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

ESTIMATED SIMPLE LINEAR REGRESSION EQUATION

$$\hat{y} = b_0 + b_1 x \qquad (12.3)$$

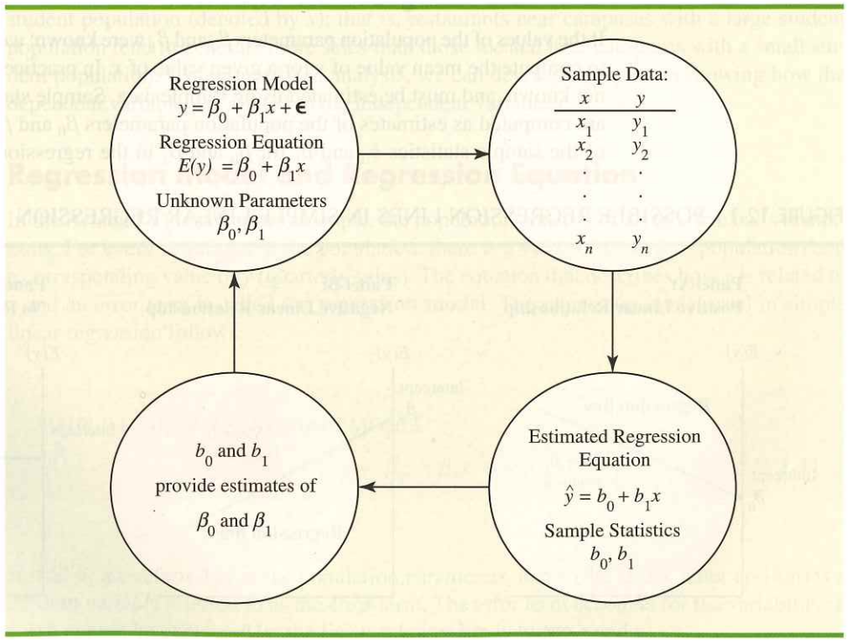
The graph of the estimated simple linear regression equation is called the *estimated regression line*; b_0 is the y intercept and b_1 is the slope. In the next section, we show how the least squares method can be used to compute the values of b_0 and b_1 in the estimated regression equation.

In general, $\hat{y}$ is the point estimator of $E(y)$, the mean value of y for a given value of x. Thus, to estimate the mean or expected value of quarterly sales for all restaurants located near campuses with 10,000 students, Armand's would substitute the value of 10,000 for x in equation (12.3). In some cases, however, Armand's may be more interested in predicting sales for one particular restaurant. For example, suppose Armand's would like to predict quarterly sales for the restaurant located near Talbot College, a school with 10,000 students. As it turns out, the best estimate of y for a given value of x is also provided by $\hat{y}$. Thus, to predict quarterly sales for the restaurant located near Talbot College, Armand's would also substitute the value of 10,000 for x in equation (12.3).

Because the value of $\hat{y}$ provides both a point estimate of $E(y)$ for a given value of x and a point estimate of an individual value of y for a given value of x, we will refer to $\hat{y}$ simply as the *estimated value of y*. Figure 12.2 provides a summary of the estimation process for simple linear regression.

FIGURE 12.2 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 population parameters of interest, and b_0 and b_1 are the sample statistics used to estimate the population 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.

12.2 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 12.1. 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 12.3 is a scatter diagram of the data in Table 12.1. 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.

