Chapter FOURTEEN

Introduction to Linear Regression and Correlation Analysis

CHAPTER OUTCOMES

After studying the material in Chapter 14, you should be able to:

1. Calculate and interpret the correlation between two variables.
2. Determine whether the correlation is significant.
3. Calculate the simple linear regression equation for a set of data and know the basic assumptions behind regression analysis.
4. Determine whether a regression model is significant.
5. Recognize regression analysis applications for purposes of prediction and description.
6. Calculate and interpret confidence intervals for the regression analysis.
7. Recognize some potential problems if regression analysis is used incorrectly.

PREPARING FOR CHAPTER FOURTEEN

- Review the methods for testing a null hypothesis using the *t*-distribution in Chapter 9.
- Review confidence intervals discussed in Chapter 8.
- Make sure you review the discussion about scatter plots in Chapter 2.
- Review the concepts associated with selecting a simple random sample in Chapter 1.
- Review the *F*-distribution and the approach for finding critical values from the *F*-table as discussed in Chapters 11 and 12.

WHY YOU NEED TO KNOW

Although some business situations involve only one variable, others require decision makers to consider the relationship between two or more variables. For example, an investment broker might be interested in the relationship between stock prices and the dividends issued by a publicly traded company. A marketing manager would be interested in examining the relationship between product sales and the amount of money spent on advertising. Finally, consider a real estate appraiser who is interested in determining the fair market value of a home or business. He would begin by collecting data on a sample of "comparable properties" that have sold recently. In addition to the selling price, he would collect data on other factors, such as the size and age of the property. He might then analyze the relationship between the price and the other variables and use this relationship to determine an appraised price for the property in question.

Simple linear regression and correlation analysis, which are introduced in this chapter, are statistical techniques the broker, marketing director, and appraiser will need in their analysis. These techniques are important to decision makers who need to determine the relationship between two variables. In Chapter 15, we will extend the discussion to include three or more variables. Regression analysis and correlation analysis are two of the most often applied statistical tools for business decision making.

14.1 Scatter Plots and Correlation

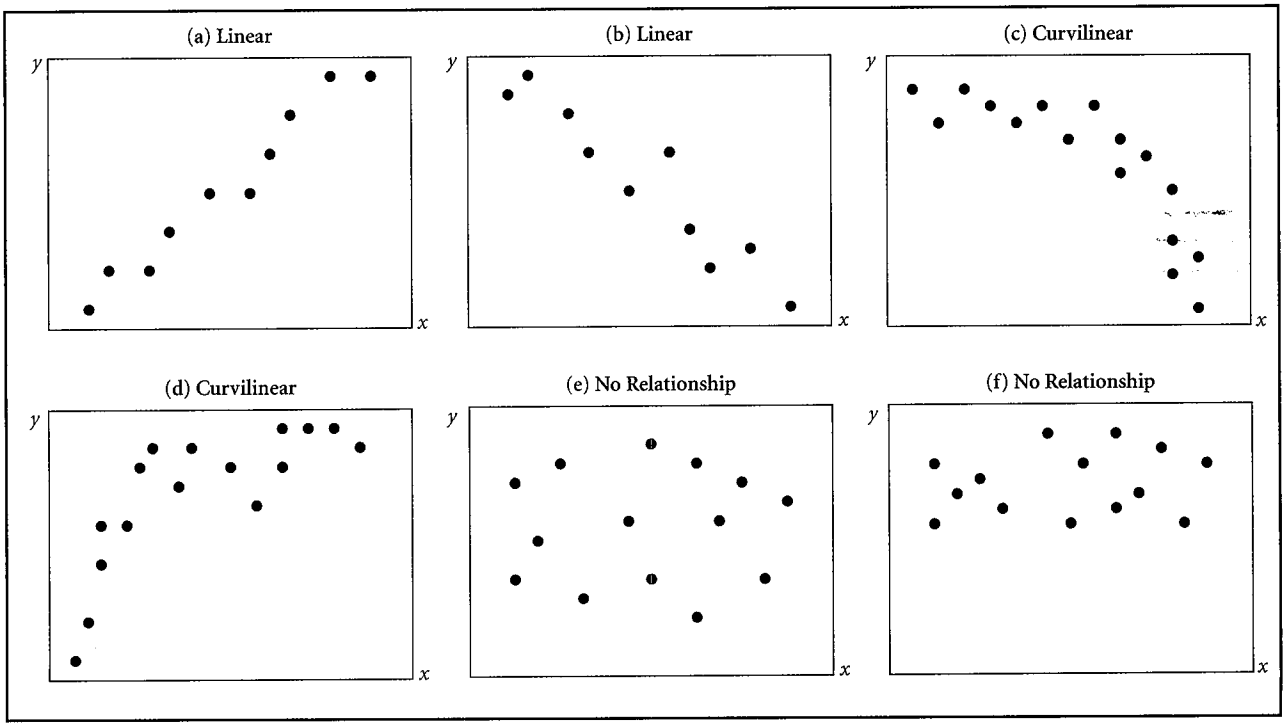
Scatter Plot
A two-dimensional plot showing the values for the joint occurrence of two quantitative variables. The scatter plot may be used to graphically represent the relationship between two variables. It is also known as a scatter diagram.

In those situations in which you are interested in analyzing the relationship between two quantitative variables, the **scatter plot**, or *scatter diagram*, introduced in Chapter 2 is very useful.

Figure 14.1 shows scatter plots that depict several potential relationships between values of a dependent variable, y, and an independent variable, x. A *dependent* (or *response*) *variable* is the variable whose variation we wish to explain. An *independent* (or *explanatory*) *variable* is a variable used to explain variation in the dependent variable. In Figure 14.1, (a) and (b) are examples of strong *linear* relationships between x and y. This means that for each unit change in the independent variable, x, the corresponding change in the dependent variable, y, will tend to be a fairly consistent amount. Note that this

FIGURE 14.1

Two-Variable Relationships



systematic change in y can be positive (y increases as x increases) or negative (y decreases as x increases). The degree of linearity exhibited depends on the degree of consistency in the change of the y variable when the independent variable, x, changes.

Figures 14.1 (c) and (d) illustrate situations in which the relationship between the x and y variable is nonlinear. There are many possible nonlinear relationships that can occur. The scatter plot is very useful for visually identifying the nature of the relationship.

Figures 14.1 (e) and (f) show examples in which there is no identifiable relationship between the two variables. This means that as x increases, y sometimes increases and sometimes decreases but with no particular pattern.

Correlation Versus Regression

In analyzing the relationship between two variables, there are two basic models that we can use, depending on the conditions under which the data are collected. These models are the subjects of this chapter. One model is referred to as the *regression model* in which the relationship between x and y assumes that the x variable takes on known values specifically selected from all the possible values for x. The y variable is a random variable observed at the different levels of x.

A second model is referred to as the *correlation model* and is used in applications in which both the x and the y variables are considered to be random variables. These two models, regression versus correlation, arise in practice by the way in which the data are obtained. Consider models that might apply to the relationship between the amount of daily sunscreen sold, y, as a function of the day's high temperature, x. We could select a random sample of 36 days and record the amount of sunscreen sold and the day's maximum temperature. In this case, the measurements obtained for both variables are observations from a joint distribution of x and y. An analysis of these data would be done using the correlation model approach.

Suppose instead that we decide to collect data for days with maximum temperatures of 75, 80, 85, 90, 95, and 100. We would measure the amount of daily sunscreen sold (y) for several randomly chosen days in which the maximum temperature is at each of these preselected temperatures. That is, we might pick six days at random from a population of days that have a maximum temperature of 75 degrees and observe the amount of sunscreen sold, and so on. Now each observation of y is from the distribution of y for a fixed x value. The analysis of these data would be done using the regression model approach.

We stress the two types of sampling because there are important differences in what can be estimated using these two methods. As we will illustrate later in this chapter, when the data have been collected at specific levels of the x variable, as was suggested in the second situation, our estimates for the y variable will be conditional on the value of x we are using.[1]

Correlation Coefficient
A quantitative measure of the strength of the linear relationship between two variables. The correlation ranges from −1.0 to +1.0. A correlation of ±1.0 indicates a perfect linear relationship, whereas a correlation of 0 indicates no linear relationship.

The Correlation Coefficient

In addition to analyzing the relationship between two variables graphically, we can also measure the strength of the linear relationship between two variables using a measure called the **correlation coefficient**.

The correlation coefficient for two variables can be estimated from sample data using Equation 14.1 or the algebraic equivalent, Equation 14.2.

CHAPTER OUTCOME #1

Sample Correlation Coefficient

$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{[\Sigma(x-\bar{x})^2][\Sigma(y-\bar{y})^2]}} \qquad \textbf{(14.1)}$$

[1] See Kutner et al., *Applied Linear Statistical Models*, 4th ed., p. 78, and N. R. Draper, and H. Smith, *Applied Regression Analysis*, 3rd ed., p. 89, for more discussion on this subject.

or the algebraic equivalent:

$$r = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{[n(\Sigma x^2) - (\Sigma x)^2][n(\Sigma y^2) - (\Sigma y)^2]}} \tag{14.2}$$

where:

r = Sample correlation coefficient
n = Sample size
x = Value of the independent variable
y = Value of the dependent variable

The sample correlation coefficient computed using Equations 14.1 and 14.2 is called the *Pearson Product Moment Correlation*. The sample correlation coefficient, r, can range from a perfect positive correlation, $+1.0$, to a perfect negative correlation, -1.0. A perfect correlation is one in which a given change in the value of the x variable is accompanied by a specific uniform amount of change in the y variable. Graphically, the x, y points will plot on a straight line. If two variables have no linear relationship, the correlation between them is 0 and there is no linear relationship between the change in x and y. Consequently, the more the correlation differs from 0.0, the stronger the linear relationship between the two variables. The sign of the correlation coefficient indicates the direction of the relationship, but it does not aid in determining the strength.

Figure 14.2 illustrates some examples of correlation between two variables. Note for the correlation coefficient to equal plus or minus 1.0, all the (x, y) points form a perfectly straight line. The more the points depart from a straight line, the weaker (closer to 0.0) the correlation is between the two variables.

Business Application

MIDWEST DISTRIBUTION COMPANY Consider the application involving Midwest Distribution which supplies soft drinks and snack foods to convenience stores in Michigan, Illinois, and Iowa. Although Midwest Distribution has been profitable, the director of

FIGURE 14.2

Correlation Between Two Variables

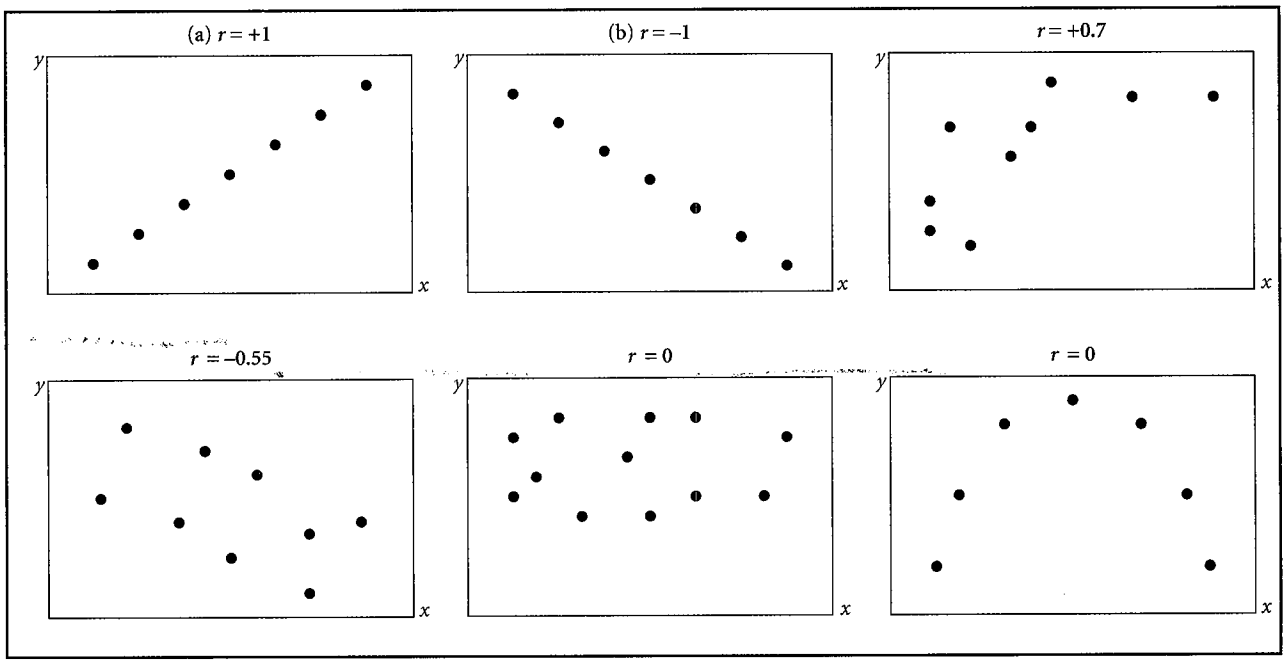
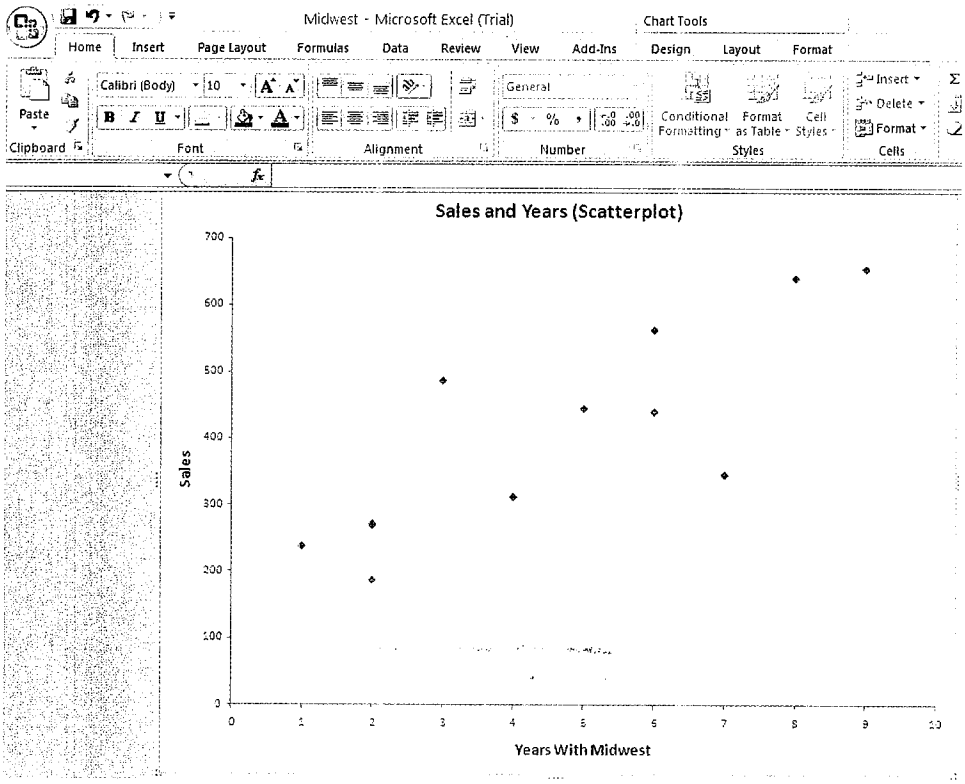


FIGURE 14.3

Excel 2007 Scatter Plot of Sales vs. Years with Midwest Distribution



Excel 2007 Instructions:

1. Open file: Midwest.xls.
2. Move the *Sales* column to the *right* of *Years with midwest* column.
3. Select data for chart.
4. On **Insert tab**, click **XY (Scatter)**, and then click the **Scatter with only Markers** Option.
5. Use the **Layout** tab of the **Chart Tools** to add titles and remove grid lines.
6. Use the **Design** tab of the **Chart Tools** to move the chart to a new worksheet.

Minitab Instructions (for similar result):

1. Open file: Midwest MTW.
2. Choose **Graph > Scatterplot.**
3. Under **Scatterplot**, choose **Simple OK.**
4. Under **Y variable**, enter *y* column.
5. In **X variable**, enter *x* column.
6. Click **OK.**

Excel and Minitab Tutorial

marketing has been concerned about the rapid turnover in her salesforce. In the course of exit interviews, she discovered a major concern with the compensation structure.

Midwest Distribution has a two-part wage structure: a base salary and a commission computed on monthly sales. Typically, about half of the total wages paid comes from the base salary, which increases with longevity with the company. This portion of the wage structure is not an issue. The concern expressed by departing employees is that new employees tend to be given parts of the sales territory previously covered by existing employees and are assigned prime customers as a recruiting inducement.

At issue, then, is the relationship between sales (on which commissions are paid) and number of years with the company. The data for a random sample of 12 sales representatives are in the file called **Midwest** on your CD-ROM. The first step is to develop a scatter plot of the data. Both Excel and Minitab have procedures for constructing a scatter plot and computing the correlation coefficient.

The scatter plot for the Midwest data is shown in Figure 14.3. Based on this plot, total sales and years with the company appear to be linearly related. However, the strength of this relationship is uncertain. That is, how close do the points come to falling on a straight line? To answer this question, we need a quantitative measure of the strength of the linear relationship between the two variables. That measure is the correlation coefficient.

Equation 14.1 is used to determine the correlation between sales and years with the company. Table 14.1 shows the manual calculations that were used to determine this correlation coefficient of 0.8325. However, because the calculations are rather tedious and long, we almost always use computer software to perform the computation, as shown in Figure 14.4. The $r = 0.8325$ indicates that there is a fairly strong, positive correlation between these two variables for the sample data.

TABLE 14.1 Correlation Coefficient Calculations for the Midwest Distribution Example

Sales	Years					
y	x	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
487	3	−1.58	82.42	−130.22	2.50	6,793.06
445	5	0.42	40.42	16.98	0.18	1,633.78
272	2	−2.58	−132.58	342.06	6.66	17,577.46
641	8	3.42	236.42	808.56	11.70	55,894.42
187	2	−2.58	−217.58	561.36	6.66	47,341.06
440	6	1.42	35.42	50.30	2.02	1,254.58
346	7	2.42	−58.58	−141.76	5.86	3,431.62
238	1	−3.58	−166.58	596.36	12.82	27,748.90
312	4	−0.58	−92.58	53.70	0.34	8,571.06
269	2	−2.58	−135.58	349.80	6.66	18,381.94
655	9	4.42	250.42	1,106.86	19.54	62,710.18
563	6	1.42	158.42	224.96	2.02	25,096.90
$\Sigma = 4{,}855$	$\Sigma = 55$			$\Sigma = 3{,}838.92$	$\Sigma = 76.92$	$\Sigma = 276{,}434.92$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{4{,}855}{12} = 404.58 \qquad \bar{x} = \frac{\Sigma x}{n} = \frac{55}{12} = 4.58$$

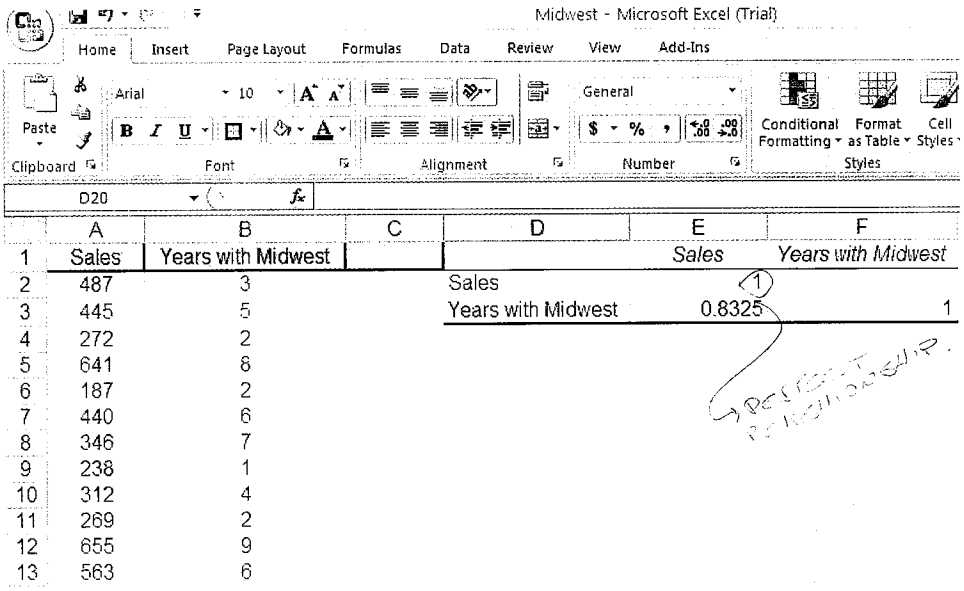
Using Equation 14.1,

$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2\,\Sigma(y-\bar{y})^2}} = \frac{3{,}838.92}{\sqrt{(76.92)(276{,}434.92)}} = 0.8325$$

FIGURE 14.4

Excel 2007 Correlation Output for Midwest Distribution

Excel 2007 Instructions:
1. Open file: Midwest.xls.
2. On the **Data** tab, click **Data Analysis**.
3. Select **Correlation**.
4. Define Data Range.
5. Click on **Labels in First Row**.
6. Specify output location.
7. Click **OK**.



Minitab Instructions (for similar results):
1. Open file: Midwest MTW.
2. Choose **Stat > Basic Statistics > Correlation**.
3. In **Variables**, enter ***Y*** and ***X*** columns.
4. Click **OK**.

Significance Test for the Correlation Although a correlation coefficient of 0.8325 seems quite high (relative to 0), you should remember that this value is based on a sample of 12 data points and is subject to sampling error. Therefore, a formal hypothesis-testing procedure is needed to determine whether the linear relationship between sales and years with the company is significant.

The null and alternative hypotheses to be tested are

H_0: $\rho = 0$ (no correlation)
H_A: $\rho \neq 0$ (correlation exists)

where the Greek symbol, ρ (rho) represents the population correlation coefficient.

We must test whether the sample data support or refute the null hypothesis. The test procedure utilizes the *t*-test statistic in Equation 14.3.

CHAPTER OUTCOME #2

Test Statistic for Correlation

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \qquad df = n-2 \tag{14.3}$$

where:

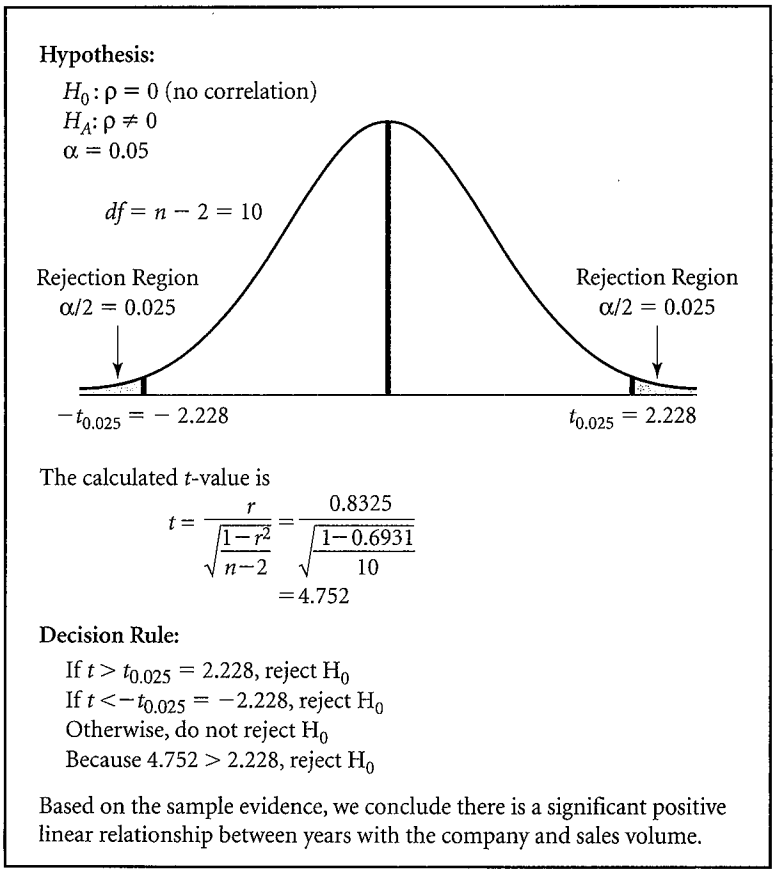
t = Number of standard errors r is from 0
r = Sample correlation coefficient
n = Sample size

The degrees of freedom for this test are $n - 2$, because we lose 1 degree of freedom for each of the two sample means that are used to estimate the population means for the two variables.

