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

$$E(y) = \beta_0 + \beta_1 x \qquad (14.2)$$

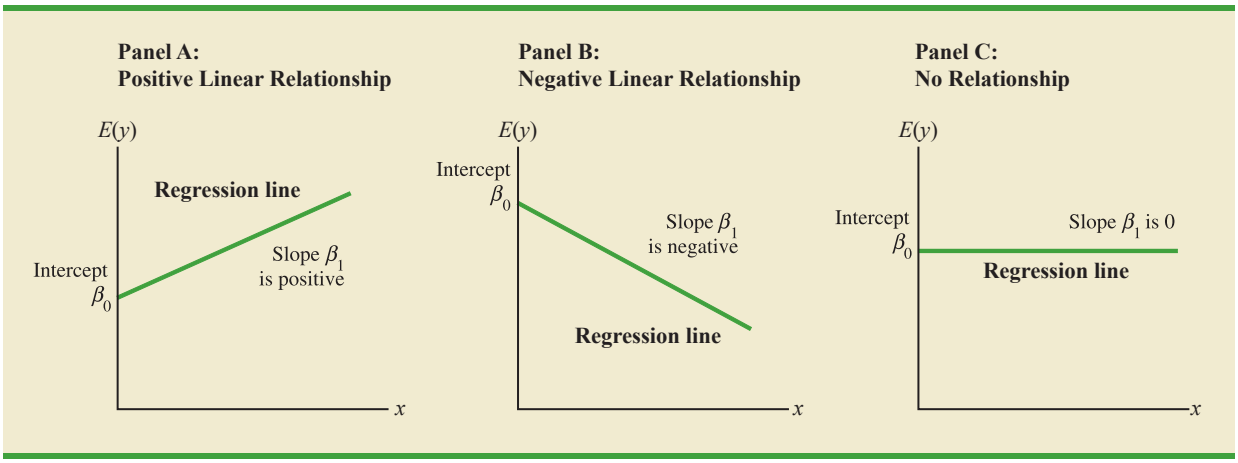
The graph of the simple linear regression equation is a straight line; β_0 is the y-intercept of the regression line, β_1 is the slope, and $E(y)$ is the mean or expected value of y for a given value of x.

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

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

Figure 14.2 provides a summary of the estimation process for simple linear regression.

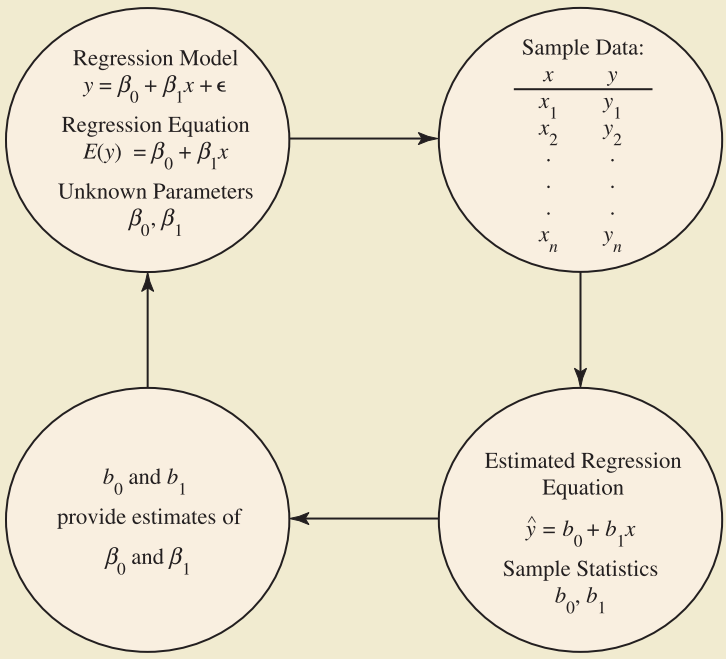
The graph of the estimated simple linear regression equation is called the *estimated regression line*; b_0 is the y-intercept and b_1 is the slope. In the next section, we show how the least squares method can be used to compute the values of b_0 and b_1 in the estimated regression equation.

The value of $\hat{y}$ provides both a point estimate of $E(y)$ for a given value of x and a prediction of an individual value of y for a given value of x.

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 they are considering building near Talbot College, a school with 10,000 students. As it turns out, the best predictor 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).

FIGURE 14.2 THE ESTIMATION PROCESS IN 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.



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.

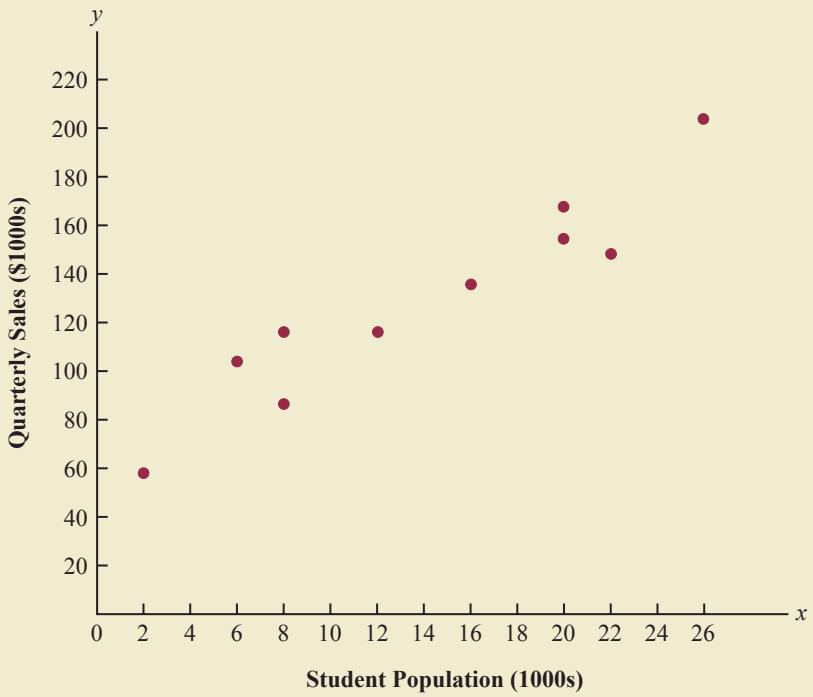
What preliminary conclusions can be drawn from Figure 14.3? Quarterly sales appear to be higher at campuses with larger student populations. In addition, for these data the relationship between the size of the student population and quarterly sales appears to be approximated by a straight line; indeed,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

$\hat{y}_i$ = predicted 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 predicted value of sales for restaurant i, every restaurant in the sample will have an observed value of sales y_i and a predicted 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 predicted 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 predicted 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
$\hat{y}_i$ = predicted 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'S 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.

