CHAPTER 14

Simple Linear Regression Analysis

Learning Objectives

After mastering the material in this chapter, you will be able to:

- **LO14-1** Explain the simple linear regression model.
- **LO14-2** Find the least squares point estimates of the slope and *y*-intercept.
- **LO14-3** Describe the assumptions behind simple linear regression and calculate the standard error.
- **LO14-4** Test the significance of the slope and *y*-intercept.
- **LO14-5** Calculate and interpret a confidence interval for a mean value and a prediction interval for an individual value.
- **LO14-6** Calculate and interpret the simple coefficients of determination and correlation.
- **LO14-7** Test hypotheses about the population correlation coefficient.
- **LO14-8** Test the significance of a simple linear regression model by using an *F*-test.
- **LO14-9** Use residual analysis to check the assumptions of simple linear regression.

Chapter Outline

- **14.1** The Simple Linear Regression Model and the Least Squares Point Estimates
- **14.2** Model Assumptions and the Standard Error
- **14.3** Testing the Significance of the Slope and *y*-Intercept
- **14.4** Confidence and Prediction Intervals
- **14.5** Simple Coefficients of Determination and Correlation (This section may be read any time after reading Section 14.1)
- **14.6** Testing the Significance of the Population Correlation Coefficient (Optional)
- **14.7** An *F*-Test for the Model
- **14.8** The QHIC Case: Developing An Advertising Strategy
- **14.9** Residual Analysis
- **14.10** Some Shortcut Formulas (Optional)

Managers often make decisions by studying the relationships between variables, and process improvements can often be made by understanding how changes in one or more variables affect the process output. **Regression analysis** is a statistical technique in which we use observed data to relate a variable of interest, which is called the **dependent** (or **response**) **variable,** to one or more **independent** (or **predictor**) **variables.** The objective is to build a **regression model,** or **prediction equation,** that can be used to **describe, predict,** and **control** the dependent variable on the basis of the independent variables. For example, a company might wish to improve its marketing process. After collecting data concerning the demand for a product, the product's price, and the advertising expenditures made to promote the product, the company might use regression analysis to develop an equation to predict demand on the basis of price and advertising expenditure. Predictions of demand for various price–advertising expenditure combinations can then be used to evaluate potential changes in the company's marketing strategies.

In the next two chapters we give a thorough presentation of regression analysis. We begin in this chapter by presenting **simple linear regression** analysis. Using this technique is appropriate when we are relating a dependent variable to a single independent variable and when *a straight-line model* describes the relationship between these two variables. We explain many of the methods of this chapter in the context of two new cases:

C

The Tasty Sub Shop Case: A business entrepreneur uses simple linear regression analysis to predict the yearly revenue for a potential restaurant site on the basis of the number of residents living near the site. The entrepreneur then uses the prediction to assess the profitability of the potential restaurant site.

The QHIC Case: The marketing department at Quality Home Improvement Center (QHIC) uses simple linear regression analysis to predict home upkeep expenditure on the basis of home value. Predictions of home upkeep expenditures are used to help determine which homes should be sent advertising brochures promoting QHIC's products and services.

14.1 The Simple Linear Regression Model and the Least Squares Point Estimates

LO14-1 Explain the simple linear regression model.

The simple linear regression model

The **simple linear regression model** assumes that the relationship between the **dependent variable, which is denoted *y*,** and the **independent variable, denoted *x*,** can be approximated by a straight line. We can tentatively decide whether there is an approximate straight-line relationship between y and x by making a **scatter diagram,** or **scatter plot,** of y versus x. First, data concerning the two variables are observed in pairs. To construct the scatter plot, each value of y is plotted against its corresponding value of x. If the y values tend to increase or decrease in a straight-line fashion as the x values increase, and if there is a scattering of the (x, y) points around the straight line, then it is reasonable to describe the relationship between y and x by using the simple linear regression model. We illustrate this in the following case study.

EXAMPLE 14.1 The Tasty Sub Shop Case: Predicting Yearly Revenue for a Potential Restaurant Site

C

Part 1: Purchasing a Tasty Sub Shop franchise

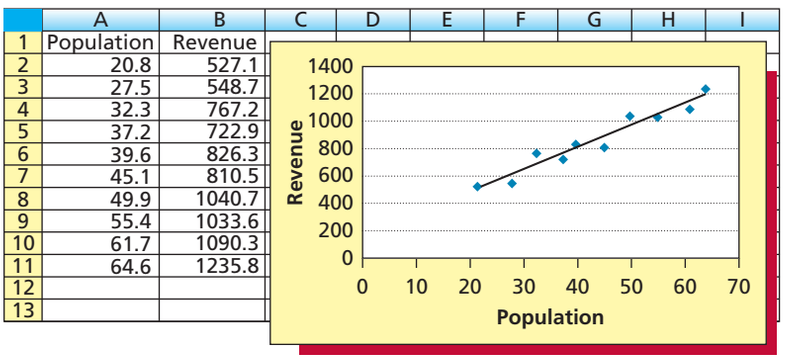
The Tasty Sub Shop is a restaurant chain that sells franchises to business entrepreneurs. Like Quiznos and Subway, the Tasty Sub Shop does not construct a standard, recognizable building to house each of its restaurants. Instead, the entrepreneur wishing to purchase a Tasty Sub franchise finds a suitable site, which consists of a suitable geographical location and suitable store space to rent. Then, when Tasty Sub approves the site, an architect and a contractor are hired to remodel the store rental space and thus "build" the Tasty Sub Shop restaurant. Franchise regulations allow Tasty Sub (and other chains) to help entrepreneurs understand the factors that affect restaurant profitability and to provide basic guidance in evaluating potential restaurant sites. However, in order to prevent restaurant chains from overpredicting profits and thus misleading potential franchise owners, these regulations

TABLE 14.1 The Tasty Sub Shop Revenue Data

TastySub1

Restaurant	Population Size, x (Thousands of Residents)	Yearly Revenue, y (Thousands of Dollars)
1	20.8	527.1
2	27.5	548.7
3	32.3	767.2
4	37.2	722.9
5	39.6	826.3
6	45.1	810.5
7	49.9	1040.7
8	55.4	1033.6
9	61.7	1090.3
10	64.6	1235.8

FIGURE 14.1 Excel Output of a Scatter Plot of y versus x



make each individual entrepreneur responsible for predicting the profits of his or her potential restaurant sites.

In this case study we consider a business entrepreneur who has found several potential sites for a Tasty Sub Shop restaurant. Similar to most existing Tasty Sub restaurant sites, each of the entrepreneur's sites is a store rental space located in an outdoor shopping area that is close to one or more residential areas. For a Tasty Sub restaurant built on such a site, yearly revenue is known to partially depend on (1) the number of residents living near the site and (2) the amount of business and shopping near the site. Referring to the number of residents living near a site as *population size* and to the yearly revenue for a Tasty Sub restaurant built on the site as *yearly revenue,* the entrepreneur will—in this chapter—try to predict the **dependent (response) variable** yearly revenue (y) on the basis of the **independent (predictor) variable** population size (x). (In the next chapter the entrepreneur will also use the amount of business and shopping near a site to help predict yearly revenue.) To predict yearly revenue on the basis of population size, the entrepreneur randomly selects 10 existing Tasty Sub restaurants that are built on sites similar to the sites that the entrepreneur is considering. The entrepreneur then asks the owner of each existing restaurant what the restaurant's revenue y was last year and estimates—with the help of the owner and published demographic information—the number of residents, or population size x, living near the site. The values of y (measured in thousands of dollars) and x (measured in thousands of residents) that are obtained are given in Table 14.1. In Figure 14.1 we give an Excel output of a scatter plot of y versus x. This plot shows (1) a tendency for the yearly revenues to increase in a straight-line fashion as the population sizes increase and (2) a scattering of points around the straight line. A **regression model** describing the relationship between y and x must represent these two characteristics. We now develop such a model.

Part 2: The simple linear regression model The **simple linear regression model** relating y to x can be expressed as follows:

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

This model says that the values of y can be represented by a *mean level* ($\mu_y = \beta_0 + \beta_1 x$) that changes in a straight line fashion as x changes, combined with random fluctuations (described by the error term ε) that cause the values of y to deviate from the mean level. Here:

1 The **mean level** $\mu_y = \beta_0 + \beta_1 x$ is the mean yearly revenue corresponding to a particular population size x. That is, noting that different Tasty Sub restaurants could potentially be built near different populations of the same size x, the mean level $\mu_y = \beta_0 + \beta_1 x$ is the mean of the yearly revenues that would be obtained by all such restaurants. In addition, because $\mu_y = \beta_0 + \beta_1 x$ is the equation of a straight line, the mean yearly revenues that correspond to increasing values of the population size x lie on a straight line. For example,

FIGURE 14.2 The Simple Linear Regression Model Relating Yearly Revenue (y) to Population (x)