Figure 14.5 shows the hypothesis test for the Midwest Distribution example using an alpha level of 0.05. Recall that the sample correlation coefficient was $r = 0.8325$. Based on

FIGURE 14.5

Correlation Significance Test for the Midwest Distribution Example



these sample data, we should conclude there is a significant, positive linear relationship in the population between years of experience and total sales for Midwest Distribution sales representatives. The implication is that the more years an employee has been with the company, the more sales that employee generates. This runs counter to the claims made by some of the departing employees. The manager will probably want to look further into the situation to see whether a problem might exist in certain regions.

The *t*-test for determining whether the population correlation is significantly different from 0 requires the following assumptions:

Assumptions

1. The data are interval- or ratio-level.
2. The two variables (y and x) are distributed as a *bivariate normal* distribution.

Although the formal mathematical representation is beyond the scope of this text, *two variables are bivariate normal if their joint distribution is normally distributed.* Although the *t*-test assumes a bivariate normal distribution, it is robust—that is, correct inferences can be reached even with slight departures from the normal-distribution assumption. (See Kutner et al., *Applied Linear Statistical Models* for further discussion of bivariate normal distributions.)

EXAMPLE 14-1 Correlation Analysis

TRY PROBLEM 14.1

Yellow Page Advertising Recently a publisher of a regional telephone book surveyed a simple random sample of 10 of its commercial yellow-page advertising customers in an attempt to determine whether the size of their advertisements, in square inches (x), were positively correlated with the proportion of calls to the businesses that were generated by the ads (y). For a one-month period, each commercial customer asked each caller to their business if they had learned about the business through the yellow pages. To determine whether there is a statistically significant correlation between the two variables, the following steps can be employed:

Step 1 Specify the population parameter of interest.
The publisher wishes to determine whether the size of an ad is positively correlated with the proportion of calls to the business that were generated by the ad. The parameter of interest is, therefore, the population correlation, ρ.

Step 2 Formulate the appropriate null and alternative hypotheses.
Because the regional phone company is interested in establishing a positive relationship between ad size and proportion of calls generated from the ad, the test will be one-tailed, as follows:

$$H_0: \rho \leq 0$$
$$H_A: \rho > 0$$

Step 3 Specify the level of significance.
A significance level of 0.05 is chosen.

Step 4 Compute the correlation coefficient and the test statistic.
Compute the sample correlation coefficient using Equation 14.1 or 14.2, or by using software such as Excel or Minitab.

The following sample data were obtained:

Square Inches	Proportion of Calls Generated by Ad
9	0.13
16	0.16
25	0.21
16	0.18
20	0.18
16	0.19
20	0.15
20	0.17
16	0.13
9	0.11

Using Equation 14.1, we get

$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{[\Sigma(x-\bar{x})^2][\Sigma(y-\bar{y})^2]}} = 0.7796$$

Compute the t-test statistic using Equation 14.3.

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.7796}{\sqrt{\frac{1-0.7796^2}{10-2}}} = 3.52$$

Step 5 Construct the rejection region and decision rule.
For an alpha level equal to 0.05, the one-tailed, upper-tail, critical value for $n - 2 = 10 - 2 = 8$ degrees of freedom is $t_{0.05} = 1.8595$. The decision rule is

If $t > 1.8595$, reject the null hypothesis:
Otherwise, do not reject the null hypothesis.

Step 6 Reach a decision.
Because

$t = 3.52 > 1.8595$, reject the null hypothesis.

Step 7 Draw a conclusion.
Because the null hypothesis is rejected, the sample data do support the contention that there is a positive linear relationship between ad size and the proportion of calls that were generated by the ad.

Cause-and-Effect Interpretations Care must be used when interpreting the correlation results. For example, even though we found a significant linear relationship between years of experience and sales for the Midwest Distribution salesforce, the correlation does not imply cause and effect. Although an increase in experience may, in fact, cause sales to change, simply because the two variables are correlated does not guarantee a cause-and-effect situation. Two seemingly unconnected variables may be highly correlated. For example, over a period of time, teachers' salaries in North Dakota might be highly correlated with the price of grapes in Spain. Yet, we doubt that a change in grape prices will *cause* a corresponding change in salaries for teachers in North Dakota, or vice versa. When a correlation exists between two seemingly unrelated variables, the correlation is said to be

a *spurious correlation*. You should take great care to avoid basing conclusions on spurious correlations.

The Midwest Distribution marketing director has a logical reason to believe that years of experience with the company and total sales are related. That is, sales theory and customer feedback hold that product knowledge is a major component in successfully marketing a product. However, a statistically significant correlation alone does not prove that this cause-and-effect relationship exists. When two seemingly unrelated variables are correlated, they may both be responding to changes in some third variable. For example, the observed correlation could be the effect of a company policy of giving better sales territories to more senior salespeople.

14-1: Exercises

Skill Development

14-1. A random sample of two variables, x and y, produced the following observations:

x	y
19	7
13	9
17	8
9	11
12	9
25	6
20	7
17	8

a. Develop a scatter plot for the two variables and describe what relationship, if any, exists.
b. Compute the correlation coefficient for these sample data.
c. Test to determine whether the population correlation coefficient is negative. Use a significance level of 0.05 for the hypothesis test.

14-2. The following data for the dependent variable, y, and the independent variable, x, have been collected using simple random sampling:

x	y
10	120
14	130
16	170
12	150
20	200
18	180
16	190
14	150
16	160
18	200

a. Construct a scatter plot for these data. Based on the scatter plot, how would you describe the relationship between the two variables?
b. Compute the correlation coefficient.

14-3. An industry study was recently conducted in which the sample correlation between units sold and marketing expenses was 0.57. The sample size for the study included 15 companies. Based on the sample results, test to determine whether there is a significant positive correlation between these two variables. Use an $\alpha = 0.05$.

14-4. You are given the following data for variables x and y:

x	y
3.0	1.5
2.0	0.5
2.5	1.0
3.0	1.8
2.5	1.2
4.0	2.2
1.5	0.4
1.0	0.3
2.0	1.3
2.5	1.0

a. Plot these variables in scatter plot format. Based on this plot, what type of relationship appears to exist between the two variables?
b. Compute the correlation coefficient for these sample data. Indicate what the correlation coefficient measures.
c. Test to determine whether the population correlation coefficient is positive. Use the $\alpha = 0.01$ level to conduct the test. Be sure to state the null and alternative hypotheses and show the test and decision rule clearly.

14-5. A random sample of the following two variables was obtained:

x	29	48	28	22	28	42	33	26	48	44
y	16	46	34	26	49	11	41	13	47	16

a. Calculate the correlation between these two variables.
b. Conduct a test of hypothesis to determine if there exists a correlation between the two variables in the population. Use a significance level of 0.10.

14-6. For each of the following circumstances, perform the indicated hypothesis tests:
a. H_A: $\rho > 0$, $r = 0.53$, and $n = 30$ with $\alpha = 0.01$, using a test-statistic approach.
b. H_A: $\rho \neq 0$, $r = -0.48$, and $n = 20$ with $\alpha = 0.05$, using a p-value approach.
c. H_A: $\rho \neq 0$, $r = 0.39$, and $n = 45$ with $\alpha = 0.02$, using a test-statistic approach.
d. H_A: $\rho < 0$, $r = 0.34$, and $n = 25$ with $\alpha = 0.05$, using a test-statistic approach.

Business Applications

14-7. A random sample of 50 high school students from a Seattle high school is selected and each student's scores on a standardized mathematics examination and a standardized English examination are recorded. Suppose the correlation coefficient for the two examination scores is 0.75.
a. Provide an explanation of the sample correlation coefficient in this context.
b. Using a level of significance of $\alpha = 0.01$, test to determine whether there is a positive linear relationship between mathematics scores and English scores for high school students at the Seattle high school.

14-8. A random sample of 50 bank accounts was selected from a local branch bank. The account balance and the number of deposits and withdrawals during the past month were the two variables recorded. The correlation coefficient for the two variables was -0.23.
a. Discuss what the $r = -0.23$ measures. Make sure to frame your discussion in terms of the two variables mentioned here.
b. Using an $\alpha = 0.10$ level, test to determine whether there is a significant linear relationship between account balance and the number of transactions in the account during the past month. State the null and alternative hypotheses and show the decision rule.
c. Consider the decision you reached in part b. Describe the type of error you could have made in the context of this problem.

14-9. Farmers National Bank issues MasterCard credit cards to its customers. A main factor in determining whether a credit card will be profitable to the bank is the average monthly balance that the customer will maintain on the card that will be subject to finance charges. Bank analysts wish to determine whether there is a relationship between the average monthly credit card balance and the income stated on the original credit card application form. The following sample data have been collected from existing credit card customers:

Income	Credit Balance
$43,000	$345
$35,000	$1,370
$47,000	$1,140
$55,000	$201
$55,000	$56
$59,000	$908
$28,000	$2,345
$43,000	$104
$54,000	$0
$36,000	$1,290
$39,000	$130
$31,000	$459
$30,000	$0
$37,000	$1,950
$39,000	$240

a. Indicate which variable is to be the independent variable and which is to be the dependent variable in the bank's analysis and indicate why.
b. Construct a scatter plot for these data and describe what, if any, relationship appears to exists between these two variables.
c. Calculate the correlation coefficient for these two variables and test to determine whether there is a significant correlation at the $\alpha = 0.05$ level.

14-10. Amazon.com has become one of the most successful online merchants. Two measures of its success are sales and net income/loss figures. In an article entitled "Amazon CEO takes long view," *USA Today* (Byron Acohido, July 6, 2005) presented these figures (in $million) for the period 1995 to 2004:

Net Income/Loss	−0.3	−5.7	−27.5	−124.5	−719.9
Sales	0.5	15.7	147.7	609.8	1,639.8
Net Income/Loss	−1,411.2	−567.3	−149.1	35.3	588.5
Sales	2,761.9	3,122.9	3,932.9	5,263.7	6,921.1

a. Produce a scatter plot for Amazon's net income/loss and sales figures for the period 1995 to 2004. Does there appear to be a linear relationship between these two variables? Explain your response.
b. Calculate the correlation coefficient between Amazon's net income/loss and sales figures for the period 1995 to 2004.
c. Conduct a hypothesis test to determine if a positive correlation exists between Amazon's net income/loss and sales figures. Use a significance level of 0.05 and assume that these figures form a random sample.

14-11. Complaints concerning excessive commercials seem to grow as the amount of "clutter," including commercials and advertisements for other television shows, steadily increases on network and cable TV. The analysis by Nielsen Monitor-Plus (Gary Levin, "Ad glut turns off viewers," *USA Today*, October 12, 2005) compares the average nonprogram minutes in an hour of prime time for both network and cable television.

Year	1996	1999	2001	2004
Network	9.88	14.00	14.65	15.80
Cable	12.77	13.88	14.50	14.92

a. Calculate the correlation coefficient for the average nonprogram minutes in an hour of prime time between network and cable television.
b. Conduct a hypothesis test to determine if a positive correlation exists between the average nonprogram minutes in an hour of prime time between network and cable television. Use a significance level of 0.05 and assume that these figures form a random sample.

Computer Database Exercises

14-12. Customers who made online purchases last quarter from an Internet retailer were randomly sampled from the retailer's database. The dollar value of each customer's quarterly purchases along with the time the customer spent shopping the company's online catalog that quarter were recorded. The sample results are contained in the file **Online**.
a. Create a scatter plot of the variables Time (x) and Purchases (y). What relationship, if any, appears to exist between the two variables?
b. Compute the correlation coefficient for these sample data. What does the correlation coefficient measure?
c. Conduct a hypothesis test to determine if there is a positive relationship between time viewing the retailer's catalog and dollar amount purchased. Use a level of significance equal to 0.025. Provide a managerial explanation of your results.

14-13. A regional accreditation board for colleges and universities is interested in determining whether a relationship exists between college student applicant verbal SAT scores and the in-state tuition costs at the university. Data have been collected on a sample of colleges and universities and are in the data file called **Colleges and Universities**.
a. Develop a scatter plot for these two variables and discuss what, if any, relationship you see between the two variables based on the scatter plot.
b. Compute the sample correlation coefficient.
c. Based on the correlation coefficient computed in part b, test to determine whether the population correlation coefficient is positive for these two variables. That is, can we expect schools that charge higher in-state tuition will attract students with higher average verbal SAT scores? Test using a 0.05 significance level.

14-14. Platinum Billiards, Inc. is a Jacksonville, Florida-based retailer of billiard supplies. It stands out among billiard suppliers because of the research it does to assure its products are top-notch. One experiment was conducted to measure the speed attained by a cue ball struck by various weighted pool cues. The conjecture is that a light cue generates faster speeds while breaking the balls at the beginning of a game of pool. Anecdotal experience has indicated that a billiard cue weighing less than 19 ounces generates faster speeds. Platinum used a robotic arm to investigate this claim. Its research generated the data given in the file entitled **Breakcue**.
a. To determine if there is a negative relationship between the weight of the pool cue and the speed attained by the cue ball, calculate a correlation coefficient.
b. Conduct a test of hypothesis to determine if there is a negative relationship between the weight of the pool cue and the speed attained by the cue ball. Use a significance level of 0.025 and a p-value approach.

14-15. As the number of air travelers with time on their hands increases, logic would indicate spending on retail purchases in airports would increase as well. A study by *Airport Revenue News* addressed the per person spending at select airports for merchandise excluding food, gifts, and news items. A file entitled **Revenues** contains sample data selected from airport retailers in 2001 and again in 2004.
a. Produce a scatter plot for the per person spending at selected airports for merchandise excluding food, gifts, and news items for the years 2001 and 2004. Does there appear to be a linear relationship between spending in 2001 and spending in 2004? Explain your response.

b. Calculate the correlation coefficient between the per person spending in 2001 and the per-person spending in 2004. Does it appear that an increase in per person spending in 2001 would be associated with an increase in spending in 2004? Support your assertion.

c. Conduct a hypothesis test to determine if a positive correlation exists between the per person spending in 2001 and that in 2004. Use a significance level of 0.05 and assume that these figures form a random sample.

14.2 Simple Linear Regression Analysis

In the Midwest Distribution application, we determined that the relationship between years of experience and total sales is linear and statistically significant, based on the correlation analysis performed in the previous section. Because hiring and training costs have been increasing, we would like to use this relationship to help formulate a more acceptable wage package for the salesforce.

Simple Linear Regression
The method of regression analysis in which a single independent variable is used to predict the dependent variable.

The statistical method we will use to analyze the relationship between years of experience and total sales is *regression analysis*. When we have only two variables—a dependent variable, such as sales, and an independent variable, such as years with the company—the technique is referred to as *simple regression analysis*. When the relationship between the dependent variable and the independent variable is linear, the technique is **simple linear regression**.

The Regression Model and Assumptions

The objective of simple linear regression (which we shall call *regression analysis*) is to represent the relationship between values of x and y with a model of the form shown in Equation 14.4.

Simple Linear Regression Model (Population Model)

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{14.4}$$

where:

y = Value of the dependent variable
x = Value of the independent variable
β_0 = Population's y intercept
β_1 = Slope of the population regression line
ε = Error term, or residual (i.e., the difference between the actual y-value and the value of y predicted by the population model)

The simple linear regression population model described in Equation 14.4 has four assumptions:

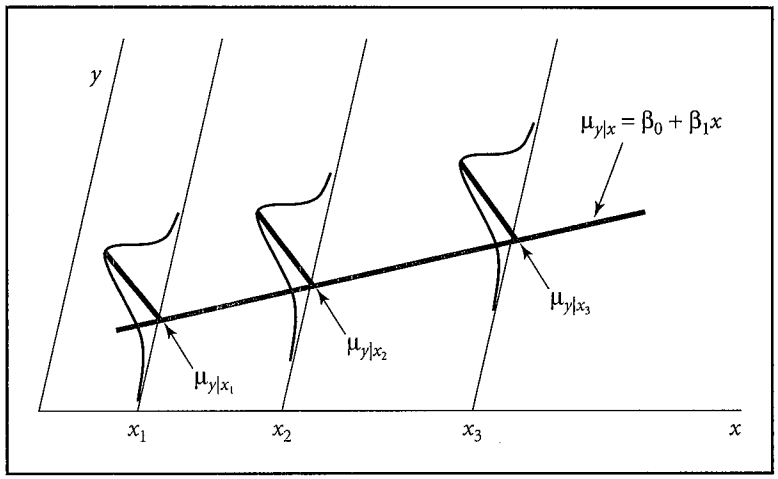
Assumptions

1. Individual values of the error terms, ε, are statistically independent of one another, and these values represent a random sample from the population of possible ε-values at each level of x.
2. For a given value of x, there can exist many values of y and therefore many values of ε. Further, the distribution of possible ε-values for any x-value is normal.
3. The distributions of possible ε-values have equal variances for all values of x.
4. The means of the dependent variable, y, for all specified values of the independent variable, ($\mu_{y|x}$), can be connected by a straight line called the population regression model.

Figure 14.6 illustrates assumptions 2, 3, and 4. The regression model (straight line) connects the average of the y-values for each level of the independent variable, x. The

FIGURE 14.6

Graphical Display of Linear Regression Assumptions



actual y-values for each level of x are normally distributed around the mean of y. Finally, observe that the spread of possible y-values is the same regardless of the level of x. The population regression line is determined by two values, β_0 and β_1. These values are known as the population *regression coefficients*. Value β_0 identifies the y intercept and β_1 the slope of the regression line. Under the regression assumptions, the coefficients define the true population model. For each observation, the actual value of the dependent variable, y, for any x is the sum of two components:

$$y = \underbrace{\beta_0 + \beta_1 x}_{\text{Linear Component}} + \underbrace{\varepsilon}_{\text{Random Error Component}}$$

The random error component, ε, may be positive, zero, or negative, depending on whether a single value of y for a given x falls above, on, or below the population regression line. Section 15.5 in Chapter 15 discusses how to check whether assumptions have been violated and the possible courses of action if the violations occur.

Meaning of the Regression Coefficients

Regression Slope Coefficient
The average change in the dependent variable for a unit change in the independent variable. The slope coefficient may be positive or negative, depending on the relationship between the two variables.

Coefficient β_1, the **regression slope coefficient** of the population regression line, measures the average change in the value of the dependent variable, y, for each unit change in x. The regression slope can be either positive, zero, or negative, depending on the relationship between x and y. For example, a positive population slope of 12 ($\beta_1 = 12$) means that for a 1-unit increase in x, we can expect an average 12-unit increase in y. Correspondingly, if the population slope is negative 12 ($\beta_1 = -12$), we can expect an average decrease of 12 units in y for a 1-unit increase in x.

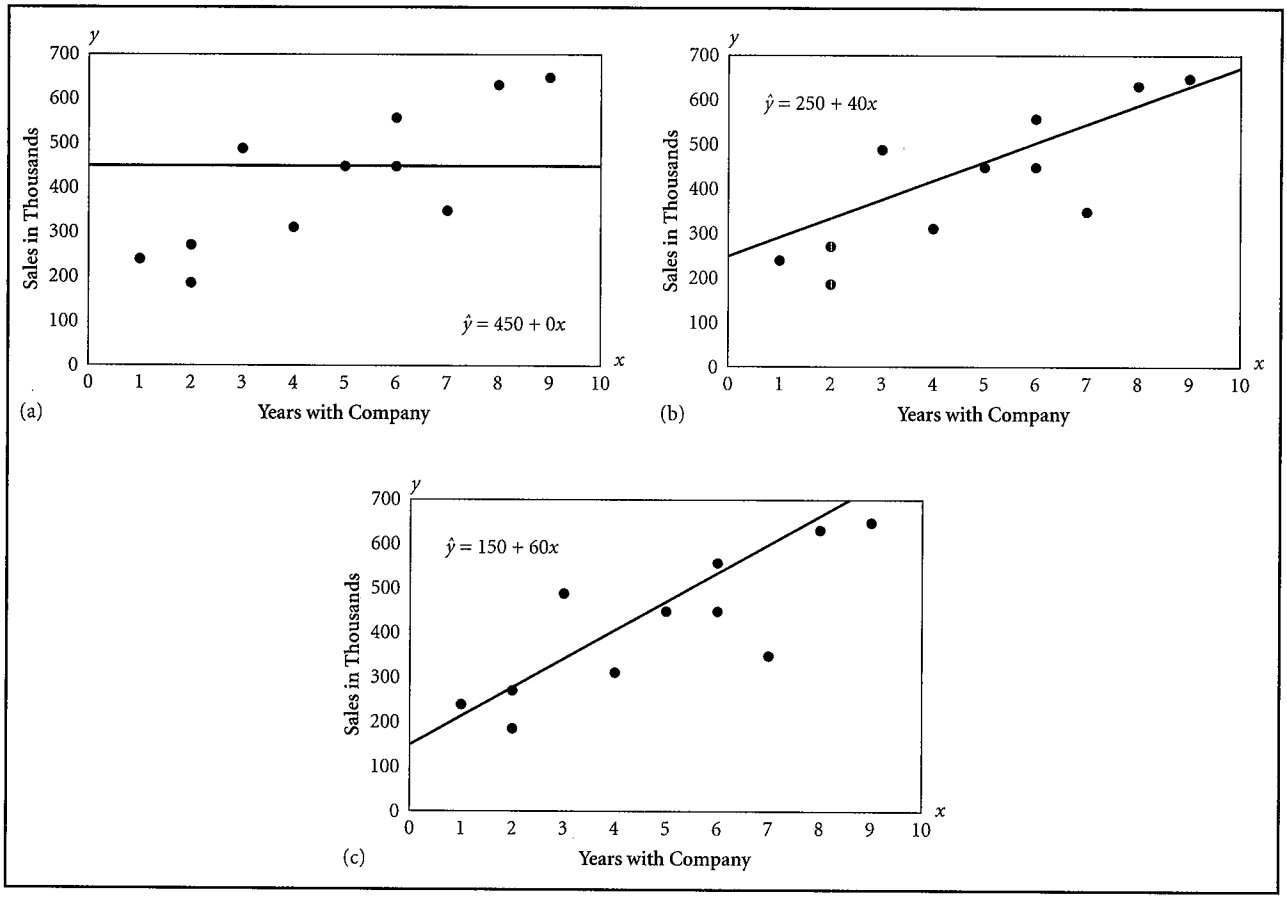
The population's y intercept, β_0, indicates the mean value of y when x is 0. However, this interpretation holds only if the population could have x values equal to 0. When this cannot occur, β_0 does not have a meaningful interpretation in the regression model.

Business Application

MIDWEST DISTRIBUTION (*CONTINUED*) The Midwest Distribution marketing manager has data for a sample of 12 sales representatives. In Section 14.1, she has established that a significant linear relationship exists between years of experience and total sales using correlation analysis. (Recall that the sample correlation between the two variables was $r = 0.8325$.) Now she would like to estimate the regression equation that defines the *true* linear relationship (that is, the population's linear relationship) between years of experience and sales. Figure 14.3 shows the scatter plot for two variables: years with the company and sales. We need to use the sample data to estimate β_0 and β_1, the true intercept and slope of the line representing the relationship between two variables. The *regression line* through the sample data is the best estimate of the population regression line. However, there are an infinite number of possible regression lines for a set of points. For example, Figure 14.7 shows three of the possible different lines that pass through the Midwest Distribution data. Which line should be used to estimate the true regression model?