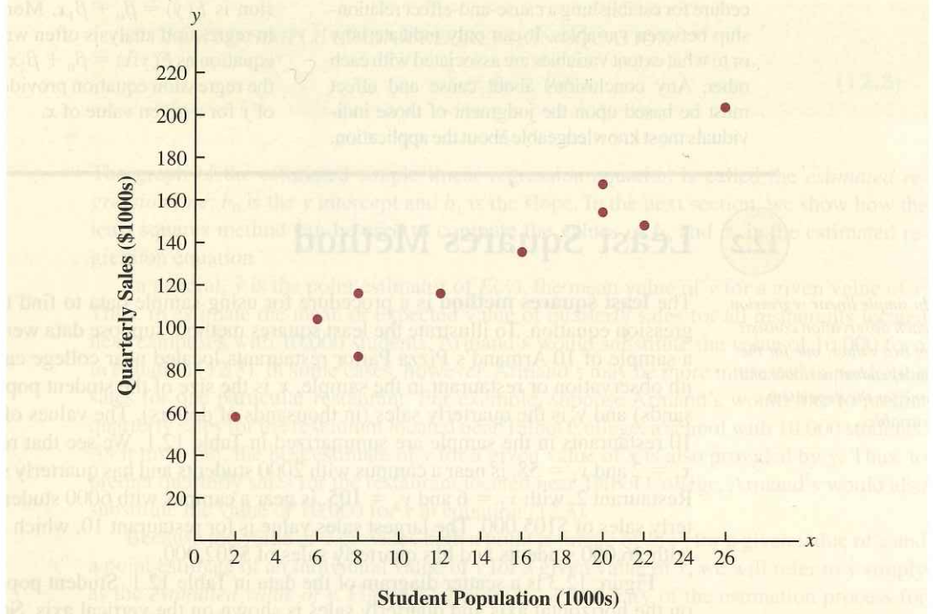
What preliminary conclusions can be drawn from Figure 12.3? Quarterly sales appear to be higher at campuses with larger student populations. In addition, for these data the relationship between the size of the student population and quarterly sales appears to be approximated by a straight line; indeed, a positive linear relationship is indicated between x

TABLE 12.1 STUDENT POPULATION AND QUARTERLY SALES DATA FOR 10 ARMAND'S PIZZA PARLORS

CD 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 12.3 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 12.1 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

$$\hat{y}_i = b_0 + b_1 x_i \qquad \textbf{(12.4)}$$

where

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

With y_i denoting the observed (actual) sales for restaurant i and $\hat{y}_i$ in equation (12.4) representing the estimated value of sales for restaurant i, every restaurant in the sample will have an observed value of sales y_i and an estimated value of sales $\hat{y}_i$. For the estimated regression line to provide a good fit to the data, we want the differences between the observed sales values and the estimated sales values to be small.

The least squares method uses the sample data to provide the values of b_0 and b_1 that minimize the *sum of the squares of the deviations* between the observed values of the dependent variable y_i and the estimated values of the dependent variable.The criterion for the least squares method is given by expression (12.5).

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

LEAST SQUARES CRITERION

$$\min \Sigma(y_i - \hat{y}_i)^2 \qquad \textbf{(12.5)}$$

where

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

Differential calculus can be used to show that the values of b_0 and b_1 that minimize expression (12.5) can be found by using equations (12.6) and (12.7).

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*

$$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} \qquad \textbf{(12.6)}$$

$$b_0 = \bar{y} - b_1\bar{x} \qquad \textbf{(12.7)}$$

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 12.2. With the sample of 10 restaurants, we have $n = 10$ observations. Because equations (12.6) and (12.7) 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 (12.6) and (12.7) and the information in Table 12.2, 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.

*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 (12.6) is often recommended when using a calculator to compute b_1.

TABLE 12.2 CALCULATIONS FOR THE LEAST SQUARES ESTIMATED REGRESSION EQUATION FOR ARMAND PIZZA PARLORS

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

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

The calculation of the y intercept (b_0) follows.

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

Thus, the estimated regression equation is

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

Figure 12.4 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.