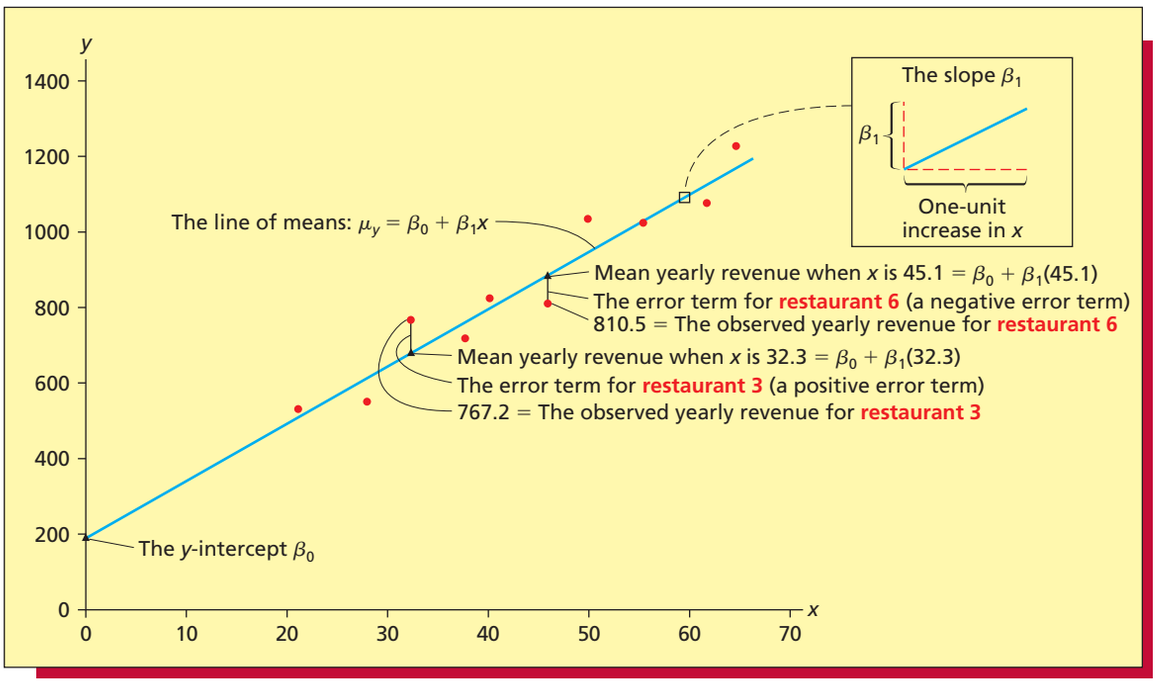


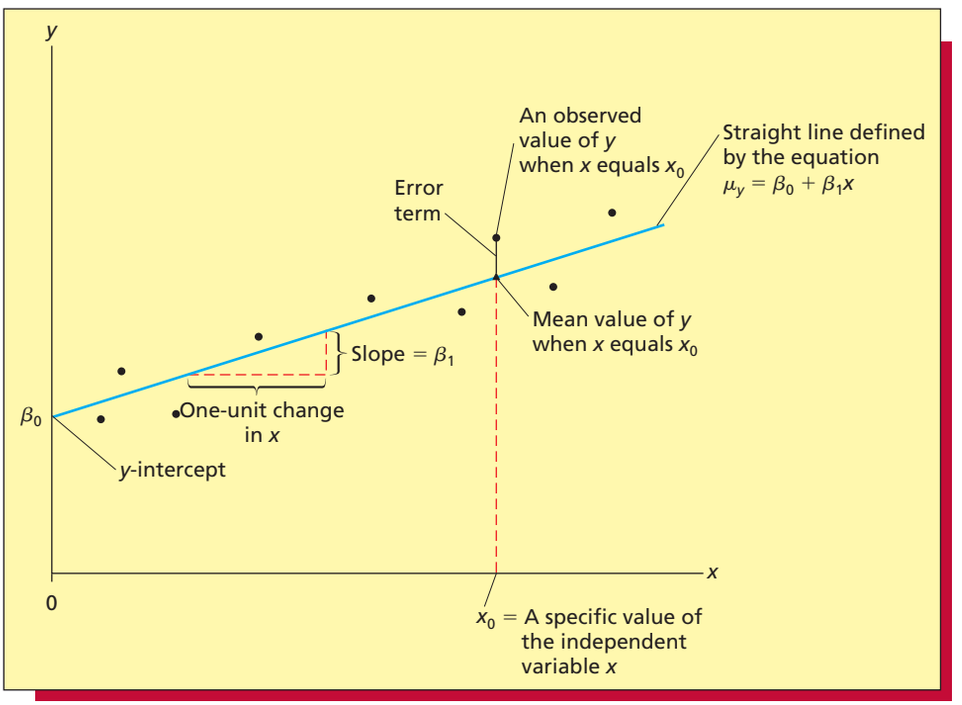
Table 14.1 tells us that 32,300 residents live near restaurant 3 and 45,100 residents live near restaurant 6. It follows that the mean yearly revenue for all Tasty Sub restaurants that could potentially be built near populations of 32,300 residents is $\beta_0 + \beta_1(32.3)$. Similarly, the mean yearly revenue for all Tasty Sub restaurants that could potentially be built near populations of 45,100 residents is $\beta_0 + \beta_1(45.1)$. Figure 14.2 depicts these two mean yearly revenues as triangles that lie on the straight line $\mu_y = \beta_0 + \beta_1 x$, which we call the **line of means.** The unknown parameters β_0 and β_1 are the ***y*-intercept** and the **slope** of the line of means. When we estimate β_0 and β_1 in the next subsection, we will be able to estimate mean yearly revenue μ_y on the basis of the population size x.

2 The ***y*-intercept** β_0 of the line of means can be understood by considering Figure 14.2. As illustrated in this figure, the y-intercept β_0 is the mean yearly revenue for all Tasty Sub restaurants that could potentially be built near populations of zero residents. However, because it is unlikely that a Tasty Sub restaurant would be built near a population of zero residents, this interpretation of β_0 is of dubious practical value. There are many regression situations where the y-intercept β_0 lacks a practical interpretation. In spite of this, statisticians have found that β_0 is almost always an important component of the line of means and thus of the simple linear regression model.

3 The **slope** β_1 of the line of means can also be understood by considering Figure 14.2. As illustrated in this figure, the slope β_1 is the change in mean yearly revenue that is associated with a one-unit increase (that is, a 1,000 resident increase) in the population size x.

4 The **error term** ε of the simple linear regression model accounts for any factors affecting yearly revenue other than the population size x. Such factors would include the amount of business and shopping near a restaurant and the skill of the owner as an operator of the restaurant. For example, Figure 14.2 shows that the error term for restaurant 3 is positive. Therefore, the observed yearly revenue $y = 767.2$ for restaurant 3 is above the corresponding mean yearly revenue for all restaurants that have $x = 32.3$. As another example, Figure 14.2 also shows that the error term for restaurant 6 is negative. Therefore, the observed yearly revenue $y = 810.5$ for restaurant 6 is below the corresponding mean yearly revenue for all restaurants that have $x = 45.1$. Of course, because we do not know the true values of β_0 and β_1, the relative positions of the quantities pictured in Figure 14.2 are only hypothetical.

FIGURE 14.3 The Simple Linear Regression Model (Here the Slope β_1 Is Positive)



With the Tasty Sub Shop example as background, we are ready to define the **simple linear regression model relating the dependent variable *y* to the independent variable *x*.** We suppose that we have gathered n observations—each observation consists of an observed value of x and its corresponding value of y. Then:

The Simple Linear Regression Model

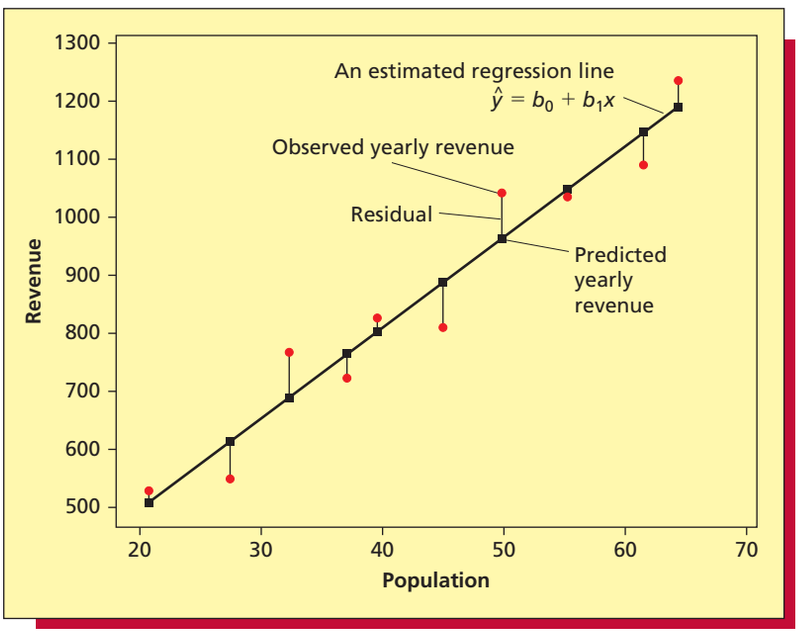
The **simple linear** (or **straight line**) **regression model** is: $y = \beta_0 + \beta_1 x + \varepsilon$

Here

1. $\mu_y = \beta_0 + \beta_1 x$ is the **mean value** of the dependent variable y when the value of the independent variable is x.
2. β_0 is the ***y*-intercept.** β_0 is the mean value of y when x equals zero.
3. β_1 is the **slope.** β_1 is the change (amount of increase or decrease) in the mean value of y associated with a one-unit increase in x. If β_1 is positive, the mean value of y increases as x increases. If β_1 is negative, the mean value of y decreases as x increases.
4. ε is an **error term** that describes the effects on y of all factors other than the value of the independent variable x.