FIGURE 14.7

Possible Regression Lines



Least Squares Criterion
The criterion for determining a regression line that minimizes the sum of squared prediction errors.

Residual
The difference between the actual value of y and the predicted value $\hat{y}$ for a given level of the independent variable, x.

We must establish a criterion for selecting the best line. The criterion used is the **least squares criterion.**[2] To understand the least squares criterion, you need to know about prediction error, or **residual**, which is the distance between the actual y coordinate of an (x, y) point and the predicted value of that y coordinate produced by the regression line. Figure 14.8 shows how the prediction error is calculated for the employee who was with Midwest for four years ($x = 4$) using one possible regression line: $\hat{y} = 150 + 60x$ (where $\hat{y}$ is the predicted sales value). The predicted sales value is

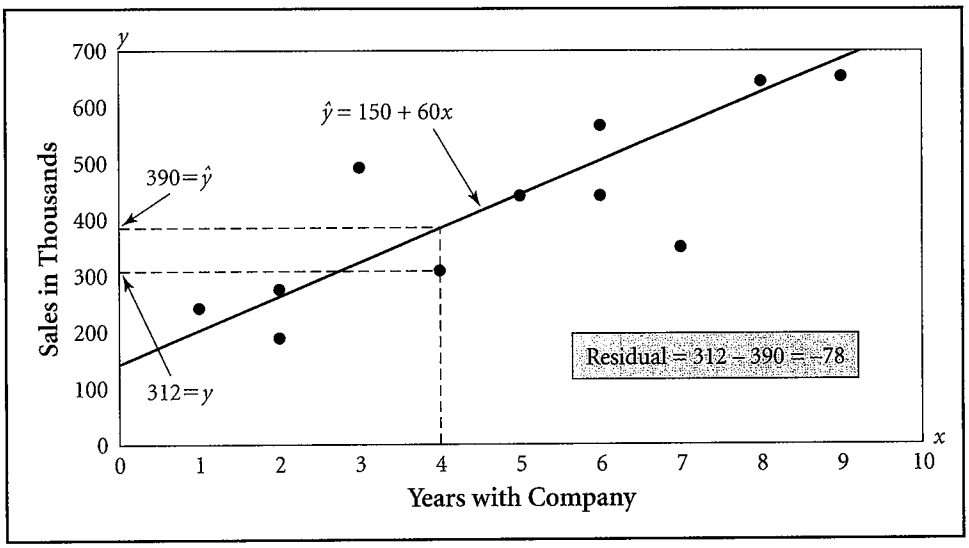
$$\hat{y} = 150 + 60(4) = 390$$

However, the actual sales (y) for this employee were 312 (see Table 14.2). Thus, when $x = 4$, the difference between the observed value, $y = 312$, and the predicted value, $\hat{y} = 390$, is $312 - 390 = -78$. The residual (or prediction error) for this case when $x = 4$ is -78. Table 14.2 shows the calculated prediction errors and sum of squared errors for each of the three regression lines shown in Figure 14.7. Of these three potential regression models, the line with the equation $\hat{y} = 150 + 60x$ has the smallest sum of squared errors. However, is this line the best of all possible lines? That is, would $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ be smaller than for any other line? One way to determine this is to calculate the sum of squared errors for all other regression lines. However, because there are an infinite number of these lines, this approach is not feasible. Fortunately, through the use of calculus,

[2] The reason we are using the sum of the squared residuals is that the sum of the residuals will be zero for the best regression line (the positive values will balance the negative values).

FIGURE 14.8

Computation of Regression Error for the Midwest Distribution Example



equations can be derived to directly determine the slope and intercept estimates such that $\Sigma_{i=1}^{n}(y_i - \hat{y}_i)^2$ is minimized.[3] This is accomplished by letting the estimated regression model be of the form shown in Equation 14.5.

CHAPTER OUTCOME 3

Estimated Regression Model (Sample Model)

$$\hat{y} = b_0 + b_1 x \tag{14.5}$$

where:

$\hat{y}$ = Estimated, or predicted, y-value
b_0 = Unbiased estimate of the regression intercept, found using Equation 14.8
b_1 = Unbiased estimate of the regression slope, found using Equation 14.6 or 14.7
x = Value of the independent variable

Equations 14.6 and 14.8 are referred to as the *least squares equations* because they provide the slope and intercept that minimize the sum of squared errors. Equation 14.7 is the algebraic equivalent of Equation 14.6 and may be easier to use when the computation is performed using a calculator.

Least Squares Equations

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

algebraic equivalent:

$$b_1 = \frac{\Sigma xy - \dfrac{\Sigma x \Sigma y}{n}}{\Sigma x^2 - \dfrac{(\Sigma x)^2}{n}} \tag{14.7}$$

and

$$b_0 = \bar{y} - b_1 \bar{x} \tag{14.8}$$

[3] The calculus derivation of the least squares equations is contained in the Kutner, et al. reference shown at the end of this chapter.

TABLE 14.2 Sum of Squared Errors for Three Linear Equations for Midwest Distribution

From Figure 14.7(a):
$\hat{y} = 450 + 0x$

			Residual	
x	$\hat{y}$	y	$y - \hat{y}$	$(y - \hat{y})^2$
3	450	487	37	1,369
5	450	445	−5	25
2	450	272	−178	31,684
8	450	641	191	36,481
2	450	187	−263	69,169
6	450	440	−10	100
7	450	346	−104	10,816
1	450	238	−212	44,944
4	450	312	−138	19,044
2	450	269	−181	32,761
9	450	655	205	42,025
6	450	563	113	12,769
				$\Sigma = 301{,}187$

From Figure 14.7(b):
$\hat{y} = 250 + 40x$

			Residual	
x	$\hat{y}$	y	$y - \hat{y}$	$(y - \hat{y})^2$
3	370	487	117	13,689
5	450	445	−5	25
2	330	272	−58	3,364
8	570	641	71	5,041
2	330	187	−143	20,449
6	490	440	−50	2,500
7	530	346	−184	33,856
1	290	238	−52	2,704
4	410	312	−98	9,604
2	330	269	−61	3,721
9	610	655	45	2,025
6	490	563	73	5,329
				$\Sigma = 102{,}307$

From Figure 14.7(c):
$\hat{y} = 150 + 60x$

			Residual	
x	$\hat{y}$	y	$y - \hat{y}$	$(y - \hat{y})^2$
3	330	487	157	24,649
5	450	445	−5	25
2	270	272	2	4
8	630	641	11	121
2	270	187	−83	6,889
6	510	440	−70	4,900
7	570	346	−224	50,176
1	210	238	28	784
4	390	312	−78	6,084
2	270	269	−1	1
9	690	655	−35	1,225
6	510	563	53	2,809
				$\Sigma = 97{,}667$

TABLE 14.3 Manual Calculations for Least Squares Regression Coefficients for the Midwest Distribution Example

y	x	xy	x^2	y^2
487	3	1,461	9	237,169
445	5	2,225	25	198,025
272	2	544	4	73,984
641	8	5,128	64	410,881
187	2	374	4	34,969
440	6	2,640	36	193,600
346	7	2,422	49	119,716
238	1	238	1	56,644
312	4	1,248	16	97,344
269	2	538	4	72,361
655	9	5,895	81	429,025
563	6	3,378	36	316,969
$\Sigma y = 4{,}855$	$\Sigma x = 55$	$\Sigma xy = 26{,}091$	$\Sigma x^2 = 329$	$\Sigma y^2 = 2{,}240{,}687$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{4{,}855}{12} = 404.58 \qquad \bar{x} = \frac{\Sigma x}{n} = \frac{55}{12} = 4.58$$

$$b_1 = \frac{\Sigma xy - \dfrac{\Sigma x \Sigma y}{n}}{\Sigma x^2 - \dfrac{(\Sigma x)^2}{n}} = \frac{26{,}091 - \dfrac{55(4{,}855)}{12}}{329 - \dfrac{(55)^2}{12}}$$

$$= 49.91$$

Then,

$$b_0 = \bar{y} - b_1\bar{x} = 404.58 - 49.91(4.58) = 175.99$$

The least squares regression line is, therefore,

$$\hat{y} = 175.99 + 49.91(x)$$

There is a slight difference between the manual calculation and the computer result due to rounding.

Table 14.3 shows the manual calculations, which are subject to rounding, for the least squares estimates for the Midwest Distribution example. However, you will almost always use a software package such as Excel or Minitab to perform these computations. (Figures 14.9a and 14.9b show the Excel and Minitab output.) In this case, the "best" regression line, given the least squares criterion, is $\hat{y} = 175.8288 + 49.9101(x)$. Figure 14.10 shows the predicted sales values along with the prediction errors and squared errors associated with this best simple linear regression line. Keep in mind that the prediction errors are also referred to as residuals. From Figure 14.10, the sum of the squared errors is 84,834.29. This is the smallest sum of squared residuals possible for this set of sample data. No other simple linear regression line through these 12 (x, y) points will produce a smaller sum of squared errors. Equation 14.9 presents a formula that can be used to calculate the sum of squared errors manually.

Sum of Squared Errors

$$SSE = \Sigma y^2 - b_0 \Sigma y - b_1 \Sigma xy \qquad \textbf{(14.9)}$$

FIGURE 14.9A

Excel 2007 Midwest Distribution Regression Results

Excel 2007 Instructions:
1. Open file: Midwest.xls.
2. On the **Data** tab, click **Data Analysis**.
3. Select **Regression Analysis**.
4. Define x (Years with Midwest) and y (Sales) variable data range.
5. Select output location.
6. Check **Labels**.
7. Click **Residuals**.
8. Click **OK**.

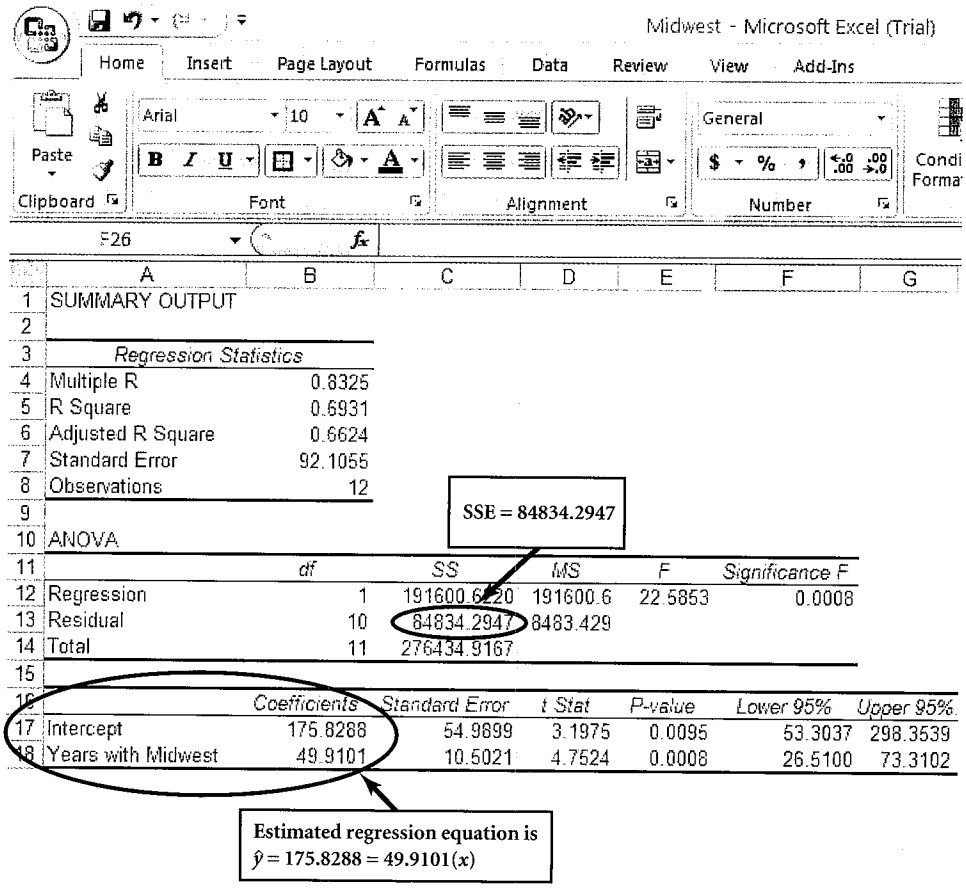


FIGURE 14.9B

Minitab Midwest Distribution Regression Results

Minitab Instructions:
1. Open file: Midwest.MTW.
2. Choose **Stat > Regression > Regression.**
3. In **Response,** enter the Y variable column.
4. In **Predictors**, enter the X variable column.
5. Click **Storage**; under **Diagnostic Measures** select **Residuals.**
6. Click **OK. OK.**

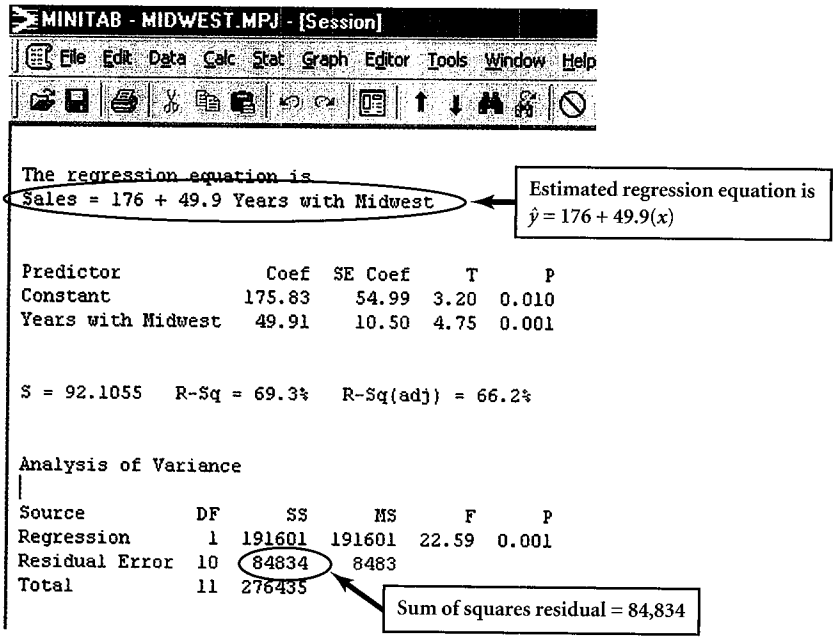


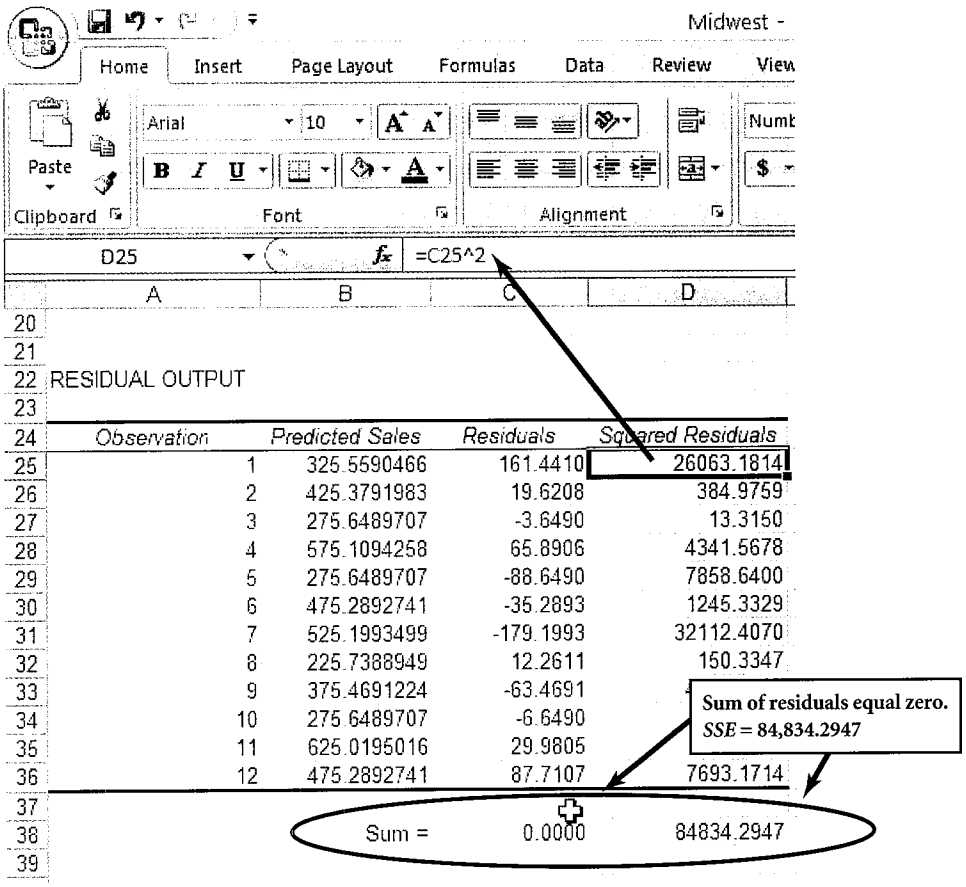
Figure 14.11 shows the scatter plot of sales and years of experience and the least squares regression line for Midwest Distribution. This line is the *best fit* for these sample data. The regression line passes through the point corresponding to $(\bar{x}, \bar{y})$. This will always be the case.

FIGURE 14.10

Residuals and Squared Residuals for the Midwest Distribution Example

Excel 2007 Instructions:
1. Create Squared Residuals using Excel formula (i.e. for cell d25, use = c25^2).
2. Sum the residuals and squared residuals columns.

Minitab Instructions (for similar results):
1. Choose **Calc > Column Statistics**.
2. Under **Statistics**, Choose **Sum**.
3. In **Input variable**, enter residual column.
4. Click **OK**.
5. Choose **Calc > Column Statistics**.
6. Under **Statistic**, choose **Sum of Squares**.
7. In **Input variable**, enter residual column.
8. Click **OK**.



Least Squares Regression Properties

Figure 14.10 illustrates several important properties of least squares regression. These are:

1. The sum of the residuals from the least squares regression line is 0 (Equation 14.10). The total underprediction by the regression model is exactly offset by the total overprediction.

Sum of Residuals

$$\sum_{i=1}^{n}(y_i - \hat{y}_i) = 0 \qquad \textbf{(14.10)}$$

2. The sum of the squared residuals is the minimum (Equation 14.11).

Sum of Squared Residuals (Errors)

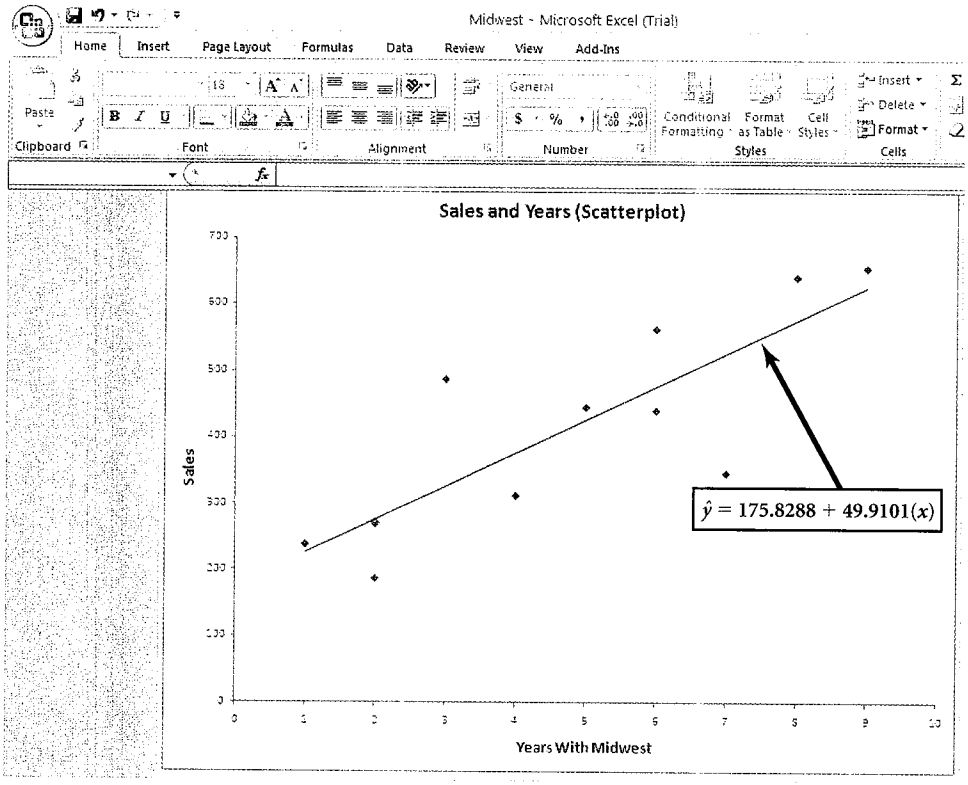
$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad \textbf{(14.11)}$$

This property provided the basis for developing the equations for b_0 and b_1.

3. The simple regression line always passes through the mean of the y variable, $\bar{y}$, and the mean of the x variable, $\bar{x}$. So, to manually draw any simple linear regression line, all you need to do is to draw a line connecting the least squares y intercept with the $(\bar{x}, \bar{y})$ point.
4. The least squares coefficients are unbiased estimates of β_0 and β_1. Thus, the expected values of b_0 and b_1 equal β_0 and β_1, respectively.

FIGURE 14.11

Least Squares Regression Line for Midwest Distribution



Excel 2007 Instructions:
1. Open file: Midwest.xls.
2. Move the *Sales* column to the right of *Years with Midwest* column.
3. Select data for chart.
4. On **Insert** tab, click **XY (Scatter)**, and then click **Scatter with only Markers** option.
5. Use the **Layout** tab of the **Chart Tools** to add titles and remove grid lines.
6. Use the **Design** tab of the **Chart Tools** to move the chart to a new worksheet.
7. Click on a chart point.
8. Right click and select **Add Trendline**.
9. Select **Linear**.

EXAMPLE 14-2 Simple Linear Regression and Correlation

TRY PROBLEM 14.20

Fitzpatrick & Associates The investment firm Fitzpatrick & Associates wants to manage the pension fund of a major Chicago retailer. For their presentation to the retailer, the Fitzpatrick analysts want to use simple linear regression to model the relationship between profits and numbers of employees for 50 Fortune 500 companies in the firm's portfolio. The data for the analysis is contained in the CD-ROM file **Fortune 50**. This analysis can be done using the following steps:

Excel and Minitab Tutorial

Step 1 Specify the independent and dependent variables.
The object in this example is to model the linear relationship between number of employees (the independent variable) and each company's profits (the dependent variable).

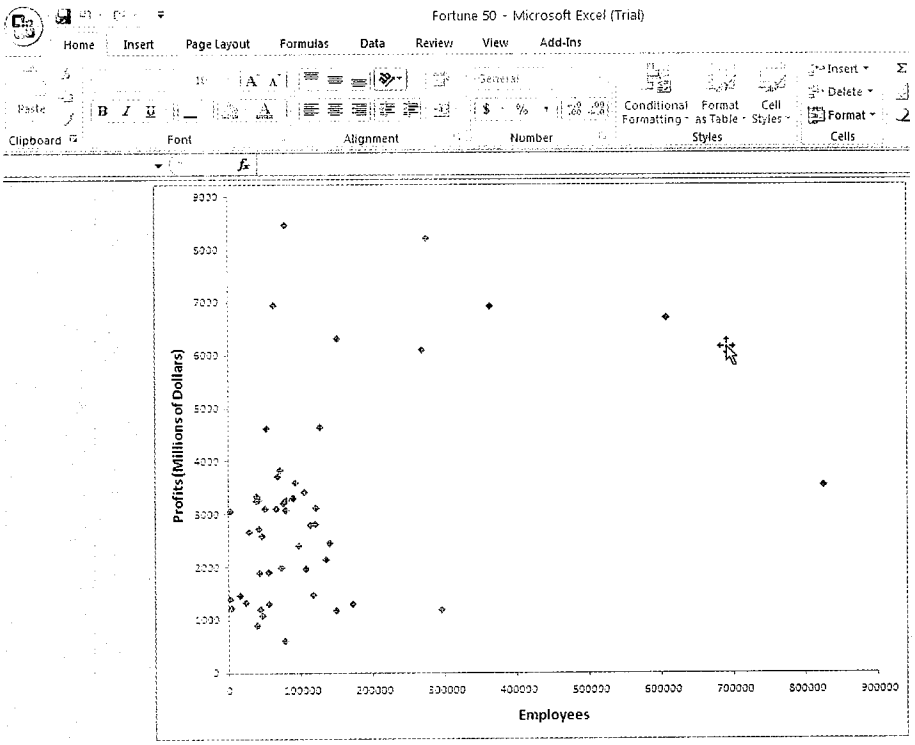
Step 2 Develop a scatter plot to graphically display the relationship between the independent and dependent variables.
Figure 14.12 shows the scatter plot, where the dependent variable, y, is company profits and the independent variable, x, is number of employees. There appears to be a slight positive linear relationship between the two variables.

