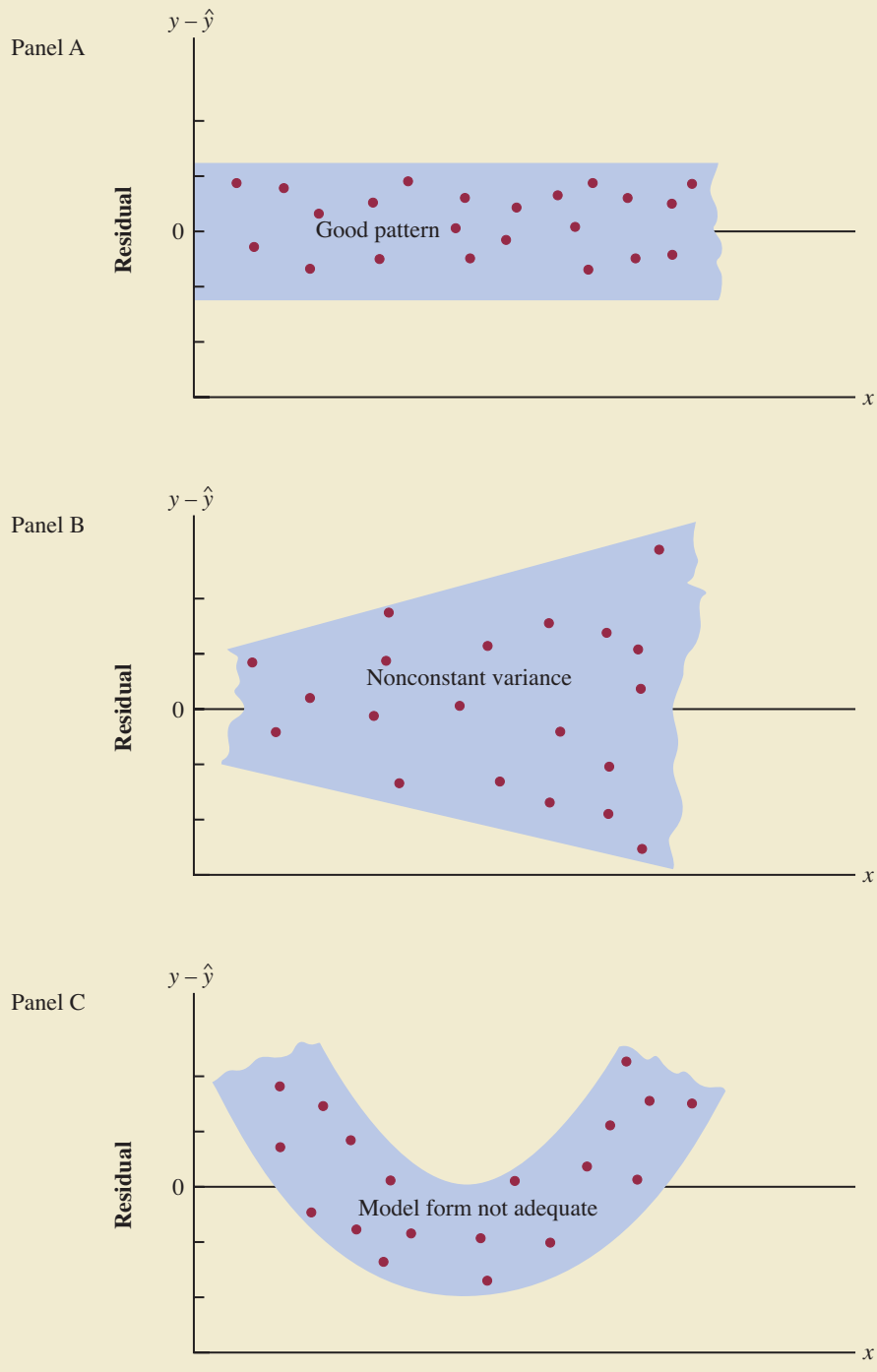
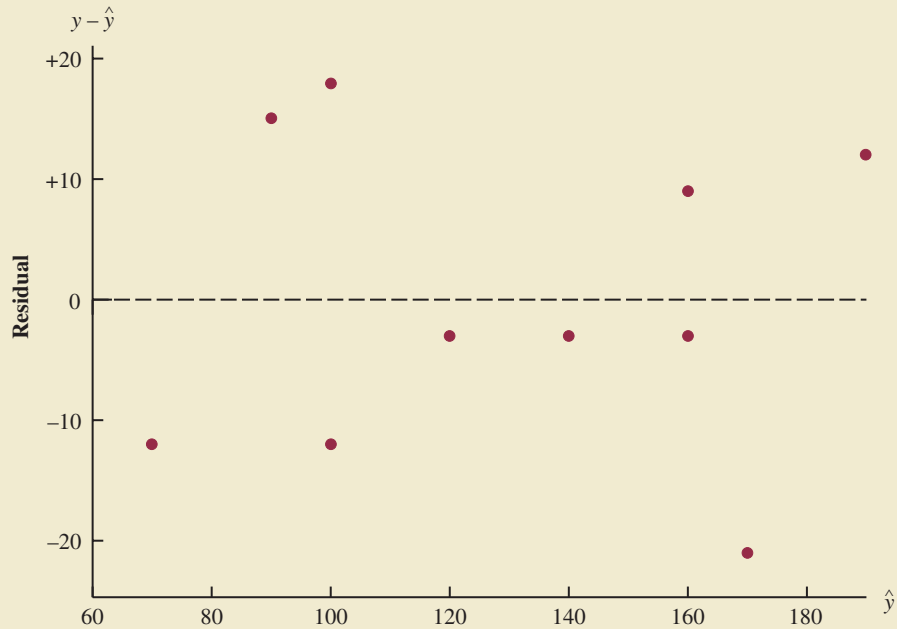


FIGURE 14.13 PLOT OF THE RESIDUALS AGAINST THE PREDICTED VALUES $\hat{y}$ FOR ARMAND'S PIZZA PARLORS



least squares method, the mean of the residuals is zero. Thus, simply dividing each residual by its standard deviation provides the **standardized residual**.

It can be shown that the standard deviation of residual i depends on the standard error of the estimate s and the corresponding value of the independent variable x_i.

STANDARD DEVIATION OF THE ith RESIDUAL[3]

$$s_{y_i - \hat{y}_i} = s\sqrt{1 - h_i} \tag{14.30}$$

where

$s_{y_i - \hat{y}_i}$ = the standard deviation of residual i
s = the standard error of the estimate

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\Sigma(x_i - \bar{x})^2} \tag{14.31}$$

Note that equation (14.30) shows that the standard deviation of the ith residual depends on x_i because of the presence of h_i in the formula.[4] Once the standard deviation of each residual is calculated, we can compute the standardized residual by dividing each residual by its corresponding standard deviation.

[3]This equation actually provides an estimate of the standard deviation of the ith residual, because s is used instead of σ.
[4]h_i is referred to as the *leverage* of observation i. Leverage will be discussed further when we consider influential observations in Section 14.9.

TABLE 14.8 COMPUTATION OF STANDARDIZED RESIDUALS FOR ARMAND'S PIZZA PARLORS

Restaurant i	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$\frac{(x_i - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}$	h_i	$s_{y_i - \hat{y}_i}$	$y_i - \hat{y}_i$	Standardized Residual
1	2	−12	144	.2535	.3535	11.1193	−12	−1.0792
2	6	−8	64	.1127	.2127	12.2709	15	1.2224
3	8	−6	36	.0634	.1634	12.6493	−12	−.9487
4	8	−6	36	.0634	.1634	12.6493	18	1.4230
5	12	−2	4	.0070	.1070	13.0682	−3	−.2296
6	16	2	4	.0070	.1070	13.0682	−3	−.2296
7	20	6	36	.0634	.1634	12.6493	−3	−.2372
8	20	6	36	.0634	.1634	12.6493	9	.7115
9	22	8	64	.1127	.2127	12.2709	−21	−1.7114
10	26	12	144	.2535	.3535	11.1193	12	1.0792
		Total	568					

Note: The values of the residuals were computed in Table 14.7.

STANDARDIZED RESIDUAL FOR OBSERVATION i

$$\frac{y_i - \hat{y}_i}{s_{y_i - \hat{y}_i}} \qquad (14.32)$$

Table 14.8 shows the calculation of the standardized residuals for Armand's Pizza Parlors. Recall that previous calculations showed $s = 13.829$. Figure 14.14 is the plot of the standardized residuals against the independent variable x.

Small departures from normality do not have a great effect on the statistical tests used in regression analysis.

The standardized residual plot can provide insight about the assumption that the error term ϵ has a normal distribution. If this assumption is satisfied, the distribution of the standardized residuals should appear to come from a standard normal probability distribution.[5] Thus, when looking at a standardized residual plot, we should expect to see approximately 95% of the standardized residuals between −2 and +2. We see in Figure 14.14 that for the Armand's example all standardized residuals are between −2 and +2. Therefore, on the basis of the standardized residuals, this plot gives us no reason to question the assumption that ϵ has a normal distribution.

Because of the effort required to compute the estimated values of $\hat{y}$, the residuals, and the standardized residuals, most statistical packages provide these values as optional regression output. Hence, residual plots can be easily obtained. For large problems computer packages are the only practical means for developing the residual plots discussed in this section.

Normal Probability Plot

Another approach for determining the validity of the assumption that the error term has a normal distribution is the **normal probability plot**. To show how a normal probability plot is developed, we introduce the concept of *normal scores*.

Suppose 10 values are selected randomly from a normal probability distribution with a mean of zero and a standard deviation of one, and that the sampling process is repeated over and over with the values in each sample of 10 ordered from smallest to largest. For now, let

[5] Because s is used instead of σ in equation (14.30), the probability distribution of the standardized residuals is not technically normal. However, in most regression studies, the sample size is large enough that a normal approximation is very good.