This model is illustrated in Figure 14.3 (note that x_0 in this figure denotes a specific value of the independent variable x). The y-intercept β_0 and the slope β_1 are called **regression parameters.** In addition, we have interpreted the slope β_1 to be the change in the mean value of y associated with a one-unit increase in x. We sometimes refer to this change as *the effect of the independent variable x on the dependent variable y*. However, we cannot prove that a *change in an independent variable causes a change in the dependent variable.* Rather, regression can be used only to establish that the two variables move together and that the independent variable contributes information for predicting the dependent variable. For instance, regression analysis might be used to establish that as liquor sales have increased over the years, college professors' salaries have also increased. However, this does not prove that increases in liquor sales cause increases in college professors' salaries. Rather, both variables are influenced by a third variable—long-run growth in the national economy.

FIGURE 14.4 An Estimated Regression Line Drawn through the Tasty Sub Shop Revenue Data



The least squares point estimates Suppose that we have gathered n observations (x_1, y_1), (x_2, y_2), . . . , (x_n, y_n), where each observation consists of a value of an independent variable x and a corresponding value of a dependent variable y. Also, suppose that a scatter plot of the n observations indicates that the simple linear regression model relates y to x. In order to estimate the y-intercept β_0 and the slope β_1 of the line of means of this model, we could visually draw a line—called an **estimated regression line**—through the scatter plot. Then, we could read the y-intercept and slope off the estimated regression line and use these values as the point estimates of β_0 and β_1. Unfortunately, if different people visually drew lines through the scatter plot, their lines would probably differ from each other. What we need is the "best line" that can be drawn through the scatter plot. Although there are various definitions of what this best line is, one of the most useful best lines is the *least squares line*.

LO14-2 Find the least squares point estimates of the slope and y-intercept.

To understand the least squares line, we let

$$\hat{y} = b_0 + b_1x$$

denote the general equation of an estimated regression line drawn through a scatter plot. Here, because we will use this line to predict y on the basis of x, we call $\hat{y}$ *the predicted value of* y when the value of the independent variable is x. In addition, b_0 is the y-intercept and b_1 is the slope of the estimated regression line. When we determine numerical values for b_0 and b_1, these values will be the point estimates of the y-intercept β_0 and the slope β_1 of the line of means. To explain which estimated regression line is the least squares line, we begin with the Tasty Sub Shop situation. Figure 14.4 shows an estimated regression line drawn through a scatter plot of the Tasty Sub Shop revenue data. In this figure the red dots represent the 10 observed yearly revenues and the black squares represent the 10 predicted yearly revenues given by the estimated regression line. Furthermore, the line segments drawn between the red dots and black squares represent *residuals,* which are the differences between the observed and predicted yearly revenues. Intuitively, if a particular estimated regression line provides a good "fit" to the Tasty Sub Shop revenue data, it will make the predicted yearly revenues "close" to the observed yearly revenues, and thus the residuals given by the line will be small. The *least squares line* is the line that minimizes the sum of squared residuals. That is, the least squares line is the line positioned on the scatter plot so as to minimize the sum of the squared vertical distances between the observed and predicted yearly revenues.

To define the least squares line in a general situation, consider an arbitrary observation (x_i, y_i) in a sample of n observations. For this observation, the **predicted value of the dependent variable y** given by an estimated regression line is

$$\hat{y}_i = b_0 + b_1 x_i$$

Furthermore, the difference between the observed and predicted values of y, $y_i - \hat{y}_i$, is the **residual** for the observation, and the **sum of squared residuals** for all n observations is

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

The **least squares line** is the line that minimizes SSE. To find this line, we find the values of the y-intercept b_0 and slope b_1 that give values of $\hat{y}_i = b_0 + b_1 x_i$ that minimize SSE. These values of b_0 and b_1 are called the **least squares point estimates** of β_0 and β_1. Using calculus, it can be shown that these estimates are calculated as follows:[1]

The Least Squares Point Estimates

For the simple linear regression model:

1 The **least squares point estimate of the slope β_1** is

$$b_1 = \frac{SS_{xy}}{SS_{xx}} \quad \text{where}$$

$$SS_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n} \quad \text{and} \quad SS_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

2 The **least squares point estimate of the y-intercept β_0** is

$$b_0 = \bar{y} - b_1 \bar{x} \quad \text{where}$$

$$\bar{y} = \frac{\sum y_i}{n} \quad \text{and} \quad \bar{x} = \frac{\sum x_i}{n}$$

Here n is the number of observations (an observation is an observed value of x and its corresponding value of y).

The following example illustrates how to calculate these point estimates and how to use these point estimates to estimate mean values and predict individual values of the dependent variable. Note that the quantities SS_{xy} and SS_{xx} used to calculate the least squares point estimates are also used throughout this chapter to perform other important calculations.

[1] In order to simplify notation, we will often drop the limits on summations in this and subsequent chapters. That is, instead of using the summation $\sum_{i=1}^{n}$ we will simply write $\sum$.

EXAMPLE 14.2 The Tasty Sub Shop Case: The Least Squares Estimates

Part 1: Calculating the least squares point estimates Again consider the Tasty Sub Shop problem. To compute the least squares point estimates of the regression parameters β_0 and β_1 we first calculate the following preliminary summations:

y_i	x_i	x_i^2	$x_i y_i$
527.1	20.8	$(20.8)^2 = 432.64$	$(20.8)(527.1) = 10963.68$
548.7	27.5	$(27.5)^2 = 756.25$	$(27.5)(548.7) = 15089.25$
767.2	32.3	$(32.3)^2 = 1{,}043.29$	$(32.3)(767.2) = 24780.56$
722.9	37.2	$(37.2)^2 = 1{,}383.84$	$(37.2)(722.9) = 26891.88$
826.3	39.6	$(39.6)^2 = 1{,}568.16$	$(39.6)(826.3) = 32721.48$
810.5	45.1	$(45.1)^2 = 2{,}034.01$	$(45.1)(810.5) = 36553.55$
1040.7	49.9	$(49.9)^2 = 2{,}490.01$	$(49.9)(1040.7) = 51930.93$
1033.6	55.4	$(55.4)^2 = 3{,}069.16$	$(55.4)(1033.6) = 57261.44$
1090.3	61.7	$(61.7)^2 = 3{,}806.89$	$(61.7)(1090.3) = 67271.51$
1235.8	64.6	$(64.6)^2 = 4{,}173.16$	$(64.6)(1235.8) = 79832.68$
$\sum y_i = 8603.1$	$\sum x_i = 434.1$	$\sum x_i^2 = 20{,}757.41$	$\sum x_i y_i = 403{,}296.96$

Using these summations, we calculate SS_{xy} and SS_{xx} as follows.

$$SS_{xy} = \sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}$$
$$= 403{,}296.96 - \frac{(434.1)(8603.1)}{10}$$
$$= 29{,}836.389$$

$$SS_{xx} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$
$$= 20{,}757.41 - \frac{(434.1)^2}{10}$$
$$= 1913.129$$

It follows that the least squares point estimate of the slope β_1 is

$$b_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{29{,}836.389}{1913.129} = 15.596$$

Furthermore, because

$$\bar{y} = \frac{\sum y_i}{n} = \frac{8603.1}{10} = 860.31 \quad \text{and} \quad \bar{x} = \frac{\sum x_i}{n} = \frac{434.1}{10} = 43.41$$

the least squares point estimate of the y-intercept β_0 is

$$b_0 = \bar{y} - b_1\bar{x} = 860.31 - (15.596)(43.41) = 183.31$$

(where we have used more decimal place accuracy than shown to obtain the result 183.31).

Because $b_1 = 15.596$, we estimate that mean yearly revenue at Tasty Sub restaurants increases by 15.596 (that is by \$15,596) for each one-unit (1,000 resident) increase in the population size x. Because $b_0 = 183.31$, we estimate that mean yearly revenue for all Tasty Sub restaurants that could potentially be built near populations of zero residents is \$183,310. However, because it is unlikely that a Tasty Sub restaurant would be built near a population of zero residents, this interpretation is of dubious practical value.

The least squares line

$$\hat{y} = b_0 + b_1 x = 183.31 + 15.596x$$

is sometimes called the *least squares prediction equation.* In Table 14.2 (on the next page) we summarize using this prediction equation to calculate the predicted yearly revenues and the

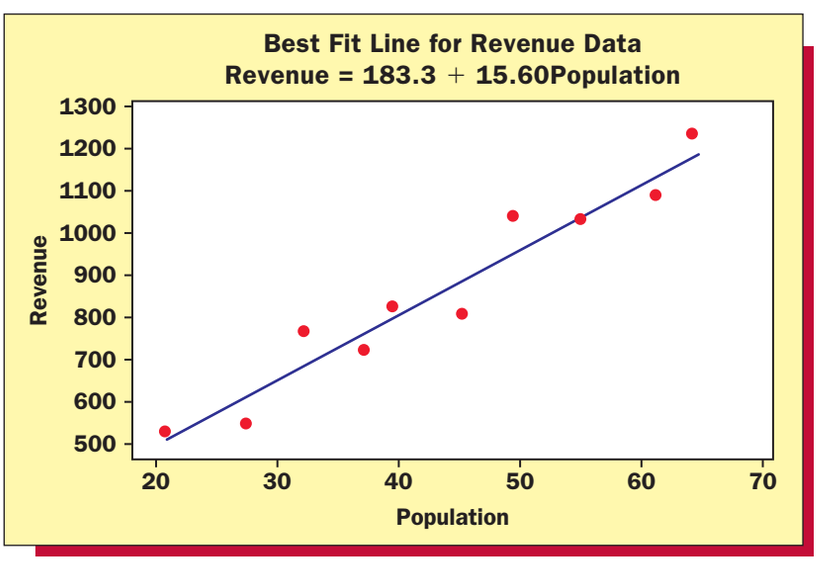
TABLE 14.2 Calculation of *SSE* Obtained by Using the Least Squares Point Estimates

y_i	x_i	$\hat{y}_i = 183.31 + 15.596x_i$	$y_i - \hat{y}_i$
527.1	20.8	183.31 + 15.596(20.8) = 507.69	19.41
548.7	27.5	183.31 + 15.596(27.5) = 612.18	−63.48
767.2	32.3	687.04	80.16
722.9	37.2	763.46	−40.56
826.3	39.6	800.89	25.41
810.5	45.1	886.67	−76.17
1040.7	49.9	961.53	79.17
1033.6	55.4	1047.30	−13.70
1090.3	61.7	1145.55	−55.25
1235.8	64.6	1190.78	45.02

$$SSE = \sum (y_i - \hat{y}_i)^2 = (19.41)^2 + (-63.48)^2 + \cdots + (45.02)^2 = 30{,}460.21$$

Note: The predictions and residuals in this table are taken from MINITAB, which uses values of b_0 and b_1 that are more precise than the rounded values we have calculated by hand. If you use the formula $\hat{y}_i = 183.31 + 15.596x_i$, your figures may differ slightly from those given here.