Step 3 Calculate the correlation coefficient and the linear regression equation.
Do either manually using Equations 14.1, 14.6, and 14.8, respectively, or by using Excel or Minitab software. Figure 14.13 shows the regression results. The sample correlation coefficient (called multiple R in Excel) is

$$r = 0.3638$$

FIGURE 14.12

Excel 2007 Scatter Plot for Fitzpatrick & Associates



Excel 2007 Instructions:

1. Open file: Fortune 50.xls.
2. Copy the *Profits* column to the immediate right of *Employees* column.
3. Select data for chart (*Employees and Profits*).
4. On **Insert** tab, click **XY (Scatter)**, and then click the **Scatter with only Markers** option.
5. Use the **Layout** tab of the **Chart Tools** to add titles and remove grid lines.
6. Use the **Design** tab of the **Chart Tools** to move the chart to a new worksheet.

Minitab Instructions (for similar result):

1. Open file: Fortune 50.MTW.
2. Choose **Graph > Character Graphs > Scatterplot**.
3. In **Y variable**, enter *y* column.
4. In **X variable**, enter *x* column.
5. Click **OK**.

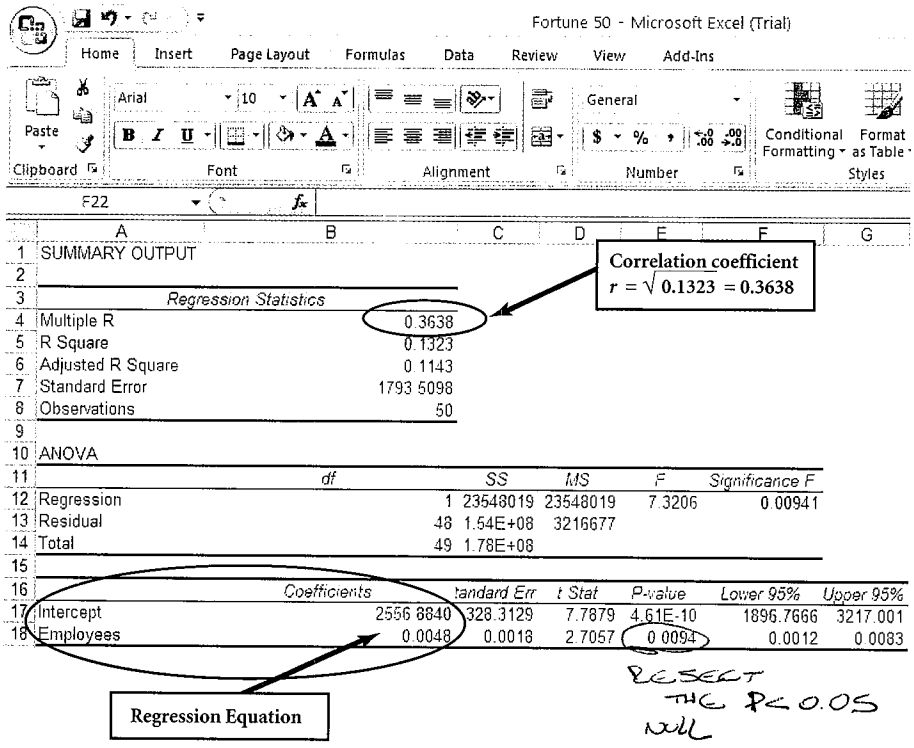
The regression equation is

$$\hat{y} = 2{,}556.88 + 0.0048x$$

The regression slope is estimated to be 0.0048, which means that for each additional employee, the average increase in company profit is 0.0048 million dollars, or $4,800. The intercept can only be interpreted when a value equal to zero for the x variable (employees) is plausible. Clearly, no company has zero employees, so the intercept in this case has no meaning other than it locates the height of the regression line for $x = 0$.

FIGURE 14.13

Excel 2007 Regression Results for Fitzpatrick & Associates



	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.3638					
5	R Square	0.1323					
6	Adjusted R Square	0.1143					
7	Standard Error	1793.5098					
8	Observations	50					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	1	23548019	23548019	7.3206	0.00941	
13	Residual	48	1.54E+08	3216677			
14	Total	49	1.78E+08				
15							
16		*Coefficients*	*tandard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	2556.8840	328.3129	7.7879	4.61E-10	1896.7666	3217.001
18	Employees	0.0048	0.0018	2.7057	0.0094	0.0012	0.0083

Excel 2007 Instructions:

1. Open file: Fortune 50.xls.
2. On the **Data** tab, click **Data Analysis**.
3. Select **Regression Analysis**.
4. Define x (Employees) and y (Profits) variable data ranges.
5. Check **Labels**.
6. Select output location.
7. Click **OK**.

Minitab Instructions (for similar result):

1. Open file: Fortune 50.MTW.
2. Choose **Stat > Regression > Regression**.
3. In **Response**, enter the Y variable column.
4. In **Predictors**, enter the X variable column.
5. Click **OK**.

CHAPTER OUTCOME #4

Significance Tests in Regression Analysis

In Section 14.1, we pointed out that the correlation coefficient computed from sample data is a point estimate of the population correlation coefficient and is subject to sampling error. We also introduced a test of significance for the correlation coefficient. Likewise, the regression coefficients developed from a sample of data are also point estimates of the true regression coefficients for the population. The regression coefficients are subject to sampling error. For example, due to sampling error the estimated slope coefficient may be positive or negative while the population slope is really 0. Therefore, we need a test procedure to determine whether the regression slope coefficient is statistically significant. As you will see in this section, the test for the simple linear regression slope coefficient is equivalent to the test for the correlation coefficient. That is, if the correlation between two variables is found to be significant, then the regression slope coefficient will also be significant.

The Coefficient of Determination, R^2

Business Application

MIDWEST DISTRIBUTION (*CONTINUED*) Recall that the Midwest Distribution marketing manager was analyzing the relationship between the number of years an employee had been with the company (independent variable) and the sales generated by the employee (dependent variable). We note when looking at the sample data for 12 employees (see Table 14.3) that sales vary among employees. Regression analysis aims to determine the extent to which an independent variable can explain this variation. In this case, does number of years with the company help explain the variation in sales from employee to employee?

The *SST* (total sum of squares) can be used in measuring the variation in the dependent variable. *SST* is computed using Equation 14.12. For Midwest Distribution, the total sum of squares for sales is provided in the output generated by Excel or Minitab, as shown in Figure 14.14a and Figure 14.14b. As you can see, the total sum of squares in sales that needs to be explained is 276,434.92. Note that the *SST* value is in squared units and has no particular meaning.

Total Sum of Squares

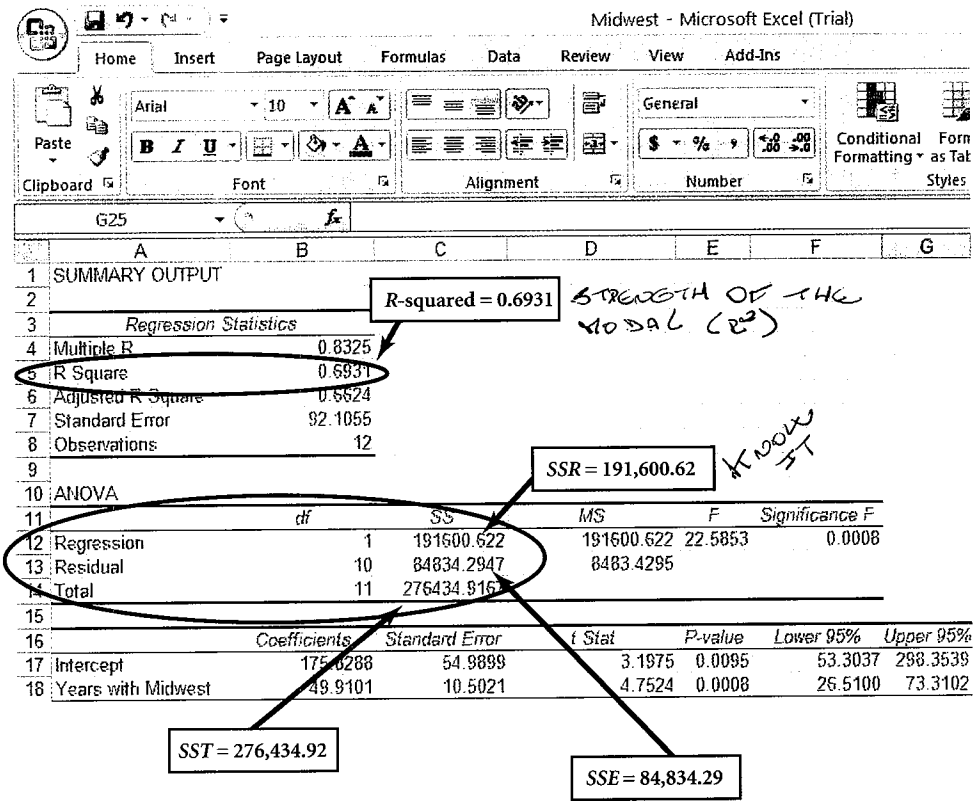
$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad (14.12)$$

where:

SST = Total sum of squares
n = Sample size
y_i = ith value of the dependent variable
$\bar{y}$ = Average value of the dependent variable

FIGURE 14.14A

Excel 2007 Regression Results for Midwest Distribution

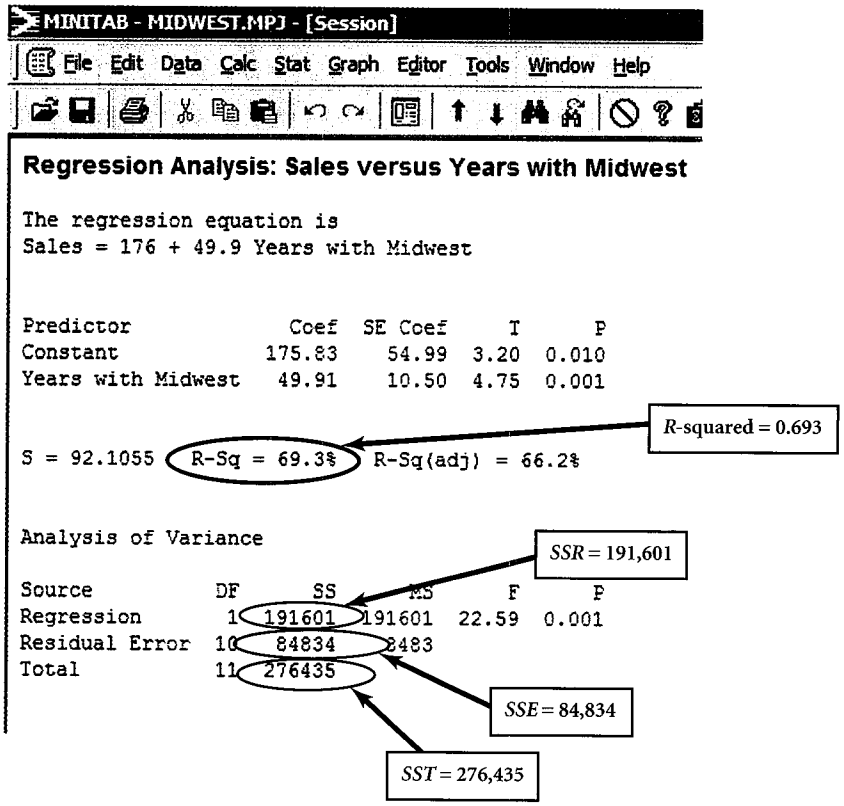


Excel 2007 Instructions:

1. Open file: Midwest.xls.
2. On the **Data** tab, click **Data Analysis**.
3. Select **Regression Analysis**.
4. Define x (Years with Midwest) and y (Sales) variable data range.
5. Click on **Labels**.
6. Specify output location.
7. Click **OK**.

FIGURE 14.14B

Minitab Regression Results for Midwest Distribution



Minitab Instructions:

1. Open file: Midwest.MTW.
2. Choose **Stat > Regression > Regression.**
3. In **Response**, enter the *y* variable column.
4. In **Predictors**, enter the *x* variable column.
5. Click **OK.**

The least squares regression line is computed so that the sum of squared residuals is minimized (recall the discussion of the least squares equations). The sum of squares residuals is also called the *sum of squares error* (*SSE*) and is defined by Equation 14.13.

Sum of Squares Error

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad \textbf{(14.13)}$$

where:

n = Sample size
y_i = ith value of the dependent variable
$\hat{y}_i$ = ith predicted value of y given the ith value of x

SSE represents the amount of the total sum of squares in the dependent variable that *is not explained* by the least squares regression line. Excel refers to *SSE* as *sum of squares residual* and Minitab refers to SSE as *residual error*. This value is contained in the regression output shown in Figure 14.14a and Figure 14.14b.

$$SSE = \Sigma(y - \hat{y})^2 = 84{,}834.29$$

Thus, of the total sum of squares (*SST* = 276,434.92), the regression model leaves *SSE* = 84,834.29 unexplained. Then, the portion of the total sum of squares that *is explained* by the regression line is called the *sum of squares regression* (*SSR*) and is calculated by Equation 14.14.

Sum of Squares Regression

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \quad \textbf{(14.14)}$$

where:

$\hat{y}_i$ = Estimated value of y for each value of x.
$\bar{y}$ = Average value of the y variable

The sum of squares regression (SSR = 191,600.62) is also provided in the regression output shown in Figure 14.14a and Figure 14.14b. You should also note that the following holds:

$$SST = SSR + SSE$$

For the Midwest Distribution example, in the Minitab output we get

$$276{,}435 = 191{,}601 + 84{,}834$$

Coefficient of Determination
The portion of the total variation in the dependent variable that is explained by its relationship with the independent variable. The coefficient of determination is also called R-squared and is denoted as R^2.

We can use these calculations to compute an important measure in regression analysis called the **coefficient of determination**. The coefficient of determination is calculated using Equation 14.15.

Coefficient of Determination, R^2

$$R^2 = \frac{SSR}{SST} \quad \textbf{(14.15)}$$

Then, for the Midwest Distribution example, the proportion of variation in sales that can be explained by its linear relationship with the years of salesforce experience is

$$R^2 = \frac{SSR}{SST} = \frac{191{,}600.62}{276{,}434.92} = 0.6931$$

This means that 69.31% of the variation in the sales data for this sample can be explained by the linear relationship between sales and years of experience. Notice that R-squared is part of the regression output in Figures 14.14a and 14.14b.

R^2 can be a value between 0 and 1.0. If there is a perfect linear relationship between two variables, then the coefficient of determination, R^2, will be 1.0. This would correspond to a situation in which the least squares regression line would pass through each of the points in the scatter plot. R^2 is the measure used by many decision makers to indicate how well the linear regression line fits the (x, y) data points. The better the fit, the closer R^2 will be to 1.0. R^2 will be close to 0 when there is a weak linear relationship.

Finally, when you are employing *simple linear regression* (a linear relationship between a single independent variable and the dependent variable), there is an alternative way of computing R^2, as shown in Equation 14.16.

Coefficient of Determination, Single Independent Variable Case

$$R^2 = r^2 \quad \textbf{(14.16)}$$

where:

R^2 = Coefficient of determination
r = Sample correlation coefficient

Therefore, by squaring the correlation coefficient, we can get R^2 for the simple regression model. Figure 14.14a shows the correlation, $r = 0.8325$, which is referred to as Multiple R in Excel. Then, using Equation 14.16, we get R^2.

$$\begin{aligned} R^2 &= r^2 \\ &= 0.8325^2 \\ &= 0.6931 \end{aligned}$$

Keep in mind that $R^2 = 0.6931$ is based on the random sample of size 12 and is subject to sampling error. Thus, just because $R^2 = 0.6931$ for the sample data does not mean that knowing the number of years an employee has worked for the company will explain 69.31% of the variation in sales for the population of all employees with the company. Likewise, just because $R^2 > 0.0$ for the sample data does not mean that the population coefficient of determination, noted as ρ^2 (rho-squared), is greater than zero.

However, a statistical test exists for testing the following null and alternative hypotheses:

$$\begin{aligned} H_0&: \rho^2 = 0 \\ H_A&: \rho^2 > 0 \end{aligned}$$

The test statistic is an F-test with the test statistic defined as shown in Equation 14.17.

Test Statistic for Significance of the Coefficient of Determination

$$F = \frac{\dfrac{SSR}{1}}{\dfrac{SSE}{(n-2)}} \quad df = (D_1 = 1, D_2 = n-2) \tag{14.17}$$

where:

SSR = Sum of squares regression
SSE = Sum of squares error

For the Midwest Distribution example, the test statistic is computed using Equation 14.17 as follows:

$$F = \frac{\dfrac{191{,}600.62}{1}}{\dfrac{84{,}834.29}{(12-2)}} = 22.58$$

The critical value from the F-distribution table in Appendix H for $\alpha = 0.05$ and for 1 and 10 degrees of freedom is 4.965. This gives the following decision rule:

If $F > 4.965$, reject the null hypothesis;
Otherwise, do not reject the null hypothesis.

Because $F = 22.58 > 4.965$, we reject the null hypothesis and conclude the population coefficient of determination (ρ^2) is greater than zero. This means the independent variable explains a significant proportion of the variation in the dependent variable.

For a simple regression model (a regression model with a single independent variable), the test for ρ^2 is equivalent to the test shown earlier for the population correlation coefficient, ρ. Refer to Figure 14.5 to see that the t-test statistic for the correlation coefficient was $t = 4.752$. If we square this t-value we get

$$t^2 = 4.752^2 = F = 22.58$$

Thus, the tests are equivalent. They will provide the same conclusions about the relationship between the x and y variables.

Significance of the Slope Coefficient For a simple linear regression model (one independent variable), there are three equivalent statistical tests.

1. Test for significance of the correlation between x and y.
2. Test for significance of the coefficient of determination.
3. Test for significance of the regression slope coefficient.

We have already introduced the first two of these tests. The third one deals specifically with the significance of the regression slope coefficient. The null and alternative hypotheses to be tested are

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

To test the significance of the simple linear regression slope coefficient, we are interested in determining whether the population regression slope coefficient is 0. A slope of 0 would imply that there is no linear relationship between x and y variables and the x variable, in its linear form is of no use in explaining the variation in y. If the linear relationship is useful, then we should reject the hypothesis that the regression slope is 0. However, because the estimated regression slope coefficient, b_1, is calculated from sample data, it is subject to sampling error. Therefore, even though b_1 is not 0, we must determine whether its difference from 0 is greater than would generally be attributed to sampling error.

If we selected several samples from the same population and for each sample determined the least squares regression line, we would likely get regression lines with different slopes and different y intercepts. This is analogous to getting different sample means from different samples. Just as the distribution of possible sample means has a standard error, the possible regression slopes have a standard error, which is given in Equation 14.18.

Simple Regression Standard Error of the Slope Coefficient (Population)

$$\sigma_{b_1} = \frac{\sigma_\varepsilon}{\sqrt{\Sigma(x - \bar{x})^2}} \tag{14.18}$$

where:

σ_{b_1} = Standard deviation of the regression slope (called the *standard error of the slope*)

σ_ε = Population standard error of the estimate

Equation 14.18 requires that we know the *standard error of the estimate*. It measures the dispersion of the dependent variable about its mean value at each value of the dependent variable in the original units of the dependent variable. However, because we are sampling from the population, we can estimate σ_ε as shown in Equation 14.19.

Simple Regression Estimator for the Standard Error of the Estimate

$$s_\varepsilon = \sqrt{\frac{SSE}{n-2}} \tag{14.19}$$

where:

SSE = Sum of squares error

n = Sample size

Equation 14.18, the standard error of the regression slope, applies when we are dealing with a population. However, in most cases, such as the Midwest Distribution example, we are dealing with a sample from the population. Thus, we need to estimate the regression slope's standard error using Equation 14.20.

Simple Regression Estimator for the Standard Deviation of the Slope

$$s_{b_1} = \frac{s_\varepsilon}{\sqrt{\Sigma(x-\bar{x})^2}} \tag{14.20}$$

where:

s_{b_1} = Estimate of the standard error of the least squares slope

$s_\varepsilon = \sqrt{\frac{SSE}{n-2}}$ = Sample standard error of the estimate (the measure of deviation of the actual y-values around the regression line)

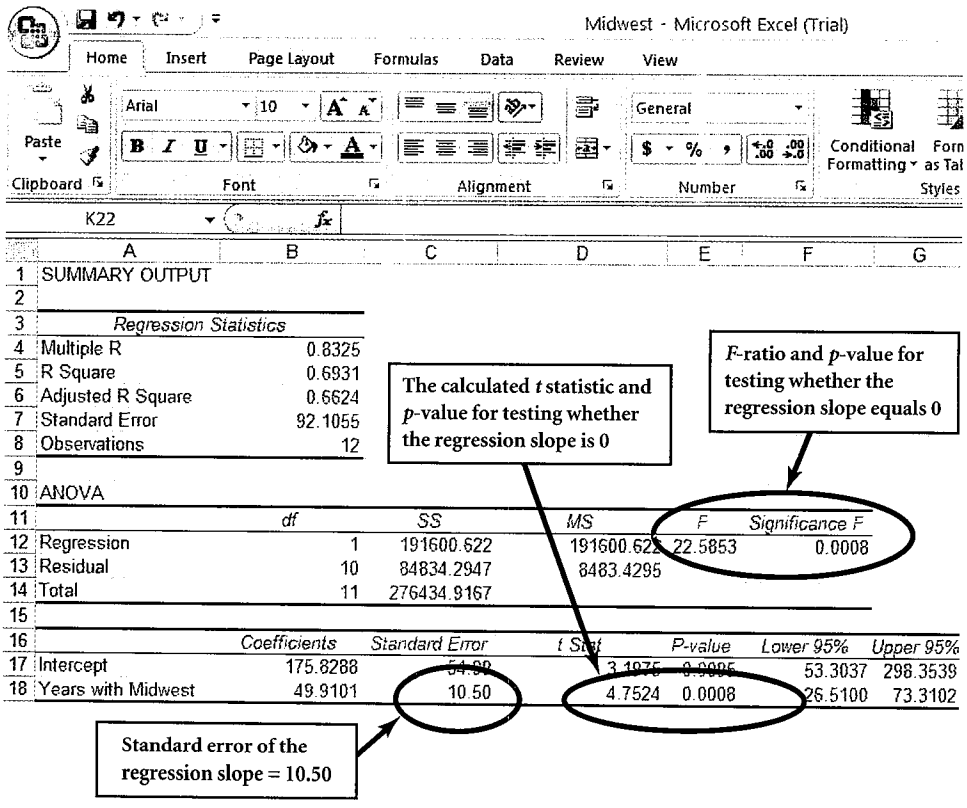
Business Application

MIDWEST DISTRIBUTION (*CONTINUED*) For Midwest Distribution, the regression outputs in Figures 14.15a and Figure 14.15b show $b_1 = 49.91$. The question is whether this value is different enough from 0 to have not been caused by sampling error. We find the answer by looking at the value of the estimate of the standard error of the slope, calculated using Equation 14.20, which is also shown in Figure 14.15a. The standard error of the slope coefficient is 10.50.