FIGURE 14.14 PLOT OF THE STANDARDIZED RESIDUALS AGAINST THE INDEPENDENT VARIABLE x FOR ARMAND'S PIZZA PARLORS

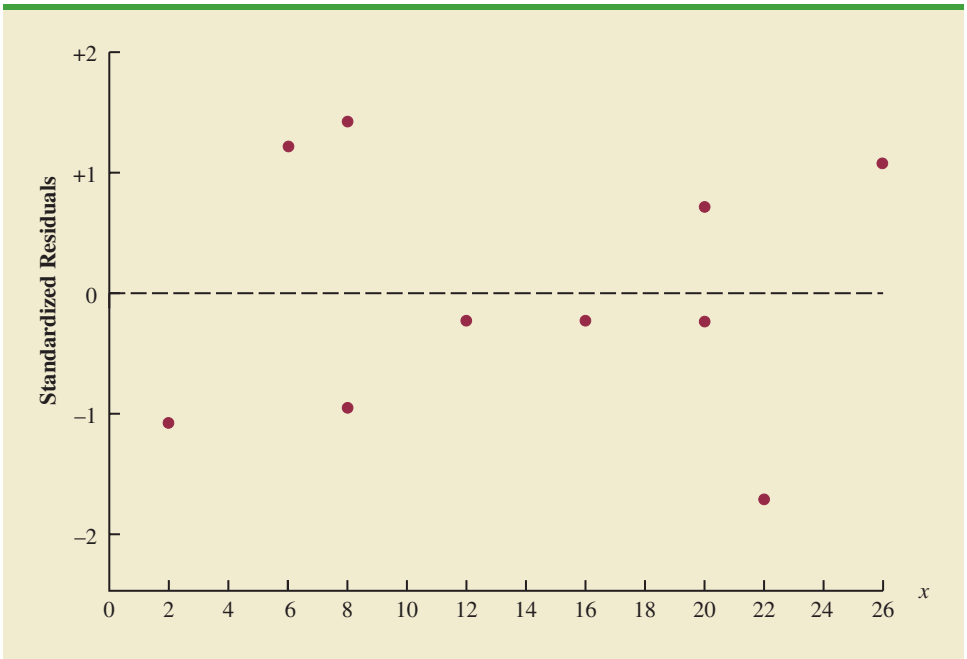


TABLE 14.9

NORMAL SCORES FOR $n = 10$

Order Statistic	Normal Score
1	−1.55
2	−1.00
3	−.65
4	−.37
5	−.12
6	.12
7	.37
8	.65
9	1.00
10	1.55

TABLE 14.10

NORMAL SCORES AND ORDERED STANDARDIZED RESIDUALS FOR ARMAND'S PIZZA PARLORS

Normal Scores	Ordered Standardized Residuals
−1.55	−1.7114
−1.00	−1.0792
−.65	−.9487
−.37	−.2372
−.12	−.2296
.12	−.2296
.37	.7115
.65	1.0792
1.00	1.2224
1.55	1.4230

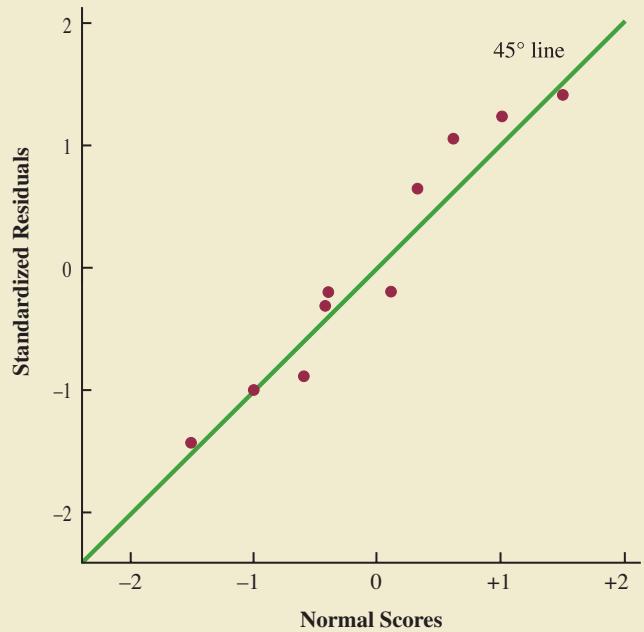
us consider only the smallest value in each sample. The random variable representing the smallest value obtained in repeated sampling is called the first-order statistic.

Statisticians show that for samples of size 10 from a standard normal probability distribution, the expected value of the first-order statistic is −1.55. This expected value is called a normal score. For the case with a sample of size $n = 10$, there are 10 order statistics and 10 normal scores (see Table 14.9). In general, a data set consisting of n observations will have n order statistics and hence n normal scores.

Let us now show how the 10 normal scores can be used to determine whether the standardized residuals for Armand's Pizza Parlors appear to come from a standard normal probability distribution. We begin by ordering the 10 standardized residuals from Table 14.8. The 10 normal scores and the ordered standardized residuals are shown together in Table 14.10. If the normality assumption is satisfied, the smallest standardized residual should be close to the smallest normal score, the next smallest standardized residual should be close to the next smallest normal score, and so on. If we were to develop a plot with the normal scores on the horizontal axis and the corresponding standardized residuals on the vertical axis, the plotted points should cluster closely around a 45-degree line passing through the origin if the standardized residuals are approximately normally distributed. Such a plot is referred to as a *normal probability plot.*

Figure 14.15 is the normal probability plot for the Armand's Pizza Parlors example. Judgment is used to determine whether the pattern observed deviates from the line enough to conclude that the standardized residuals are not from a standard normal probability distribution. In Figure 14.15, we see that the points are grouped closely about the line. We therefore conclude that the assumption of the error term having a normal probability distribution is reasonable. In general, the more closely the points are clustered about the 45-degree line, the stronger the evidence supporting the normality assumption. Any substantial curvature in the normal probability plot is evidence that the residuals have not come from a normal distribution. Normal scores and the associated normal probability plot can be obtained easily from statistical packages such as Minitab.

FIGURE 14.15 NORMAL PROBABILITY PLOT FOR ARMAND'S PIZZA PARLORS



NOTES AND COMMENTS

1. We use residual and normal probability plots to validate the assumptions of a regression model. If our review indicates that one or more assumptions are questionable, a different regression model or a transformation of the data should be considered. The appropriate corrective action when the assumptions are violated must be based on good judgment; recommendations from an experienced statistician can be valuable.
2. Analysis of residuals is the primary method statisticians use to verify that the assumptions associated with a regression model are valid. Even if no violations are found, it does not necessarily follow that the model will yield good predictions. However, if additional statistical tests support the conclusion of significance and the coefficient of determination is large, we should be able to develop good estimates and predictions using the estimated regression equation.

Exercises

Methods

SELF test

45. Given are data for two variables, x and y.

x_i	6	11	15	18	20
y_i	6	8	12	20	30

a. Develop an estimated regression equation for these data.
b. Compute the residuals.

c. Develop a plot of the residuals against the independent variable x. Do the assumptions about the error terms seem to be satisfied?
d. Compute the standardized residuals.
e. Develop a plot of the standardized residuals against $\hat{y}$. What conclusions can you draw from this plot?

46. The following data were used in a regression study.

Observation	x_i	y_i	Observation	x_i	y_i
1	2	4	6	7	6
2	3	5	7	7	9
3	4	4	8	8	5
4	5	6	9	9	11
5	7	4			

a. Develop an estimated regression equation for these data.
b. Construct a plot of the residuals. Do the assumptions about the error term seem to be satisfied?

Applications

SELF test

47. Data on advertising expenditures and revenue (in thousands of dollars) for the Four Seasons Restaurant follow.

Advertising Expenditures	Revenue
1	19
2	32
4	44
6	40
10	52
14	53
20	54

a. Let x equal advertising expenditures and y equal revenue. Use the method of least squares to develop a straight line approximation of the relationship between the two variables.
b. Test whether revenue and advertising expenditures are related at a .05 level of significance.
c. Prepare a residual plot of $y - \hat{y}$ versus $\hat{y}$. Use the result from part (a) to obtain the values of $\hat{y}$.
d. What conclusions can you draw from residual analysis? Should this model be used, or should we look for a better one?

48. Refer to exercise 7, where an estimated regression equation relating years of experience and annual sales was developed.
a. Compute the residuals and construct a residual plot for this problem.
b. Do the assumptions about the error terms seem reasonable in light of the residual plot?

49. Recent family home sales in San Antonio provided the following data (San Antonio Realty Watch website, November 2008).

WEB file
HomePrices

Square Footage	Price ($)
1580	142,500
1572	145,000
1352	115,000
2224	155,900
1556	95,000
1435	128,000
1438	100,000
1089	55,000
1941	142,000
1698	115,000
1539	115,000
1364	105,000
1979	155,000
2183	132,000
2096	140,000
1400	85,000
2372	145,000
1752	155,000
1386	80,000
1163	100,000

a. Develop the estimated regression equation that can be used to predict the sales prices given the square footage.
b. Construct a residual plot of the standardized residuals against the independent variable.
c. Do the assumptions about the error term and model form seem reasonable in light of the residual plot?

14.9 Residual Analysis: Outliers and Influential Observations

In Section 14.8 we showed how residual analysis could be used to determine when violations of assumptions about the regression model occur. In this section, we discuss how residual analysis can be used to identify observations that can be classified as outliers or as being especially influential in determining the estimated regression equation. Some steps that should be taken when such observations occur are discussed.

Detecting Outliers

Figure 14.16 is a scatter diagram for a data set that contains an **outlier**, a data point (observation) that does not fit the trend shown by the remaining data. Outliers represent observations that are suspect and warrant careful examination. They may represent erroneous data; if so, the data should be corrected. They may signal a violation of model assumptions; if so, another model should be considered. Finally, they may simply be unusual values that occurred by chance. In this case, they should be retained.

To illustrate the process of detecting outliers, consider the data set in Table 14.11; Figure 14.17 is a scatter diagram. Except for observation 4 ($x_4 = 3$, $y_4 = 75$), a pattern suggesting a negative linear relationship is apparent. Indeed, given the pattern of the rest of the data, we would expect y_4 to be much smaller and hence would identify the corresponding observation as an outlier. For the case of simple linear regression, one can often detect outliers by simply examining the scatter diagram.