FIGURE 14.5 The MINITAB Output of the Least Squares Line



residuals for the 10 observed Tasty Sub restaurants. For example, because the population size for restaurant 1 was 20.8, the predicted yearly revenue for restaurant 1 is

$$\hat{y}_1 = 183.31 + 15.596(20.8) = 507.69$$

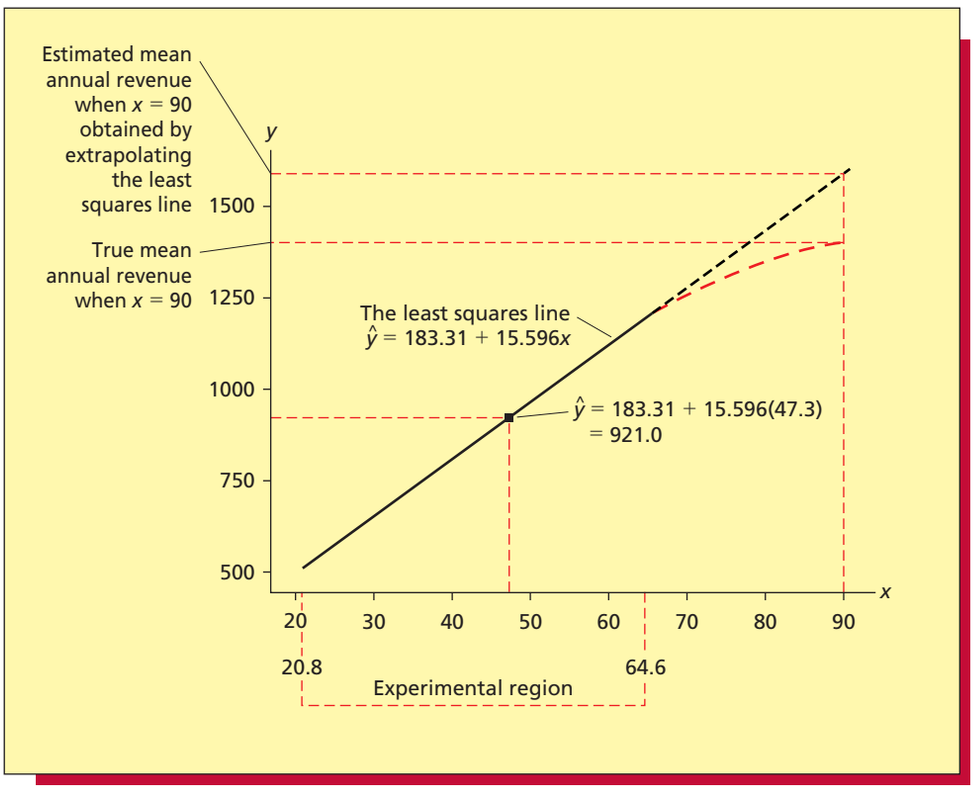
It follows, because the observed yearly revenue for restaurant 1 was $y_1 = 527.1$, that the residual for restaurant 1 is

$$y_1 - \hat{y}_1 = 527.1 - 507.69 = 19.41$$

If we consider all of the residuals in Table 14.2 and add their squared values, we find that *SSE*, the sum of squared residuals, is 30,460.21. This *SSE* value will be used throughout this chapter. Figure 14.5 gives the MINITAB output of the least squares line. Note that this output gives (within rounding) the least squares estimates we have calculated ($b_0 = 183.3$ and $b_1 = 15.60$). In general, we will rely on Excel and MINITAB to compute the least squares estimates (and to perform many other regression calculations).

Part 2: Estimating a mean yearly revenue and predicting an individual yearly revenue We define the **experimental region** to be the range of the previously observed population sizes. Referring to Table 14.2, we see that the experimental region consists of the range

FIGURE 14.6 Point Estimation and Point Prediction, and the Danger of Extrapolation



of population sizes from 20.8 to 64.6. The simple linear regression model relates yearly revenue y to population size x for values of x that are in the experimental region. For such values of x, the least squares line is the estimate of the line of means. It follows that the point on the least squares line corresponding to a population size of x

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

is the point estimate of $\beta_0 + \beta_1 x$, the mean yearly revenue for all Tasty Sub restaurants that could potentially be built near populations of size x. In addition, we predict the error term ε to be 0. Therefore, $\hat{y}$ is also the *point prediction* of an *individual value* $y = \beta_0 + \beta_1 x + \varepsilon$, which is the yearly revenue for a single (individual) Tasty Sub restaurant that is built near a population of size x. Note that the reason we predict the error term ε to be zero is that, because of several *regression assumptions* to be discussed in the next section, ε has a 50 percent chance of being positive and a 50 percent chance of being negative.

For example, suppose that one of the business entrepreneur's potential restaurant sites is near a population of 47,300 residents. Because $x = 47.3$ is in the experimental region,

$$\begin{aligned}\hat{y} &= 183.31 + 15.596(47.3) \\ &= 921.0 \text{ (that is, \$921,000)}\end{aligned}$$

is

1. The **point estimate** of the mean yearly revenue for all Tasty Sub restaurants that could potentially be built near populations of 47,300 residents.
2. The **point prediction** of the yearly revenue for a single Tasty Sub restaurant that is built near a population of 47,300 residents.

Figure 14.6 illustrates $\hat{y} = 921.0$ as a square on the least squares line. Moreover, suppose that the yearly rent and other fixed costs for the entrepreneur's potential restaurant will be \$257,550 and that (according to Tasty Sub corporate headquarters) the yearly food and other variable costs for the restaurant will be 60 percent of the yearly revenue. Because we predict that the yearly revenue

for the restaurant will be \$921,000, it follows that we predict that the yearly total operating cost for the restaurant will be \$257,550 + .6(\$921,000) = \$810,150. In addition, if we subtract this predicted yearly operating cost from the predicted yearly revenue of \$921,000, we predict that the yearly profit for the restaurant will be \$110,850. Of course, these predictions are point predictions. In Section 14.4 we will predict the restaurant's yearly revenue and profit *with confidence*.

To conclude this example, note that Figure 14.6 illustrates the potential danger of using the least squares line to predict outside the experimental region. In the figure, we extrapolate the least squares line beyond the experimental region to obtain a prediction for a population size of $x = 90$. As shown in Figure 14.6, for values of x in the experimental region (that is, between 20.8 and 64.6) the observed values of y tend to increase in a straight-line fashion as the values of x increase. However, for population sizes greater than $x = 64.6$, we have no data to tell us whether the relationship between y and x continues as a straight-line relationship or, possibly, becomes a curved relationship. If, for example, this relationship becomes the sort of curved relationship shown in Figure 14.6, then extrapolating the straight-line prediction equation to obtain a prediction for $x = 90$ would overestimate mean yearly revenue (see Figure 14.6).

The previous example illustrates that when we are using a least squares regression line, we should not estimate a mean value or predict an individual value unless the corresponding value of x is in the **experimental region**—the range of the previously observed values of x. Often the value $x = 0$ is not in the experimental region. In such a situation, it would not be appropriate to interpret the y-intercept b_0 as the estimate of the mean value of y when x equals 0. For example, consider the Tasty Sub Shop problem. Figure 14.6 illustrates that the population size $x = 0$ is not in the experimental region. Therefore, it would not be appropriate to use $b_0 = 183.31$ as the point estimate of the mean yearly revenue for all Tasty Sub restaurants that could potentially be built near populations of zero residents. Because it is not meaningful to interpret the y-intercept in many regression situations, we often omit such interpretations.

We now present a general procedure for estimating a mean value and predicting an individual value:

Point Estimation and Point Prediction in Simple Linear Regression

Let b_0 and b_1 be the least squares point estimates of the y-intercept β_0 and the slope β_1 in the simple linear regression model, and suppose that x_0, a specified value of the independent variable x, is inside the experimental region. Then

$$\hat{y} = b_0 + b_1 x_0$$

1. is the **point estimate** of the **mean value of the dependent variable** when the value of the independent variable is x_0.
2. is the **point prediction** of an **individual value of the dependent variable** when the value of the independent variable is x_0. Here we predict the error term to be 0.