If the standard error of the slope σ_{b_1} is large, then the value of b_1 will be quite variable from sample to sample. Conversely, if σ_{b_1} is small, the slope values will be less variable. However, regardless of the standard error of the slope, the average value of b_1 will equal β_1, the true regression slope, if the assumptions of the regression analysis are satisfied. Figure 14.16 illustrates what this means. Notice that when the standard error of the slope is large, the sample slope can take on values *much* different from the true population slope. As Figure 14.16(a) shows, a sample slope and the true population slope can even have different signs. However, when σ_{b_1} is small, the sample regression lines will cluster closely around the true population line [Figure 14.16(b)].

FIGURE 14.15A

Excel 2007 Regression Results for Midwest Distribution

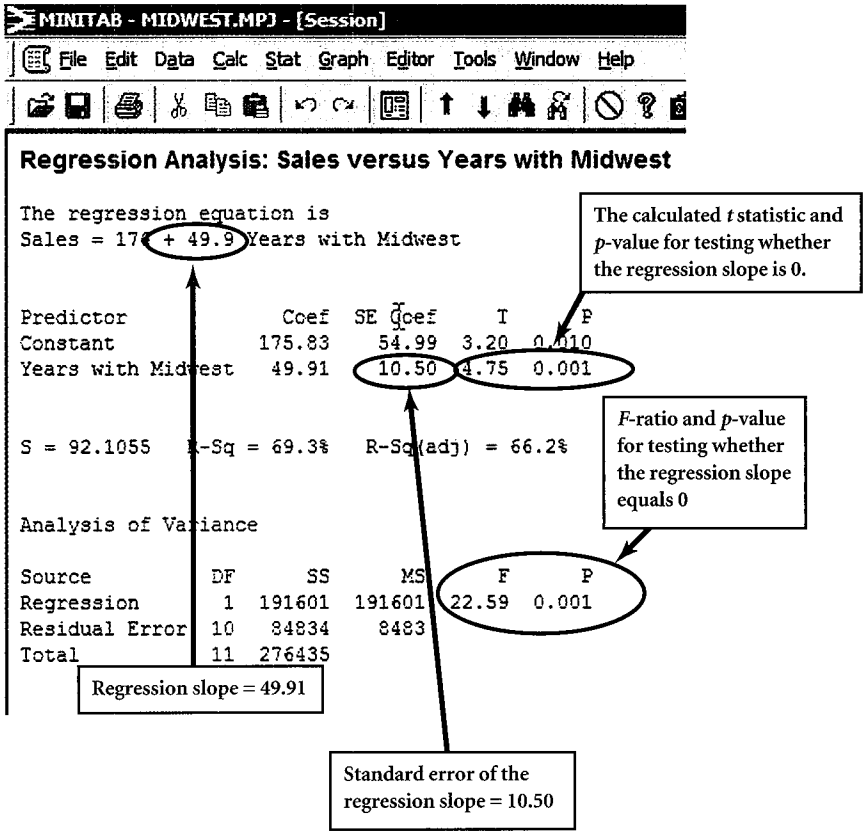


Excel 2007 Instructions:

1. Open file: Midwest.xls.
2. On the **Data** tab, click **Data Analysis**.
3. Select **Regression Analysis**.
4. Define x (Years with Midwest) and y (Sales) variable data range.
5. Click on **Labels**.
6. Specify output location.
7. Click **OK**.

FIGURE 14.15B

Minitab Regression Results for Midwest Distribution



Minitab Instructions:

1. Open file: Midwest MTW.
2. Choose **Stat > Regression > Regression.**
3. In **Response**, enter the y variable column.
4. In **Predictors** enter the x variable column.
5. Click **OK.**

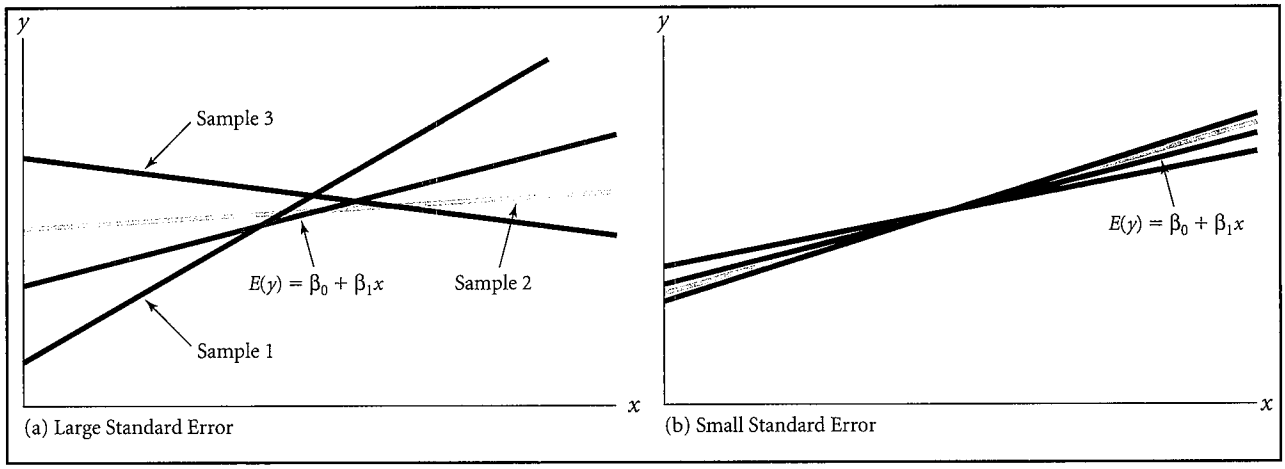
Because the sample regression slope will most likely not equal the true population slope, we must test to determine whether the true slope could possibly be 0. A slope of 0 in the linear model means that the independent variable will not explain any variation in the dependent variable, nor will it be useful in predicting the dependent variable. The null and alternative hypotheses to be tested at the 0.05 level of significance are

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

To test the significance of a slope coefficient, we use the t-test value in Equation 14.21.

FIGURE 14.16

Standard Error of the Slope



CHAPTER OUTCOME #4

Simple Linear Regression Test Statistic for Test of the Significance of the Slope

$$t = \frac{b_1 - \beta_1}{s_{b_1}} \qquad df = n - 2 \tag{14.21}$$

where:

b_1 = Sample regression slope coefficient
β_1 = Hypothesized slope (usually $\beta_1 = 0$)
s_{b_1} = Estimator of the standard error of the slope

Figure 14.17 illustrates this test for the Midwest Distribution example. The calculated *t*-value of 4.752 exceeds the critical value from the *t*-distribution with 10 degrees of freedom and $\alpha/2 = 0.025$. This indicates that we should reject the hypothesis that the true regression slope is 0. Thus, years of experience can be used to help explain the variation in an individual representative's sales. (Note that the calculated *t* is the same value that we found in Figure 14.6 for the test of the correlation coefficient. Thus, this test is equivalent to the tests for ρ and ρ^2 presented earlier.

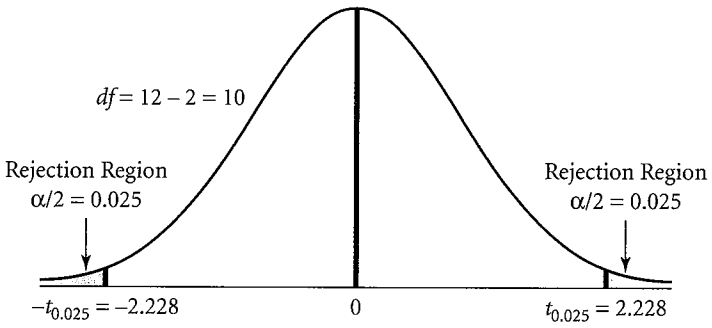
The output shown in Figures 14.15a and 14.15b also contains the calculated *t* statistic. The *p*-value for the calculated *t* statistic is also provided. As with other situations involving two-tailed hypothesis tests, if the *p*-value is less than α, the null hypothesis is rejected. In this case, because *p*-value = 0.0008 < 0.05, we reject the null hypothesis.

FIGURE 14.17

Significance Test of the Regression Slope for Midwest Distribution

Hypotheses:

H_0: $\beta_1 = 0$
H_A: $\beta_1 \neq 0$
$\alpha = 0.05$



The calculated *t* is

$$t = \frac{b_1 - \beta_1}{s_{b_1}} = \frac{49.91 - 0}{10.50} = 4.752$$

Decision Rule:

If $t > t_{0.025} = 2.228$, reject H_0.
If $t < -t_{0.025} = -2.228$, reject H_0.
Otherwise, do not reject H_0.

Because 4.752 > 2.228, we should reject the null hypothesis and conclude that the true slope is not 0. Thus, the simple linear relationship that utilizes the independent variable, years with the company, is useful in explaining the variation in the dependent variable, sales volume.

SUMMARY Simple Linear Regression Analysis

The following steps outline the process that can be used in developing a simple linear regression model and the various hypotheses tests used to determine the significance of a simple linear regression model.

1. Define the independent (x) and dependent (y) variables and select a simple random sample of pairs of (x, y) values.
2. Develop a scatter plot of y and x. You are looking for a linear relationship between the two variables.
3. Compute the correlation coefficient for the sample data.
4. Calculate the least squares regression line for the sample data and the coefficient of determination, R^2. The coefficient of determination measures the proportion of variation in the dependent variable explained by the independent variable.
5. Conduct any of the following tests for determining whether the regression model is statistically significant.
   a. Test to determine whether the true regression slope is 0. The test statistic with $df = n - 2$ is
   $$t = \frac{b_1 - \beta_1}{s_{b_1}} = \frac{b_1 - 0}{s_{b_1}}$$
   b. Test to see whether ρ is significantly different from 0. The test statistic is
   $$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$
   c. Test to see whether ρ^2 is significantly greater than 0. The test statistic is
   $$F = \frac{\frac{SSR}{1}}{\frac{SSE}{(n-2)}}$$
6. Reach a decision.
7. Draw a conclusion.

EXAMPLE 14-3 Simple Linear Regression Analysis

Excel and Minitab Tutorial

TRY PROBLEM 14.17

Vantage Electronic Systems Consider the example involving Vantage Electronic Systems in Deerfield, Michigan, which started out supplying electronic equipment for the automobile industry but in recent years has ventured into other areas. One area is visibility sensors that are used by airports to provide takeoff and landing information and by transportation departments to detect low visibility on roadways during fog and snow. The recognized leader in the visibility sensor business is the SCR Company, which makes a sensor called the Scorpion. The R&D department at Vantage has recently performed a test on its new unit by locating a Vantage sensor and a Scorpion sensor side-by-side. Various data, including visibility measurements, were collected at randomly selected points in time over a two-week period. These data are contained in a CD-ROM file called **Vantage**.

Step 1 Define the independent (x) and dependent (y) variables.
The analysis included a simple linear regression using the Scorpion visibility measurement as the dependent variable, y, and the Vantage visibility measurement as the independent variable, x.

Step 2 Develop a scatter plot of y and x.
The scatter plot is shown in Figure 14.18. There does not appear to be a strong linear relationship.

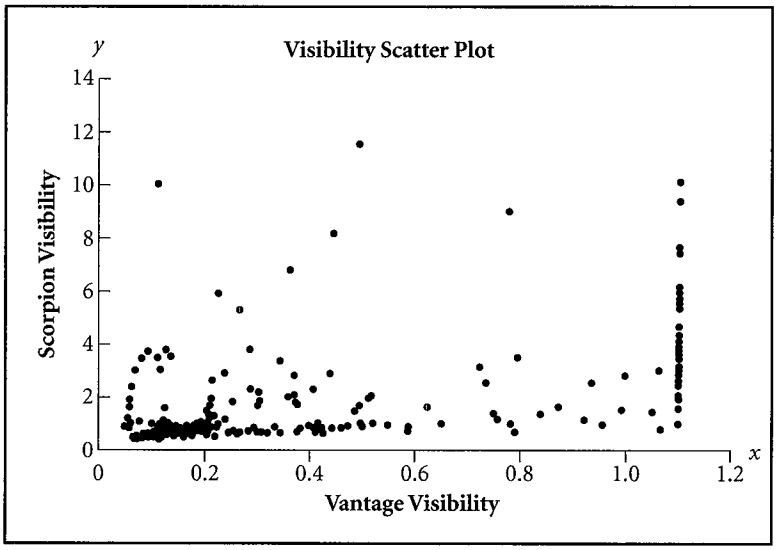
Step 3 Compute the correlation coefficient for the sample data.
Equation 14.1 or 14.2 can be used for manual computation, or we can use Excel or Minitab. The sample correlation coefficient is

$$r = 0.5778$$

Step 4 Calculate the least squares regression line for the sample data and the coefficient of determination, R^2.
Equations 14.7 and 14.8 can be used to manually compute the regression slope coefficient and intercept, respectively, and Equation 14.15 or 14.16

FIGURE 14.18

Scatter Plot—Example 14-3



can be used to manually compute R^2. Excel and Minitab can also be used to eliminate the computational burden. The coefficient of determination is

$$R^2 = r^2 = 0.5778^2 = 0.3339$$

Thus, approximately 33% of the variation in the Scorpion visibility measures is explained by knowing the corresponding Vantage system visibility measure. The least squares regression equation is

$$\hat{y} = 0.586 + 3.017x$$

Step 5 Conduct a test to determine whether the regression model is statistically significant (or whether the population correlation is equal to 0).

The null and alternative hypotheses to test the correlation coefficient are

$$H_0: \rho = 0$$
$$H_A: \rho \neq 0$$

The t-test statistic using Equation 14.3 is

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.5778}{\sqrt{\frac{1-0.5778^2}{280-2}}} = 11.8$$

The $t = 11.8$ exceeds the critical t for any reasonable level of α for 278 degrees of freedom, so the null hypothesis is rejected and we conclude that there is a statistically significant linear relationship between visibility measures for the two visibility sensors.

Alternatively, the null and alternative hypotheses to test the regression slope coefficient are

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

The t-test statistic is

$$t = \frac{b_1 - \beta_1}{s_{b_1}} = \frac{3.017 - 0}{0.2557} = 11.8$$

Step 6 **Reach a decision.**
The t-test statistic of 11.8 exceeds the t-critical for any reasonable level of α for 278 degrees of freedom.

Step 7 **Draw a conclusion.**
The regression slope coefficient is not equal to 0. This means that knowing the Vantage visibility reading provides useful help in knowing what the Scorpion visibility reading will be.

14-2: Exercises

Skill Development

14-16. You are given the following sample data for variables x and y:

x (independent)	y (dependent)
1	16
7	50
3	22
8	59
11	63
5	46
4	43

a. Construct a scatter plot for these data and describe what, if any, relationship appears to exist.
b. Compute the regression equation based on these sample data and interpret the regression coefficients.
c. Based on the sample data, what percentage of the total variation in the dependent variable can be explained by the independent variable?
d. Test the significance of the overall regression model using a significance level of 0.01.
e. Test to determine whether the true regression slope coefficient is equal to 0. Use a significance level of 0.01.

14-17. The following data for the dependent variable, y, and the independent variable, x, have been collected using simple random sampling:

x	y
10	120
14	130
16	170
12	150
20	200
18	180
16	190
14	150
16	160
18	200

a. Develop a simple linear regression equation for these data.
b. Calculate the sum of squared residuals, the total sum of squares, and the coefficient of determination.
c. Calculate the standard error of the estimate.
d. Calculate the standard deviation for the regression slope.
e. Conduct the hypothesis test to determine whether the regression slope coefficient is equal to 0. Test using $\alpha = 0.02$.

14-18. You are given the following sample data for variables y and x:

y	140.1	120.3	80.8	100.7	130.2	90.6	110.5	120.2	130.4	130.3	100.1
x	5	3	2	4	5	4	4	5	6	5	4

a. Develop a scatter plot for these data and describe what, if any, relationship exists.

b. (1) Compute the correlation coefficient. (2) Test to determine whether the correlation is significant at the significance level of 0.05. Conduct this hypothesis test using the p-value approach. (3) Compute the regression equation based on these sample data and interpret the regression coefficients.
c. Use the F-test approach to test the significance of the overall regression model using significance level equal to 0.05.

14-19. Consider the following sample data for the variables y and x:

x	30.3	4.8	15.2	24.9	8.6	20.1	9.3	11.2
y	14.6	27.9	17.6	15.3	19.8	13.2	25.6	19.4

a. Calculate the linear regression equation for these data.
b. Determine the predicted y-value when $x = 10$.
c. Estimate the change in the y variable resulting from the increase in the x variable of 10 units.
d. Conduct a hypothesis test to determine if an increase of 1 unit in the x variable will result in the decrease of the average value of the y variable. Use a significance of 0.025.

14-20. Examine the following sample data for the variables y and x:

x	1	2	3	4	5
y	4	2	5	8	9

a. Construct a scatter plot of these data. Describe the relationship between x and y.
b. Calculate the sum of squares error for the following equations: (1) $\hat{y} = 0.8 + 1.60x$, (2) $\hat{y} = 1 + 1.50x$, and (3) $\hat{y} = 0.7 + 1.60x$.
c. Which of these equations provides the "best" fit of these data? Describe the criterion you used to determine "best" fit.
d. Determine the regression line that minimizes the sum of squares error.

Business Applications

14-21. A shipping company believes that the variation in the cost of a customer's shipment can be explained by differences in the weight of the package being shipped. In order to investigate whether this relationship is useful, a random sample of 20 customer shipments was selected and the weight (in lb) and the cost (in dollars, rounded to the nearest dollar) for each shipment were recorded. The following results were obtained:

Weight (x)	Cost (y)
8	11
6	8
5	11
7	11
12	17
9	11
17	27
13	16
8	9
18	25
17	21
17	24
10	16
20	24
9	21
5	10
13	21
6	16
6	11
12	20

a. Construct a scatter plot for these data. What, if any, relationship appears to exist between the two variables?
b. Compute the linear regression model based on the sample data. Interpret the slope and intercept coefficients.
c. Test the significance of the overall regression model using a significance level equal to 0.05.
d. What percentage of the total variation in shipping cost can be explained by the regression model you developed in part b?

14-22. The Skelton Manufacturing Company recently did a study of its customers. A random sample of 50 customer accounts was pulled from the computer records. Two variables were observed:

y = Total dollar volume of business this year
x = Miles customer is from corporate headquarters

The following statistics were computed:

$$\hat{y} = 2{,}140.23 - 10.12x$$
$$s_{b_1} = 3.12$$

a. Interpret the regression slope coefficient.
b. Using a significance level of 0.01, test to determine whether it is true that the further a customer is from the corporate headquarters, the smaller is the total dollar volume of business.

14-23. College tuition has risen at a pace faster than inflation for more than two decades according to an

article in *USA Today* (Andriene Lewis, "Tuition outstrips inflation," January 5, 2004). The following data indicate the average college tuition (in 2003 dollars) for private and public colleges:

Period	1983–84	1988–89	1993–94	1998–99	2003–04
Private	9,202	12,146	13,844	16,454	19,710
Public	2,074	2,395	3,188	3,632	4,694

a. Conduct a simple linear regression analysis of these data in which the average tuition for private colleges is predicted by the average public college tuition. Test the significance of the model using an $\alpha = 0.10$.
b. How much does the average private college tuition increase when the average public college tuition increases by $100?
c. When the average public college tuition reaches $7,500, how much would you expect the average private college tuition to be?

Computer Database Exercises

14-24. The consumer price index (CPI) is a measure of the average change in prices over time in a fixed market basket of goods and services typically purchased by consumers. The CPI for all urban consumers covers about 80% of the total population. It is prepared and published by the Bureau of Labor Statistics of the Department of Labor, which measures average changes in prices of goods and services. The Consumer Price Index is one way the government measures the general level of inflation. The annual percentage change in the value of this index is one way of measuring the annual inflation rate. The file entitled **CPI** contains the monthly CPI and inflation rate for the period 2000–2005.
a. Construct a scatter plot of the CPI versus inflation for the period 2000–2005. Describe the relationship that appears to exist between these two variables.
b. Conduct a hypothesis test to confirm your preconception of the relationship between the CPI and the inflation rate. Use $\alpha = 0.05$.
c. Does it appear that the CPI and the inflation rate are measuring the same component of our economy? Support your assertion with statistical reasoning.

14-25. The National Football League is arguably the most successful professional sports league in the United States. Following the 2005 season, the NFL commissioner's office staff performed an analysis in which a simple linear regression model was developed with average home attendance used as the dependent variable and the total number of games won during the season as the independent variable. The staff was interested in determining whether games won could be used as a predictor for average attendance. Develop the simple linear regression model. (*Data source*: http://sports.espn.go.com.) The data are in the file called **NFL-2005**.
a. What percentage of total variation in average home attendance is explained by knowing the number of games the team won?
b. What is the standard error of the estimate for this regression model?
c. Using $\alpha = 0.05$, test to determine whether the regression slope coefficient is significantly different from 0.
d. After examining the regression analysis results, what should the NFL staff conclude about how the average attendance is related to the number of games the team won during the 2005 season?

14-26. The file **Online** contains a random sample of 48 customers who made purchases last quarter from an online retailer. The file contains information related to the time each customer spent viewing the online catalog and the dollar amount of purchases made. The retailer would like to analyze the sample data to determine whether a relationship exists between the time spent viewing the online catalog and the dollar amount of purchases.
a. Compute the regression equation based on these sample data and interpret the regression coefficients.
b. Compute the coefficient of determination and interpret its meaning.
c. Test the significance of the overall regression model using a significance level of 0.01.
d. Test to determine whether the true regression slope coefficient is equal to 0. Use a significance level of 0.01 to conduct the hypothesis test.

14-27. In a press release entitled "College Board Offers Glimpse of New SAT with Writing for Upcoming Class of '06," August 30, 2005, the College Board announced SAT scores for students in the class of 2005, the last to take the former version of the SAT featuring math and verbal sections. The board indicated that for the class of 2005, the average SAT math scores continued their strong upward trend, increasing from 518 in 2004 to 520 this year, 14 points above 10 years ago and an all-time high. The file entitled **MathSAT** contains the math SAT scores for the interval 1967 to 2005.
a. Produce a scatter plot of the average SAT math scores versus the year they were taken for all students (male and female) during the last 10 years (1996–2005).
b. Construct a regression equation to predict the average math scores with the year as the predictor.

c. Use regression to determine if the College Board's assertion concerning the improvement in SAT average math test scores over the last 10 years is overly optimistic.

14-28. One of the editors of a major automobile publication has collected data on 30 of the best-selling cars in the United States. The data are in a file called **Automobiles**. The editor is particularly interested in the relationship between highway mileage and curb weight of the vehicles.

a. Develop a scatter plot for these data. Discuss what the plot implies about the relationship between the two variables. Assume that you wish to predict highway mileage by using vehicle curb weight.

b. Compute the correlation coefficient for the two variables and test to determine whether there is a linear relationship between the curb weight and the highway mileage of automobiles.

c. (1) Compute the linear regression equation based on the sample data. (2) Cadillac's 1999 Sedan DeVille weighs approximately 4,012 pounds. Provide an estimate of the average highway mileage you would expect to obtain from this model.

14-29. *The Insider View of Las Vegas* Web site (www.insidervlv.com) furnishes information and facts concerning Las Vegas. A set of data published by them provides the amount of gaming revenue for various portions of Clark County, Nevada. The file entitled **VEGAS** provides the gaming revenue for the year 2005.

a. Compute the linear regression equation to predict the gaming revenue for Clark County based on the gaming revenue of the Las Vegas Strip.

b. Conduct a hypothesis test to determine if the gaming revenue from the Las Vegas Strip can be used to predict the gaming revenue for all of Clark County.

c. Estimate the increased gaming revenue that would accrue to all of Clark County if the gaming revenue on the Las Vegas Strip were to increase by a million dollars.