The standardized residuals can also be used to identify outliers. If an observation deviates greatly from the pattern of the rest of the data (e.g., the outlier in Figure 14.16), the corresponding standardized residual will be large in absolute value. Many computer packages

FIGURE 14.16 DATA SET WITH AN OUTLIER

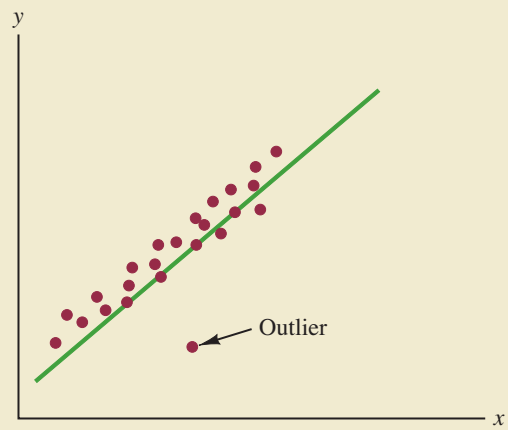


TABLE 14.11
DATA SET ILLUSTRATING THE EFFECT OF AN OUTLIER

x_i	y_i
1	45
1	55
2	50
3	75
3	40
3	45
4	30
4	35
5	25
6	15

automatically identify observations with standardized residuals that are large in absolute value. In Figure 14.18 we show the Minitab output from a regression analysis of the data in Table 14.11. The next to last line of the output shows that the standardized residual for observation 4 is 2.67. Minitab provides a list of each observation with a standardized residual of less than -2 or greater than $+2$ in the Unusual Observation section of the output; in such cases, the observation is printed on a separate line with an R next to the standardized residual, as shown in Figure 14.18. With normally distributed errors, standardized residuals should be outside these limits approximately 5% of the time.

In deciding how to handle an outlier, we should first check to see whether it is a valid observation. Perhaps an error was made in initially recording the data or in entering the data into the computer file. For example, suppose that in checking the data for the outlier in Table 14.17, we find an error; the correct value for observation 4 is $x_4 = 3$, $y_4 = 30$. Figure 14.19 is the Minitab output obtained after correction of the value of y_4. We see that

FIGURE 14.17 SCATTER DIAGRAM FOR OUTLIER DATA SET

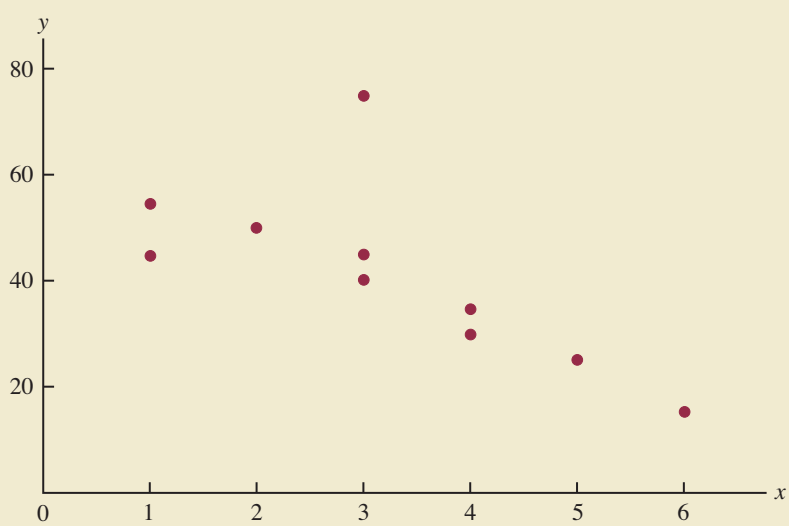


FIGURE 14.18 MINITAB OUTPUT FOR REGRESSION ANALYSIS OF THE OUTLIER DATA SET

```
The regression equation is
y = 65.0 - 7.33 x

Predictor    Coef  SE Coef      T      p
Constant   64.958    9.258   7.02  0.000
X          -7.331    2.608  -2.81  0.023

S = 12.6704   R-sq = 49.7%   R-sq(adj) = 43.4%

Analysis of Variance

SOURCE          DF      SS      MS     F      p
Regression       1  1268.2  1268.2  7.90  0.023
Residual Error   8  1284.3   160.5
Total            9  2552.5

Unusual Observations
Obs     x      y    Fit  SE Fit  Residual  St Resid
  4  3.00  75.00  42.97    4.04     32.03      2.67R

R denotes an observation with a large standardized residual.
```

FIGURE 14.19 MINITAB OUTPUT FOR THE REVISED OUTLIER DATA SET

```
The regression equation is
Y = 59.2 - 6.95 X

Predictor    Coef  SE Coef      T      p
Constant   59.237    3.835  15.45  0.000
X          -6.949    1.080  -6.43  0.000

S = 5.24808   R-sq = 83.8%   R-sq(adj) = 81.8%

Analysis of Variance

SOURCE          DF      SS      MS      F      p
Regression       1  1139.7  1139.7  41.38  0.000
Residual Error   8   220.3    27.5
Total            9  1360.0
```

using the incorrect data value substantially affected the goodness of fit. With the correct data, the value of R-sq increased from 49.7% to 83.8% and the value of b_0 decreased from 64.958 to 59.237. The slope of the line changed from -7.331 to -6.949. The identification of the outlier enabled us to correct the data error and improve the regression results.

Detecting Influential Observations

Sometimes one or more observations exert a strong influence on the results obtained. Figure 14.20 shows an example of an **influential observation** in simple linear regression. The estimated regression line has a negative slope. However, if the influential observation were

FIGURE 14.20 DATA SET WITH AN INFLUENTIAL OBSERVATION

dropped from the data set, the slope of the estimated regression line would change from negative to positive and the y-intercept would be smaller. Clearly, this one observation is much more influential in determining the estimated regression line than any of the others; dropping one of the other observations from the data set would have little effect on the estimated regression equation.

Influential observations can be identified from a scatter diagram when only one independent variable is present. An influential observation may be an outlier (an observation with a y value that deviates substantially from the trend), it may correspond to an x value far away from its mean (e.g., see Figure 14.20), or it may be caused by a combination of the two (a somewhat off-trend y value and a somewhat extreme x value).

Because influential observations may have such a dramatic effect on the estimated regression equation, they must be examined carefully. We should first check to make sure that no error was made in collecting or recording the data. If an error occurred, it can be corrected and a new estimated regression equation can be developed. If the observation is valid, we might consider ourselves fortunate to have it. Such a point, if valid, can contribute to a better understanding of the appropriate model and can lead to a better estimated regression equation. The presence of the influential observation in Figure 14.20, if valid, would suggest trying to obtain data on intermediate values of x to understand better the relationship between x and y.

Observations with extreme values for the independent variables are called **high leverage points**. The influential observation in Figure 14.20 is a point with high leverage. The leverage of an observation is determined by how far the values of the independent variables are from their mean values. For the single-independent-variable case, the leverage of the ith observation, denoted h_i, can be computed by using equation (14.33).

TABLE 14.12
DATA SET WITH A HIGH LEVERAGE OBSERVATION

x_i	y_i
10	125
10	130
15	120
20	115
20	120
25	110
70	100

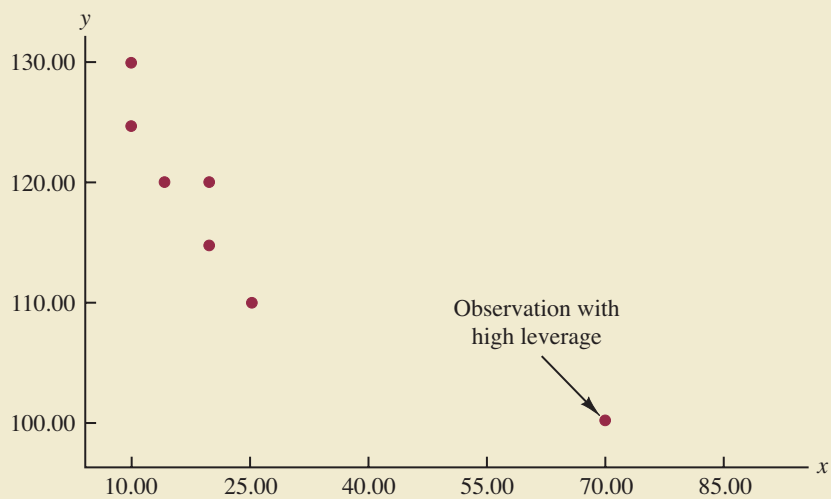
LEVERAGE OF OBSERVATION i

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\Sigma(x_i - \bar{x})^2} \qquad \textbf{(14.33)}$$

From the formula, it is clear that the farther x_i is from its mean $\bar{x}$, the higher the leverage of observation i.

Many statistical packages automatically identify observations with high leverage as part of the standard regression output. As an illustration of how the Minitab statistical package identifies points with high leverage, let us consider the data set in Table 14.12.

FIGURE 14.21 SCATTER DIAGRAM FOR THE DATA SET WITH A HIGH LEVERAGE OBSERVATION



From Figure 14.21, a scatter diagram for the data set in Table 14.12, it is clear that observation 7 ($x = 70$, $y = 100$) is an observation with an extreme value of x. Hence, we would expect it to be identified as a point with high leverage. For this observation, the leverage is computed by using equation (14.33) as follows.

$$h_7 = \frac{1}{n} + \frac{(x_7 - \bar{x})^2}{\Sigma(x_i - \bar{x})^2} = \frac{1}{7} + \frac{(70 - 24.286)^2}{2621.43} = .94$$

Computer software packages are essential for performing the computations to identify influential observations. Minitab's selection rule is discussed here.