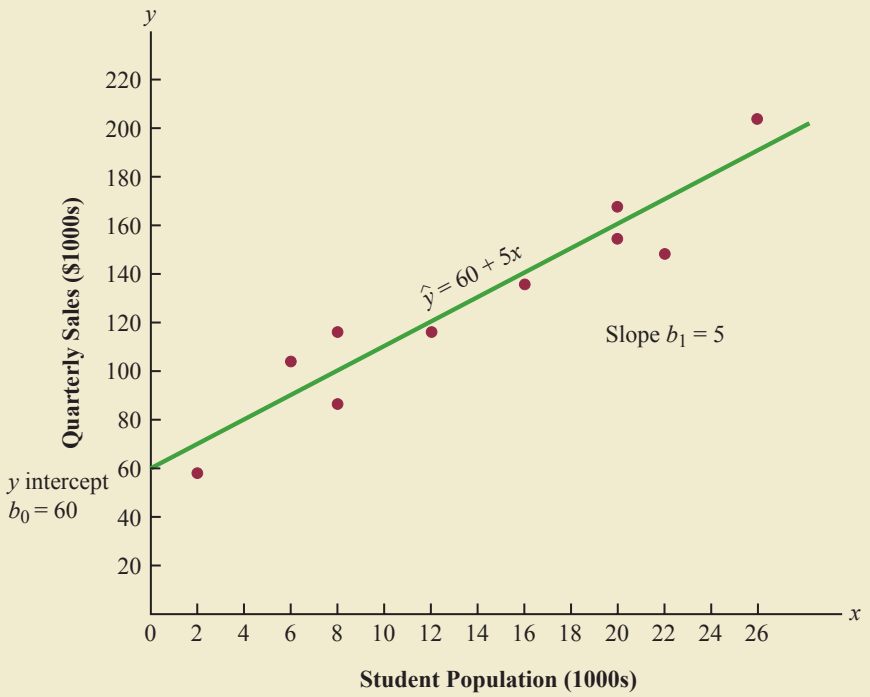
Using the estimated regression equation to make predictions outside the range of the values of the independent variable should be done with caution because outside that range we cannot be sure that the same relationship is valid.

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 predicted 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 give the percentage of women working in five companies in the retail and trade industry. The percentage of management jobs held by women in each company is also shown.

%Working	67	45	73	54	61
% Management	49	21	65	47	33

a. Develop a scatter diagram for these data with the percentage of women working in the company 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 the percentage of women working in the company and the percentage of management jobs held by women in that company.
d. Develop the estimated regression equation by computing the values of b_0 and b_1.
e. Predict the percentage of management jobs held by women in a company that has 60% women employees.

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 National Football League (NFL) records a variety of performance data for individuals and teams. To investigate the importance of passing on the percentage of games won by a team, the following data show the average number of passing yards per attempt (Yds/Att) and the percentage of games won (WinPct) for a random sample of 10 NFL teams for the 2011 season (NFL website, February 12, 2012).

WEB file
NFLPassing

Team	Yds/Att	WinPct
Arizona Cardinals	6.5	50
Atlanta Falcons	7.1	63
Carolina Panthers	7.4	38
Chicago Bears	6.4	50
Dallas Cowboys	7.4	50
New England Patriots	8.3	81
Philadelphia Eagles	7.4	50
Seattle Seahawks	6.1	44
St. Louis Rams	5.2	13
Tampa Bay Buccaneers	6.2	25

a. Develop a scatter diagram with the number of passing yards per attempt on the horizontal axis and the percentage of games won 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 percentage of games won given the average number of passing yards per attempt.
d. Provide an interpretation for the slope of the estimated regression equation.

e. For the 2011 season, the average number of passing yards per attempt for the Kansas City Chiefs was 6.2. Use the estimated regression equation developed in part (c) to predict the percentage of games won by the Kansas City Chiefs. (*Note:* For the 2011 season the Kansas City Chiefs record was 7 wins and 9 losses.) Compare your prediction to the actual percentage of games won by the Kansas City Chiefs.

7. A sales manager collected the following data on annual sales for new customer accounts and the number of years of experience for a sample of 10 salespersons.

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. The American Association of Individual Investors (AAII) On-Line Discount Broker Survey polls members on their experiences with discount brokers. As part of the survey, members were asked to rate the quality of the speed of execution with their broker as well as provide an overall satisfaction rating for electronic trades. Possible responses (scores) were no opinion (0), unsatisfied (l), somewhat satisfied (2), satisfied (3), and very satisfied (4). For each broker summary scores were computed by calculating a weighted average of the scores provided by each respondent. A portion of the survey results follow (AAII website, February 7, 2012).

WEB file
BrokerRatings

Brokerage	Speed	Satisfaction
Scottrade, Inc.	3.4	3.5
Charles Schwab	3.3	3.4
Fidelity Brokerage Services	3.4	3.9
TD Ameritrade	3.6	3.7
E*Trade Financial	3.2	2.9
Vanguard Brokerage Services	3.8	2.8
USAA Brokerage Services	3.8	3.6
Thinkorswim	2.6	2.6
Wells Fargo Investments	2.7	2.3
Interactive Brokers	4.0	4.0
Zecco.com	2.5	2.5

a. Develop a scatter diagram for these data with the speed of execution 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. Provide an interpretation for the slope of the estimated regression equation.
e. Suppose Zecco.com developed new software to increase their speed of execution rating. If the new software is able to increase their speed of execution rating from the current value of 2.5 to the average speed of execution rating for the other 10 brokerage firms that were surveyed, what value would you predict for the overall satisfaction rating?

9. Using a global-positioning-system (GPS)-based navigator for your car, you enter a destination and the system will plot a route, give spoken turn-by-turn directions, and show your progress along the route. Today, even budget units include features previously available only on more expensive models. *Consumer Reports* conducted extensive tests of GPS-based navigators and developed an overall rating based on factors such as ease of use, driver information, display, and battery life. The following data show the price and rating for a sample of 20 GPS units with a 4.3-inch screen that *Consumer Reports* tested (*Consumer Reports* website, April 17, 2012).

WEB file
GPS

Brand and Model	Price ($)	Rating
Garmin Nuvi 3490LMT	400	82
Garmin Nuvi 3450	330	80
Garmin Nuvi 3790T	350	77
Garmin Nuvi 3790LMT	400	77
Garmin Nuvi 3750	250	74
Garmin Nuvi 2475LT	230	74
Garmin Nuvi 2455LT	160	73
Garmin Nuvi 2370LT	270	71
Garmin Nuvi 2360LT	250	71
Garmin Nuvi 2360LMT	220	71
Garmin Nuvi 755T	260	70
Motorola Motonab TN565t	200	68
Motorola Motonab TN555	200	67
Garmin Nuvi 1350T	150	65
Garmin Nuvi 1350LMT	180	65
Garmin Nuvi 2300	160	65
Garmin Nuvi 1350	130	64
Tom Tom VIA 1435T	200	62
Garmin Nuvi 1300	140	62
Garmin Nuvi 1300LM	180	62

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. Predict the rating for a GPS system with a 4.3-inch screen that has a price of $200.

10. On March 31, 2009, Ford Motor Company's shares were trading at a 26-year low of $2.63. Ford's board of directors gave the CEO a grant of options and restricted shares with an estimated value of $16 million. On April 26, 2011, the price of a share of Ford had increased to $15.58, and the CEO's grant was worth $202.8 million, a gain in value of $186.8 million. The following table shows the share price in 2009 and 2011 for 10 companies, the stock-option and share grants to the CEOs in late 2008 and 2009, and the value of the options and grants in 2011. Also shown are the percentage increases in the stock price and the percentage gains in the options values (*The Wall Street Journal,* April 27, 2011).