14.3 Uses for Regression Analysis

CHAPTER OUTCOME #5

Regression analysis is a statistical tool that is used for two main purposes: description and prediction. This section discusses these two applications.

Regression Analysis for Description

Business Application

CAR MILEAGE In the summer of 2006, gasoline prices soared to record levels in the United States, heightening motor vehicle customers' concern for fuel economy. Analysts at a major automobile company collected data on a variety of variables for a sample of 30 different cars and small trucks. Included among those data were the EPA highway mileage rating and the horsepower of each vehicle. The analysts were interested with the relationship between horsepower (x) and highway mileage (y). The data are contained in the file **Automobiles** on the CD-ROM.

Excel and Minitab Tutorial

A simple linear regression model can be developed using Excel or Minitab. The Excel output is shown in Figure 14.19. For these sample data, the coefficient of determination, $R^2 = 0.3016$, indicates that knowing the horsepower of the vehicle explains 30.16% of the variation in the highway mileage. The estimated regression equation is

$$\hat{y} = 31.1658 - 0.0286x$$

Before the analysts attempt to describe the relationship between horsepower and highway mileage, they first need to test whether there is a statistically significant linear relationship between the two variables. To do this, they can apply the t-test described in Section 14.2 to test the following null and alternative hypotheses:

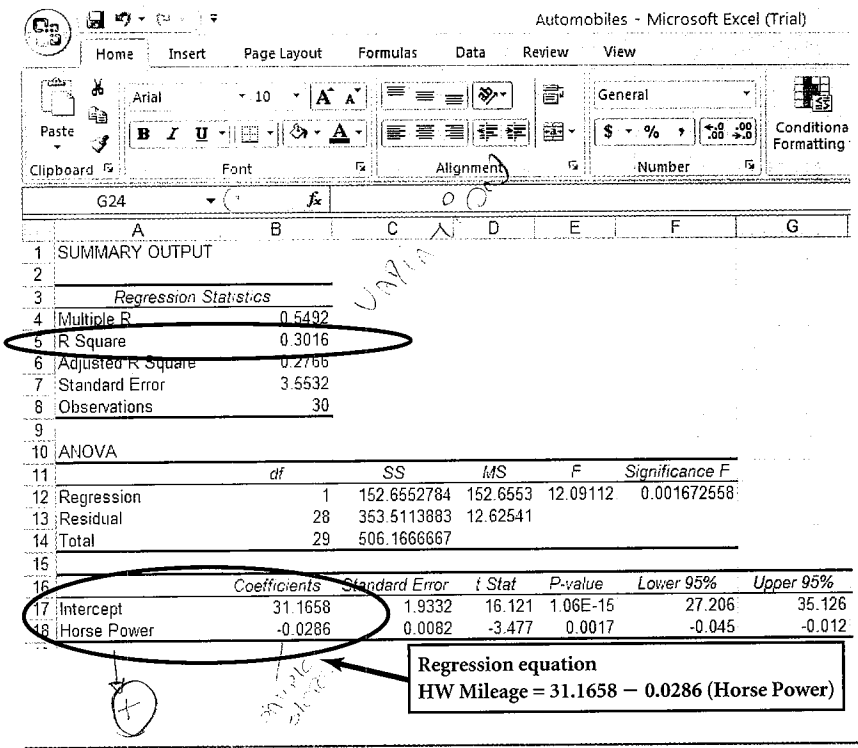
$$H_0: \beta_1 = 0$$
$$H_A: \beta_1 \neq 0$$

at the significance level

$$\alpha = 0.05$$

FIGURE 14.19

Excel 2007 Regression Results for the Automobile Mileage Study



Excel 2007 Instructions:
1. Open file: Automobiles.xls.
2. Click on **Data** tab.
3. Select **Data Analysis > Regression**.
4. Define *y* variable range (Highway Mileage) and *x* variable range (Horse Power).
5. Check **Labels**.
6. Specify Output Location.
7. Click **OK**.

Minitab Instructions (for similar results):
1. Open file: Automobiles.MTW.
2. Choose **Stat > Regression > Regression.**
3. In **Response**, enter the *y* variable column.
4. In **Predictors,** enter the *x* variable column.
5. Click **OK**.

The calculated t statistic and the corresponding p-value are shown in Figure 14.19. Because the

$$p\text{-value (Significance F)} = 0.0017 < 0.05$$

the null hypothesis, H_0, is rejected and the analysts can conclude that the population regression slope is not equal to 0.

The sample slope, b_1, equals -0.0286. This means that for each 1-unit increase in horsepower, the highway mileage decreases by an average of 0.0286 miles per gallon. However, b_1 is subject to sampling error and is considered a *point estimate* for the true regression slope coefficient. From earlier discussions about point estimates in Chapters 8 and 10, we expect that $b_1 \neq \beta_1$. Therefore, to fully describe the relationship between the independent variable, horsepower, and the dependent variable, highway miles per gallon, we need to develop a *confidence interval estimate* for β_1. Equation 14.22 is used to do this.

Confidence Interval Estimate for the Regression Slope, Simple Linear Regression

$$b_1 \pm t s_{b_1} \qquad \textbf{(14.22)}$$

or equivalently,

$$b_1 \pm t \frac{s_\varepsilon}{\sqrt{\Sigma(x-\bar{x})^2}} \qquad df = n-2$$

here:

s_{b_1} = Standard error of the regression slope coefficient

s_ε = Standard error of the estimate

The regression output shown in Figure 14.19 contains the 95% confidence interval estimate for the slope coefficient, which is

$$-0.045 \text{ ---------- } -0.012$$

Thus, at the 95% confidence level, based on the sample data, the analysts for the car company can conclude that a 1-unit increase in horsepower will result in a drop in mileage by an average amount between 0.012 and 0.045 miles per gallon.

There are many other situations in which the prime purpose of regression analysis is description. Economists use regression analysis for descriptive purposes as they search for a way of explaining the economy. Market researchers also use regression analysis, among other techniques, in an effort to describe the factors that influence the demand for products.

EXAMPLE 14-4 Developing a Confidence Interval Estimate for the Regression Slope

TRY PROBLEM 14.31

Home Prices Home values are determined by a variety of factors. One factor is the size of the house (square feet). Recently, a study was conducted by First City Real Estate aimed at estimating the average value of each additional square foot of space in a house. A simple random sample of 319 homes that were sold within the past year was collected. The data are in a file called **First-City** on the CD-ROM. Here are the steps required to compute a confidence interval estimate for the regression slope coefficient:

Step 1 Define the y (dependent) and x (independent) variables.
The dependent variable is sales price, and the independent variable is square feet.

Step 2 Obtain the sample data.
The study consists of sales prices and corresponding square feet for a random sample of 319 homes.

Step 3 Compute the regression equation and the standard error of the slope coefficient.
These computations can be performed manually using Equations 14.7 and 14.8 for the regression model and Equation 14.20 for the standard error of the slope. Alternatively, we can use Excel or Minitab to obtain these values.

	Coefficients	Standard Error
Intercept (b_0)	39,838.48	7,304.95
Square Feet (b_1)	75.70	3.78

The point estimate for the regression slope coefficient is \$75.70. Thus, for a 1-square-foot increase in the size of a house, house prices increase by an average of \$75.70. This is a point estimate and is subject to sampling error.

Step 4 Construct and interpret the confidence interval estimate for the regression slope using Equation 14.22.
The confidence interval estimate is

$$b_1 \pm ts_{b_1}$$

where the degrees of freedom for the critical t is $319 - 2 = 317$. The critical t for a 95% confidence interval estimate is approximately 1.96, and the interval estimate is

$$\$75.70 \pm 1.96(\$3.78)$$
$$\$75.70 \pm \$7.41$$
$$\$68.29 \text{ ---------- } \$83.11$$

So, for a 1-square-foot increase in house size, at the 95% confidence level, homes increase in price by an average of between \$68.29 and \$83.11.

Regression Analysis for Prediction

Business Application

CHAPTER OUTCOME #5

FREEDOM HOSPITAL One of the main uses of regression analysis is *prediction*. You may need to predict the value of the dependent variable based on the value of the independent variable. Consider the administrator for Freedom Hospital, who has been asked by the hospital's board of directors to develop a model to predict the total charges for a geriatric patient. The CD-ROM file **Patients** contains the data that the administrator has collected.

Although the Regression tool in Excel works well for generating the simple linear regression equation and other useful information, it does not provide predicted values for the dependent variable. However, both Minitab and the PHStat add-ins do provide predictions. We will illustrate the Minitab output, which is formatted somewhat differently from the Excel output but contains the same basic information.

The administrator is attempting to construct a simple linear regression model, with total charges as the dependent (y) variable and length of stay as the independent (x) variable. Figure 14.20 shows the Minitab regression output. The least squares regression equation is

$$\hat{y} = 528 + 1{,}353x$$

As shown in the figure, the regression slope coefficient is significantly different from 0 (t = 14.17; p-value = 0.000). The model explains 59.6% of the variation in the total charges (R-squared = 59.6%). Notice in Figure 14.20 that Minitab has rounded the regression coefficient. The more precise values are provided in the column headed "Coef" and are

$$\hat{y} = 527.6 + 1{,}352.80x$$

The administrator could use this equation to predict total charges by substituting the length of stay into the regression equation for x. For example, suppose a patient has a five-day stay. The predicted total charges are

$$\hat{y} = 527.6 + 1{,}352.80(5)$$
$$\hat{y} = \$7{,}291.60$$

FIGURE 14.20

Minitab Regression Output for Freedom Hospital

```
MINITAB - Patients.MPJ - [Session]
File Edit Data Calc Stat Graph Editor Tools Window Help

The regression equation is
Total Chges = 528 + 1353 Length of Stay

Predictor          Coef  SE Coef      T      P
Constant          527.6    483.8   1.09  0.277
Length of Stay  1352.80    95.44  14.17  0.000

S = 2894.78   R-Sq = 59.6%   R-Sq(adj) = 59.3%

Analysis of Variance

Source           DF          SS          MS       F      P
Regression        1  1683440143  1683440143  200.89  0.000
Residual Error  136  1139647631     8379762
Total           137  2823087774
```

Minitab Instructions:

1. Open file: Patients.MTW.
2. Choose **Stat > Regression > Regression**.
3. In **Response**, enter the y variable column.
4. In **Predictors**, enter the x variable column.
5. Click **OK**.

Note that this predicted value is a *point estimate* of the actual charges for this patient. The true charges will be either higher or lower than this amount. The administrator can develop a prediction interval, which is similar to the confidence interval estimates developed in Chapter 8.

Excel and Minitab Tutorial

Confidence Interval for the Average *y*, Given *x* The marketing manager might like a 95% confidence interval for *average* charges for all patients who stay in the hospital five days. The confidence interval for the expected value of a dependent variable, given a specific level of the independent variable, is determined by Equation 14.23. Observe that the specific value of x used to provide the prediction is denoted as x_p.

CHAPTER OUTCOME #6

Confidence Interval for $E(y)|x_p$

$$\hat{y} \pm ts_\varepsilon \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x - \bar{x})^2}} \qquad \textbf{(14.23)}$$

where:

$\hat{y}$ = Point estimate of the dependent variable
t = Critical value with $n - 2\,df$
n = Sample size
x_p = Specific value of the independent variable
$\bar{x}$ = Mean of the independent variable observations in the sample
s_ε = Estimate of the standard error of the estimate

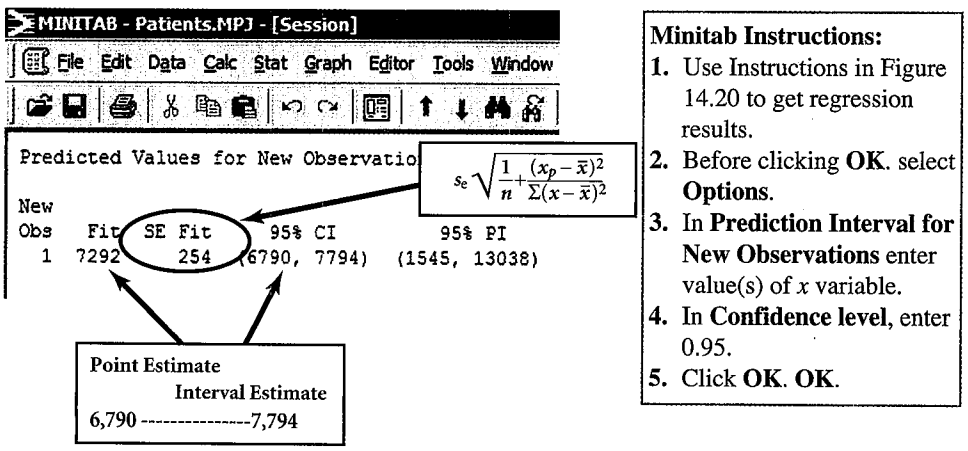
Although the confidence interval estimate can be manually computed using Equation 14.23, using your computer is much easier. Both PHStat and Minitab have built-in options to generate the confidence interval estimate for the dependent variable for a given value of the x variable. Figure 14.21 shows the Minitab results when length of stay, x, equals five days. Given this length of stay, the point estimate for the mean total charges is rounded by Minitab to \$7,292, and at the 95% confidence level, the administrators believe the mean total charges will be in the interval \$6,790 to \$7,794.

Excel and Minitab Tutorial

Prediction Interval for a Particular *y*, Given *x* The confidence interval shown in Figure 14.21 is for the average value of y given x_p. The administrator might also be interested in predicting the total charges for a *particular* patient with a five-day stay, rather than the average of the charges for all patients staying five days. Developing this 95% prediction interval requires only a slight modification to Equation 14.23. This prediction interval is given by Equation 14.24.

FIGURE 14.21

Minitab Output: Freedom Hospital Confidence Interval Estimate



Prediction Interval for $y|x_p$

$$\hat{y} \pm ts_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x - \bar{x})^2}} \qquad \textbf{(14.24)}$$

As was the case with the confidence interval application discussed previously, the manual computations required to use Equation 14.24 can be onerous. We recommend using your computer and software such as Minitab or PHStat to find the prediction interval. Figure 14.22 shows the PHStat results. Note that the same PHStat process generates both the prediction and confidence interval estimates.

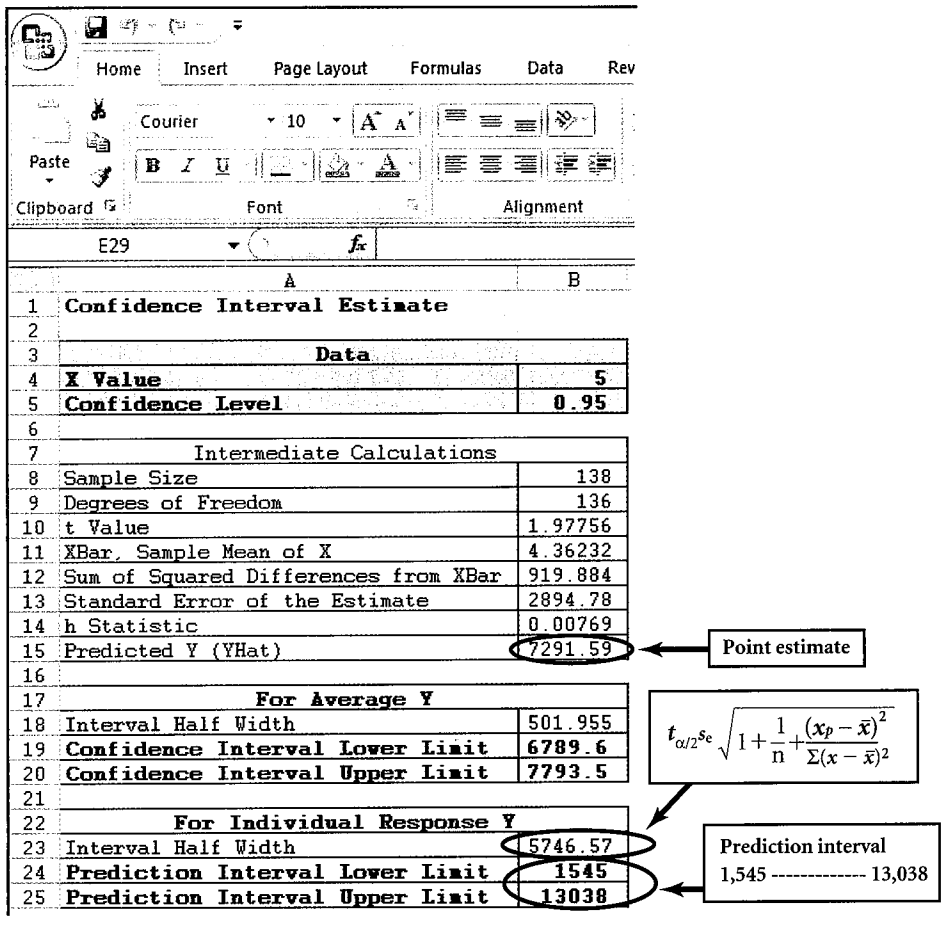
Based on this regression model, at the 95% confidence level, the hospital administrators can predict total charges for any patient with length of stay of five days to be between \$1,545 and \$13,038. As you can see, this prediction has extremely poor precision. We doubt any hospital administrator will use a prediction interval that is so wide. Although the regression model explains a significant proportion of variation in the dependent variable, it is relatively imprecise for predictive purposes. To improve the precision, we might decrease the confidence requirements or increase the sample size and redevelop the model.

The prediction interval for a specific value of the dependent variable is wider (less precise) than the confidence interval for predicting the average value of the dependent variable. This will always be the case, as seen in Equations 14.23 and 14.24. From an intuitive viewpoint, we should expect to come closer to predicting an average value than a single value.

Note, the term $(x_p - \bar{x})^2$ has a particular effect on the confidence interval determined by both Equations 14.23 and 14.24. The farther x_p (the value of the independent variable used to predict y), is from $\bar{x}$, the greater $(x_p - \bar{x})^2$ becomes. Figure 14.23 shows two regression lines developed from two samples with the same set of x-values. We have made both lines pass through the same $(\bar{x}, \bar{y})$ point; however, they have different slopes and intercepts.

FIGURE 14.22

Excel 2007 (PHStat) Prediction Interval for Freedom Hospital



	A	B
1	**Confidence Interval Estimate**	
2		
3	**Data**	
4	**X Value**	**5**
5	**Confidence Level**	**0.95**
6		
7	Intermediate Calculations	
8	Sample Size	138
9	Degrees of Freedom	136
10	t Value	1.97756
11	XBar, Sample Mean of X	4.36232
12	Sum of Squared Differences from XBar	919.884
13	Standard Error of the Estimate	2894.78
14	h Statistic	0.00769
15	Predicted Y (YHat)	7291.59
16		
17	**For Average Y**	
18	Interval Half Width	501.955
19	**Confidence Interval Lower Limit**	**6789.6**
20	**Confidence Interval Upper Limit**	**7793.5**
21		
22	**For Individual Response Y**	
23	Interval Half Width	5746.57
24	**Prediction Interval Lower Limit**	**1545**
25	**Prediction Interval Upper Limit**	**13038**

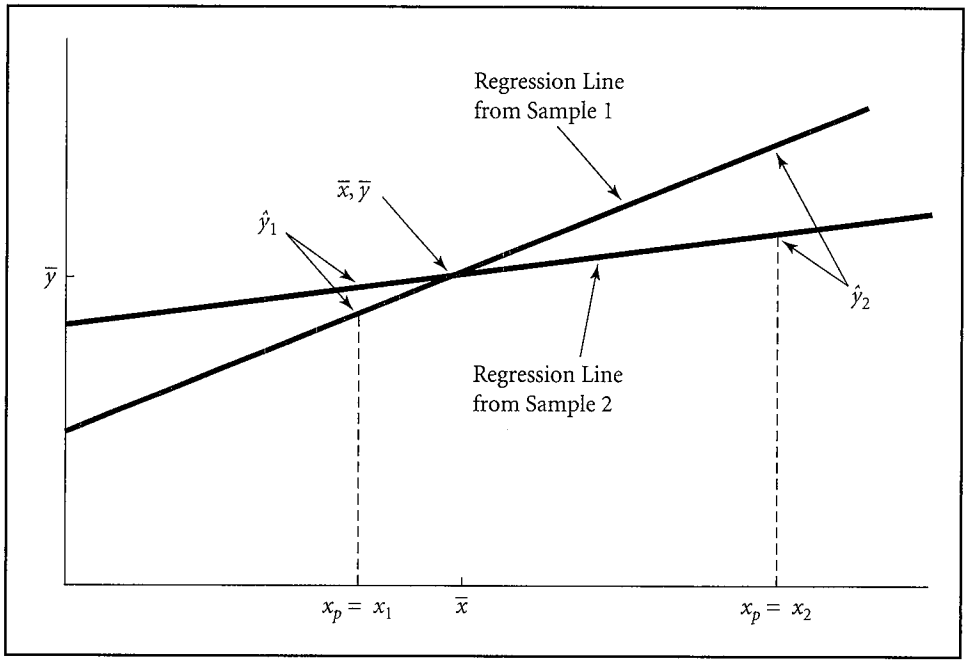
$$t_{\alpha/2} s_e \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x - \bar{x})^2}}$$

Excel 2007 (PHStat) Instructions:

1. Open file: Patients.xls.
2. Click on **Add-Ins > PHStat.**
3. Select **Regression > Simple Linear Regression.**
4. Define y variable range (Total Charges) and x variable range (Length of Stay).
5. Select **Confidence and Prediction Interval** – set $x = 5$ and 95% confidence.

FIGURE 14.23

Regression Lines Illustrating the Increase in Potential Variation in y as x_p Moves Farther from $\bar{x}$



At $x_p = x_1$, the two regression lines give predictions of y that are close to each other. However, for $x_p = x_2$, the predictions of y are quite different. Thus, when x_p is close to $\bar{x}$, the problems caused by variations in regression slopes are not as great as when x_p is far from $\bar{x}$. Figure 14.24 shows the prediction intervals over the range of possible x_p values. The band around the estimated regression line bends away from the regression line as x_p moves in either direction from $\bar{x}$.

Residual Analysis

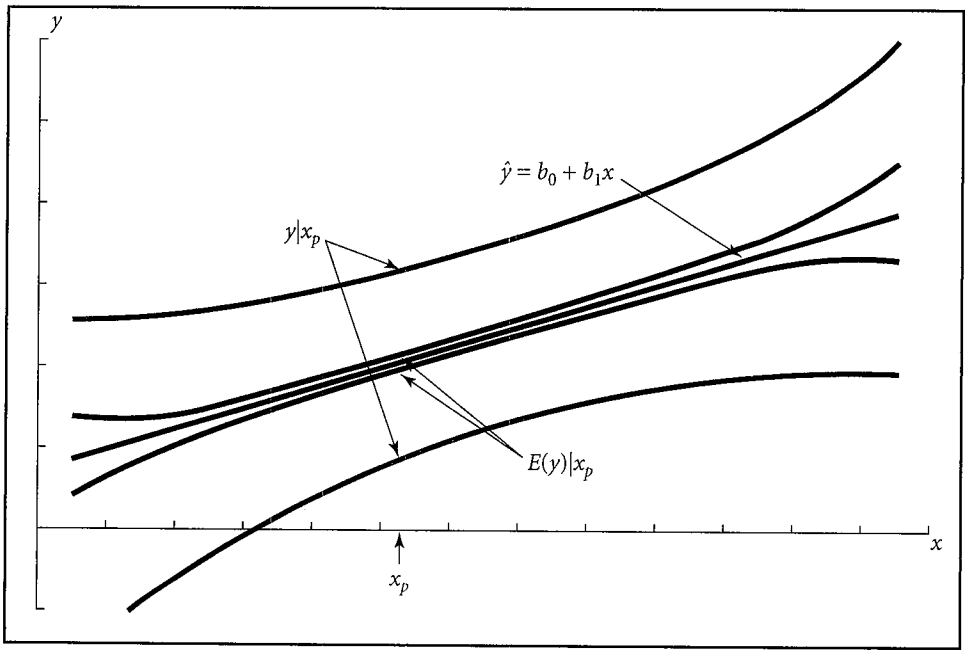
Recall two important assumptions associated with linear regression analysis.