For the case of simple linear regression, Minitab identifies observations as having high leverage if $h_i > 6/n$ or .99, whichever is smaller. For the data set in Table 14.12, $6/n = 6/7 = .86$. Because $h_7 = .94 > .86$, Minitab will identify observation 7 as an observation whose x value gives it large influence. Figure 14.22 shows the Minitab output for a regression analysis of this data set. Observation 7 ($x = 70$, $y = 100$) is identified as having large influence; it is printed on a separate line at the bottom, with an X in the right margin.

Influential observations that are caused by an interaction of large residuals and high leverage can be difficult to detect. Diagnostic procedures are available that take both into account in determining when an observation is influential. One such measure, called Cook's D statistic, will be discussed in Chapter 15.

NOTES AND COMMENTS

Once an observation is identified as potentially influential because of a large residual or high leverage, its impact on the estimated regression equation should be evaluated. More advanced texts discuss diagnostics for doing so. However, if one is not familiar with the more advanced material, a simple procedure is to run the regression analysis with and without the observation. This approach will reveal the influence of the observation on the results.

FIGURE 14.22 MINITAB OUTPUT FOR THE DATA SET WITH A HIGH LEVERAGE OBSERVATION

```
The regression equation is
y = 127 - 0.425 x

Predictor      Coef  SE Coef      T      p
Constant    127.466    2.961  43.04  0.000
X          -0.42507  0.09537  -4.46  0.007

S = 4.88282   R-sq = 79.9%   R-sq(adj) = 75.9%

Analysis of Variance

SOURCE          DF      SS      MS      F      p
Regression       1  473.65  473.65  19.87  0.007
Residual Error   5  119.21   23.84
Total            6  592.86

Unusual Observations
Obs     x       y    Fit  SE Fit  Residual  St Resid
  7  70.0  100.00  97.71    4.73      2.29      1.91 X

X denotes an observation whose X value gives it large influence.
```

Exercises

Methods

SELF test

50. Consider the following data for two variables, x and y.

x_i	135	110	130	145	175	160	120
y_i	145	100	120	120	130	130	110

a. Compute the standardized residuals for these data. Do the data include any outliers? Explain.
b. Plot the standardized residuals against $\hat{y}$. Does this plot reveal any outliers?
c. Develop a scatter diagram for these data. Does the scatter diagram indicate any outliers in the data? In general, what implications does this finding have for simple linear regression?

51. Consider the following data for two variables, x and y.

x_i	4	5	7	8	10	12	12	22
y_i	12	14	16	15	18	20	24	19

a. Compute the standardized residuals for these data. Do the data include any outliers? Explain.
b. Compute the leverage values for these data. Do there appear to be any influential observations in these data? Explain.
c. Develop a scatter diagram for these data. Does the scatter diagram indicate any influential observations? Explain.

Applications

SELF test

52. The following data show the media expenditures ($ millions) and the shipments in bbls. (millions) for 10 major brands of beer.

WEB file
Beer

Brand	Media Expenditures ($ millions)	Shipments
Budweiser	120.0	36.3
Bud Light	68.7	20.7
Miller Lite	100.1	15.9
Coors Light	76.6	13.2
Busch	8.7	8.1
Natural Light	0.1	7.1
Miller Genuine Draft	21.5	5.6
Miller High Life	1.4	4.4
Busch Light	5.3	4.3
Milwaukee's Best	1.7	4.3

a. Develop the estimated regression equation for these data.
b. Use residual analysis to determine whether any outliers and/or influential observations are present. Briefly summarize your findings and conclusions.

53. Health experts recommend that runners drink 4 ounces of water every 15 minutes they run. Runners who run three to eight hours need a larger-capacity hip-mounted or over-the-shoulder hydration system. The following data show the liquid volume (fl oz) and the price for 26 Ultimate Direction hip-mounted or over-the-shoulder hydration systems (*Trail Runner Gear Guide,* 2003).

WEB file
Hydration2

Model	Volume (fl oz)	Price ($)
Fastdraw	20	10
Fastdraw Plus	20	12
Fitness	20	12
Access	20	20
Access Plus	24	25
Solo	20	25
Serenade	20	35
Solitaire	20	35
Gemini	40	45
Shadow	64	40
SipStream	96	60
Express	20	30
Lightning	28	40
Elite	40	60
Extender	40	65
Stinger	32	65
GelFlask Belt	4	20
GelDraw	4	7
GelFlask Clip-on Holster	4	10
GelFlask Holster SS	4	10
Strider (W)	20	30
Walkabout (W)	230	40
Solitude I.C.E.	20	35
Getaway I.C.E.	40	55
Profile I.C.E.	64	50
Traverse I.C.E.	64	60

a. Develop the estimated regression equation that can be used to predict the price of a hydration system given its liquid volume.
b. Use residual analysis to determine whether any outliers or influential observations are present. Briefly summarize your findings and conclusions.

54. The following data show the annual revenue ($ millions) and the estimated team value ($ millions) for the 32 teams in the National Football League (Forbes website, February 2009).

WEB file
NFLValues

Team	Revenue ($ millions)	Value ($ millions)
Arizona Cardinals	203	914
Atlanta Falcons	203	872
Baltimore Ravens	226	1062
Buffalo Bills	206	885
Carolina Panthers	221	1040
Chicago Bears	226	1064
Cincinnati Bengals	205	941
Cleveland Browns	220	1035
Dallas Cowboys	269	1612
Denver Broncos	226	1061
Detroit Lions	204	917
Green Bay Packers	218	1023
Houston Texans	239	1125
Indianapolis Colts	203	1076
Jacksonville Jaguars	204	876
Kansas City Chiefs	214	1016
Miami Dolphins	232	1044
Minnesota Vikings	195	839
New England Patriots	282	1324
New Orleans Saints	213	937
New York Giants	214	1178
New York Jets	213	1170
Oakland Raiders	205	861
Philadelphia Eagles	237	1116
Pittsburgh Steelers	216	1015
San Diego Chargers	207	888
San Francisco 49ers	201	865
Seattle Seahawks	215	1010
St. Louis Rams	206	929
Tampa Bay Buccaneers	224	1053
Tennessee Titans	216	994
Washington Redskins	327	1538

a. Develop a scatter diagram with Revenue on the horizontal axis and Value on the vertical axis. Looking at the scatter diagram, does it appear that there are any outliers and/or influential observations in the data?
b. Develop the estimated regression equation that can be used to predict team value given the value of annual revenue.
c. Use residual analysis to determine whether any outliers and/or influential observations are present. Briefly summarize your findings and conclusions.

Summary

In this chapter we showed how regression analysis can be used to determine how a dependent variable y is related to an independent variable x. In simple linear regression, the regression model is $y = \beta_0 + \beta_1 x + \epsilon$. The simple linear regression equation $E(y) = \beta_0 + \beta_1 x$ describes how the mean or expected value of y is related to x. We used sample data and the least squares

method to develop the estimated regression equation $\hat{y} = b_0 + b_1 x$. In effect, b_0 and b_1 are the sample statistics used to estimate the unknown model parameters β_0 and β_1.

The coefficient of determination was presented as a measure of the goodness of fit for the estimated regression equation; it can be interpreted as the proportion of the variation in the dependent variable y that can be explained by the estimated regression equation. We reviewed correlation as a descriptive measure of the strength of a linear relationship between two variables.

The assumptions about the regression model and its associated error term ϵ were discussed, and t and F tests, based on those assumptions, were presented as a means for determining whether the relationship between two variables is statistically significant. We showed how to use the estimated regression equation to develop confidence interval estimates of the mean value of y and prediction interval estimates of individual values of y.

The chapter concluded with a section on the computer solution of regression problems and two sections on the use of residual analysis to validate the model assumptions and to identify outliers and influential observations.

Glossary

Dependent variable The variable that is being predicted or explained. It is denoted by y.

Independent variable The variable that is doing the predicting or explaining. It is denoted by x.

Simple linear regression Regression analysis involving one independent variable and one dependent variable in which the relationship between the variables is approximated by a straight line.

Regression model The equation that describes how y is related to x and an error term; in simple linear regression, the regression model is $y = \beta_0 + \beta_1 x + \epsilon$.

Regression equation The equation that describes how the mean or expected value of the dependent variable is related to the independent variable; in simple linear regression, $E(y) = \beta_0 + \beta_1 x$.

Estimated regression equation The estimate of the regression equation developed from sample data by using the least squares method. For simple linear regression, the estimated regression equation is $\hat{y} = b_0 + b_1 x$.

Least squares method A procedure used to develop the estimated regression equation. The objective is to minimize $\Sigma(y_i - \hat{y}_i)^2$.

Scatter diagram A graph of bivariate data in which the independent variable is on the horizontal axis and the dependent variable is on the vertical axis.

Coefficient of determination A measure of the goodness of fit of the estimated regression equation. It can be interpreted as the proportion of the variability in the dependent variable y that is explained by the estimated regression equation.

***i*th residual** The difference between the observed value of the dependent variable and the value predicted using the estimated regression equation; for the ith observation the ith residual is $y_i - \hat{y}_i$.

Correlation coefficient A measure of the strength of the linear relationship between two variables (previously discussed in Chapter 3).

Mean square error The unbiased estimate of the variance of the error term σ^2. It is denoted by MSE or s^2.

Standard error of the estimate The square root of the mean square error, denoted by s. It is the estimate of σ, the standard deviation of the error term ϵ.

ANOVA table The analysis of variance table used to summarize the computations associated with the F test for significance.

Confidence interval The interval estimate of the mean value of y for a given value of x.

Prediction interval The interval estimate of an individual value of y for a given value of x.

Residual analysis The analysis of the residuals used to determine whether the assumptions made about the regression model appear to be valid. Residual analysis is also used to identify outliers and influential observations.
Residual plot Graphical representation of the residuals that can be used to determine whether the assumptions made about the regression model appear to be valid.
Standardized residual The value obtained by dividing a residual by its standard deviation.
Normal probability plot A graph of the standardized residuals plotted against values of the normal scores. This plot helps determine whether the assumption that the error term has a normal probability distribution appears to be valid.
Outlier A data point or observation that does not fit the trend shown by the remaining data.
Influential observation An observation that has a strong influence or effect on the regression results.
High leverage points Observations with extreme values for the independent variables.