WEB file
CEOGrants

Company	Stock Price 2009 ($)	Stock Price 2011 ($)	% Increase in Stock Price	Options and Grants Value 2009 ($ millions)	Options and Grants Value 2011 ($ millions)	% Gain in Options Value
Ford Motor	2.63	15.58	492	16.0	202.8	1168
Abercrombie & Fitch	23.80	70.47	196	46.2	196.1	324
Nabors Industries	9.99	32.06	221	37.2	132.2	255
Starbucks	9.99	32.06	221	12.4	75.9	512
Salesforce.com	32.73	137.61	320	7.8	67.0	759
Starwood Hotels	12.70	60.28	375	5.8	57.1	884
Caterpillar	27.96	111.94	300	4.0	47.5	1088
Oracle	18.07	34.97	94	61.9	97.5	58
Capital One	12.24	54.61	346	6.0	40.6	577
Dow Chemical	8.43	39.97	374	5.0	38.8	676

a. Develop a scatter diagram for these data with the percentage increase in the stock price 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. Provide an interpretation for the slope of the estimated regression equation.
e. Do the rewards for the CEO appear to be based on performance increases as measured by the stock price?

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. Concur Technologies, Inc., is a large expense-management company located in Redmond, Washington. *The Wall Street Journal* asked Concur to examine the data from 8.3 million expense reports to provide insights regarding business travel expenses. Their analysis of the data showed that New York was the most expensive city, with an average daily hotel room rate of $198 and an average amount spent on entertainment, including group meals and tickets for shows, sports, and other events, of $172. In comparison, the U.S. averages for these two categories were $89 for the room rate and $99 for entertainment. The following table shows the average daily hotel room rate and the amount spent on entertainment for a random sample of 9 of the 25 most visited U.S. cities (*The Wall Street Journal,* August 18, 2011).

WEB file
BusinessTravel

City	Room Rate ($)	Entertainment ($)
Boston	148	161
Denver	96	105
Nashville	91	101
New Orleans	110	142
Phoenix	90	100
San Diego	102	120
San Francisco	136	167
San Jose	90	140
Tampa	82	98

a. Develop a scatter diagram for these data with the room rate 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. Provide an interpretation for the slope of the estimated regression equation.
e. The average room rate in Chicago is $128, considerably higher than the U.S. average. Predict the entertainment expense per day for Chicago.

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. Predict the 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. Predict 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 predicted 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

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

The value of SSE is a measure of the error in using the estimated regression equation to predict 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

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

regression equation, we find that the predicted value of quarterly sales for restaurant 1 is $\hat{y}_1 = 60 + 5(2) = 70$. Thus, the error in using $\hat{y}_1$ to predict 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 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 predict 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 predict sales. The corresponding sum of squares, called the *total sum of squares,* is denoted SST.

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	5184
2	6	105	−25	625
3	8	88	−42	1764
4	8	118	−12	144
5	12	117	−13	169
6	16	137	7	49
7	20	157	27	729
8	20	169	39	1521
9	22	149	19	361
10	26	202	72	5184
			SST =	15,730

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* $y = \bar{y}$.

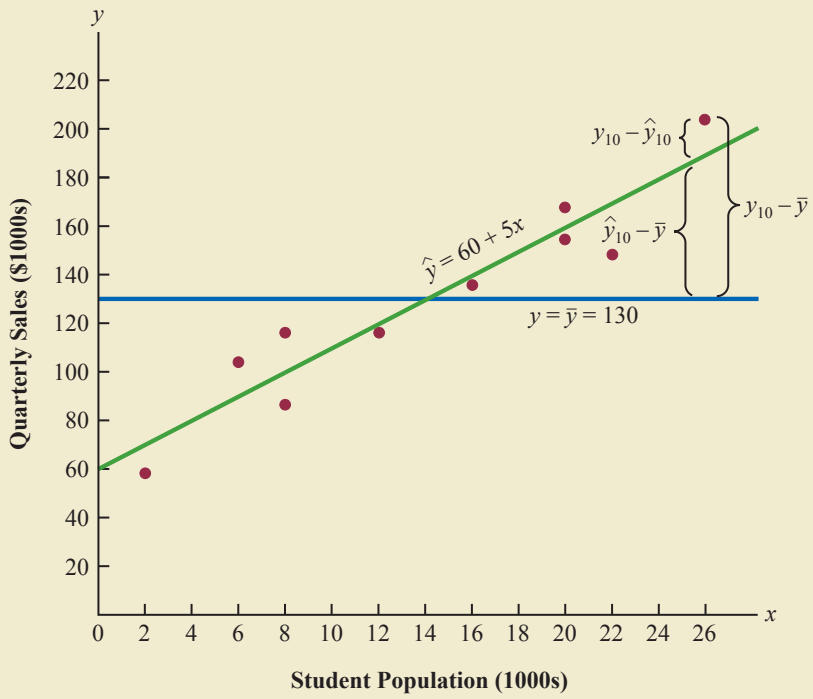
In Figure 14.5 we show the estimated regression line $\hat{y} = 60 + 5x$ and the line corresponding to $\bar{y} = 130$. Note that the points cluster more closely around the estimated regression line than they do about the line $\bar{y} = 130$. For example, for the 10th restaurant in the sample we see that the error is much larger when $\bar{y} = 130$ is used to predict 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

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

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

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

where

b_1 = the slope of the estimated regression equation $\hat{y} = b_0 + b_1x$

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 show the brand, price ($), and the overall score for six stereo headphones that were tested by *Consumer Reports* (*Consumer Reports* website, March 5, 2012). The overall score is based on sound quality and effectiveness of ambient noise reduction. Scores range from 0 (lowest) to 100 (highest). The estimated regression equation for these data is $\hat{y} = 23.194 + .318x$, where x = price ($) and y = overall score.

Brand	Price ($)	Score
Bose	180	76
Skullcandy	150	71
Koss	95	61
Phillips/O'Neill	70	56
Denon	70	40
JVC	35	26

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. *Bicycling,* the world's leading cycling magazine, reviews hundreds of bicycles throughout the year. Their "Road-Race" category contains reviews of bikes used by riders primarily interested in racing. One of the most important factors in selecting a bike for racing is the weight of the bike. The following data show the weight (pounds) and price ($) for 10 racing bikes reviewed by the magazine (*Bicycling* website, March 8, 2012).

WEB file
RacingBicycles

Brand	Weight	Price ($)
FELT F5	17.8	2100
PINARELLO Paris	16.1	6250
ORBEA Orca GDR	14.9	8370
EDDY MERCKX EMX-7	15.9	6200
BH RC1 Ultegra	17.2	4000
BH Ultralight 386	13.1	8600
CERVELO S5 Team	16.2	6000
GIANT TCR Advanced 2	17.1	2580
WILIER TRIESTINA Gran Turismo	17.6	3400
SPECIALIZED S-Works Amira SL4	14.1	8000

a. Use the data to develop an estimated regression equation that could be used to estimate the price for a bike given the weight.
b. Compute r^2. Did the estimated regression equation provide a good fit?
c. Predict the price for a bike that weighs 15 pounds.