Assumptions

1. The model errors are normally distributed.
2. The model errors have a constant variance at all levels of the independent variable.

FIGURE 14.24

Confidence Intervals for $y|x_p$ and $E(y)|x_p$



Before using a regression model for description or prediction, you should check to see if these assumptions are satisfied. One way to do this is by examining graphs called *residual plots*. Both Excel and Minitab can be used to generate residual plots.

Excel and Minitab Tutorial

Business Application

FREEDOM HOSPITAL (*CONTINUED*) Previously we showed the regression model constructed by the administrator at Freedom Hospital. He wanted to predict the total patient charges by knowing the patient length of stay. The resulting model was statistically significant. However, before the hospital actually uses this model, the administrator might develop two different residual plots. The first is a *residual frequency histogram*, which is shown in Figure 14.25. As you can see, the histogram closely resembles a normal distribution, which is one indication that the normality assumption is satisfied.

The second residual plot charts the residuals against the x variable, as shown in Figure 14.26. Chapter 15 discusses this type of plot more fully. For now, you will be looking for a result in which the residuals have approximately the same spread at all levels of x. In Figure 14.26, the plot illustrates that for short lengths of stay, the spread in the residuals is less than when stays are longer. This implies that the assumption of equal variances in the residuals is violated. We will discuss this in more detail in Chapter 15 and suggest possible steps for improving the regression model.

Common Problems Using Regression Analysis

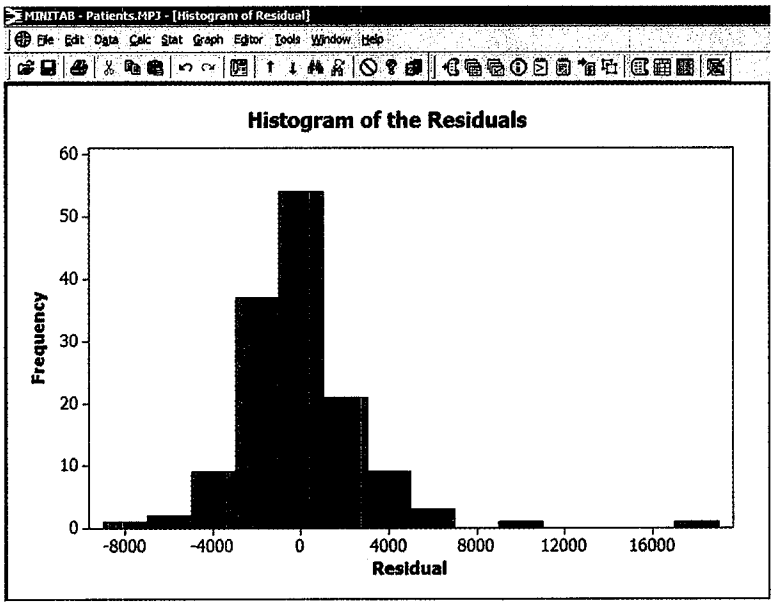
CHAPTER OUTCOME #7

Regression is perhaps the most widely used statistical tool other than descriptive statistical techniques. Because it is so widely used, you need to be aware of the common problems found when the technique is employed.

One potential problem occurs when decision makers apply regression analysis for predictive purposes. The conclusions and inferences made from a regression line are statistically valid only over the range of the data contained in the sample used to develop the regression line. For instance, in the Midwest Distribution example, we analyzed the

FIGURE 14.25

Minitab Residual Frequency Histogram for Freedom Hospital

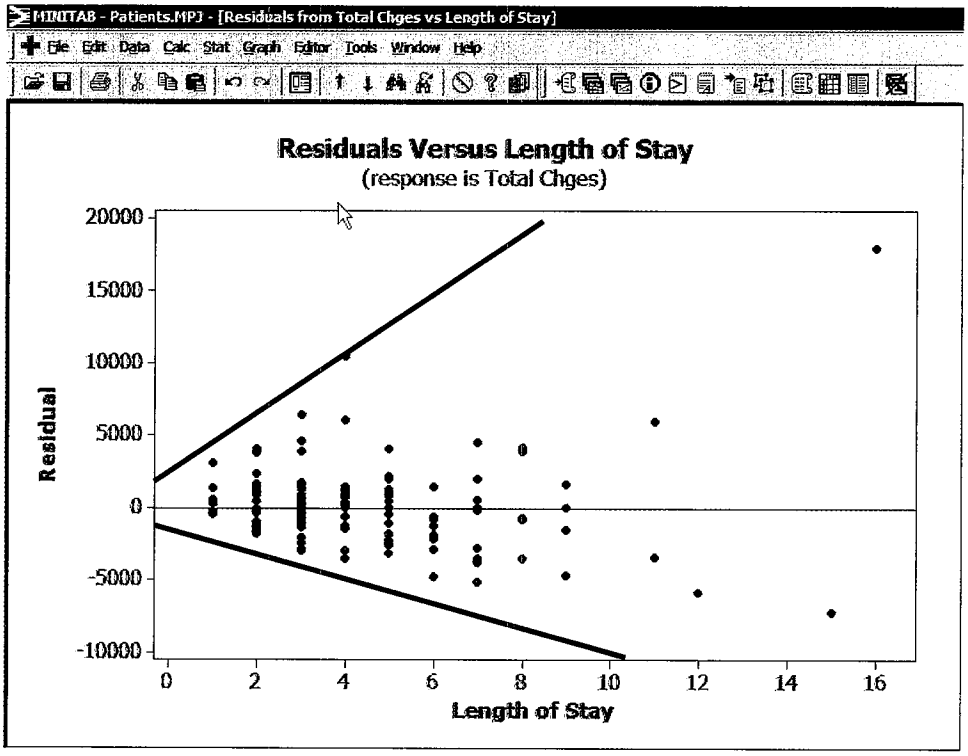


Minitab Instructions:

1. Open file: Patients.MTW.
2. Choose **Stat > Regression > Regression.**
3. In **Response**, enter the y variable column.
4. In **Predictors**, enter the x variable column.
5. Click **Storage**, under **Diagnostic Measures** select **Residuals.**
6. Click **OK.**
7. Choose **Graphs.**
8. Under **Residuals for Plots**, select **Regular.**
9. Select **Individual Plots.**
10. Choose **Histogram of Residuals.**
11. Click **OK. OK.**

FIGURE 14.26

Minitab Residual Plot for Freedom Hospital Example



Excel 2007 Instructions (for Similar results):

1. Open file: Patients.xls.
2. Choose **Data > Data Analysis > Regression**.
3. Select data range for *y* variable.
4. Select data range for *x* variable.
5. Check boxes for residuals and residual plots.

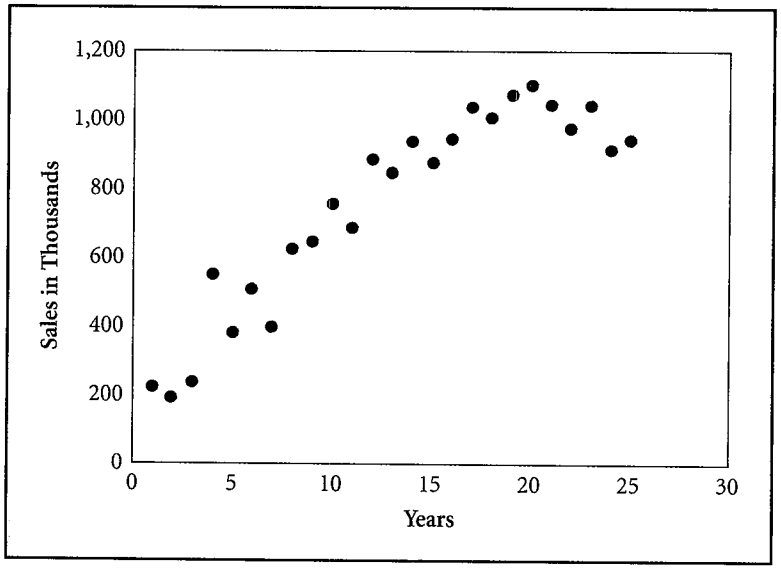
Minitab Instructions:

1. Open file: Patients.MTW.
2. Choose **Stat > Regression > Regression**.
3. In **Response** enter the *y* **variable** column.
4. **In Predictors,** enter the *x* variable **column**.
5. Select **Graphs**.
6. In **Residuals versus the variables,** enter the *x* variable column.
7. Click **OK. OK.**

performance of sales representatives with 1 to 9 years of experience. Therefore, predicting sales levels for employees with one to nine years of experience would be justified. However, if we were to try to predict the sales performance of someone with more than nine years of experience, the relationship between sales and experience might be different. Because no observations were taken for experience levels beyond the 1- to 9-year range, we have no information about what might happen outside that range. Figure 14.27 shows a

FIGURE 14.27

Graph for a Sales Peak at 20 Years



case in which the true relationship between sales and experience reaches a peak value at about 20 years and then starts to decline. If a linear regression equation were used to predict sales based on experience levels beyond the relevant range of data, large prediction errors could occur.

A second important consideration, one that was discussed earlier, involves correlation and causation. The fact that a significant linear relationship exists between two variables does not imply that one variable causes the other. Although there may be a cause-and-effect relationship, you should not infer that such a relationship is present based only on regression and/or correlation analysis. You should also recognize that a cause-and-effect relationship between two variables is not necessary for regression analysis to be an effective tool. What matters is that the regression model accurately reflects the relationship between the two variables and that the relationship remains stable.

Many users of regression analysis mistakenly believe that a high coefficient of determination (R^2) guarantees that the regression model will be a good predictor. You should remember that R^2 is a measure of the variation in the dependent variable explained by the independent variable. Although the least squares criterion assures us that R^2 will be maximized (because the sum of squares error is a minimum) for the given set of sample data, the value applies only to those data used to develop the model. Thus, R^2 measures the fit of the regression line to the sample data. There is no guarantee that there will be an equally good fit with new data. The only true test of a regression model's predictive ability is how well the model actually predicts.

Finally, we should mention that you might find a large R^2 with a large standard error. This can happen if total sum of squares is large in comparison to the *SSE*. Then, even though R^2 is relatively large, so too is the estimate of the model's standard error. Thus, confidence and prediction errors may be simply too wide for the model to be used in many situations. This is discussed more fully in Chapter 15.

14-3: Exercises

Skill Development

Problems 14-30 and 14-31 refer to the following output for a simple linear regression model:

Summary Output

Regression Statistics	
Multiple *R*	0.1027
R-Square	0.0105
Adjusted *R*-Square	−0.0030
Standard Error	9.8909
Observations	75

Anova

	df	*SS*	*MS*	*F*	Significance *F*
Regression	1	76.124	76.12	0.778	0.3806
Residual	73	7141.582	97.83		
Total	74	7217.706			

	Coefficents	Standard Error	*t*-Stat	*p*-value	Lower 95%	Upper 95%
Intercept	4.0133	3.878	1.035	0.3041	−3.715	11.742
x	0.0943	0.107	0.882	0.3806	−0.119	0.307

14-30. Referring to the displayed regression model, what percent of variation in the *y* variable is explained by the *x* variable in the model?

14-31. Construct and interpret a 90% confident interval estimate for the regression slope coefficient.

14-32. The following data have been collected by an accountant who is performing an audit of paper products at a large office supply company. The dependent variable, *y*, is the time taken (in minutes) by the accountant to count the units. The independent variable, *x*, is the number of units on the computer inventory record.

y	23.1	100.5	242.9	56.4	178.7	10.5	94.2	200.4	44.2	128.7	180.5
x	24	120	228	56	190	13	85	190	32	120	230

a. Develop a scatter plot for these data.

b. Determine the regression equation representing the data. Is the model significant? Test using a

significance level of 0.10 and the p-value approach.

c. Develop a 90% confidence interval estimate for the true regression slope and interpret this interval estimate. Based on this interval, could you conclude the accountant takes an additional minute to count each additional unit?

14-33. You are given the following sample data:

x	y
10	3
6	7
9	3
3	8
2	9
8	5
3	7

a. Develop a scatter plot for these data.

b. Determine the regression equation for the data.

c. Develop a 95% confidence interval estimate for the true regression slope and interpret this interval estimate.

d. Provide a 95% prediction interval estimate for a particular y, given $x_p = 7$.

14-34. You are given the following summary statistics from a regression analysis:

$$\hat{y} = 200 + 150x$$
$$SSE = 25.25$$
$$SSX = \text{Sum of squares } X = \Sigma(x-\bar{x})^2 = 99{,}645$$
$$n = 18$$
$$\bar{x} = 52.0$$

a. Determine the point estimate for y if $x_p = 48$ is used.

b. Provide a 95% confidence interval estimate for the average y, given $x_p = 48$. Interpret this interval.

c. Provide a 95% prediction interval estimate for a particular y, given $x_p = 48$. Interpret.

d. Discuss the difference between the estimates provided in parts b and c.

14-35. The sales manager at Sun City Real Estate Company in Tempe, Arizona, is interested in describing the relationship between condo sales prices and the number of weeks the condo is on the market before it sells. He has collected a random sample of 17 low-end condos that have sold within the past three months in the Tempe area. These data are shown as follows:

Weeks on the Market	Selling Price
23	$76,500
48	$102,000
9	$53,000
26	$84,200
20	$73,000
40	$125,000
51	$109,000
18	$60,000
25	$87,000
62	$94,000
33	$76,000
11	$90,000
15	$61,000
26	$86,000
27	$70,000
56	$133,000
12	$93,000

a. Develop a simple linear regression model to explain the variation in selling price based on the number of weeks the condo is on the market.

b. Test to determine whether the regression slope coefficient is significantly different from 0 using a significance level equal to 0.05.

c. Construct and interpret a 95% confidence interval estimate for the regression slope coefficient.

14-36. A sample of 10 yields the following data:

x	10	8	11	7	10	11	6	7	15	9
y	103	85	115	73	97	102	65	75	155	95

a. Provide a 95% confidence interval for the average y when $x_p = 9.4$.

b. Provide a 95% confidence interval for the average y when $x_p = 10$.

c. Obtain the margin of errors for both part a and part b. Explain why the margin of error obtained in part b is larger than that in part a.

14-37. A regression analysis from a sample of 15 produced the following:

$$\Sigma(x_i - \bar{x})(y_i - \bar{y}) = 156.4$$
$$\Sigma(x_i - \bar{x})^2 = 173.5$$
$$\Sigma(y_i - \bar{y})^2 = 181.6$$
$$\Sigma(y_i - \hat{y})^2 = 40.621$$
$$\bar{x} = 13.4 \text{ and } \bar{y} = 56.4$$

a. Produce the regression line.
b. Determine if there is a linear relationship between the dependent and independent variables. Use a significance level of 0.05 and a p-value approach.
c. Calculate a 90% confidence interval for the amount the dependent variable changes when the independent variable increases by 1 unit.

Business Applications

14-38. Gym Outfitters sells and services exercise equipment such as treadmills, ellipticals, and stair climbers to gymnasiums and recreational centers. The company's management would like to determine if there is a relationship between the number of minutes required to complete a routine service call and the number of machines serviced. A random sample of 15 records revealed the following information concerning the number of machines serviced and the time (in minutes) to complete the routine service call:

Number of Machines	Service Time (minutes)
11	115
8	60
9	80
10	90
7	55
6	65
8	70
4	33
10	95
5	50
5	40
12	110

a. Estimate the least squares regression equation.
b. If a gymnasium had 6 machines, how many minutes should Gym Outfitters expect a routine service call to require?
c. Provide a 90% confidence interval for the average amount of time required to complete a routine service call when the number of machines being serviced is 9.
d. Provide a 90% prediction interval for the time required to complete a particular routine service call for a gymnasium that has 7 machines.

14-39. The National Association of Realtors Existing-Home Sales Series provides a measurement of the residential real estate market. On or about the 25th of each month, NAR releases statistics on sales and prices of condos and co-ops, in addition to existing single-family homes, for the nation and the four regions. The data presented here indicate the number of (thousands of) existing-home sales as well as condo/co-op sales:

Year		Single-Family Sales	Condo/ Co-op Sales
2005	Apr	6,270	895
	May	6,230	912
	Jun	6,330	943
	Jul	6,220	914
	Aug	6,280	928
	Sept	6,290	908
	Oct	6,180	867
	Nov	6,150	876
	Dec	5,860	885
2006	Jan	5,790	781
	Feb	6,050	852
	Mar	6,040	862
	Apr	5,920	839

a. Construct the regression equation that would predict the number of condo/co-op sales using the number of single-family sales.
b. One might conjecture that these two markets (single-family sales and condo/co-op sales) would be competing for the same audience. Therefore, we would expect that as the number of single-family sales increases, the number of condo/co-op sales would decrease. Conduct a hypothesis test to determine this using a significance level of 0.05.
c. Provide a prediction interval for the number of condo/co-op sales when the number of single-family sales is 6,000 (thousands). Use a confidence level of 95%.

14-40. The following data have been collected by an accountant who is performing an audit of account balances for a major retail company. The population from which the data were collected represented those accounts for which the customer had indicated the balance was incorrect. The dependent variable, y, is the actual account balance as verified by the accountant. The independent variable, x, is the computer account balance.

y	233	10	24	56	78	102	90	200	344	120	18
x	245	12	22	56	90	103	85	190	320	120	23

a. Compute the least squares regression equation.
b. If the computer account balance were 100, what would you expect to be the actual account balance as verified by the accountant?

c. The computer balance for Timothy Jones is listed as 100 in the computer account record. Provide a 90% interval estimate for Mr. Jones's actual account balance.
d. Provide a 90% interval estimate for the average of all customers' actual account balances in which a computer account balance is the same as that of Mr. Jones. (Part c.) Interpret.

14-41. J.D. Power and Associates conducts an initial quality study (IQS) each year to determine the quality of newly manufactured automobiles. IQS measures 135 attributes across nine categories, including ride/handling/braking, engine and transmission, and a broad range of quality problem symptoms reported by vehicle owners. The 2005 Initial Quality Study was based on responses from more than 62,000 purchasers and lessees of new 2005 model-year cars and trucks, who were surveyed after 90 days of ownership. The data given here portray the initial quality industry average of the number of reported problems per 100 vehicles for 1998–2005.

Year	1998	1999	2000	2001	2002	2003	2004	2005
Problems	176	167	154	147	133	133	119	118

a. Construct a scatter plot of the number of reported problems per 100 vehicles as a function of the year.
b. Determine if the average number of reported problems per 100 vehicles declines from year to year. Use a significance level of 0.01 and a *p*-value approach.
c. Assume the relationship between the number of reported problems per 100 vehicles and the year continues into the future. Provide a 95% prediction interval for the initial quality industry average of the number of reported problems per 100 vehicles for 2010.

Computer Database Exercises

14-42. A manufacturer produces a wash-down motor for the food service industry. The company manufactures the motors to order by modifying a base model to meet the specifications requested by the customer. The motors are produced in a batch environment with the batch size equal to the number ordered. The manufacturer has recently sampled 27 customer orders. The motor manufacturer would like to determine if there is a relationship between the cost of producing the order and the order size so that it could estimate the cost of producing a particular size order. The sampled data are contained in the file **Washdown Motors**.

a. Use the sample data to estimate the least squares regression model.
b. Provide an interpretation of the regression coefficients.
c. Test the significance of the overall regression model using a significance level of 0.01.
d. The company has just received an order for 30 motors. Use the regression model developed in part a to estimate the cost of producing this particular order.
e. Referring to part d, what is the 90% confidence interval for an average cost of an order of 30 motors?

14-43. Each month, the Bureau of Labor Statistics (BLS) of the U.S. Department of Labor announces the total number of employed and unemployed persons in the United States for the previous month. At the same time, it also publishes the inflation rate, which is the rate of change in the price of goods and services from one month to the next. It seems quite plausible that there should be some relationship between these two indicators. The file entitled **CPI** provides the monthly unemployment and inflation rates for the period 2000–2005.

a. Construct a scatter plot of the unemployment rate versus inflation rate for the period 2000–2005. Describe the relationship that appears to exist between these two variables.
b. Produce a 95% prediction interval for the unemployment rate for the maximum inflation rate in the period 2000–2005. Interpret the interval.
c. Produce a 95% prediction interval for the unemployment rate when the inflation rate is 0.00.
d. Which of the prediction intervals in parts b and c has the larger margin of error? Explain why this is the case.

14-44. The National Highway Transportation Safety Administration's National Center for Statistics and Analysis released its Vehicle Survivability Travel Mileage Schedules in January 2006. One item investigated was the relationship between the annual vehicle miles traveled (VMT) as a function of vehicle age for passenger cars up to 25 years old. The VMT data were collected by asking consumers to estimate the number of miles driven in a given year. The data were collected over a 14-month period, starting in March 2001 and ending in May 2002. The file entitled **Miles** contains this data.

a. Produce a regression equation modeling the relationship between VMT and the age of the vehicle. Estimate how many more annual vehicle miles would be traveled for a

vehicle that is 10 years older than another vehicle.

b. Provide a 90% interval estimate for the average annual vehicle miles traveled when the age of the vehicle is 15 years.

c. Determine if it is plausible for a vehicle that is 10 years old to travel 12,000 miles in a year. Support your answer with statistical reasoning.

Summary and Conclusions

Correlation and regression analysis are two of the most frequently used statistical techniques by business decision makers. This chapter has introduced the basics of these two topics. The discussion of regression analysis has been limited to situations in which you have one dependent variable and one independent variable. In these cases, the technique for modeling the linear relationship between the two variables is referred to as simple linear regression analysis.

If two variables are significantly correlated, then they are said to be linearly related. When that's the case, the resulting simple linear regression model will be statistically significant, which means that the fraction of variation in the dependent variable that is explained by the independent variable (R-squared) is significant and the predictions for the y variable based on values of x will be superior to using the mean of y as the predictor.

This chapter introduced the methods used to test whether a correlation is 0 and whether a regression slope coefficient is 0. We also introduced you to the uses of regression for descriptive and predictive purposes and showed how to construct confidence interval estimates for the true regression slope coefficient and prediction intervals.

Chapter 15 will extend the discussion of regression analysis by showing how two or more independent variables are included in the analysis. The focus of that chapter will be on building a model for explaining the variation in the dependent variable. However, the basic concepts presented in this chapter will be carried forward.