Key Formulas

Simple Linear Regression Model

$$y = \beta_0 + \beta_1 x + \epsilon \tag{14.1}$$

Simple Linear Regression Equation

$$E(y) = \beta_0 + \beta_1 x \tag{14.2}$$

Estimated Simple Linear Regression Equation

$$\hat{y} = b_0 + b_1 x \tag{14.3}$$

Least Squares Criterion

$$\min \Sigma(y_i - \hat{y}_i)^2 \tag{14.5}$$

Slope and *y*-Intercept for the Estimated Regression Equation

$$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} \tag{14.6}$$

$$b_0 = \bar{y} - b_1 \bar{x} \tag{14.7}$$

Sum of Squares Due to Error

$$\text{SSE} = \Sigma(y_i - \hat{y}_i)^2 \tag{14.8}$$

Total Sum of Squares

$$\text{SST} = \Sigma(y_i - \bar{y})^2 \tag{14.9}$$

Sum of Squares Due to Regression

$$\text{SSR} = \Sigma(\hat{y}_i - \bar{y})^2 \tag{14.10}$$

Relationship Among SST, SSR, and SSE

$$\text{SST} = \text{SSR} + \text{SSE} \tag{14.11}$$

Coefficient of Determination

$$r^2 = \frac{\text{SSR}}{\text{SST}} \tag{14.12}$$

Sample Correlation Coefficient

$$r_{xy} = (\text{sign of } b_1)\sqrt{\text{Coefficient of determination}} = (\text{sign of } b_1)\sqrt{r^2} \quad (14.13)$$

Mean Square Error (Estimate of σ^2)

$$s^2 = \text{MSE} = \frac{\text{SSE}}{n-2} \quad (14.15)$$

Standard Error of the Estimate

$$s = \sqrt{\text{MSE}} = \sqrt{\frac{\text{SSE}}{n-2}} \quad (14.16)$$

Standard Deviation of b_1

$$\sigma_{b_1} = \frac{\sigma}{\sqrt{\Sigma(x_i - \bar{x})^2}} \quad (14.17)$$

Estimated Standard Deviation of b_1

$$s_{b_1} = \frac{s}{\sqrt{\Sigma(x_i - \bar{x})^2}} \quad (14.18)$$

***t* Test Statistic**

$$t = \frac{b_1}{s_{b_1}} \quad (14.19)$$

Mean Square Regression

$$\text{MSR} = \frac{\text{SSR}}{\text{Number of independent variables}} \quad (14.20)$$

***F* Test Statistic**

$$F = \frac{\text{MSR}}{\text{MSE}} \quad (14.21)$$

Estimated Standard Deviation of $\hat{y}_p$

$$s_{\hat{y}_p} = s\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}} \quad (14.23)$$

Confidence Interval for $E(y_p)$

$$\hat{y}_p \pm t_{\alpha/2} s_{\hat{y}_p} \quad (14.24)$$

Estimated Standard Deviation of an Individual Value

$$s_{\text{ind}} = s\sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}} \quad (14.26)$$

Prediction Interval for y_p

$$\hat{y}_p \pm t_{\alpha/2} s_{\text{ind}} \quad (14.27)$$

Residual for Observation *i*

$$y_i - \hat{y}_i \tag{14.28}$$

Standard Deviation of the *i*th Residual

$$s_{y_i - \hat{y}_i} = s\sqrt{1 - h_i} \tag{14.30}$$

Standardized Residual for Observation *i*

$$\frac{y_i - \hat{y}_i}{s_{y_i - \hat{y}_i}} \tag{14.32}$$

Leverage of Observation *i*

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\Sigma(x_i - \bar{x})^2} \tag{14.33}$$

Supplementary Exercises

55. Does a high value of r^2 imply that two variables are causally related? Explain.
56. In your own words, explain the difference between an interval estimate of the mean value of y for a given x and an interval estimate for an individual value of y for a given x.
57. What is the purpose of testing whether $\beta_1 = 0$? If we reject $\beta_1 = 0$, does it imply a good fit?
58. The data in the following table show the number of shares selling (millions) and the expected price (average of projected low price and projected high price) for 10 selected initial public stock offerings.

WEB file
IPO

Company	Shares Selling (millions)	Expected Price ($)
American Physician	5.0	15
Apex Silver Mines	9.0	14
Dan River	6.7	15
Franchise Mortgage	8.75	17
Gene Logic	3.0	11
International Home Foods	13.6	19
PRT Group	4.6	13
Rayovac	6.7	14
RealNetworks	3.0	10
Software AG Systems	7.7	13

   a. Develop an estimated regression equation with the number of shares selling as the independent variable and the expected price as the dependent variable.
   b. At the .05 level of significance, is there a significant relationship between the two variables?
   c. Did the estimated regression equation provide a good fit? Explain.
   d. Use the estimated regression equation to estimate the expected price for a firm considering an initial public offering of 6 million shares.

59. The following data show Morningstar's Fair Value estimate and the Share Price for 28 companies. Fair Value is an estimate of a company's value per share that takes into account estimates of the company's growth, profitability, riskiness, and other factors over the next five years (*Morningstar Stocks 500*, 2008 edition).

WEB file
Stocks500

Company	Fair Value ($)	Share Price ($)
Air Products and Chemicals	80	98.63
Allied Waste Industries	17	11.02
America Mobile	83	61.39
AT&T	35	41.56
Bank of America	70	41.26
Barclays PLC	68	40.37
Citigroup	53	29.44
Costco Wholesale Corp.	75	69.76
Covidien, Ltd.	58	44.29
Darden Restaurants	52	27.71
Dun & Bradstreet	87	88.63
Equifax	42	36.36
Gannett Co.	38	39.00
Genuine Parts	48	46.30
GlaxoSmithKline PLC	57	50.39
Iron Mountain	33	37.02
ITT Corporation	83	66.04
Johnson & Johnson	80	66.70
Las Vegas Sands	98	103.05
Macrovision	23	18.33
Marriott International	39	34.18
Nalco Holding Company	29	24.18
National Interstate	25	33.10
Portugal Telecom	15	13.02
Qualcomm	48	39.35
Royal Dutch Shell Ltd.	87	84.20
SanDisk	60	33.17
Time Warner	42	27.60

a. Develop the estimated regression equation that could be used to estimate the Share Price given the Fair Value.
b. At the .05 level of significance, is there a significant relationship between the two variables?
c. Use the estimated regression equation to estimate the Share Price for a company that has a Fair Value of $50.
d. Do you believe the estimated regression equation would provide a good prediction of the share price? Use r^2 to support your answer.

60. One of the biggest changes in higher education in recent years has been the growth of online universities. The Online Education Database is an independent organization whose mission is to build a comprehensive list of the top accredited online colleges. The following table shows the retention rate (%) and the graduation rate (%) for 29 online colleges (Online Education Database website, January 2009).

College	Retention Rate (%)	Graduation Rate (%)
Western International University	7	25
South University	51	25
University of Phoenix	4	28
American InterContinental University	29	32
Franklin University	33	33
Devry University	47	33

WEB file
OnlineEdu

College	Retention Rate (%)	Graduation Rate (%)
Tiffin University	63	34
Post University	45	36
Peirce College	60	36
Everest University	62	36
Upper Iowa University	67	36
Dickinson State University	65	37
Western Governors University	78	37
Kaplan University	75	38
Salem International University	54	39
Ashford University	45	41
ITT Technical Institute	38	44
Berkeley College	51	45
Grand Canyon University	69	46
Nova Southeastern University	60	47
Westwood College	37	48
Everglades University	63	50
Liberty University	73	51
LeTourneau University	78	52
Rasmussen College	48	53
Keiser University	95	55
Herzing College	68	56
National University	100	57
Florida National College	100	61

a. Develop a scatter diagram with retention rate as the independent variable. What does the scatter diagram indicate about the relationship between the two variables?
b. Develop the estimated regression equation.
c. Test for a significant relationship. Use $\alpha = .05$.
d. Did the estimated regression equation provide a good fit?
e. Suppose you were the president of South University. After reviewing the results, would you have any concerns about the performance of your university as compared to other online universities?
f. Suppose you were the president of the University of Phoenix. After reviewing the results, would you have any concerns about the performance of your university as compared to other online universities?

61. Jensen Tire & Auto is in the process of deciding whether to purchase a maintenance contract for its new computer wheel alignment and balancing machine. Managers feel that maintenance expense should be related to usage, and they collected the following information on weekly usage (hours) and annual maintenance expense (in hundreds of dollars).

WEB file
Jensen

Weekly Usage (hours)	Annual Maintenance Expense
13	17.0
10	22.0
20	30.0
28	37.0
32	47.0
17	30.5
24	32.5
31	39.0
40	51.5
38	40.0

a. Develop the estimated regression equation that relates annual maintenance expense to weekly usage.
b. Test the significance of the relationship in part (a) at a .05 level of significance.
c. Jensen expects to use the new machine 30 hours per week. Develop a 95% prediction interval for the company's annual maintenance expense.
d. If the maintenance contract costs $3000 per year, would you recommend purchasing it? Why or why not?

62. In a manufacturing process the assembly line speed (feet per minute) was thought to affect the number of defective parts found during the inspection process. To test this theory, managers devised a situation in which the same batch of parts was inspected visually at a variety of line speeds. They collected the following data.

Line Speed	Number of Defective Parts Found
20	21
20	19
40	15
30	16
60	14
40	17

a. Develop the estimated regression equation that relates line speed to the number of defective parts found.
b. At a .05 level of significance, determine whether line speed and number of defective parts found are related.
c. Did the estimated regression equation provide a good fit to the data?
d. Develop a 95% confidence interval to predict the mean number of defective parts for a line speed of 50 feet per minute.

63. A sociologist was hired by a large city hospital to investigate the relationship between the number of unauthorized days that employees are absent per year and the distance (miles) between home and work for the employees. A sample of 10 employees was chosen, and the following data were collected.