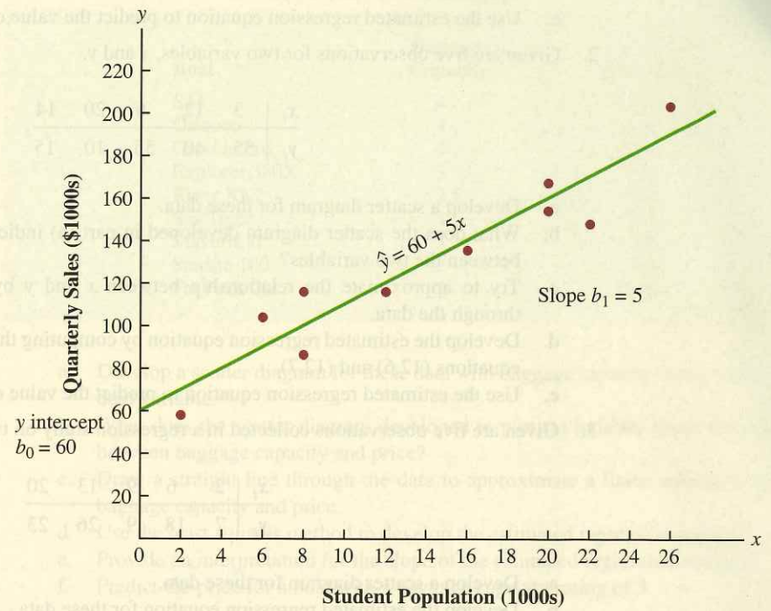
Using the estimated regression equation to make predictions outside the range of the values of the independent variable should be done with caution because outside that range we cannot be sure that the same relationship is valid.

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



NOTES AND COMMENTS

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

Exercises

Methods

SELF test

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

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

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

c. Try to approximate the relationship between x and y by drawing a straight line through the data.
d. Develop the estimated regression equation by computing the values of b_0 and b_1 using equations (12.6) and (12.7).
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 (12.6) and (12.7).
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 $x = 4$.

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. Technological advances helped make inflatable paddlecraft suitable for backcountry use. These blow-up rubber boats, which can be rolled into a bundle not much bigger than a golf bag, are large enough to accommodate one or two paddlers and their camping gear. *Canoe & Kayak* magazine tested boats from nine manufacturers to determine how they would perform on a three-day wilderness paddling trip. One of the criteria in their evaluation was the

baggage capacity of the boat, evaluated using a 4-point rating scale from 1 (lowest rating) to 4 (highest rating). The following data show the baggage capacity rating and the price of the boat (*Canoe & Kayak,* March 2003).

CD file
Boats

Boat	Baggage Capacity	Price ($)
S14	4	1595
Orinoco	4	1399
Outside Pro	4	1890
Explorer 380X	3	795
River XK2	2.5	600
Sea Tiger	4	1995
Maverik II	3	1205
Starlite 100	2	583
Fat Pack Cat	3	1048

a. Develop a scatter diagram for these data with baggage capacity rating as the independent variable.
b. What does the scatter diagram developed in part (a) indicate about the relationship between baggage capacity and price?
c. Draw a straight line through the data to approximate a linear relationship between baggage capacity and price.
d. Use the least squares method to develop the estimated regression equation.
e. Provide an interpretation for the slope of the estimated regression equation.
f. Predict the price for a boat with a baggage capacity rating of 3.

6. The following data show the annual advertising expenditure in millions of dollars and the market share for six automobile companies (*Advertising Age,* June 23, 2006).

CD file
MktShare

Company	Advertising ($ millions)	Market Share (%)
DaimlerChrysler	1590	14.9
Ford Motor Co.	1568	18.6
General Motors Corp.	3004	26.2
Honda Motor Co.	854	8.6
Nissan Motor Co.	1023	6.3
Toyota Motor Corp.	1075	13.3

a. Develop a scatter diagram for these data with the advertising expenditure as the independent variable and the market share as the dependent 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. Suppose that Honda Motor Co. believes that the estimated regression equation developed in part (c) is applicable for developing an estimate of market share for next year. Predict Honda's market share if they decide to increase their advertising expenditure to $1200 million next year.

7. Would you expect more reliable cars to cost more? *Consumer Reports* rated 15 upscale sedans. Reliability was rated on a 5-point scale: poor (1), fair (2), good (3), very good (4), and excellent (5). The price and reliability rating for each of the 15 cars are shown (*Consumer Reports,* February 2004).