Equations

Sample Correlation Coefficient

$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{[\Sigma(x-\bar{x})^2][\Sigma(y-\bar{y})^2]}} \quad (14.1)$$

or the algebraic equivalent:

$$r = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{[n(\Sigma x^2)-(\Sigma x)^2][n(\Sigma y^2)-(\Sigma y)^2]}} \quad (14.2)$$

Test Statistic for Correlation

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \qquad df = n-2 \quad (14.3)$$

Simple Linear Regression Model (Population Model)

$$y = \beta_0 + \beta_1 x + \varepsilon \quad (14.4)$$

Estimated Regression Model (Sample Model)

$$\hat{y} = b_0 + b_1 x \quad (14.5)$$

Least Squares Equations (Sample Values)

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

or the algebraic equivalent:

$$b_1 = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{\Sigma x^2 - \frac{(\Sigma x)^2}{n}} \quad (14.7)$$

and

$$b_0 = \bar{y} - b_1 \bar{x} \quad (14.8)$$

Sum of Squared Errors

$$SSE = \Sigma y^2 - b_0 \Sigma y - b_1 \Sigma xy \quad (14.9)$$

Sum of Residuals

$$\sum_{i=1}^{n}(y_i - \hat{y}_i) = 0 \quad (14.10)$$

Sum of Squared Residuals

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (14.11)$$

Total Sum of Squares

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 \quad (14.12)$$

Sum of Squares Error

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (14.13)$$

Sum of Squares Regression

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \quad \textbf{(14.14)}$$

Coefficient of Determination, R^2

$$R^2 = \frac{SSR}{SST} \quad \textbf{(14.15)}$$

Coefficient of Determination, Single Independent Variable Case

$$R^2 = r^2 \quad \textbf{(14.16)}$$

Test Statistic for Significance of the Coefficient of Determination

$$F = \frac{SSR/1}{SSE/(n-2)} \qquad d.f. = D_1 = 1,\ D_2 = n-2 \quad \textbf{(14.17)}$$

Simple Regression Standard Deviation of the Slope Coefficient (Population)

$$\sigma_{b_1} = \frac{\sigma_\varepsilon}{\sqrt{\Sigma(x-\bar{x})^2}} \quad \textbf{(14.18)}$$

Simple Regression Estimator for the Standard Error of the Estimate

$$s_\varepsilon = \sqrt{\frac{SSE}{n-2}} \quad \textbf{(14.19)}$$

Simple Regression Estimator for the Standard Deviation of the Slope

$$s_{b_1} = \frac{s_\varepsilon}{\sqrt{\Sigma(x-\bar{x})^2}} \quad \textbf{(14.20)}$$

Simple Linear Regression Test Statistic for Test of the Significance of the Slope

$$t = \frac{b_1 - \beta_1}{s_{b_1}} \qquad df = n-2 \quad \textbf{(14.21)}$$

Confidence Interval Estimate for the Regression Slope, Simple Linear Regression

$$b_1 \pm ts_{b_1} \quad \textbf{(14.22)}$$

or equivalently,

$$b_1 \pm t\frac{s_\varepsilon}{\sqrt{\Sigma(x-\bar{x})^2}} \qquad \text{with } df = n-2$$

Confidence Interval for $E(y)|x_p$

$$\hat{y} \pm ts_\varepsilon\sqrt{\frac{1}{n} + \frac{(x_p-\bar{x})^2}{\Sigma(x-\bar{x})^2}} \quad \textbf{(14.23)}$$

Prediction Interval for $y|x_p$

$$\hat{y} \pm ts_\varepsilon\sqrt{1 + \frac{1}{n} + \frac{(x_p-\bar{x})^2}{\Sigma(x-\bar{x})^2}} \quad \textbf{(14.24)}$$

Key Terms

Coefficient of determination 646
Correlation coefficient 623
Dependent variable 622
Independent variable 622
Least squares criterion 635
Regression coefficients 634
Regression slope coefficient 634
Residual 635
Scatter plot 622
Simple linear regression 633
Spurious correlation 630
Standard error of the estimate 648

Chapter Exercises

Conceptual Questions

14-45. If we select a random sample of data for two variables and, after computing the correlation coefficient, conclude that the two variables may have zero correlation, can we say that there is no relationship between the two variables? Discuss.

14-46. Discuss why prediction intervals that attempt to predict a particular y-value are less precise than confidence intervals for predicting an average y.

14-47. A statistics student was recently working on a class project that required him to compute a correlation coefficient for two variables. After careful work he arrived at a correlation coefficient of 0.45. Interpret this correlation coefficient for the student who did the calculations.

14-48. Referring to the previous problem, another student in the same class computed a regression equation relating the same two variables. The slope of the equation was found to be −0.735. After trying

several times and always coming up with the same result, she felt that she must have been doing something wrong since the value was negative and she knew that this could not be right. Comment on this student's conclusion.

14-49. Consider the two following scenarios:

a. The number of new workers hired per week in your county has a high positive correlation with the average weekly temperature. Can you conclude that an increase in temperature causes an increase in the number of new hires? Discuss.

b. Suppose the stock price and the common dividends declared for a certain company have a high positive correlation. Are you safe in concluding on the basis of the correlation coefficient that an increase in the common dividends declared causes an increase in the stock price? Present other reasons than the correlation coefficient that might lead you to conclude that an increase in common dividends declared causes an increase in the stock price.

14-50. Consider the following set of data:

x	48	27	34	24	49	29	39	38	46	32
y	47	23	31	20	50	48	47	47	42	47

a. Calculate the correlation coefficient of these two variables.

b. Multiply each value of the variable x by 5 and add 10 to the resulting products. Now multiply each value of the variable y by 3 and subtract 7 from the resulting products. Finally, calculate the correlation coefficient of the new x and y variables.

c. Describe the principle that the example developed in parts a and b demonstrates.

14-51. Go to the library and locate an article in a journal related to your major (*Journal of Marketing*, *Journal of Finance*, etc.) that uses linear regression. Discuss:

a. How the author chose the dependent and independent variables.

b. How the data were gathered.

c. What statistical tests the author used.

d What conclusions the analysis allowed the author to draw.

Business Applications

14-52. The Farmington City Council recently commissioned a study of park users in their community. Data were collected on the age of the person surveyed and the amount of hours he or she has spent in the park in the past month. The data collected were as follows:

Time in Park	Age
7.2	16
3.5	15
6.6	28
5.4	16
1.5	29
2.3	38
4.4	48
8.8	18
4.9	24
5.1	33
1.0	56

a. Draw a scatter plot for these data and discuss what, if any, relationship appears to be present between the two variables.

b. Compute the correlation coefficient between age and the amount of time spent in the park. Provide an explanation to the Farmington City Council of what the correlation measures.

c. Test to determine whether the amount of time spent in the park increases with the age of the park user. Use a significance level of 0.10. Use a p-value approach to conduct this hypothesis test.

14-53. At State University, a study was done to establish whether a relationship existed between student graduating GPA and the SAT verbal score when the student originally entered the university. The sample data are reported as follows:

GPA	2.5	3.2	3.5	2.8	3.0	2.4	3.4	2.9	2.7	3.8
SAT	640	700	550	540	620	490	710	600	505	710

a. Develop a scatter plot for these data and describe what, if any, relationship exists between the two variables, GPA and SAT score.

b. (1) Compute the correlation coefficient.
(2) Does it appear that the success of students at State University is related to the SAT verbal scores of those students? Conduct a statistical procedure to answer this question. Use a significance level of 0.01.

c. (1) Compute the regression equation based on these sample data if you wish to predict the university GPA using the student SAT score.
(2) Interpret the regression coefficients.

14-54. The Smithfield Organic Milk Company recently studied a random sample of 30 of its distributors

and found the correlation between sales and advertising dollars to be 0.67.

a. Is there a significant linear relationship between sales and advertising? If so, is it fair to conclude that advertising causes sales to increase?
b. If a regression model were developed using sales as the dependent variable and advertising as the independent variable, determine the proportion of the variation in sales that would be explained by its relationship to advertising. Discuss what this says about the usefulness of using advertising to predict sales.

14-55. A previous exercise citing an article in *USA Today* (Andriene Lewis, "Tuition outstrips inflation," January 5, 2004) discussed the relationship between the average college tuition (in 2003 dollars) for private and public colleges. The data indicated in the article follows:

Period	1983–84	1988–89	1993–94	1998–99	2003–04
Private	9,202	12,146	13,844	16,454	19,710
Public	2,074	2,395	3,188	3,632	4,694

a. Construct the regression equation that would predict the average college tuition for private colleges using that of the public colleges.
b. Determine if there is a linear tendency for the average college tuition for private colleges to increase when the average college tuition for public colleges increases. Use a significance level of 0.05 and a p-value approach.
c. Provide a 95% confidence interval for the average college tuition for private colleges when the average college tuition for public colleges reaches $5,000.
d. Is it plausible that the average college tuition for private colleges would be larger than $25,000 when the average college tuition for public colleges reaches $5,000? Support you assertion with statistical reasoning.

14-56. An American airline company recently performed a customer survey in which it asked a random sample of 100 passengers to indicate their income and the total cost of the airfares they have purchased for pleasure trips during the past year. A regression model was developed for the purposes of determining whether income could be used as a variable to explain the variation in number of times individuals fly on airlines in a year. The following regression results were obtained:

$$\hat{y} = 0.25 + 0.0150x$$
$$s_\varepsilon = 721.44$$
$$R^2 = 0.65$$
$$s_{b_1} = 0.0000122$$

a. Produce an estimate of the maximum and minimum amount of difference in the amounts allocated to purchase airline tickets by two families that have a difference of $20,000 in family income. Assume that you wish to use a 90% confidence level.
b. Can the intercept of the regression equation be interpreted in this case, assuming that no one who was surveyed had an income of 0 dollars? Explain.
c. Use the information provided to perform an F-test for the significance of the regression model. Discuss your results, assuming the test is performed at the significance level of 0.05.

14-57. One of the advances that have helped to diminish carpal tunnel syndrome is ergonomic keyboards. The ergonomic keyboards may also increase typing speed. Ten administrative assistants were chosen to type on both standard and ergonomic keyboards. The resulting typing speeds follow:

Ergonomic:	69	80	60	71	73	64	63	70	63	74
Standard:	70	68	54	56	58	64	62	51	64	53

a. Produce a scatter plot of the typing speed of administrative assistants using ergonomic and standard keyboards. Does there appear to be a linear relationship between these two variables? Explain your response.
b. Calculate the correlation coefficient of the typing speed of administrative assistants using ergonomic and standard keyboards.
c. Conduct a hypothesis test to determine if a positive correlation exists between administrative assistants using ergonomic and standard keyboards. Use a significance level of 0.05.

14-58. A company is considering recruiting new employees from a particular college and plans to place a great deal of emphasis on the student's college grade point average (GPA). However, the company is aware that not all schools have the same grading standards, so it is possible that a student at this school might have a lower (or higher) GPA than a student from another school, yet really be on par with the other student. To make this comparison between schools, the company has devised a test that it has administered utilizing a sample size of 400 students. With the results of the test, it has developed a regression model that it uses to predict student GPA. The following equation represents the model:

$$\hat{y} = 1.0 + 0.028x$$

The R^2 for this model is 0.88 and the standard error of the estimate is 0.20, based on the sample data used to develop the model. Note that the dependent

variable is the GPA and the independent variable is test score, where this score can range from 0 to 100. For the sample data used to develop the model, the following values are known:

$$\bar{y} = 2.76$$
$$\bar{x} = 68$$
$$\Sigma(x - \bar{x})^2 = 148{,}885.73$$

a. Based on the information contained in this problem, can you conclude that as the test score increases, the GPA will also increase, using a significance level of 0.05?
b. Suppose a student interviews with this company, takes the company test, and scores 80 correct. What is the 90% prediction interval estimate for this student's GPA? Interpret the interval.
c. Suppose the student in part b actually has a 2.90 GPA at this school. Based on this evidence, what might be concluded about this person's actual GPA compared with other students with the same GPA at other schools? Discuss the limitations you might place on this conclusion.
d. Suppose a second student with a 2.45 GPA took the test and scored 65 correct. What is the 90% prediction interval for this student's "real" GPA? Interpret.

Computer Database Exercises

14-59. The Briggs Bank and Trust recently performed a study of its checking account customers. One objective of the study was to determine whether it is possible to explain the variation in average checking account balance by knowing the number of checks written per month. The sample data selected are contained in the data file named **Briggs**.
a. Draw a scatter plot for these data.
b. Develop the least squares regression equation for these data.
c. Develop the 90% confidence interval estimate for the change in the average checking account balance when a person who formerly wrote 25 checks a month doubled the number of checks used.
d. Test to determine if an increase in the number of checks written by an individual can be used to predict the checking account balance of that individual. Use $\alpha = 0.05$. Comment on this result and the result of part c.

14-60. An economist for the state government of Mississippi recently collected the data contained in the file called **Mississippi** on the percentage of people unemployed in the state at randomly selected points in time over the past 25 years and the interest rate of Treasury bills offered by the federal government at that point in time.
a. (1) Develop a plot showing the relationship between the two variables. (2) Describe the relationship as being either linear or curvilinear.
b. (1) Develop a simple linear regression model with unemployment rate as the dependent variable. (2) Write a short report describing the model and indicating the important measures.

14-61. The Cooley Service Center polishes and cleans automobiles. It has major accounts, such as the Bayview Taxi Service and Bayview Police Department, and also does work for the general public by appointment. Recently, the manager decided to survey customers to determine how satisfied they were with the work performed by the Cooley Service Center. He devised a rating scale between 0 and 100, with 0 being poor and 100 being excellent service. He selected a random sample of 14 customers and asked the customers when they picked up their cars to rate the service. He also recorded the amount of time spent on each customer's car. These data are in the data file named **Cooley**.
a. (1) Draw a scatter plot showing these two variables, with the y variable on the vertical axis and the x variable on the horizontal axis. (2) Describe the relationship between these two variables.
b. (1) Develop a linear regression model to explain the variation in the service rating. (2) Write a short report describing the model and showing the results of pertinent hypothesis tests, using a significance level of 0.10.

14-62. In a press release entitled "College Board Offers Glimpse of New SAT with Writing for Upcoming Class of '06," August 30, 2005, the College Board announced SAT scores for students in the class of 2005, the last to take the former version of the SAT featuring math and verbal sections. The file entitled **MathSAT** contains the math SAT scores for the interval 1967 to 2005. One point of interest concerning the data is the relationship between the average scores of male and female students.
a. Produce a scatter plot depicting the relationship between the average math SAT score of males (the dependent variable) and females (independent variable) over the period 1967 to 2005. Describe the relationship between these two variables.
b. Is there a linear relationship between the average score for males and females over the period 1967 to 2005? Use a significance level of 0.05 and the p-value approach to determine this.

14-63. The National Association of Home Builders published an article ("Housing Facts, Figures, and

Trends," August 2005) summarizing characteristics of the housing market in the United States. Data are listed for the average and median selling prices for houses during the period 1995 to 2004 and also January to July of 2005. The file entitled **House** contains the data. Assume the data can be viewed as samples of the relevant populations.

a. Determine the linear relationship that could be used to predict the average selling prices for January to July of 2005 using the median selling prices for that period.
b. Conduct a hypothesis test to determine if the median selling prices in the period January to July of 2005 could be used to determine the average selling prices in that period. Use a significance level of 0.05 and the *p*-value approach to conduct the test.
c. Provide an interval estimate of the average selling price of homes in the period 1995 to 2004 for a year in which the median selling price was $195,000. Use a 90% confidence interval.

14-64. The Grinfield Service Company marketing director is interested in analyzing the relationship between her company's sales and the advertising dollars spent. In the course of her analysis, she selected a random sample of 20 weeks and recorded the sales for each week and the amount spent on advertising. These data are contained in the data file called **Grinfield**:

a. Identify the independent and dependent variables.
b. Draw a scatter plot with the dependent variable on the vertical axis and the independent variable on the horizontal axis.
c. The marketing director wishes to know if increasing the amount spent on advertising increases sales. As a first attempt, use a statistical test that will provide the required information. Use a significance level of 0.025. Upon careful consideration, the marketing manager realizes that it takes a certain amount of time for the effect of advertising to register in terms of increased sales. She, therefore, asks you to calculate a correlation coefficient for sales of the current week against amount of advertising spent in the previous week and conduct a hypothesis test to determine if, under this model, increasing the amount spent on advertising increases sales. Again, use a significance level of 0.025.

14-65. Refer to the Grinfield Service Company discussed in Problem 14-64.

a. Develop the least squares regression equation for these variables. Plot the regression line on the scatter plot.
b. Develop a 95% confidence interval estimate for the increase in sales resulting from increasing the advertising budget by $50. Interpret the interval.
c. Discuss whether it is appropriate to interpret the intercept value in this model. Under what conditions is it appropriate? Discuss.
d. Develop a 90% confidence interval for the mean sales amount achieved during all weeks in which advertising is $200 for the week.
e. Suppose you are asked to use this regression model to predict the weekly sales when advertising is to be set at $100. What would you reply to the request? Discuss.

CASE 14.1

A & A Industrial Products

Alex Court, the cost accountant for A & A Industrial Products, was puzzled by the repair cost analysis report he had just reviewed. This was the third consecutive report where unscheduled plant repair costs were out of line with the repair cost budget allocated to each plant. A & A budgets for both scheduled maintenance and unscheduled repair costs for its plants' equipment, mostly large industrial machines. Budgets for scheduled maintenance activities are easy to estimate and are based on the equipment manufacturer's recommendations. The unscheduled repair costs, however, are harder to determine. Historically, A & A Industrial Products has estimated unscheduled maintenance using a formula based on the average number of hours of operation between major equipment failures at a plant. Specifically, plants were given a budget of $65.00 per hour of operation between major failures. Alex had arrived at this amount by dividing aggregate historical repair costs by the total number of hours between failures. Then plant averages would be used to estimate unscheduled repair cost. For example, if a plant averaged 450 hours of run time before a major repair occurred, the plant would be allocated a repair budget of 450 × $65 = $29,250 per repair. If the plant was expected to be in operation 3,150 hours per year, the company would anticipate 7 unscheduled repairs (3,150/450) annually and budget $204,750 for annual unscheduled repair costs.

Alex was becoming more and more convinced that this approach was not working. Not only was upper management upset about the variance between predicted and actual costs of repair, plant managers believed that the model did

not account for potential differences among the company's three plants when allocating dollars for unscheduled repairs. At the weekly management meeting, Alex was informed that he needed to analyze his cost projections further and produce a report that provided a more reliable method for predicting repair costs. Upon leaving the meeting, Alex had his assistant randomly pull 64 unscheduled repair reports. The data are in the file **A & A Costs**. The management team is anxiously waiting for Alex's analysis.

CASE 14.2

Sapphire Coffee—Part 1

Jennie Garcia could not believe that her career had moved so far so fast. When she left graduate school with a master's degree in anthropology, she intended to work at a local coffee shop until something else came along that was more related to her academic background. But after a few months she came to enjoy the business, and in a little over a year she was promoted to store manager. When the company for whom she worked continued to grow, Jennie was given oversight of a few stores.

Now, eight years after she started as a barista, Jennie was in charge of operations and planning for the company's southern region. As a part of her responsibilities, Jennie tracks store revenues and forecasts coffee demand. Historically, Sapphire Coffee would base its demand forecast on the number of stores, believing that each store sold approximately the same amount of coffee. This approach seemed to work well when the company had shops of similar size and layout, but as the company grew, stores became more varied. Now, some stores had drive-thru windows, a feature that top management added to some stores believing that it would increase coffee sales for customers who wanted a cup of coffee on their way to work but who were too rushed to park and enter the store to place an order.

Jennie noticed that weekly sales seemed to be more variable across stores in her region and was wondering what, if anything, might explain the differences. The company's financial vice president had also noticed the increased differences in sales across stores and was wondering what might be happening. In an e-mail to Jennie he stated that weekly store sales are expected to average $5.00 per square foot. Thus, a 1,000-square-foot store would have average weekly sales of $5,000. He asked that Jennie analyze the stores in her region to see if this rule of thumb was a reliable measure of a store's performance.

The VP of finance was expecting the analysis to be completed by the weekend. Jennie decided to randomly select weekly sales records for 53 stores. The data are in the file **Sapphire Coffee-1**. A full analysis would need to be sent to the corporate office by Friday.

CASE 14.3

Alamar Industries

While driving home in northern Kentucky at 8:00 P.M., Juan Alamar wondered whether his father had done him any favor by retiring early and letting him take control of the family machine-tool-restoration business. When his father started the business of overhauling machine tools (both for resale and on a contract basis), American companies dominated the tool manufacturing market. During the past 30 years, however, the original equipment industry had been devastated, first by competition from Germany and then from Japan. Although foreign competition had not yet invaded the overhaul segment of the business, Juan had heard about foreign companies establishing operations on the West Coast.

The foreign competitors were apparently stressing the high-quality service and operations that had been responsible for their great inroads into the original equipment market. Last week Juan had attended a daylong conference on total quality management that had discussed the advantages of competing for the Baldrige Award, the national quality award established in 1987. Presenters from past Baldrige winners, including Xerox, Federal Express, Cadillac, and Motorola, stressed the positive effects on their companies of winning and said similar effects would be possible for any company. This assertion of only positive effects was what Juan questioned. He was certain that the effect on his remaining free time would not be positive.

The Baldrige Award considers seven corporate dimensions of quality. Although the award is not based on a

numerical score, an overall score is calculated. The maximum score is 1,000, with most recent winners scoring about 800. Juan did not doubt the award was good for the winners, but he wondered about the nonwinners. In particular, he wondered about any relationship between attempting to improve quality according to the Baldrige dimensions and company profitability. Individual company scores are not released, but Juan was able to talk to one of the conference presenters, who shared some anonymous data, such as companies' scores in the year they applied, their returns on investment in the year applied, and returns on investment in the year after application. Juan decided to commit the company to a total quality management process if the data provided evidence that the process would lead to increased profitability.

Baldrige Score	ROI Application Year	ROI Next Year
470	11%	13%
520	10	11
660	14	15
540	12	12
600	15	16
710	16	16
580	11	12
600	12	13
740	16	16
610	11	14
570	12	13
660	17	19

CASE 14.4

Continental Trucking

Norm Painter is the newly hired cost analyst for Continental Trucking. Continental is a nationwide trucking firm, and, until recently, most of its routes were driven under regulated rates. These rates were set to allow small trucking firms to earn an adequate profit, leaving little incentive to work to reduce costs by efficient management techniques. In fact, the greatest effort was made to try to influence regulatory agencies to grant rate increases.

A recent rash of deregulation moves has made the long-distance trucking industry more competitive. Norm has been hired to analyze Continental's whole expense structure. As part of this study, Norm is looking at truck repair costs. Because the trucks are involved in long hauls, they inevitably break down. In the past, little preventive maintenance was done, and if a truck broke down in the middle of a haul, either a replacement tractor was sent or an independent contractor finished the haul. The truck was then repaired at the nearest local shop. Norm is sure this procedure has led to more expense than if major repairs had been made before the trucks failed.

Norm thinks that some method should be found for determining when preventive maintenance is needed. He believes that fuel consumption is a good indicator of possible breakdowns, as trucks begin to run badly, they will consume more fuel. Unfortunately, the major determinants of fuel consumption are the weight of a truck and headwinds. Norm picks a sample of a single truck model and gathers data relating fuel consumption to truck weight. All trucks in the sample are in good condition. He separates the data by direction of the haul, realizing that winds tend to blow predominantly out of the west.

East-West Haul		West-East Haul	
Miles/Gallon	Haul Weight	Miles/Gallon	Haul Weight
4.1	41,000 lb	4.3	40,000 lb
4.7	36,000	4.5	37,000
3.9	37,000	4.8	36,000
4.3	38,000	5.2	38,000
4.8	32,000	5.0	35,000
5.1	37,000	4.7	42,000
4.3	46,000	4.9	37,000
4.6	35,000	4.5	36,000
5.0	37,000	5.2	42,000
		4.8	41,000

Although he can rapidly gather future data on fuel consumption and haul weight, now that Norm has these data, he is not quite sure what to do with them.

References

Berenson, Mark L., and David M. Levine, *Basic Business Statistics: Concepts and Applications*, 10th ed. (Upper Saddle River, NJ: Prentice Hall, 2006).

Cryer, Jonathan D., and Robert B. Miller, *Statistics for Business: Data Analysis and Modeling*, 2nd ed. (Belmont, CA: Duxbury Press, 1994).

Dielman, Terry E., *Applied Regression Analysis—A Second Course in Business and Economic Statistics*, 4th ed. (Belmont, CA: Duxbury Press, 2005).

Draper, Norman R., and Harry Smith, *Applied Regression Analysis*, 3rd ed. (New York: John Wiley and Sons, 1998).

Frees, Edward W., *Data Analysis Using Regression Models: The Business Perspective* (Upper Saddle River, NJ: Prentice Hall, 1996).

Kleinbaum, David G., Lawrence L. Kupper, Keith E. Muller, and Azhar Nizam, *Applied Regression Analysis and Other Multivariable Methods*, 3rd ed. (Belmont, CA: Duxbury Press, 1998).

Kutner, Michael H., Christopher J. Nachtsheim, John Neter, and William Li, *Applied Linear Statistical Models*, 5th ed. (New York: McGraw-Hill Irwin, 2005).

Microsoft Excel 2007 (Redmond, WA: Microsoft Corp., 2007).

Minitab for Windows Version 14 (State College, PA: Minitab, 2005).