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. Predict the 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

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

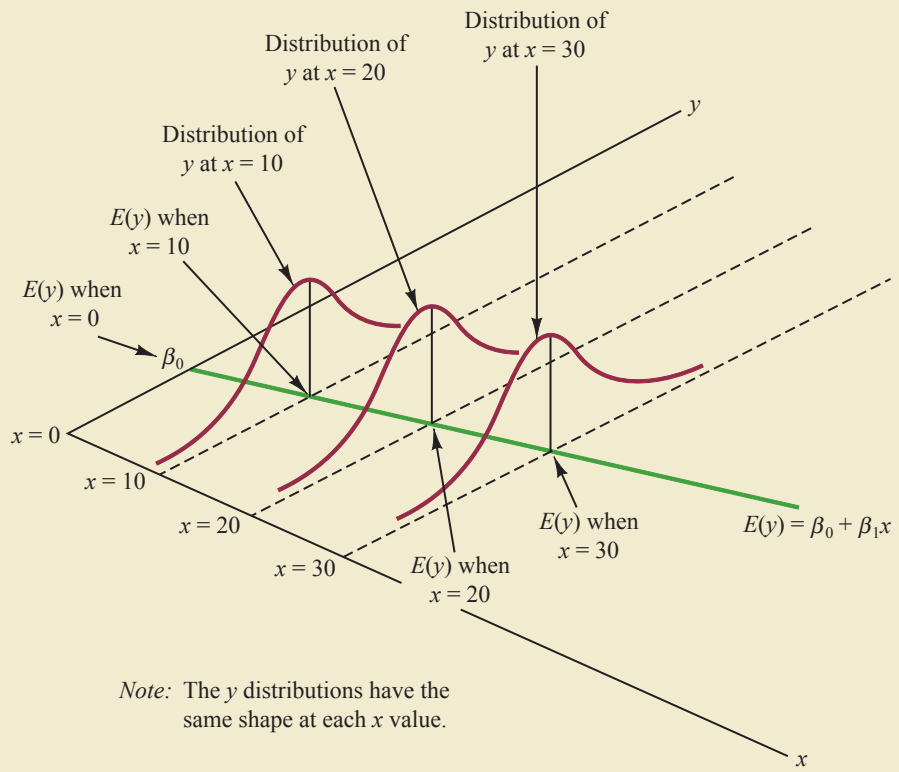
1. The error term ϵ is a random variable with a mean or expected value of zero; that is, $E(\epsilon) = 0$.
   Implication: β_0 and β_1 are constants, therefore $E(\beta_0) = \beta_0$ and $E(\beta_1) = \beta_1$; thus, for a given value of x, the expected value of y is

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

   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 for all values of x.
   Implication: Because y is a linear function of ϵ, y is also a normally distributed random variable for all values of x.

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 $E(y)$.

FIGURE 14.6 ASSUMPTIONS FOR THE REGRESSION MODEL



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 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)

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

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

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

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 (Table 2 of Appendix D) 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

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

REJECTION RULE

p-value approach:	Reject H_0 if p-value α
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

$$F = \frac{\text{MSR}}{\text{MSE}} \tag{14.21}$$

REJECTION RULE

p-value approach:	Reject H_0 if *p*-value α
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

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.

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

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	$n - 1$			

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			

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 $F = \text{MSR}/\text{MSE}$, 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.

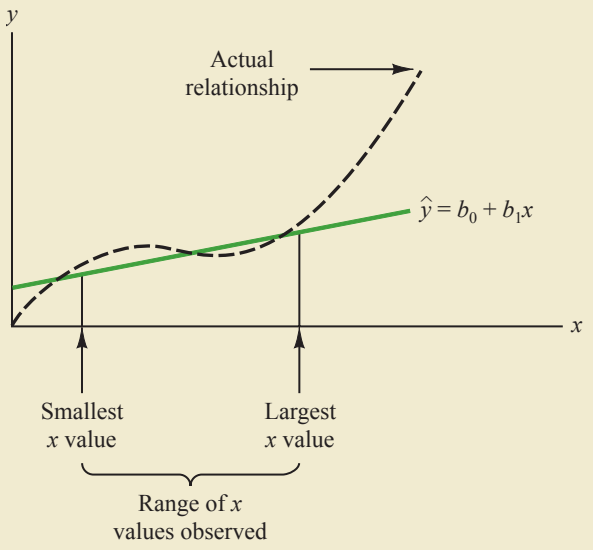
Some Cautions About the Interpretation of Significance Tests

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

FIGURE 14.7 EXAMPLE OF A LINEAR APPROXIMATION OF A NONLINEAR RELATIONSHIP



linear approximation provided by $\hat{y} = b_0 + b_1 x$ 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 price ($) and the overall score for six stereo headphones tested by *Consumer Reports* were as follows (*Consumer Reports* website, March 5, 2012).

Brand	Price ($)	Score
Bose	180	76
Skullcandy	150	71
Koss	95	61
Phillips/O'Neill	70	56
Denon	70	40
JVC	35	26

a. Does the t test indicate a significant relationship between price and the overall score? 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 for these data.

27. The number of megapixels in a digital camera is one of the most important factors in determining picture quality. But, do digital cameras with more megapixels cost more? The following data show the number of megapixels and the price ($) for 10 digital cameras (*Consumer Reports,* March 2009).

WEB file
DigitalCameras

Brand and Model	Megapixels	Price ($)
Canon PowerShot SD1100 IS	8	180
Casio Exilim Card EX-510	10	200
Sony Cyber-shot DSC-T70	7	230
Pentax Optio M50	8	120
Canon PowerShot G10	15	470
Canon PowerShot A590 IS	8	140
Canon PowerShot E1	10	180
Fujifilm FinePix F00FD	12	310
Sony Cyber-shot DSC-W170	10	250
Canon PowerShot A470	7	110

a. Use these data to develop an estimated regression equation that can be used to predict the price of a digital camera given the number of megapixels.
b. At the .05 level of significance, are the number of megapixels and the price related? Explain.
c. Would you feel comfortable using the estimated regression equation developed in part (a) to predict the price of a digital camera given the number of megapixels? Explain.
d. The Canon Power Shot S95 digital camera has 10 megapixels. Predict the price of this camera using the estimated regression equation developed in part (a).

WEB file
BrokerRatings

28. In exercise 8 ratings data on x = the quality of the speed of execution and y = overall satisfaction with electronic trades provided the estimated regression equation $\hat{y} = .2046 + .9077x$. At the .05 level of significance, test whether speed of execution and overall satisfaction 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?

WEB file
RacingBicycles

31. In exercise 20, data on x = weight (pounds) and y = price ($) for 10 road-racing bikes provided the estimated regression equation $\hat{y} = 28{,}574 - 1439x$. (*Bicycling* website, March 8, 2012). For these data SSE = 7,102,922.54 and SST = 52,120,800. Use the F test to determine whether the weight for a bike and the price 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.

For the Armand's Pizza Parlors example, the estimated regression equation is $\hat{y} = 60 + 5x$. At the end of Section 14.1 we stated that $\hat{y}$ can be used as a *point estimator* of $E(y)$, the mean or expected value of y for a given value of x, and as a predictor of an individual value of y. For example, suppose Armand's managers want to estimate 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 $x = 10$ (10,000 students), $\hat{y} = 60 + 5(10) = 110$. Thus, a *point estimate* of the mean quarterly sales for all restaurant locations near campuses with 10,000 students is \$110,000. In this case we are using $\hat{y}$ as the point estimator of the mean value of y when $x = 10$.

We can also use the estimated regression equation to *predict* an individual value of y for a given value of x. For example, to predict quarterly sales for a new restaurant Armand's is considering building near Talbot College, a campus with 10,000 students, we would compute $\hat{y} = 60 + 5(10) = 110$. Hence, we would predict quarterly sales of \$110,000 for such a new restaurant. In this case, we are using $\hat{y}$ as the *predictor* of y for a new observation when $x = 10$.