WEB file
Absent

Distance to Work (miles)	Number of Days Absent
1	8
3	5
4	8
6	7
8	6
10	3
12	5
14	2
14	4
18	2

a. Develop a scatter diagram for these data. Does a linear relationship appear reasonable? Explain.
b. Develop the least squares estimated regression equation.
c. Is there a significant relationship between the two variables? Use $\alpha = .05$.
d. Did the estimated regression equation provide a good fit? Explain.
e. Use the estimated regression equation developed in part (b) to develop a 95% confidence interval for the expected number of days absent for employees living 5 miles from the company.

64. The regional transit authority for a major metropolitan area wants to determine whether there is any relationship between the age of a bus and the annual maintenance cost. A sample of 10 buses resulted in the following data.

WEB file
AgeCost

Age of Bus (years)	Maintenance Cost ($)
1	350
2	370
2	480
2	520
2	590
3	550
4	750
4	800
5	790
5	950

a. Develop the least squares estimated regression equation.
b. Test to see whether the two variables are significantly related with $\alpha = .05$.
c. Did the least squares line provide a good fit to the observed data? Explain.
d. Develop a 95% prediction interval for the maintenance cost for a specific bus that is 4 years old.

65. A marketing professor at Givens College is interested in the relationship between hours spent studying and total points earned in a course. Data collected on 10 students who took the course last quarter follow.

WEB file
HoursPts

Hours Spent Studying	Total Points Earned
45	40
30	35
90	75
60	65
105	90
65	50
90	90
80	80
55	45
75	65

a. Develop an estimated regression equation showing how total points earned is related to hours spent studying.
b. Test the significance of the model with $\alpha = .05$.
c. Predict the total points earned by Mark Sweeney. He spent 95 hours studying.
d. Develop a 95% prediction interval for the total points earned by Mark Sweeney.

66. Reuters reported the market beta for Xerox was 1.22 (Reuters website, January 30, 2009). Market betas for individual stocks are determined by simple linear regression. For each stock, the dependent variable is its quarterly percentage return (capital appreciation plus dividends) minus the percentage return that could be obtained from a risk-free investment (the Treasury Bill rate is used as the risk-free rate). The independent variable is the quarterly percentage return (capital appreciation plus dividends) for the stock market (S&P 500) minus the percentage return from a risk-free investment. An estimated regression equation is developed with quarterly data; the market beta for the stock is the slope of the estimated regression equation (b_1). The value of the market beta is often interpreted as a measure of the risk associated with the stock. Market betas greater than 1 indicate that the

stock is more volatile than the market average; market betas less than 1 indicate that the stock is less volatile than the market average. Suppose that the following figures are the differences between the percentage return and the risk-free return for 10 quarters for the S&P 500 and Horizon Technology.

WEB file
MktBeta

S&P 500	Horizon
1.2	−0.7
−2.5	−2.0
−3.0	−5.5
2.0	4.7
5.0	1.8
1.2	4.1
3.0	2.6
−1.0	2.0
.5	−1.3
2.5	5.5

a. Develop an estimated regression equation that can be used to determine the market beta for Horizon Technology. What is Horizon Technology's market beta?
b. Test for a significant relationship at the .05 level of significance.
c. Did the estimated regression equation provide a good fit? Explain.
d. Use the market betas of Xerox and Horizon Technology to compare the risk associated with the two stocks.

67. The Transactional Records Access Clearinghouse at Syracuse University reported data showing the odds of an Internal Revenue Service audit. The following table shows the average adjusted gross income reported and the percent of the returns that were audited for 20 selected IRS districts.

WEB file
IRSAudit

District	Adjusted Gross Income ($)	Percent Audited
Los Angeles	36,664	1.3
Sacramento	38,845	1.1
Atlanta	34,886	1.1
Boise	32,512	1.1
Dallas	34,531	1.0
Providence	35,995	1.0
San Jose	37,799	0.9
Cheyenne	33,876	0.9
Fargo	30,513	0.9
New Orleans	30,174	0.9
Oklahoma City	30,060	0.8
Houston	37,153	0.8
Portland	34,918	0.7
Phoenix	33,291	0.7
Augusta	31,504	0.7
Albuquerque	29,199	0.6
Greensboro	33,072	0.6
Columbia	30,859	0.5
Nashville	32,566	0.5
Buffalo	34,296	0.5

a. Develop the estimated regression equation that could be used to predict the percent audited given the average adjusted gross income reported.
b. At the .05 level of significance, determine whether the adjusted gross income and the percent audited are related.
c. Did the estimated regression equation provide a good fit? Explain.

d. Use the estimated regression equation developed in part (a) to calculate a 95% confidence interval for the expected percent audited for districts with an average adjusted gross income of $35,000.

68. The Australian Public Service Commission's State of the Service Report 2002–2003 reported job satisfaction ratings for employees. One of the survey questions asked employees to choose the five most important workplace factors (from a list of factors) that most affected how satisfied they were with their job. Respondents were then asked to indicate their level of satisfaction with their top five factors. The following data show the percentage of employees who nominated the factor in their top five, and a corresponding satisfaction rating measured using the percentage of employees who nominated the factor in the top five and who were "very satisfied" or "satisfied" with the factor in their current workplace (*www.apsc.gov.au/stateoftheservice*).

WEB file
JobSat

Workplace Factor	Top Five (%)	Satisfaction Rating (%)
Appropriate workload	30	49
Chance to be creative/innovative	38	64
Chance to make a useful contribution to society	40	67
Duties/expectations made clear	40	69
Flexible working arrangements	55	86
Good working relationships	60	85
Interesting work provided	48	74
Opportunities for career development	33	43
Opportunities to develop my skills	46	66
Opportunities to utilize my skills	50	70
Regular feedback/recognition for effort	42	53
Salary	47	62
Seeing tangible results from my work	42	69

a. Develop a scatter diagram with Top Five (%) on the horizontal axis and Satisfaction Rating (%) on the vertical axis.
b. What does the scatter diagram developed in part (a) indicate about the relationship between the two variables?
c. Develop the estimated regression equation that could be used to predict the Satisfaction Rating (%) given the Top Five (%).
d. Test for a significant relationship at the .05 level of significance.
e. Did the estimated regression equation provide a good fit? Explain.
f. What is the value of the sample correlation coefficient?

Case Problem 1 Measuring Stock Market Risk

One measure of the risk or volatility of an individual stock is the standard deviation of the total return (capital appreciation plus dividends) over several periods of time. Although the standard deviation is easy to compute, it does not take into account the extent to which the price of a given stock varies as a function of a standard market index, such as the S&P 500. As a result, many financial analysts prefer to use another measure of risk referred to as *beta*.

Betas for individual stocks are determined by simple linear regression. The dependent variable is the total return for the stock and the independent variable is the total return for the stock market.* For this case problem we will use the S&P 500 index as the measure of

*Various sources use different approaches for computing betas. For instance, some sources subtract the return that could be obtained from a risk-free investment (e.g., T-bills) from the dependent variable and the independent variable before computing the estimated regression equation. Some also use different indexes for the total return of the stock market; for instance, *Value Line* computes betas using the New York Stock Exchange composite index.

WEB file
Beta

the total return for the stock market, and an estimated regression equation will be developed using monthly data. The beta for the stock is the slope of the estimated regression equation (b_1). The data contained in the file named Beta provides the total return (capital appreciation plus dividends) over 36 months for eight widely traded common stocks and the S&P 500.

The value of beta for the stock market will always be 1; thus, stocks that tend to rise and fall with the stock market will also have a beta close to 1. Betas greater than 1 indicate that the stock is more volatile than the market, and betas less than 1 indicate that the stock is less volatile than the market. For instance, if a stock has a beta of 1.4, it is 40% *more* volatile than the market, and if a stock has a beta of .4, it is 60% *less* volatile than the market.

Managerial Report

You have been assigned to analyze the risk characteristics of these stocks. Prepare a report that includes but is not limited to the following items.

a. Compute descriptive statistics for each stock and the S&P 500. Comment on your results. Which stocks are the most volatile?

b. Compute the value of beta for each stock. Which of these stocks would you expect to perform best in an up market? Which would you expect to hold their value best in a down market?

c. Comment on how much of the return for the individual stocks is explained by the market.

Case Problem 2 U.S. Department of Transportation

As part of a study on transportation safety, the U.S. Department of Transportation collected data on the number of fatal accidents per 1000 licenses and the percentage of licensed drivers under the age of 21 in a sample of 42 cities. Data collected over a one-year period follow. These data are contained in the file named Safety.

WEB file
Safety

Percent Under 21	Fatal Accidents per 1000 Licenses	Percent Under 21	Fatal Accidents per 1000 Licenses
13	2.962	17	4.100
12	0.708	8	2.190
8	0.885	16	3.623
12	1.652	15	2.623
11	2.091	9	0.835
17	2.627	8	0.820
18	3.830	14	2.890
8	0.368	8	1.267
13	1.142	15	3.224
8	0.645	10	1.014
9	1.028	10	0.493
16	2.801	14	1.443
12	1.405	18	3.614
9	1.433	10	1.926
10	0.039	14	1.643
9	0.338	16	2.943
11	1.849	12	1.913
12	2.246	15	2.814
14	2.855	13	2.634
14	2.352	9	0.926
11	1.294	17	3.256

Managerial Report

1. Develop numerical and graphical summaries of the data.
2. Use regression analysis to investigate the relationship between the number of fatal accidents and the percentage of drivers under the age of 21. Discuss your findings.
3. What conclusion and recommendations can you derive from your analysis?

Case Problem 3 Alumni Giving

Alumni donations are an important source of revenue for colleges and universities. If administrators could determine the factors that influence increases in the percentage of alumni who make a donation, they might be able to implement policies that could lead to increased revenues. Research shows that students who are more satisfied with their contact with teachers are more likely to graduate. As a result, one might suspect that smaller class sizes and lower student-faculty ratios might lead to a higher percentage of satisfied graduates, which in turn might lead to increases in the percentage of alumni who make a donation. Table 14.13 shows data for 48 national universities (*America's Best Colleges,* Year 2000 ed.). The column labeled % of Classes Under 20 shows the percentage of classes offered with fewer than 20 students. The column labeled Student/Faculty Ratio is the number of students enrolled divided by the total number of faculty. Finally, the column labeled Alumni Giving Rate is the percentage of alumni that made a donation to the university.