CD file Cars

Make and Model	Reliability	Price ($)
Acura TL	4	33,150
BMW 330i	3	40,570
Lexus IS300	5	35,105
Lexus ES330	5	35,174
Mercedes-Benz C320	1	42,230
Lincoln LS Premium (V6)	3	38,225
Audi A4 3.0 Quattro	2	37,605
Cadillac CTS	1	37,695
Nissan Maxima 3.5 SE	4	34,390
Infiniti I35	5	33,845
Saab 9-3 Aero	3	36,910
Infiniti G35	4	34,695
Jaguar X-Type 3.0	1	37,995
Saab 9-5 Arc	3	36,955
Volvo S60 2.5T	3	33,890

a. Develop a scatter diagram for these data with the reliability rating as the independent variable.
b. Develop the least squares estimated regression equation.
c. Based upon your analysis, do you think more reliable cars cost more? Explain.
d. Estimate the price for an upscale sedan that has a good reliability rating.

8. According to *Advertising Age*'s annual salary review, Mark Hurd, the 49-year-old chairman, president, and CEO of Hewlett-Packard Co., received an annual salary of $817,000, a bonus of more than $5 million, and other compensation exceeding $17 million. His total compensation was slightly better than the average CEO total pay of $14.4 million. The following table shows the age and annual salary (in thousands of dollars) for Mark Hurd and 14 other executives who led publicly held companies (*Advertising Age,* December 5, 2006).

CD file ExecSalary

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

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

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

CD 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.

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

CD file SleepingBags

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

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

11. Although delays at major airports are now less frequent, it helps to know which airports are likely to throw off your schedule. In addition, if your plane is late arriving at a particular airport where you must make a connection, how likely is it that the departure will be late and thus increase your chances of making the connection? The following data show the percentage of late arrivals and departures during August for 13 airports (*Business 2.0,* February 2002).

CD file Airport

Airport	Late Arrivals (%)	Late Departures (%)
Atlanta	24	22
Charlotte	20	20
Chicago	30	29
Cincinnati	20	19
Dallas	20	22
Denver	23	23
Detroit	18	19
Houston	20	16
Minneapolis	18	18
Phoenix	21	22
Pittsburgh	25	22
Salt Lake City	18	17
St. Louis	16	16

a. Develop a scatter diagram for these data with the percentage of late arrivals as the independent variable.
b. What does the scatter diagram developed in part (a) indicate about the relationship between late arrivals and late departures?
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. Suppose the percentage of late arrivals at the Philadelphia airport for August was 22%. What is an estimate of the percentage of late departures?

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

CD file JetSki

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

a. Develop a scatter diagram for these data with weight as the independent variable.
b. What does the scatter diagram developed in part (a) indicate about the relationship between weight and price?
c. Use the least squares method to develop the estimated regression equation.
d. Predict the price for a three-seater PWC with a weight of 750 pounds.
e. The Honda AquaTrax F-12 weighs 750 pounds and has a price of $9500. Shouldn't the predicted price you developed in part (d) for a PWC with a weight of 750 pounds also be $9500?

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

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

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

a. Develop a scatter diagram for these data with adjusted gross income as the independent variable.
b. Use the least squares method to develop the estimated regression equation.
c. Estimate a reasonable level of total itemized deductions for a taxpayer with an adjusted gross income of $52,500. If this taxpayer claimed itemized deductions of $20,400, would the IRS agent's request for an audit appear justified? Explain.

14. Starting salaries for accountants and auditors in Rochester, New York, trail those of many U.S. cities. The following data show the starting salary (in thousands of dollars) and the cost of living index for Rochester and nine other metropolitan areas (*Democrat and Chronicle,* September 1, 2002). The cost of living index, based on a city's food, housing, taxes, and other costs, ranges from 0 (most expensive) to 100 (least expensive).

CD file Salaries

Metropolitan Area	Index	Salary ($1000s)
Oklahoma City	82.44	23.9
Tampa/St. Petersburg/Clearwater	79.89	24.5
Indianapolis	55.53	27.4
Buffalo/Niagara Falls	41.36	27.7
Atlanta	39.38	27.1
Rochester	28.05	25.6
Sacramento	25.50	28.7
Raleigh/Durham/Chapel Hill	13.32	26.7
San Diego	3.12	27.8
Honolulu	0.57	28.3

a. Develop a scatter diagram for these data with the cost of living index as the independent variable.
b. Develop the estimated regression equation relating the cost of living index to the starting salary.
c. Estimate the starting salary for a metropolitan area with a cost of living index of 50.