When we are using the estimated regression equation to estimate the mean value of y or to predict an individual value of y, it is clear that the estimate or prediction depends on the given value of x. For this reason, as we discuss in more depth the issues concerning estimation and prediction, the following notation will help clarify matters.

x^* = the given value of the independent variable x

y^* = the random variable denoting the possible values of the dependent variable y when $x = x^*$

$E(y^*)$ = the mean or expected value of the dependent variable y when $x = x^*$

$\hat{y}^* = b_0 + b_1 x^*$ = the point estimator of $E(y^*)$ and the predictor of an individual value of y^* when $x = x^*$

To illustrate the use of this notation, suppose we want to estimate the mean value of quarterly sales for *all* Armand's restaurants located near a campus with 10,000 students. For this case, $x^* = 10$ and $E(y^*)$ denotes the unknown mean value of quarterly sales for all restaurants where $x^* = 10$. Thus, the point estimate of $E(y^*)$ is provided by $\hat{y}^* = 60 + 5(10) = 110$, or \$110,000. But, using this notation, $\hat{y}^* = 110$ is also the predictor of quarterly sales for the new restaurant located near Talbot College, a school with 10,000 students.

Interval Estimation

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

Point estimators and predictors do not provide any information about the precision associated with the estimate and/or prediction. For that we must develop confidence intervals and prediction intervals. A **confidence interval** is an interval estimate of the *mean value of y* for a given value of x. A **prediction interval** is used whenever we want to *predict an individual value of y* for a new observation corresponding to a given value of x. Although the predictor of y for a given value of x is the same as the point estimator of the mean value of y for a given value of x, the interval estimates we obtain for the two cases are different. As we will show, the margin of error is larger for a prediction interval. We begin by showing how to develop an interval estimate of the mean value of y.

Confidence Interval for the Mean Value of y

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

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

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

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

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

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

The general expression for a confidence interval follows.

CONFIDENCE INTERVAL FOR $E(y^*)$

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

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

The margin of error associated with this confidence interval is $t_{\alpha/2}s_{\hat{y}^*}$.

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 $\hat{y}^* = 110$ and a margin of error of $t_{\alpha/2}s_{\hat{y}^*} = 2.306(4.95) = 11.415$, the 95% confidence interval estimate is

$$110 \pm 11.415$$

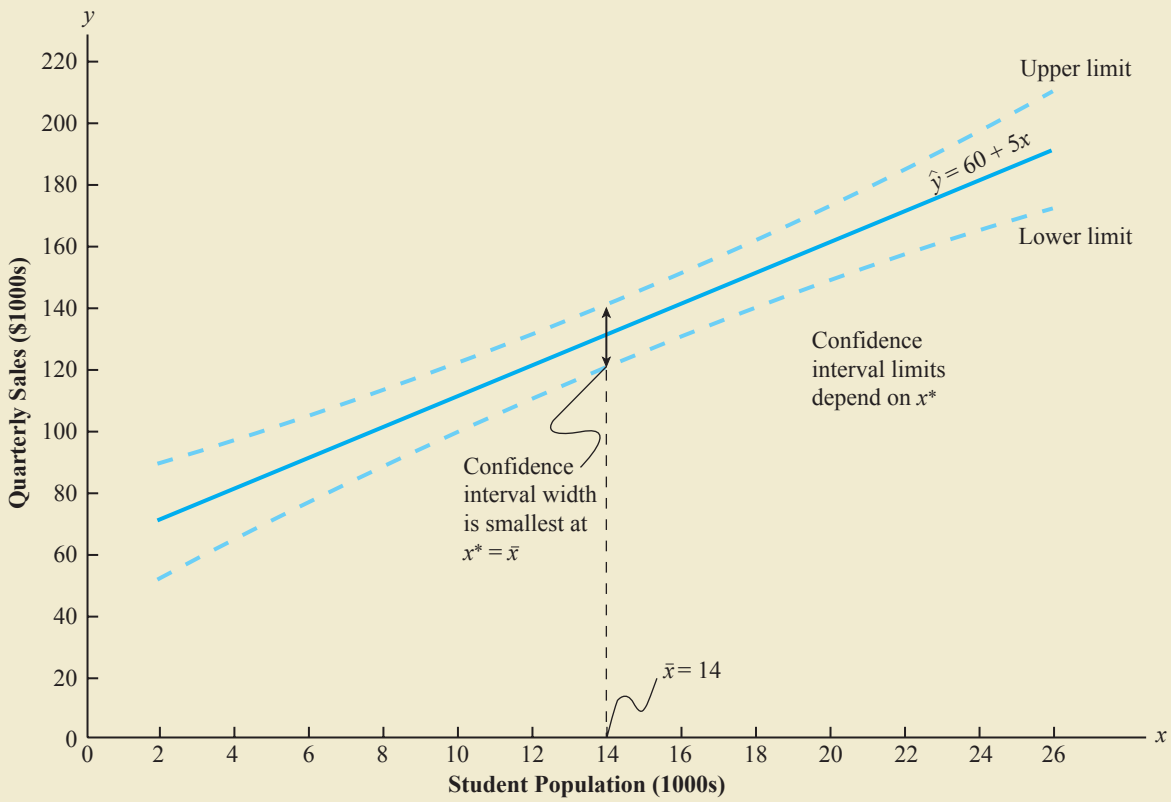
In dollars, the 95% confidence interval for the mean quarterly sales of all restaurants near campuses with 10,000 students is \$110,000 ± \$11,415. Therefore, the 95% confidence interval for the mean quarterly sales when the student population is 10,000 is \$98,585 to \$121,415.

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

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

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

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



Prediction Interval for an Individual Value of y

Instead of estimating the mean value of quarterly sales for all Armand's restaurants located near campuses with 10,000 students, suppose we want to predict quarterly sales for a new restaurant Armand's is considering building near Talbot College, a campus with 10,000 students. As noted previously, the predictor of y^*, the value of y corresponding to the given x^*, is $\hat{y}^* = b_0 + b_1x^*$. For the new restaurant located near Talbot College, $x^* = 10$ and the prediction of quarterly sales is $\hat{y}^* = 60 + 5(10) = 110$, or \$110,000. Note that the prediction of quarterly sales for the new Armand's restaurant near Talbot College is the same as the point estimate of the mean sales for all Armand's restaurants located near campuses with 10,000 students.

To develop a prediction interval, let us first determine the variance associated with using $\hat{y}^*$ as a predictor of y when $x = x^*$. This variance is made up of the sum of the following two components.

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

The formula for estimating the variance corresponding to the prediction of the value of y when $x = x^*$, denoted s^2_{pred}, is

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

Hence, an estimate of the standard deviation corresponding to the prediction of the value of y^* is

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

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

$$\begin{aligned} s_{\text{pred}} &= 13.829\sqrt{1 + \frac{1}{10} + \frac{(10 - 14)^2}{568}} \\ &= 13.829\sqrt{1.282} \\ &= 14.69 \end{aligned}$$

The general expression for a prediction interval follows.