Managerial Report

1. Develop numerical and graphical summaries of the data.
2. Use regression analysis to develop an estimated regression equation that could be used to predict the alumni giving rate given the percentage of classes with fewer than 20 students.
3. Use regression analysis to develop an estimated regression equation that could be used to predict the alumni giving rate given the student-faculty ratio.
4. Which of the two estimated regression equations provides the best fit? For this estimated regression equation, perform an analysis of the residuals and discuss your findings and conclusions.
5. What conclusions and recommendations can you derive from your analysis?

Case Problem 4 PGA Tour Statistics

The Professional Golfers Association (PGA) maintains data on performance and earnings for members of the PGA Tour. The top 125 players based on total earnings in PGA Tour events are exempt for the following season. Making the top 125 money list is important because a player who is "exempt" has qualified to be a full-time member of the PGA tour for the following season.

WEB file
PGATour

During recent years on the PGA Tour there have been significant advances in the technology of golf balls and golf clubs, and this technology has been one of the major reasons for the increase in the average driving distance of PGA Tour players. In 1992, the average driving distance was 260 yards, but in 2003 this increased to 286 yards. PGA Tour pros are hitting the ball farther than ever before, but how important is driving distance in terms of a player's performance? And what effect has this increased distance had on the players'

TABLE 14.13 DATA FOR 48 NATIONAL UNIVERSITIES

WEB file
Alumni

	% of Classes Under 20	Student/Faculty Ratio	Alumni Giving Rate
Boston College	39	13	25
Brandeis University	68	8	33
Brown University	60	8	40
California Institute of Technology	65	3	46
Carnegie Mellon University	67	10	28
Case Western Reserve Univ.	52	8	31
College of William and Mary	45	12	27
Columbia University	69	7	31
Cornell University	72	13	35
Dartmouth College	61	10	53
Duke University	68	8	45
Emory University	65	7	37
Georgetown University	54	10	29
Harvard University	73	8	46
Johns Hopkins University	64	9	27
Lehigh University	55	11	40
Massachusetts Inst. of Technology	65	6	44
New York University	63	13	13
Northwestern University	66	8	30
Pennsylvania State Univ.	32	19	21
Princeton University	68	5	67
Rice University	62	8	40
Stanford University	69	7	34
Tufts University	67	9	29
Tulane University	56	12	17
U. of California–Berkeley	58	17	18
U. of California–Davis	32	19	7
U. of California–Irvine	42	20	9
U. of California–Los Angeles	41	18	13
U. of California–San Diego	48	19	8
U. of California–Santa Barbara	45	20	12
U. of Chicago	65	4	36
U. of Florida	31	23	19
U. of Illinois–Urbana Champaign	29	15	23
U. of Michigan–Ann Arbor	51	15	13
U. of North Carolina–Chapel Hill	40	16	26
U. of Notre Dame	53	13	49
U. of Pennsylvania	65	7	41
U. of Rochester	63	10	23
U. of Southern California	53	13	22
U. of Texas–Austin	39	21	13
U. of Virginia	44	13	28
U. of Washington	37	12	12
U. of Wisconsin–Madison	37	13	13
Vanderbilt University	68	9	31
Wake Forest University	59	11	38
Washington University–St. Louis	73	7	33
Yale University	77	7	50

accuracy? To investigate these issues, year-end performance data for the 125 players who had the highest total earnings in PGA Tour events for 2008 are contained in the file named PGA-Tour (PGA Tour website, 2009). Each row of the data set corresponds to a PGA Tour player, and the data have been sorted based upon total earnings. Descriptions for the data follow.

Money: Total earnings in PGA Tour events.

Scoring Average: The average number of strokes per completed round.

DrDist (Driving Distance): DrDist is the average number of yards per measured drive. On the PGA Tour, driving distance is measured on two holes per round. Care is taken to select two holes which face in opposite directions to counteract the effect of wind. Drives are measured to the point at which they come to rest regardless of whether they are in the fairway or not.

DrAccu (Driving Accuracy): The percentage of time a tee shot comes to rest in the fairway (regardless of club). Driving accuracy is measured on every hole, excluding par 3's.

GIR (Greens in Regulation): The percentage of time a player was able to hit the green in regulation. A green is considered hit in regulation if any portion of the ball is touching the putting surface after the GIR stroke has been taken. The GIR stroke is determined by subtracting 2 from par (first stroke on a par 3, second on a par 4, third on a par 5). In other words, a green is considered hit in regulation if the player has reached the putting surface in par minus two strokes.

Managerial Report

1. Develop numerical and graphical summaries of the data.
2. Use regression analysis to investigate the relationship between Scoring Average and DrDist. Does it appear that players who drive the ball farther have lower average scores?
3. Use regression analysis to investigate the relationship between Scoring Average and DrAccu. Does it appear that players who are more accurate in hitting the fairway have lower average scores?
4. Use regression analysis to investigate the relationship between Scoring Average and GIR. Does it appear that players who are more accurate in hitting greens in regulation have lower average scores?
5. Which of the three variables (DrDist, DrAccu, and GIR) appears to be the most significant factor in terms of a player's average score?
6. Treating DrDist as the independent variable and DrAccu as the dependent variable, investigate the relationship between driving distance and driving accuracy.

Appendix 14.1 Calculus-Based Derivation of Least Squares Formulas

As mentioned in the chapter, the least squares method is a procedure for determining the values of b_0 and b_1 that minimize the sum of squared residuals. The sum of squared residuals is given by

$$\Sigma(y_i - \hat{y}_i)^2$$

Substituting $\hat{y}_i = b_0 + b_1 x_i$, we get

$$\Sigma(y_i - b_0 - b_1 x_i)^2 \qquad \textbf{(14.34)}$$

as the expression that must be minimized.

To minimize expression (14.34), we must take the partial derivatives with respect to b_0 and b_1, set them equal to zero, and solve. Doing so, we get

$$\frac{\partial \Sigma(y_i - b_0 - b_1 x_i)^2}{\partial b_0} = -2\Sigma(y_i - b_0 - b_1 x_i) = 0 \quad \textbf{(14.35)}$$

$$\frac{\partial \Sigma(y_i - b_0 - b_1 x_i)^2}{\partial b_1} = -2\Sigma x_i(y_i - b_0 - b_1 x_i) = 0 \quad \textbf{(14.36)}$$

Dividing equation (14.35) by two and summing each term individually yields

$$-\Sigma y_i + \Sigma b_0 + \Sigma b_1 x_i = 0$$

Bringing Σy_i to the other side of the equal sign and noting that $\Sigma b_0 = nb_0$, we obtain

$$nb_0 + (\Sigma x_i)b_1 = \Sigma y_i \quad \textbf{(14.37)}$$

Similar algebraic simplification applied to equation (14.36) yields

$$(\Sigma x_i)b_0 + (\Sigma x_i^2)b_1 = \Sigma x_i y_i \quad \textbf{(14.38)}$$

Equations (14.37) and (14.38) are known as the *normal equations.* Solving equation (14.37) for b_0 yields

$$b_0 = \frac{\Sigma y_i}{n} - b_1 \frac{\Sigma x_i}{n} \quad \textbf{(14.39)}$$

Using equation (14.39) to substitute for b_0 in equation (14.38) provides

$$\frac{\Sigma x_i \Sigma y_i}{n} - \frac{(\Sigma x_i)^2}{n} b_1 + (\Sigma x_i^2)b_1 = \Sigma x_i y_i \quad \textbf{(14.40)}$$

By rearranging the terms in equation (14.40), we obtain

$$b_1 = \frac{\Sigma x_i y_i - (\Sigma x_i \Sigma y_i)/n}{\Sigma x_i^2 - (\Sigma x_i)^2/n} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} \quad \textbf{(14.41)}$$

Because $\bar{y} = \Sigma y_i/n$ and $\bar{x} = \Sigma x_i/n$, we can rewrite equation (14.39) as

$$b_0 = \bar{y} - b_1 \bar{x} \quad \textbf{(14.42)}$$

Equations (14.41) and (14.42) are the formulas (14.6) and (14.7) we used in the chapter to compute the coefficients in the estimated regression equation.

Appendix 14.2 A Test for Significance Using Correlation

Using the sample correlation coefficient r_{xy}, we can determine whether the linear relationship between x and y is significant by testing the following hypotheses about the population correlation coefficient ρ_{xy}.

$$H_0: \rho_{xy} = 0$$
$$H_a: \rho_{xy} \neq 0$$

If H_0 is rejected, we can conclude that the population correlation coefficient is not equal to zero and that the linear relationship between the two variables is significant. This test for significance follows.

A TEST FOR SIGNIFICANCE USING CORRELATION

$$H_0: \rho_{xy} = 0$$
$$H_a: \rho_{xy} \neq 0$$

TEST STATISTIC

$$t = r_{xy}\sqrt{\frac{n-2}{1-r_{xy}^2}} \qquad (14.43)$$

REJECTION RULE

p-value approach: Reject H_0 if p-value $\leq \alpha$

Critical value approach: Reject H_0 if $t \leq -t_{\alpha/2}$ or if $t \geq t_{\alpha/2}$

where $t_{\alpha/2}$ is based on a t distribution with $n - 2$ degrees of freedom.