Exercises for Section 14.1

CONCEPTS

14.1 What is the least squares regression line, and what are the least squares point estimates?

14.2 Why is it dangerous to extrapolate outside the experimental region?

METHODS AND APPLICATIONS

In Exercises 14.3 through 14.8 we present six data sets involving a dependent variable y and an independent variable x. For each data set, assume that the simple linear regression model

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

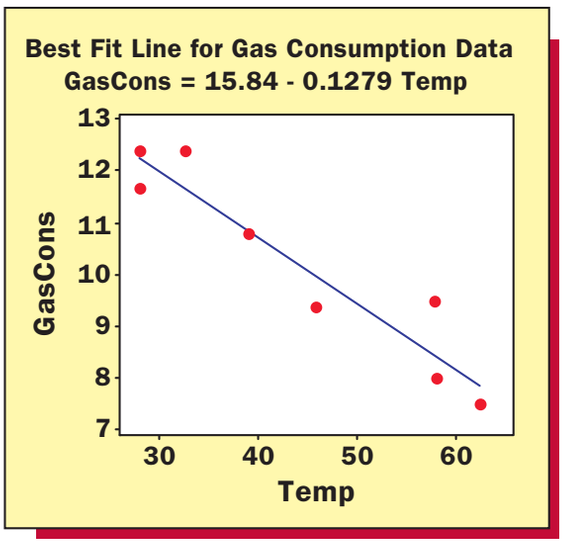
relates y to x.

14.3 THE NATURAL GAS CONSUMPTION CASE GasCon1

On the next page we give the average hourly outdoor temperature (x) in a city during a week and the city's natural gas consumption (y) during the week for each of eight weeks (the temperature readings are expressed in degrees Fahrenheit and the natural gas consumptions are expressed in

Week	Average Hourly Temperature, x (°F)	Natural Gas Consumption, y (MMcf)
1	28.0	12.4
2	28.0	11.7
3	32.5	12.4
4	39.0	10.8
5	45.9	9.4
6	57.8	9.5
7	58.1	8.0
8	62.5	7.5

DS GasCon1



millions of cubic feet of natural gas—denoted MMcf). The output to the right of the data is obtained when MINITAB is used to fit a least squares line to the natural gas consumption data.

a Find the least squares point estimates b_0 and b_1 on the computer output and report their values. Interpret b_0 and b_1. Is an average hourly temperature of 0°F in the experimental region? What does this say about the interpretation of b_0?

b Use the facts that $SS_{xy} = -179.6475$; $SS_{xx} = 1{,}404.355$; $\bar{y} = 10.2125$; and $\bar{x} = 43.98$ to hand calculate (within rounding) b_0 and b_1.

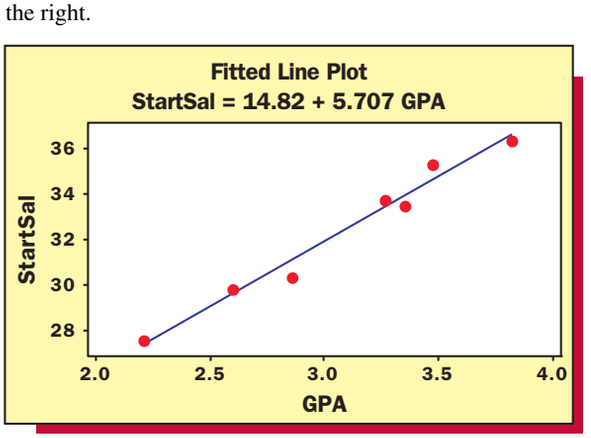
c Use the least squares line to compute a point estimate of the mean natural gas consumption for all weeks having an average hourly temperature of 40°F and a point prediction of the natural gas consumption for an individual week having an average hourly temperature of 40°F.

14.4 THE STARTING SALARY CASE DS StartSal

The chairman of the marketing department at a large state university undertakes a study to relate starting salary (y) after graduation for marketing majors to grade point average (GPA) in major courses. To do this, records of seven recent marketing graduates are randomly selected, and the data shown below on the left are obtained. The MINITAB output obtained by fitting a least squares regression line to the data is below on the right.

Marketing Graduate	GPA, x	Starting Salary, y (Thousands of Dollars)
1	3.26	33.8
2	2.60	29.8
3	3.35	33.5
4	2.86	30.4
5	3.82	36.4
6	2.21	27.6
7	3.47	35.3

DS StartSal



a Find the least squares point estimates b_0 and b_1 on the computer output and report their values. Interpret b_0 and b_1. Does the interpretation of b_0 make practical sense?

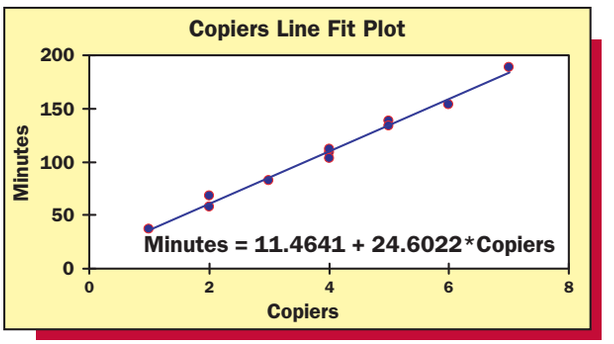
b Use the least squares line to compute a point estimate of the mean starting salary for all marketing graduates having a grade point average of 3.25 and a point prediction of the starting salary for an individual marketing graduate having a grade point average of 3.25.

14.5 THE SERVICE TIME CASE DS SrvcTime

Accu-Copiers, Inc., sells and services the Accu-500 copying machine. As part of its standard service contract, the company agrees to perform routine service on this copier. To obtain information about the time it takes to perform routine service, Accu-Copiers has collected data for 11 service calls. The data and Excel output from fitting a least squares regression line to the data follow on the next page.

Service Call	Number of Copiers Serviced, x	Number of Minutes Required, y
1	4	109
2	2	58
3	5	138
4	7	189
5	1	37
6	3	82
7	4	103
8	5	134
9	2	68
10	4	112
11	6	154

DS SrvcTime



a Find the least squares point estimates b_0 and b_1 on the computer output and report their values. Interpret b_0 and b_1. Does the interpretation of b_0 make practical sense?

b Use the least squares line to compute a point estimate of the mean time to service four copiers and a point prediction of the time to service four copiers on a single call.

14.6 THE FRESH DETERGENT CASE DS Fresh

Enterprise Industries produces Fresh, a brand of liquid laundry detergent. In order to study the relationship between price and demand for the large bottle of Fresh, the company has gathered data concerning demand for Fresh over the last 30 sales periods (each sales period is four weeks). Here, for each sales period,

y = demand for the large bottle of Fresh (in hundreds of thousands of bottles) in the sales period, and

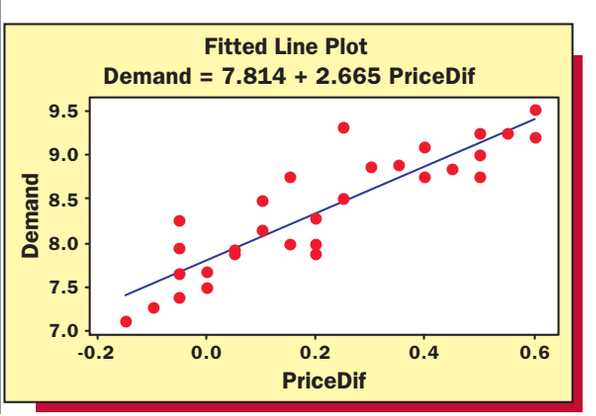
x = the difference between the average industry price (in dollars) of competitors' similar detergents and the price (in dollars) of Fresh as offered by Enterprise Industries in the sales period.

The data and MINITAB output from fitting a least squares regression line to the data follow below.

Fresh Detergent Demand Data

Sales Period	y	x	Sales Period	y	x
1	7.38	−.05	24	8.50	.10
2	8.51	.25	25	8.75	.50
3	9.52	.60	26	9.21	.60
4	7.50	0	27	8.27	−.05
5	9.33	.25	28	7.67	0
6	8.28	.20	29	7.93	.05
7	8.75	.15	30	9.26	.55
8	7.87	.05			
9	7.10	−.15			
10	8.00	.15			
11	7.89	.20			
12	8.15	.10			
13	9.10	.40			
14	8.86	.45			
15	8.90	.35			
16	8.87	.30			
17	9.26	.50			
18	9.00	.50			
19	8.75	.40			
20	7.95	−.05			
21	7.65	−.05			
22	7.27	−.10			
23	8.00	.20			

DS Fresh



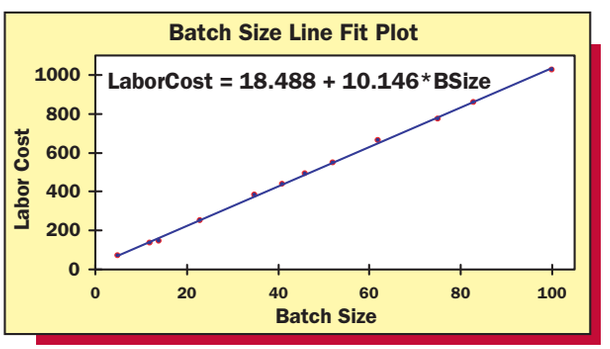
a Find the least squares point estimates b_0 and b_1 on the computer output and report their values.
b Interpret b_0 and b_1. Does the interpretation of b_0 make practical sense?
c Write the equation of the least squares line.
d Use the least squares line to compute a point estimate of the mean demand in all sales periods when the price difference is .10 and a point prediction of the actual demand in an individual sales period when the price difference is .10.

14.7 THE DIRECT LABOR COST CASE DirLab

An accountant wishes to predict direct labor cost (y) on the basis of the batch size (x) of a product produced in a job shop. Data for 12 production runs are given in the table below, along with the Excel output from fitting a least squares regression line to the data.