PREDICTION INTERVAL FOR y^*

$$\hat{y}^* \pm t_{\alpha/2}s_{\text{pred}} \qquad \textbf{(14.27)}$$

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

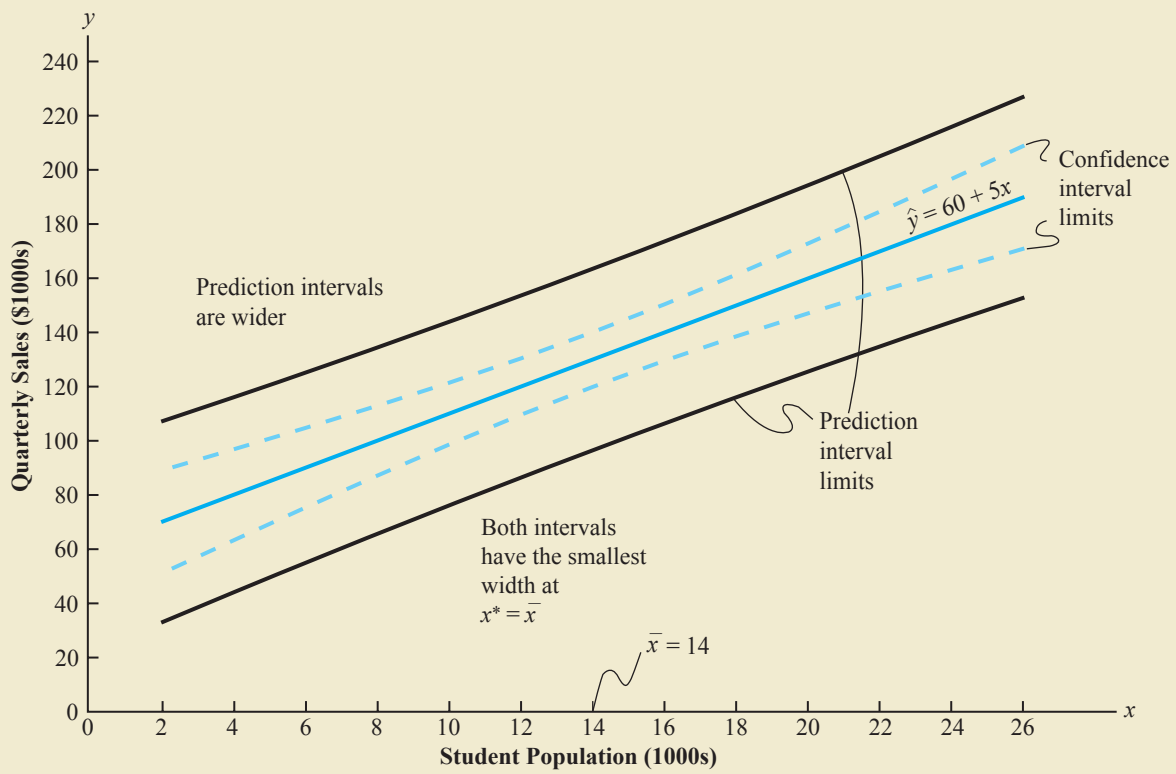
The margin of error associated with this prediction interval is $t_{\alpha/2}s_{\text{pred}}$.

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

$$110 \pm 33.875$$

In dollars, this prediction interval is \$110,000 ± \$33,875 or \$76,125 to \$143,875. Note that the prediction interval for the new restaurant located near Talbot College, a campus with 10,000 students, is wider than the confidence interval for the mean quarterly 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 predict an individual value of y.

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



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

Confidence intervals and prediction intervals are both more precise when the value of the independent variable x^* is closer to $\bar{x}$. The general shapes of confidence intervals and the wider prediction intervals are shown together in Figure 14.9.

NOTES AND COMMENTS

A prediction interval is used to predict the value of the dependent variable y for a *new observation*. As an illustration, we showed how to develop a prediction interval of quarterly sales for a new restaurant that Armand's is considering building near Talbot College, a campus with 10,000 students. The fact that the value of $x = 10$ is not one of the values of student population for the Armand's sample data in Table 14.1 is not meant to imply that prediction intervals cannot be developed for values of x in the sample data. But, for the ten restaurants that make up the data in Table 14.1, developing a prediction interval for quarterly sales for *one of these restaurants* does not make any sense because we already know the value of quarterly sales for each of these restaurants. In other words, a prediction interval only has meaning for something new, in this case a new observation corresponding to a particular value of x that may or may not equal one of the values of x in the sample.

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 $\hat{y}^*$ 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 $\hat{y}^*$ 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. 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.

GPA	Monthly Salary ($)
2.6	3600
3.4	3900
3.6	4300
3.2	3800
3.5	4200
2.9	3900

The estimated regression equation for these data is $\hat{y} = 2090.5 + 581.1x$ and MSE = 21,284.
a. Develop a point estimate of the starting salary for a student with a GPA of 3.0.
b. Develop a 95% confidence interval for the mean starting salary for all students with a 3.0 GPA.
c. Develop a 95% prediction interval for Ryan Dailey, a student with a GPA of 3.0.
d. Discuss the differences in your answers to parts (b) and (c).

WEB file
Sales

36. In exercise 7, the data on y = annual sales ($1000s) for new customer accounts and x = number of years of experience for a sample of 10 salespersons provided the estimated regression equation $\hat{y} = 80 + 4x$. For these data $\bar{x} = 7$, $\sum(x_i - \bar{x})^2 = 142$, and $s = 4.6098$.
    a. Develop a 95% confidence interval for the mean annual sales for all salespersons with nine years of experience.
    b. The company is considering hiring Tom Smart, a salesperson with nine years of experience. Develop a 95% prediction interval of annual sales for Tom Smart.
    c. Discuss the differences in your answers to parts (a) and (b).

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.

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

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

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. Predict 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. Out-of-state tuition and fees at the top graduate schools of business can be very expensive, but the starting salary and bonus paid to graduates from many of these schools can be substantial. The following data show the out-of-state tuition and fees (rounded to the nearest $1000) and the average starting salary and bonus paid to recent graduates (rounded to the nearest $1000) for a sample of 20 graduate schools of business (*U.S. News & World Report 2009 Edition America's Best Graduate Schools*).

WEB file
BusinessSchools

School	Tuition & Fees ($1000s)	Salary & Bonus ($1000s)
Arizona State University	28	98
Babson College	35	94
Cornell University	44	119
Georgetown University	40	109
Georgia Institute of Technology	30	88
Indiana University—Bloomington	35	105
Michigan State University	26	99
Northwestern University	44	123
Ohio State University	35	97
Purdue University—West Lafayette	33	96
Rice University	36	102
Stanford University	46	135
University of California—Davis	35	89
University of Florida	23	71
University of Iowa	25	78
University of Minnesota—Twin Cities	37	100
University of Notre Dame	36	95
University of Rochester	38	99
University of Washington	30	94
University of Wisconsin—Madison	27	93

a. Develop a scatter diagram with salary and bonus as the dependent variable.
b. Does there appear to be any relationship between these variables? Explain.
c. Develop an estimated regression equation that can be used to predict the starting salary and bonus paid to graduates given the cost of out-of-state tuition and fees at the school.
d. Test for a significant relationship at the .05 level of significance. What is your conclusion?
e. Did the estimated regression equation provide a good fit? Explain.
f. Suppose that we randomly select a recent graduate of the University of Virginia graduate school of business. The school has an out-of-state tuition and fees of $43,000. Predict the starting salary and bonus for this graduate.

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

(*continued*)

Helmet	Weight (oz)	Price ($)
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 predicted value of the dependent variable ($\hat{y}_i$).