In Section 14.3, we found that the sample with $n = 10$ provided the sample correlation coefficient for student population and quarterly sales of $r_{xy} = .9501$. The test statistic is

$$t = r_{xy}\sqrt{\frac{n-2}{1-r_{xy}^2}} = .9501\sqrt{\frac{10-2}{1-(.9501)^2}} = 8.61$$

The t distribution table shows that with $n - 2 = 10 - 2 = 8$ degrees of freedom, $t = 3.355$ provides an area of .005 in the upper tail. Thus, the area in the upper tail of the t distribution corresponding to the test statistic $t = 8.61$ must be less than .005. Because this test is a two-tailed test, we double this value to conclude that the p-value associated with $t = 8.61$ must be less than 2(.005) = .01. Excel or Minitab show the p-value = .000. Because the p-value is less than $\alpha = .01$, we reject H_0 and conclude that ρ_{xy} is not equal to zero. This evidence is sufficient to conclude that a significant linear relationship exists between student population and quarterly sales.

Note that except for rounding, the test statistic t and the conclusion of a significant relationship are identical to the results obtained in Section 14.5 for the t test conducted using Armand's estimated regression equation $\hat{y} = 60 + 5x$. Performing regression analysis provides the conclusion of a significant relationship between x and y and in addition provides the equation showing how the variables are related. Most analysts therefore use modern computer packages to perform regression analysis and find that using correlation as a test of significance is unnecessary.

Appendix 14.3 Regression Analysis with Minitab

WEB file Armand's

In Section 14.7 we discussed the computer solution of regression problems by showing Minitab's output for the Armand's Pizza Parlors problem. In this appendix, we describe the steps required to generate the Minitab computer solution. First, the data must be entered in a Minitab worksheet. Student population data are entered in column C1 and quarterly sales data are entered in column C2. The variable names Pop and Sales are entered as the column headings on the worksheet. In subsequent steps, we refer to the data by using the variable

names Pop and Sales or the column indicators C1 and C2. The following steps describe how to use Minitab to produce the regression results shown in Figure 14.10.

Step 1. Select the **Stat** menu
Step 2. Select the **Regression** menu
Step 3. Choose **Regression**
Step 4. When the Regression dialog box appears:
Enter Sales in the **Response** box
Enter Pop in the **Predictors** box
Click the **Options** button
When the Regression-Options dialog box appears:
Enter 10 in the **Prediction intervals for new observations** box
Click **OK**
When the Regression dialog box reappears:
Click **OK**

The Minitab regression dialog box provides additional capabilities that can be obtained by selecting the desired options. For instance, to obtain a residual plot that shows the predicted value of the dependent variable $\hat{y}$ on the horizontal axis and the standardized residual values on the vertical axis, step 4 would be as follows:

Step 4. When the Regression dialog box appears:
Enter Sales in the **Response** box
Enter Pop in the **Predictors** box
Click the **Graphs** button
When the Regression-Graphs dialog box appears:
Select **Standardized** under Residuals for Plots
Select **Residuals versus fits** under Residual Plots
Click **OK**
When the Regression dialog box reappears:
Click **OK**

Appendix 14.4 Regression Analysis with Excel

WEB file
Armand's

In this appendix we will illustrate how Excel's Regression tool can be used to perform the regression analysis computations for the Armand's Pizza Parlors problem. Refer to Figure 14.23 as we describe the steps involved. The labels Restaurant, Population, and Sales are entered into cells A1:C1 of the worksheet. To identify each of the 10 observations, we entered the numbers 1 through 10 into cells A2:A11. The sample data are entered into cells B2:C11. The following steps describe how to use Excel to produce the regression results.

Step 1. Click the **Data** tab on the Ribbon
Step 2. In the **Analysis** group, click **Data Analysis**
Step 3. Choose **Regression** from the list of Analysis Tools
Step 4. Click **OK**
Step 5. When the Regression dialog box appears:
Enter C1:C11 in the **Input Y Range** box
Enter B1:B11 in the **Input X Range** box
Select **Labels**
Select **Confidence Level**
Enter 99 in the **Confidence Level** box
Select **Output Range**
Enter A13 in the **Output Range** box
(Any upper-left-hand corner cell indicating where the output is to begin may be entered here.)
Click **OK**

FIGURE 14.23 EXCEL SOLUTION TO THE ARMAND'S PIZZA PARLORS PROBLEM

	A	B	C	D	E	F	G	H	I	J
1	**Restaurant**	**Population**	**Sales**							
2	1	2	58							
3	2	6	105							
4	3	8	88							
5	4	8	118							
6	5	12	117							
7	6	16	137							
8	7	20	157							
9	8	20	169							
10	9	22	149							
11	10	26	202							
12										
13	SUMMARY OUTPUT									
14										
15	*Regression Statistics*									
16	Multiple R	0.9501								
17	R Square	0.9027								
18	Adjusted R Square	0.8906								
19	Standard Error	13.8293								
20	Observations	10								
21										
22	ANOVA									
23		*df*	*SS*	*MS*	*F*	*Significance F*				
24	Regression	1	14200	14200	74.2484	2.55E-05				
25	Residual	8	1530	191.25						
26	Total	9	15730							
27										
28		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*	
29	Intercept	60	9.2260	6.5033	0.0002	38.7247	81.2753	29.0431	90.9569	
30	Population	5	0.5803	8.6167	2.55E-05	3.6619	6.3381	3.0530	6.9470	
31										
32										
33										
34										

The first section of the output, titled *Regression Statistics,* contains summary statistics such as the coefficient of determination (R Square). The second section of the output, titled ANOVA, contains the analysis of variance table. The last section of the output, which is not titled, contains the estimated regression coefficients and related information. We will begin our discussion of the interpretation of the regression output with the information contained in cells A28:I30.

Interpretation of Estimated Regression Equation Output

The y intercept of the estimated regression line, $b_0 = 60$, is shown in cell B29, and the slope of the estimated regression line, $b_1 = 5$, is shown in cell B30. The label Intercept in cell A29 and the label Population in cell A30 are used to identify these two values.

In Section 14.5 we showed that the estimated standard deviation of b_1 is $s_{b_1} = .5803$. Note that the value in cell C30 is .5803. The label Standard Error in cell C28 is Excel's way of indicating that the value in cell C30 is the standard error, or standard deviation, of b_1. Recall that the t test for a significant relationship required the computation of the t statistic, $t = b_1/s_{b_1}$. For the Armand's data, the value of t that we computed was $t = 5/.5803 = 8.62$. The label in cell D28, *t Stat,* reminds us that cell D30 contains the value of the t test statistic.

The value in cell E30 is the *p*-value associated with the *t* test for significance. Excel has displayed the *p*-value in cell E30 using scientific notation. To obtain the decimal value, we move the decimal point 5 places to the left, obtaining a value of .0000255. Because the *p*-value = .0000255 $< \alpha = .01$, we can reject H_0 and conclude that we have a significant relationship between student population and quarterly sales.

The information in cells F28:I30 can be used to develop confidence interval estimates of the *y* intercept and slope of the estimated regression equation. Excel always provides the lower and upper limits for a 95% confidence interval. Recall that in step 4 we selected Confidence Level and entered 99 in the Confidence Level box. As a result, Excel's Regression tool also provides the lower and upper limits for a 99% confidence interval. The value in cell H30 is the lower limit for the 99% confidence interval estimate of β_1 and the value in cell I30 is the upper limit. Thus, after rounding, the 99% confidence interval estimate of β_1 is 3.05 to 6.95. The values in cells F30 and G30 provide the lower and upper limits for the 95% confidence interval. Thus, the 95% confidence interval is 3.66 to 6.34.

Interpretation of ANOVA Output

The information in cells A22:F26 is a summary of the analysis of variance computations. The three sources of variation are labeled Regression, Residual, and Total. The label *df* in cell B23 stands for degrees of freedom, the label *SS* in cell C23 stands for sum of squares, and the label *MS* in cell D23 stands for mean square.

In Section 14.5 we stated that the mean square error, obtained by dividing the error or residual sum of squares by its degrees of freedom, provides an estimate of σ^2. The value in cell D25, 191.25, is the mean square error for the Armand's regression output. In Section 14.5 we showed that an *F* test could also be used to test for significance in regression. The value in cell F24, .0000255, is the *p*-value associated with the *F* test for significance. Because the *p*-value = .0000255 $< \alpha = .01$, we can reject H_0 and conclude that we have a significant relationship between student population and quarterly sales. The label Excel uses to identify the *p*-value for the *F* test for significance, shown in cell F23, is *Significance F.*

The label Significance F may be more meaningful if you think of the value in cell F24 as the observed level of significance for the F test.

Interpretation of Regression Statistics Output

The coefficient of determination, .9027, appears in cell B17; the corresponding label, R Square, is shown in cell A17. The square root of the coefficient of determination provides the sample correlation coefficient of .9501 shown in cell B16. Note that Excel uses the label Multiple R (cell A16) to identify this value. In cell A19, the label Standard Error is used to identify the value of the standard error of the estimate shown in cell B19. Thus, the standard error of the estimate is 13.8293. We caution the reader to keep in mind that in the Excel output, the label Standard Error appears in two different places. In the Regression Statistics section of the output, the label Standard Error refers to the estimate of σ. In the Estimated Regression Equation section of the output, the label *Standard Error* refers to s_{b_1}, the standard deviation of the sampling distribution of b_1.

Appendix 14.5 Regression Analysis Using StatTools

WEB file Armand's

In this appendix we show how StatTools can be used to perform the regression analysis computations for the Armand's Pizza Parlors problem. Begin by using the Data Set Manager to create a StatTools data set for these data using the procedure described in the appendix in Chapter 1. The following steps describe how StatTools can be used to provide the regression results.

Step 1. Click the **StatTools** tab on the Ribbon
Step 2. In the **Analyses** group, click **Regression and Classification**
Step 3. Choose the **Regression** option
Step 4. When the StatTools—Regression dialog box appears:
Select **Multiple** in the **Regression Type** box
In the **Variables** section,
Click the **Format button** and select **Unstacked**
In the column labeled **I** select **Population**
In the column labeled **D** select **Sales**
Click **OK**

The regression analysis output will appear in a new worksheet.

Note that in step 4 we selected Multiple in the Regression Type box. In StatTools, the Multiple option is used for both simple linear regression and multiple regression. The StatTools—Regression dialog box contains a number of more advanced options for developing prediction interval estimates and producing residual plots. The StatTools Help facility provides information on using all of these options.