Direct Labor Cost Data DirLab

Direct Labor Cost, y ($100s)	Batch Size, x
71	5
663	62
381	35
138	12
861	83
145	14
493	46
548	52
251	23
1024	100
435	41
772	75



a By using the formulas illustrated in Example 14.2 (see page 493) and the data provided, verify that (within rounding) $b_0 = 18.488$ and $b_1 = 10.146$, as shown on the Excel output.
b Interpret the meanings of b_0 and b_1. Does the interpretation of b_0 make practical sense?
c Write the least squares prediction equation.
d Use the least squares line to obtain a point estimate of the mean direct labor cost for all batches of size 60 and a point prediction of the direct labor cost for an individual batch of size 60.

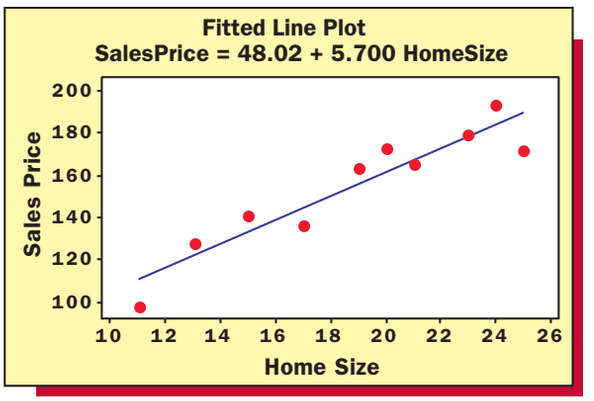
14.8 THE REAL ESTATE SALES PRICE CASE RealEst

A real estate agency collects data concerning y = the sales price of a house (in thousands of dollars), and x = the home size (in hundreds of square feet). The data are given in the table below. The MINITAB output from fitting a least squares regression line to the data is on the next page.

Real Estate Sales Price Data RealEst

Sales Price (y)	Home Size (x)	Sales Price (y)	Home Size (x)
180	23	165.9	21
98.1	11	193.5	24
173.1	20	127.8	13
136.5	17	163.5	19
141	15	172.5	25

Source: Reprinted with permission from *The Real Estate Appraiser and Analyst*, Spring 1986 issue. Copyright 1986 by the Appraisal Institute, Chicago, Illinois.



a By using the formulas illustrated in Example 14.2 (see page 493) and the data provided, verify that (within rounding) $b_0 = 48.02$ and $b_1 = 5.700$, as shown on the MINITAB output.
b Interpret the meanings of b_0 and b_1. Does the interpretation of b_0 make practical sense?
c Write the least squares prediction equation.
d Use the least squares line to obtain a point estimate of the mean sales price of all houses having 2,000 square feet and a point prediction of the sales price of an individual house having 2,000 square feet.

LO14-3 Describe the assumptions behind simple linear regression and calculate the standard error.

14.2 Model Assumptions and the Standard Error

Model assumptions In order to perform hypothesis tests and set up various types of intervals when using the simple linear regression model

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

we need to make certain assumptions about the error term ε. At any given value of x, there is a population of error term values that could potentially occur. These error term values describe the different potential effects on y of all factors other than the value of x. Therefore, these error term values explain the variation in the y values that could be observed when the independent variable is x. Our statement of the simple linear regression model assumes that μ_y, the mean of the population of all y values that could be observed when the independent variable is x, is $\beta_0 + \beta_1 x$. This model also implies that $\varepsilon = y - (\beta_0 + \beta_1 x)$, so this is equivalent to assuming that the mean of the corresponding population of potential error term values is 0. In total, we make four assumptions (called the **regression assumptions**) about the simple linear regression model. These assumptions can be stated in terms of potential y values or, equivalently, in terms of potential error term values. Following tradition, we begin by stating these assumptions in terms of potential error term values:

The Regression Assumptions

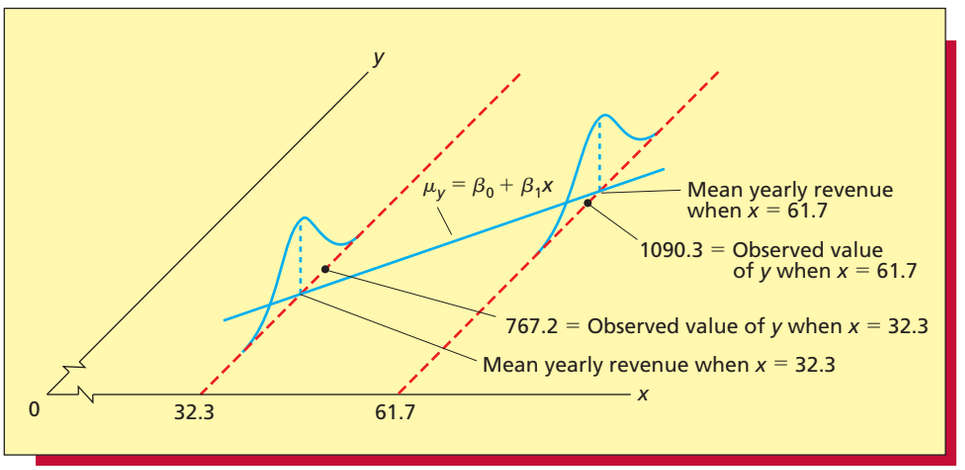
1 At any given value of x, the population of potential error term values has a **mean equal to 0.**

2 **Constant Variance Assumption**
At any given value of x, the population of potential error term values has a variance that does not depend on the value of x. That is, the different populations of potential error term values corresponding to different values of x have **equal variances.** We denote the **constant variance as σ^2.**

3 **Normality Assumption**
At any given value of x, the population of potential error term values has a **normal distribution.**

4 **Independence Assumption**
Any one value of the error term ε is **statistically independent** of any other value of ε. That is, the value of the error term ε corresponding to an observed value of y is statistically independent of the value of the error term corresponding to any other observed value of y.

FIGURE 14.7 An Illustration of the Model Assumptions



Taken together, the first three assumptions say that, at any given value of x, the population of potential error term values is **normally distributed** with **mean zero** and a **variance σ^2 that does not depend on the value of x.** Because the potential error term values cause the variation in the potential y values, these assumptions imply that the population of all y values that could be observed when the independent variable is x is **normally distributed** with **mean $\beta_0 + \beta_1 x$** and **a variance σ^2 that does not depend on x.** These three assumptions are illustrated in Figure 14.7 in the context of the Tasty Sub Shop problem. Specifically, this figure depicts the populations of yearly revenues corresponding to two values of the population size x—32.3 and 61.7. Note that these populations are shown to be normally distributed with different means (each of which is on the line of means) and with the same variance (or spread).

The independence assumption is most likely to be violated when time series data are being utilized in a regression study. For example, the natural gas consumption data in Exercise 14.3 are time series data. Intuitively, the independence assumption says that there is no pattern of positive error terms being followed (in time) by other positive error terms, and there is no pattern of positive error terms being followed by negative error terms. That is, there is no pattern of higher-than-average y values being followed by other higher-than-average y values, and there is no pattern of higher-than-average y values being followed by lower-than-average y values.

It is important to point out that the regression assumptions very seldom, if ever, hold exactly in any practical regression problem. However, it has been found that regression results are not extremely sensitive to mild departures from these assumptions. In practice, only pronounced departures from these assumptions require attention. In Section 14.9 we show how to check the regression assumptions. Prior to doing this, we will suppose that the assumptions are valid in our examples.

In Section 14.1 we stated that, when we predict an individual value of the dependent variable, we predict the error term to be 0. To see why we do this, note that the regression assumptions state that, at any given value of the independent variable, the population of all error term values that can potentially occur is normally distributed with a mean equal to 0. Because we also assume that successive error terms (observed over time) are statistically independent, each error term has a 50 percent chance of being positive and a 50 percent chance of being negative. Therefore, it is reasonable to predict any particular error term value to be 0.

The mean square error and the standard error To present statistical inference formulas in later sections, we need to be able to compute point estimates of σ^2 and σ, the constant variance and standard deviation of the error term populations. The point estimate of σ^2 is called the **mean square error** and the point estimate of σ is called the **standard error.** In the following box, we show how to compute these estimates:

The Mean Square Error and the Standard Error

If the regression assumptions are satisfied and *SSE* is the sum of squared residuals:

1 The point estimate of σ^2 is the **mean square error**

$$s^2 = \frac{SSE}{n-2}$$

2 The point estimate of σ is the **standard error**

$$s = \sqrt{\frac{SSE}{n-2}}$$

In order to understand these point estimates, recall that σ^2 is the variance of the population of y values (for a given value of x) around the mean value μ_y. Because $\hat{y}$ is the point estimate of this mean, it seems natural to use

$$SSE = \sum (y_i - \hat{y}_i)^2$$

to help construct a point estimate of σ^2. We divide *SSE* by $n-2$ because it can be proven that doing so makes the resulting s^2 an unbiased point estimate of σ^2. Here we call $n-2$ the **number of degrees of freedom** associated with *SSE*.

EXAMPLE 14.3 The Tasty Sub Shop Case: The Standard Error

Consider the Tasty Sub Shop situation, and recall that in Table 14.2 (page 494) we have calculated the sum of squared residuals to be $SSE = 30{,}460.21$. It follows, because we have observed $n = 10$ yearly revenues, that the point estimate of σ^2 is the mean square error

$$s^2 = \frac{SSE}{n-2} = \frac{30{,}460.21}{10-2} = 3807.526$$

This implies that the point estimate of σ is the standard error

$$s = \sqrt{s^2} = \sqrt{3807.526} = 61.7052$$

To conclude this section, note that in optional Section 14.10 we present a shortcut formula for calculating *SSE*. The reader may study Section 14.10 now or at any later point.