RESIDUAL FOR OBSERVATION i

$$y_i - \hat{y}_i \qquad \textbf{(14.28)}$$

where

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

$$y = \beta_0 + \beta_1 x + \epsilon \qquad \textbf{(14.29)}$$

TABLE 14.7 RESIDUALS FOR ARMAND'S PIZZA PARLORS

Student Population x_i	Sales y_i	Predicted 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.

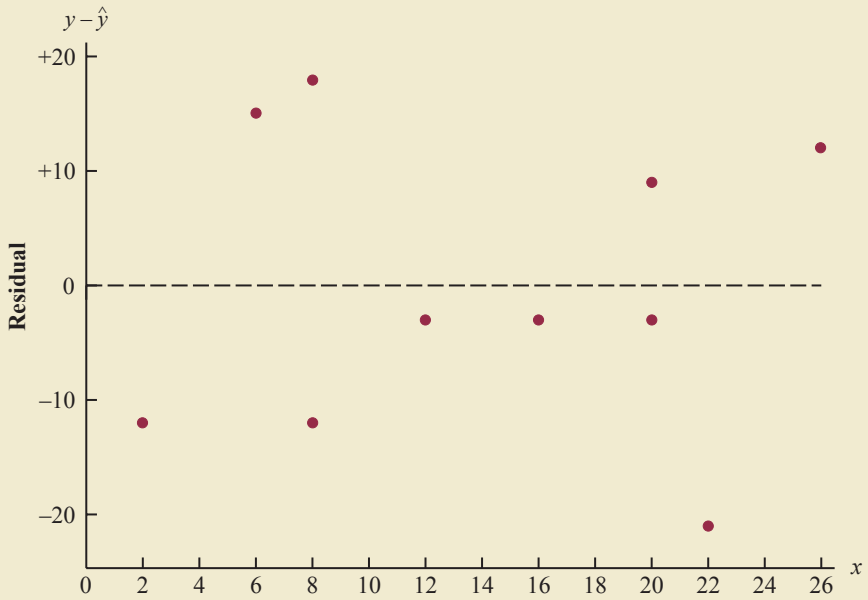
1. A plot of the residuals against values of the independent variable x
2. A plot of residuals against the predicted values of the dependent variable y
3. A standardized residual plot
4. A normal probability plot

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

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



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.

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

FIGURE 14.12 RESIDUAL PLOTS FROM THREE REGRESSION STUDIES

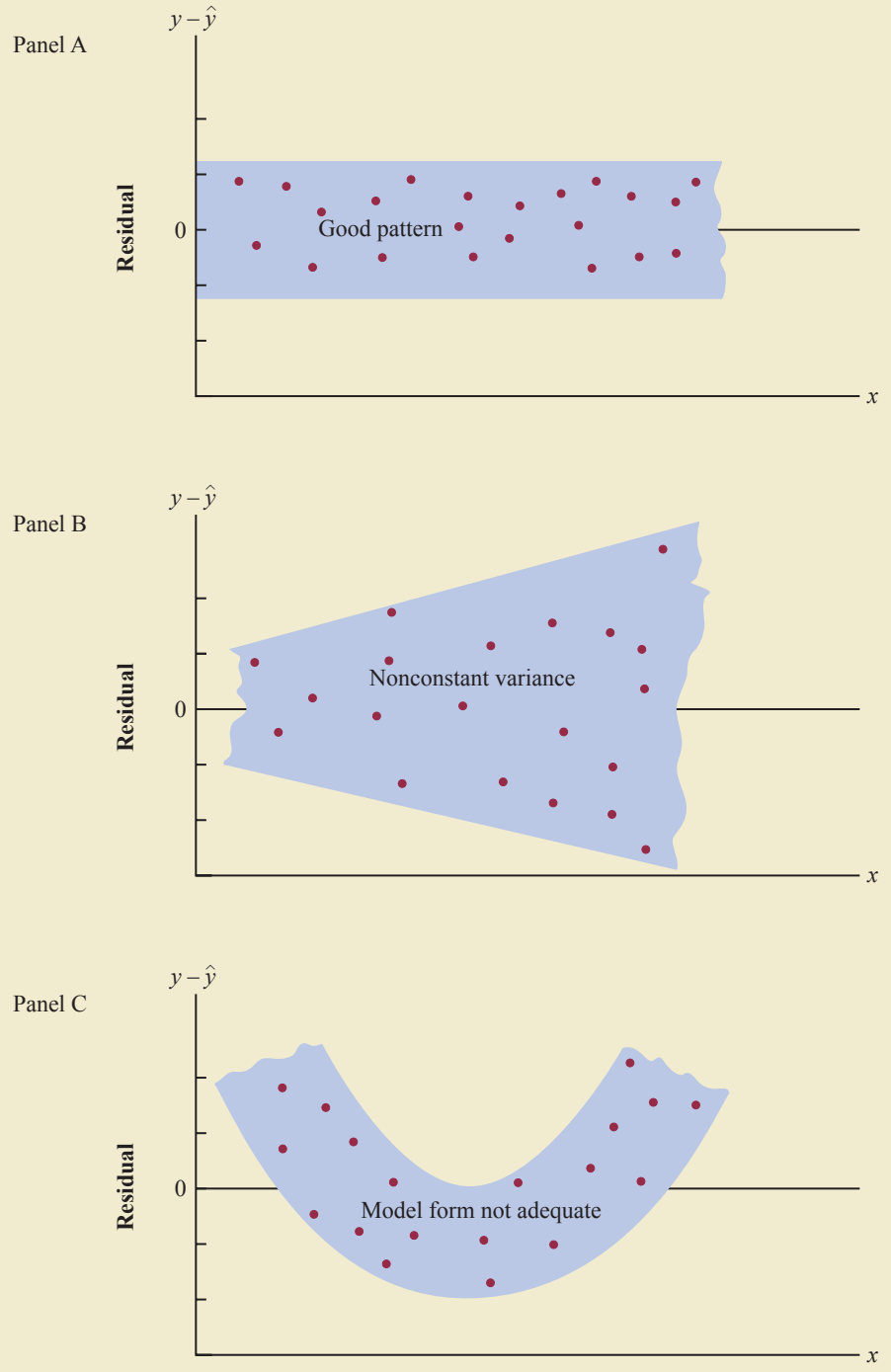
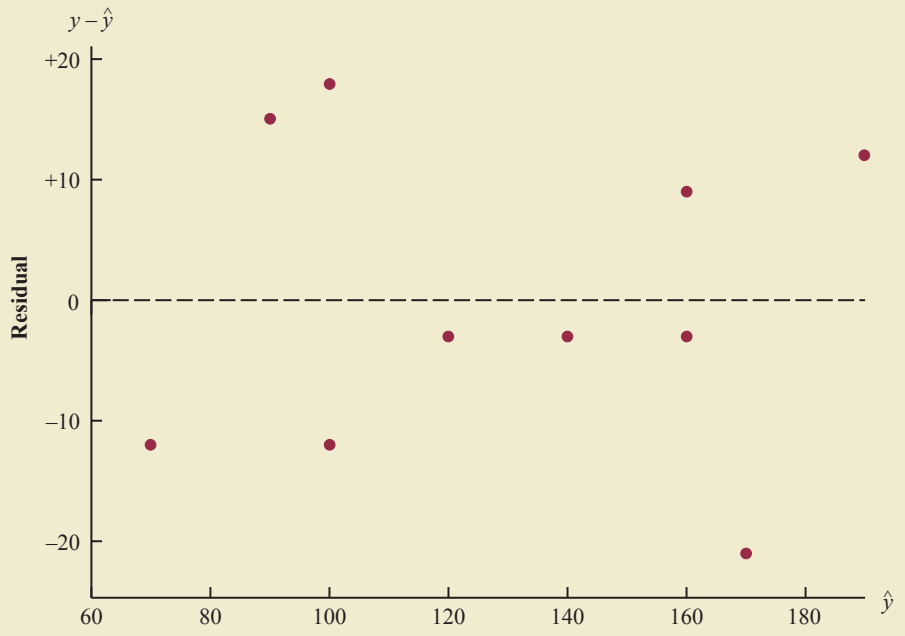


FIGURE 14.13 PLOT OF THE RESIDUALS AGAINST THE PREDICTED VALUES y FOR ARMAND'S PIZZA PARLORS



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 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]

$$s_{y_i - \hat{y}_i} = s\sqrt{1 - h_i} \qquad \textbf{(14.30)}$$

[3]This equation actually provides an estimate of the standard deviation of the ith residual, because s is used instead of σ.

where

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

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

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.