12.3 Coefficient of Determination

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

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

SUM OF SQUARES DUE TO ERROR

$$\text{SSE} = \Sigma(y_i - \hat{y}_i)^2 \qquad (12.8)$$

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

In Table 12.3 we show the calculations required to compute the sum of squares due to error for the Armand's Pizza Parlors example. For instance, for restaurant 1 the values of the independent and dependent variables are $x_1 = 2$ and $y_1 = 58$. Using the estimated regression equation, we find that the estimated value of quarterly sales for restaurant 1 is $\hat{y}_1 = 60 + 5(2) = 70$. Thus, the error in using $\hat{y}_1$ to estimate y_1 for restaurant 1 is $y_1 - \hat{y}_1 = 58 - 70 = -12$. The squared error, $(-12)^2 = 144$, is shown in the last column of Table 12.3. 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 12.2 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 12.4 we show the sum of squared deviations obtained by using the sample mean $\bar{y} = 130$ to estimate the value of quarterly sales for each restaurant in the sample. For the ith restaurant in the sample, the difference $y_i - \bar{y}$ provides a measure of the error involved in using $\bar{y}$ to estimate sales. The corresponding sum of squares, called the *total sum of squares,* is denoted SST.

TABLE 12.3 CALCULATION OF SSE FOR ARMAND'S PIZZA PARLORS

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

TABLE 12.4 COMPUTATION OF THE TOTAL SUM OF SQUARES FOR ARMAND'S PIZZA PARLORS

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

TOTAL SUM OF SQUARES

$$\text{SST} = \Sigma(y_i - \bar{y})^2 \qquad (12.9)$$

The sum at the bottom of the last column in Table 12.4 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* $y = \bar{y}$.

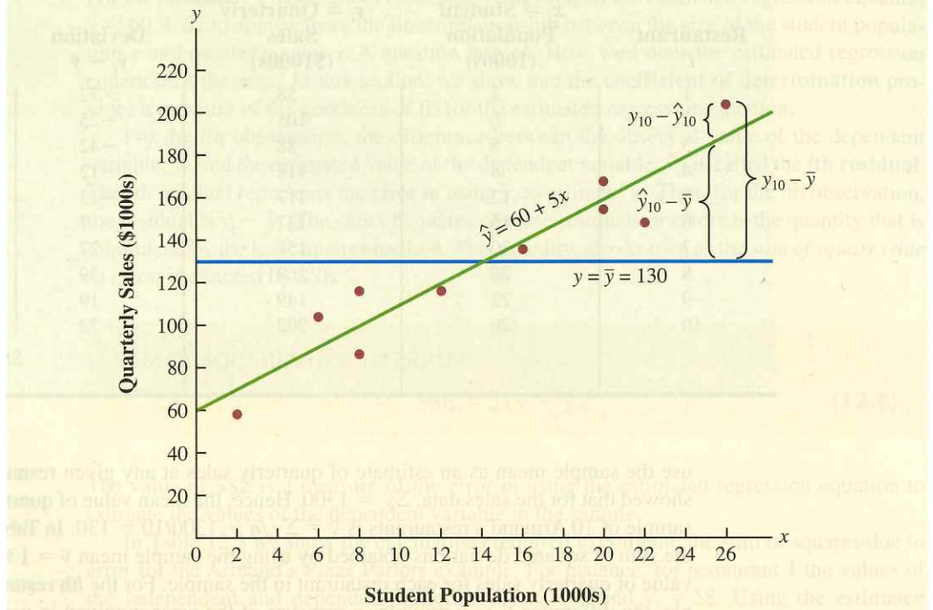
In Figure 12.5 we show the estimated regression line $\hat{y} = 60 + 5x$ and the line corresponding to $\bar{y} = 130$. Note that the points cluster more closely around the estimated regression line than they do about the line $\bar{y} = 130$. For example, for the 10th restaurant in the sample we see that the error is much larger when $\bar{y} = 130$ is used as an estimate of y_{10} than when $\hat{y}_{10} = 60 + 5(26) = 190$ is used. We can think of SST as a measure of how well the observations cluster about the $\bar{y}$ line and SSE as a measure of how well the observations cluster about the $\hat{y}$ line.