Exercises for Section 14.2

connect

CONCEPTS

14.9 What four assumptions do we make about the simple linear regression model?

14.10 What is estimated by the mean square error, and what is estimated by the standard error?

METHODS AND APPLICATIONS

14.11 THE NATURAL GAS CONSUMPTION CASE DS GasCon1

When a least squares line is fit to the 8 observations in the natural gas consumption data, we obtain $SSE = 2.568$. Calculate s^2 and s.

14.12 THE STARTING SALARY CASE DS StartSal

When a least squares line is fit to the 7 observations in the starting salary data, we obtain $SSE = 1.438$. Calculate s^2 and s.

14.13 THE SERVICE TIME CASE DS SrvcTime

When a least squares line is fit to the 11 observations in the service time data, we obtain $SSE = 191.7017$. Calculate s^2 and s.

14.14 THE FRESH DETERGENT CASE DS Fresh

When a least squares line is fit to the 30 observations in the Fresh detergent data, we obtain $SSE = 2.806$. Calculate s^2 and s.

14.15 THE DIRECT LABOR COST CASE DS DirLab

When a least squares line is fit to the 12 observations in the labor cost data, we obtain $SSE = 746.7624$. Calculate s^2 and s.

14.16 THE REAL ESTATE SALES PRICE CASE DS RealEst

When a least squares line is fit to the 10 observations in the real estate sales price data, we obtain $SSE = 896.8$. Calculate s^2 and s.

14.17 Ten sales regions of equal sales potential for a company were randomly selected. The advertising expenditures (in units of \$10,000) in these 10 sales regions were purposely set during July of last year at, respectively, 5, 6, 7, 8, 9, 10, 11, 12, 13 and 14. The sales volumes (in units of \$10,000) were then recorded for the 10 sales regions and found to be, respectively, 89, 87, 98, 110, 103, 114, 116, 110, 126, and 130. Assuming that the simple linear regression model is appropriate, it can be shown that $b_0 = 66.2121$, $b_1 = 4.4303$, and $SSE = 222.8242$. Calculate s^2 and s. DS SalesPlot

14.3 Testing the Significance of the Slope and *y*-Intercept

LO14-4 Test the significance of the slope and *y*-intercept.

Testing the significance of the slope A simple linear regression model is not likely to be useful unless there is a **significant relationship between *y* and *x*.** In order to judge the significance of the relationship between y and x, we test the null hypothesis

$$H_0: \beta_1 = 0$$

which says that there is no change in the mean value of y associated with an increase in x, versus the alternative hypothesis

$$H_a: \beta_1 \neq 0$$

which says that there is a (positive or negative) change in the mean value of y associated with an increase in x. It would be reasonable to conclude that x is significantly related to y if we can be quite certain that we should reject H_0 in favor of H_a.

In order to test these hypotheses, recall that we compute the least squares point estimate b_1 of the true slope β_1 by using a sample of n observed values of the dependent variable y. Different samples of n observed y values would yield different values of the least squares point estimate b_1. It can be shown that, if the regression assumptions hold, then the population of all possible values of b_1 is normally distributed with a mean of β_1 and with a standard deviation of

$$\sigma_{b_1} = \frac{\sigma}{\sqrt{SS_{xx}}}$$

The standard error s is the point estimate of σ, so it follows that a point estimate of σ_{b_1} is

$$s_{b_1} = \frac{s}{\sqrt{SS_{xx}}}$$

which is called the **standard error of the estimate b_1.** Furthermore, if the regression assumptions hold, then the population of all values of

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

has a t distribution with $n - 2$ degrees of freedom. It follows that, if the null hypothesis $H_0: \beta_1 = 0$ is true, then the population of all possible values of the test statistic

$$t = \frac{b_1}{s_{b_1}}$$

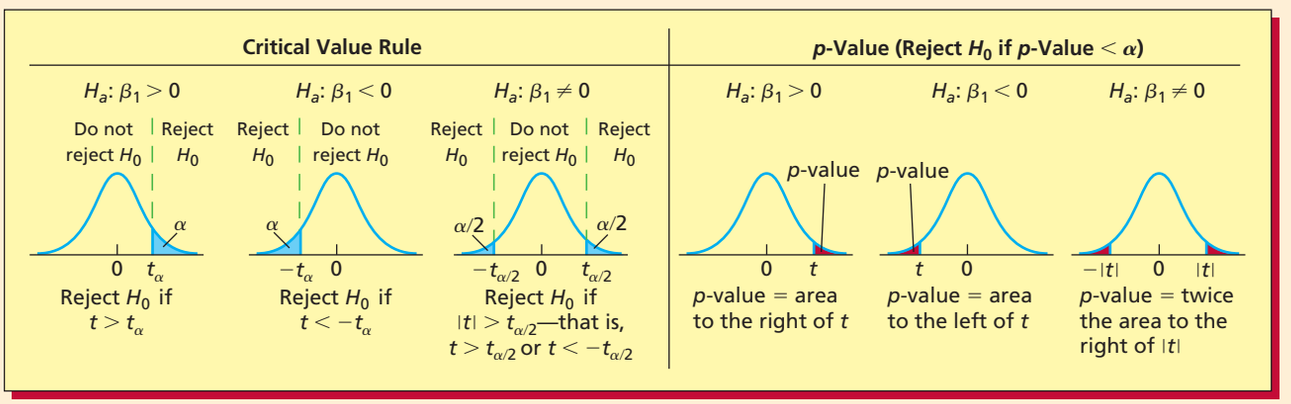
has a t distribution with $n - 2$ degrees of freedom. Therefore, we can test the significance of the regression relationship as follows:

Testing the Significance of the Regression Relationship: Testing the Significance of the Slope

Null Hypothesis $H_0: \beta_1 = 0$

Test Statistic $t = \dfrac{b_1}{s_{b_1}}$ where $s_{b_1} = \dfrac{s}{\sqrt{SS_{xx}}}$

Assumptions The regression assumptions



Here $t_{\alpha/2}$, t_{α}, and all p-values are based on $n - 2$ degrees of freedom. **If we can reject H_0: $\beta_1 = 0$ at a given value of α, then we conclude that the slope (or, equivalently, the regression relationship) is significant at the α level.**

We usually use the two-sided alternative H_a: $\beta_1 \neq 0$ for this test of significance. However, sometimes a one-sided alternative is appropriate. For example, in the Tasty Sub Shop problem we can say that if the slope β_1 is not 0, then it must be positive. A positive β_1 would say that mean yearly revenue increases as the population size x increases. Because of this, it would be appropriate to decide that x is significantly related to y if we can reject H_0: $\beta_1 = 0$ in favor of the one-sided alternative H_a: $\beta_1 > 0$. Although this test would be slightly more effective than the usual two tailed test, there is little practical difference between using the one tailed or two tailed test. Furthermore, computer packages (such as Excel and MINITAB) present results for the two tailed test. For these reasons we will emphasize the two tailed test in future discussions.

It should also be noted that

1. **If we can decide that the slope is significant at the .05 significance level,** then we have concluded that x is significantly related to y by using a test that allows only a .05 probability of concluding that x is significantly related to y when it is not. **This is usually regarded as strong evidence that the regression relationship is significant.**
2. **If we can decide that the slope is significant at the .01 significance level, this is usually regarded as very strong evidence that the regression relationship is significant.**
3. **The smaller the significance level α at which H_0 can be rejected, the stronger is the evidence that the regression relationship is significant.**

EXAMPLE 14.4 The Tasty Sub Shop Case: Testing the Significance of the Slope C

Again consider the Tasty Sub Shop revenue model. For this model $SS_{xx} = 1913.129$, $b_1 = 15.596$, and $s = 61.7052$ [see Examples 14.2 (page 493) and 14.3 (page 502)]. Therefore,

$$s_{b_1} = \frac{s}{\sqrt{SS_{xx}}} = \frac{61.7052}{\sqrt{1913.129}} = 1.411$$

and

$$t = \frac{b_1}{s_{b_1}} = \frac{15.596}{1.411} = 11.05$$

FIGURE 14.8 **Excel and MINITAB Outputs of a Simple Linear Regression Analysis of the Tasty Sub Shop Revenue Data**

(a) The Excel Output

Regression Statistics

Multiple R	0.9688
R Square	0.9386 [9]
Adjusted R Square	0.9309
Standard Error	61.7052 [8]
Observations	10

ANOVA	df	SS	MS	F	Significance F
Regression	1	465316.3004 [10]	465316.3004	122.2096 [13]	0.0000 [14]
Residual	8	30460.2086 [11]	3807.5261		
Total	9	495776.5090 [12]			

	Coefficients	Standard Error	t Stat	P-value [7]	Lower 95%	Upper 95%
Intercept	183.3051 [1]	64.2741 [3]	2.8519 [5]	0.0214	35.0888	331.5214
Population	15.5956 [2]	1.4107 [4]	11.0548 [6]	0.0000	12.3424 [19]	18.8488 [19]

(b) The MINITAB Output

```
The regression equation is
Revenue = 183 + 15.6 Population

Predictor            Coef          SE Coef              T                P [7]
Constant          183.31 [1]       64.27 [3]        2.85 [5]            0.021
Population        15.596 [2]       1.411 [4]       11.05 [6]            0.000

S = 61.7052 [8]                R-Sq = 93.9% [9]        R-Sq(adj) = 93.1%

Analysis of Variance
Source                DF          SS          MS             F          P-value
Regression             1      465316 [10]    465316      122.21 [13]     0.000 [14]
Residual Error         8       30460 [11]      3808
Total                  9      495777 [12]

 Predicted Values for New Observations
     New Obs          Fit [15]     SE Fit [16]          95% CI [17]          95% PI [18]
           1         921.0           20.3          (874.2, 967.7)       (771.2, 1070.7)

Values of Predictors for New Observations
     New Obs    Population
           1          47.3
```