STANDARDIZED RESIDUAL FOR OBSERVATION i

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

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

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.

[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.

[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

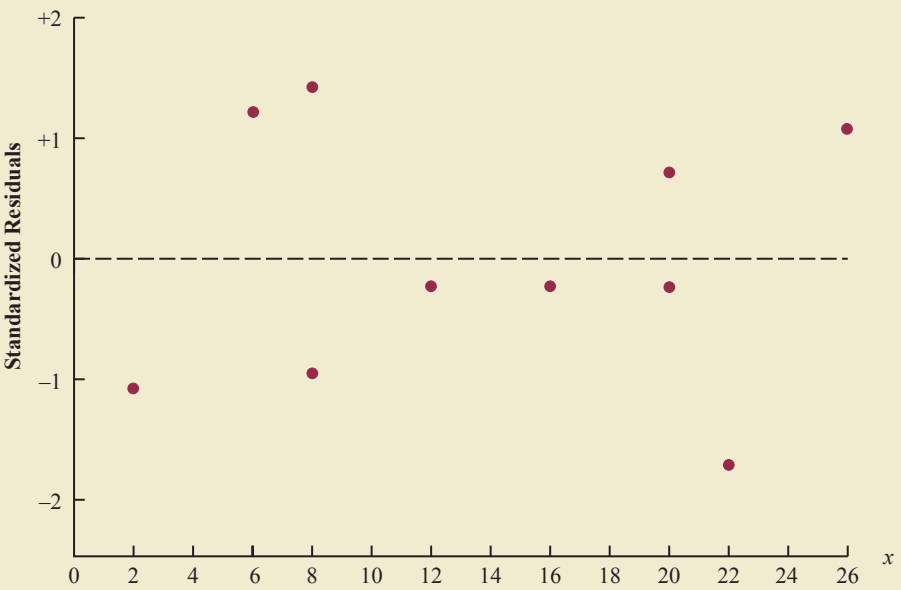


TABLE 14.9
NORMAL SCORES FOR $n = 10$

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

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

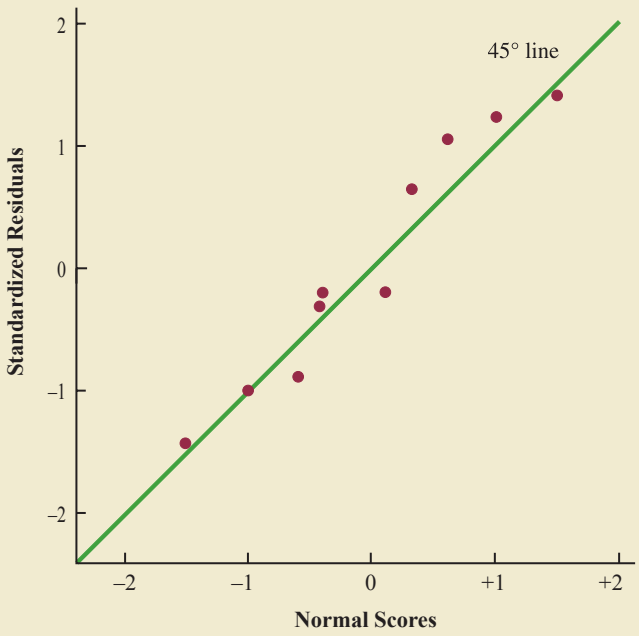
Another approach for determining the validity of the assumption that the error term has a normal distribution is the **normal probability plot**. To show how a normal probability plot is developed, we introduce the concept of *normal scores*.

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 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

FIGURE 14.15 NORMAL PROBABILITY PLOT FOR ARMAND'S PIZZA PARLORS



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.

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.

FIGURE 14.16 DATA SET WITH AN OUTLIER

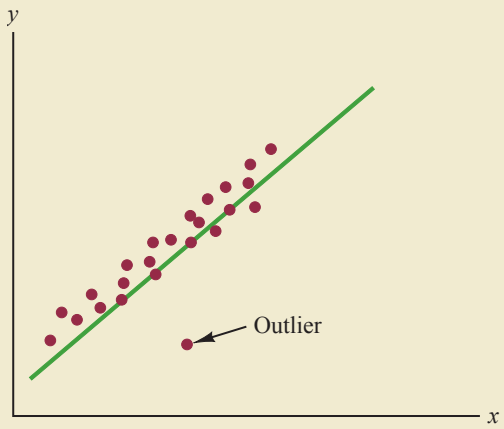


TABLE 14.11
DATA SET ILLUSTRATING THE EFFECT OF AN OUTLIER

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

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 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

FIGURE 14.17 SCATTER DIAGRAM FOR OUTLIER DATA SET

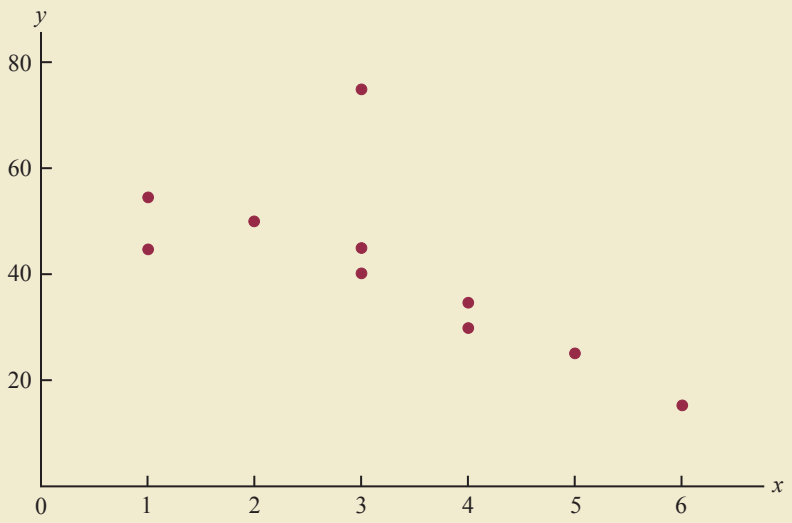


FIGURE 14.18 MINITAB OUTPUT FOR REGRESSION ANALYSIS OF THE OUTLIER DATA SET

```
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.
```

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 14.19 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 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).

FIGURE 14.20 DATA SET WITH AN INFLUENTIAL OBSERVATION