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

SUM OF SQUARES DUE TO REGRESSION

$$\text{SSR} = \Sigma(\hat{y}_i - \bar{y})^2 \qquad (12.10)$$

FIGURE 12.5 DEVIATIONS ABOUT THE ESTIMATED REGRESSION LINE AND THE LINE $y = \bar{y}$ FOR ARMAND'S PIZZA PARLORS



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

RELATIONSHIP AMONG SST, SSR, AND SSE

$$\text{SST} = \text{SSR} + \text{SSE} \qquad (12.11)$$

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

where

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

Equation (12.11) 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 (12.11), 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 (12.11), 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

$$r^2 = \frac{\text{SSR}}{\text{SST}} \qquad (12.12)$$

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

$$r_{xy} = (\text{sign of } b_1)\sqrt{\text{Coefficient of determination}} = (\text{sign of } b_1)\sqrt{r^2} \qquad (12.13)$$

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 (12.13) 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 (12.8), (12.9), and (12.10).
b. Compute the coefficient of determination r^2. Comment on the goodness of fit.
c. Compute the sample correlation coefficient.

16. The data from exercise 2 follow.

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

The estimated regression equation for these data is $\hat{y} = 68 - 3x$.

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

17. The data from exercise 3 follow.

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

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

Applications

SELF test

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

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

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

19. The data from exercise 7 follow.

CD file
Cars

Make and Model	x = Reliability	y = Price ($)
Acura TL	4	33,150
BMW 330i	3	40,570
Lexus IS300	5	35,105
Lexus ES330	5	35,174
Mercedes-Benz C320	1	42,230
Lincoln LS Premium (V6)	3	38,225
Audi A4 3.0 Quattro	2	37,605
Cadillac CTS	1	37,695
Nissan Maxima 3.5 SE	4	34,390
Infiniti I35	5	33,845
Saab 9-3 Aero	3	36,910
Infiniti G35	4	34,695
Jaguar X-Type 3.0	1	37,995
Saab 9-5 Arc	3	36,955
Volvo S60 2.5T	3	33,890

The estimated regression equation for these data is $\hat{y} = 40{,}639 - 1301.2x$. What percentage of the total sum of squares can be accounted for by the estimated regression equation? Comment on the goodness of fit. What is the sample correlation coefficient?

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

CD file PlasmaTV

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

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

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

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

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

22. *PC World* provided ratings for the top five small-office laser printers and five corporate laser printers (*PC World,* February 2003). The highest rated small-office laser printer was the Minolta-QMS PagePro 1250W, with an overall rating of 91. The highest rated corporate

laser printer, the Xerox Phaser 4400/N, had an overall rating of 83. The following data show the speed for plain text printing in pages per minute (ppm) and the price for each printer.

CD file Printers

Name	Type	Speed (ppm)	Price ($)
Minolta-QMS PagePro 1250W	Small Office	12	199
Brother HL-1850	Small Office	10	499
Lexmark E320	Small Office	12.2	299
Minolta-QMS PagePro 1250E	Small Office	10.3	299
HP Laserjet 1200	Small Office	11.7	399
Xerox Phaser 4400/N	Corporate	17.8	1850
Brother HL-2460N	Corporate	16.1	1000
IBM Infoprint 1120n	Corporate	11.8	1387
Lexmark W812	Corporate	19.8	2089
Oki Data B8300n	Corporate	28.2	2200

a. Develop the estimated regression equation with speed as the independent variable.
b. Compute r^2. What percentage of the variation in price can be explained by the printing speed?
c. What is the sample correlation coefficient between speed and price? Does it reflect a strong or weak relationship between printing speed and cost?

12.4 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

$$E(y) = \beta_0 + \beta_1 x \qquad \textbf{(12.14)}$$

(*continued*)