[1] b_0 = point estimate of the y-intercept [2] b_1 = point estimate of the slope [3] s_{b_0} = standard error of the estimate b_0 [4] s_{b_1} = standard error of the estimate b_1 [5] t for testing significance of the y-intercept [6] t for testing significance of the slope [7] p-values for t statistics [8] s = standard error [9] r^2 [10] Explained variation [11] SSE = Unexplained variation [12] Total variation [13] F(model) statistic [14] p-value for F(model) [15] $\hat{y}$ = point prediction when $x = 47.3$ [16] $s_{\hat{y}}$ = standard error of the estimate $\hat{y}$ [17] 95% confidence interval when $x = 47.3$ [18] 95% prediction interval when $x = 47.3$ [19] 95% confidence interval for the slope β_1

Figure 14.8 presents the Excel and MINITAB outputs of a simple linear regression analysis of the Tasty Sub Shop revenue data. Note that b_0 (labeled as [1] on the outputs), b_1 (labeled [2]), s (labeled [8]), s_{b_1} (labeled [4]), and t (labeled [6]) are given on each of these outputs. (The other quantities on the outputs will be discussed later.) In order to test $H_0: \beta_1 = 0$ versus $H_a: \beta_1 \neq 0$ at the $\alpha = .05$ level of significance, we compare $|t| = 11.05$ with $t_{\alpha/2} = t_{.025} = 2.306$, which is based on $n - 2 = 10 - 2 = 8$ degrees of freedom. Because $|t| = 11.05$ is greater than $t_{.025} = 2.306$, we reject $H_0: \beta_1 = 0$ and conclude that there is strong evidence that the slope (regression relationship) is significant. The p-value for testing H_0 versus H_a is twice the area to the right of $|t| = 11.05$ under the curve of the t distribution having $n - 2 = 8$ degrees of freedom. Both the Excel and MINITAB outputs in Figure 14.8 tell us that this p-value is less than .001 (see [7] on the outputs). It follows that we can reject H_0 in favor of H_a at level of significance .05, .01, or .001, which implies that we have extremely strong evidence that the regression relationship between x and y is significant.

A Confidence Interval for the Slope

If the regression assumptions hold, a **100(1 − α) percent confidence interval for the true slope β_1** is $[b_1 \pm t_{\alpha/2} s_{b_1}]$. Here $t_{\alpha/2}$ is based on $n - 2$ degrees of freedom.

EXAMPLE 14.5 The Tasty Sub Shop Case: A Confidence Interval for the Slope

The Excel and MINITAB outputs in Figure 14.8 tell us that $b_1 = 15.596$ and $s_{b_1} = 1.411$. Thus, for instance, because $t_{.025}$ based on $n - 2 = 10 - 2 = 8$ degrees of freedom equals 2.306, a 95 percent confidence interval for β_1 is

$$[b_1 \pm t_{.025} s_{b_1}] = [15.596 \pm 2.306(1.411)]$$
$$= [12.342, 18.849]$$

(where we have used more decimal place accuracy than shown to obtain the final result). This interval says we are 95 percent confident that, if the population size increases by one thousand residents, then mean yearly revenue will increase by at least \$12,342 and by at most \$18,849. Also, because the 95 percent confidence interval for β_1 does not contain 0, we can reject H_0: $\beta_1 = 0$ in favor of H_a: $\beta_1 \neq 0$ at level of significance .05. Note that the 95 percent confidence interval for β_1 is given on the Excel output but not on the MINITAB output (see Figure 14.8).

Testing the significance of the *y*-intercept We can also test the significance of the *y*-intercept β_0. We do this by testing the null hypothesis H_0: $\beta_0 = 0$ versus the alternative hypothesis H_a: $\beta_0 \neq 0$. **If we can reject H_0 in favor of H_a by setting the probability of a Type I error equal to α, we conclude that the intercept β_0 is significant at the α level.** To carry out the hypothesis test, we use the test statistic

$$t = \frac{b_0}{s_{b_0}} \quad \text{where} \quad s_{b_0} = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}}}$$

Here the critical value and *p*-value conditions for rejecting H_0 are the same as those given previously for testing the significance of the slope, except that t is calculated as b_0/s_{b_0}. For example, if we consider the Tasty Sub Shop problem and the Excel and MINITAB outputs in Figure 14.8, we see that $b_0 = 183.31$, $s_{b_0} = 64.27$, $t = 2.85$, and *p*-value = .021. Because $t = 2.85 > t_{.025} = 2.306$ and *p*-value $< .05$, we can reject H_0: $\beta_0 = 0$ in favor of H_a: $\beta_0 \neq 0$ at the .05 level of significance. This provides strong evidence that the *y*-intercept β_0 of the line of means does not equal 0 and thus is significant. Therefore, we should include β_0 in the Tasty Sub Shop revenue model.

In general, if we fail to conclude that the intercept is significant at a level of significance of .05, it might be reasonable to drop the *y*-intercept from the model. However, it is common practice to include the *y*-intercept whether or not H_0: $\beta_0 = 0$ is rejected. In fact, experience suggests that it is definitely safest, when in doubt, to include the intercept β_0.

Exercises for Section 14.3

CONCEPTS

connect

14.18 What do we conclude if we can reject H_0: $\beta_1 = 0$ in favor of H_a: $\beta_1 \neq 0$ by setting
a α equal to .05? **b** α equal to .01?

14.19 Give an example of a practical application of the confidence interval for β_1.

METHODS AND APPLICATIONS

In Exercises 14.20 through 14.25, we refer to Excel and MINITAB outputs of simple linear regression analyses of the data sets related to the six case studies introduced in the exercises for Section 14.1. Using the appropriate output for each case study,

a Find the least squares point estimates b_0 and b_1 of β_0 and β_1 on the output and report their values.
b Find *SSE* and s on the computer output and report their values.

c Find s_{b_1} and the t statistic for testing the significance of the slope on the output and report their values. Show (within rounding) how t has been calculated by using b_1 and s_{b_1} from the computer output.

d Using the t statistic and an appropriate critical value, test H_0: $\beta_1 = 0$ versus H_a: $\beta_1 \neq 0$ by setting α equal to .05. Is the slope (regression relationship) significant at the .05 level?

e Using the t statistic and an appropriate critical value, test H_0: $\beta_1 = 0$ versus H_a: $\beta_1 \neq 0$ by setting α equal to .01. Is the slope (regression relationship) significant at the .01 level?

f Find the p-value for testing H_0: $\beta_1 = 0$ versus H_a: $\beta_1 \neq 0$ on the output and report its value. Using the p-value, determine whether we can reject H_0 by setting α equal to .10, .05, .01, and .001. How much evidence is there that the slope (regression relationship) is significant?

g Calculate the 95 percent confidence interval for β_1 using numbers on the output. Interpret the interval.

h Calculate the 99 percent confidence interval for β_1 using numbers on the output.

i Find s_{b_0} and the t statistic for testing the significance of the y intercept on the output and report their values. Show (within rounding) how t has been calculated by using b_0 and s_{b_0} from the computer output.

j Find the p-value for testing H_0: $\beta_0 = 0$ versus H_a: $\beta_0 \neq 0$ on the computer output and report its value. Using the p-value, determine whether we can reject H_0 by setting α equal to .10, .05, .01, and .001. What do you conclude about the significance of the y intercept?

k Using the data set and s from the computer output, hand calculate (within rounding) SS_{xx}, s_{b_0}, and s_{b_1}.

14.20 THE NATURAL GAS CONSUMPTION CASE GasCon1

The Excel and MINITAB outputs of a simple linear regression analysis of the data set for this case (see Exercise 14.3 on pages 496 and 497) are given in Figures 14.9 and 14.10. Recall that labeled Excel and MINITAB outputs are on page 505. Note that gas consumption is called FuelCons on the MINITAB output of Figure 14.10.

14.21 THE STARTING SALARY CASE StartSal

The MINITAB output of a simple linear regression analysis of the data set for this case (see Exercise 14.4 on page 497) is given in Figure 14.11. Recall that a labeled MINITAB regression output is on page 505.

FIGURE 14.9 Excel Output of a Simple Linear Regression Analysis of the Natural Gas Consumption Data

Regression Statistics	
Multiple R	0.9484
R Square	0.8995
Adjusted R Square	0.8827
Standard Error	0.6542
Observations	8

ANOVA	df	SS	MS	F	Significance F
Regression	1	22.9808	22.9808	53.6949	0.0003
Residual	6	2.5679	0.4280		
Total	7	25.5488			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	15.8379	0.8018	19.7535	1.09E-06	13.8760	17.7997
TEMP	–0.1279	0.0175	–7.3277	0.0003	–0.1706	–0.0852

FIGURE 14.10 MINITAB Output of a Simple Linear Regression Analysis of the Natural Gas Consumption Data

```
The regression equation is
FuelCons = 15.8 - 0.128 Temp

Predictor            Coef        SE Coef            T           P
Constant          15.8379         0.8018        19.75       0.000
Temp             -0.12792        0.01746        -7.33       0.000

S = 0.654209         R-Sq = 89.9%         R-Sq(adj) = 88.3%

Analysis of Variance
Source               DF            SS           MS            F             P
Regression            1        22.981       22.981        53.69         0.000
Residual Error        6         2.568        0.428
Total                 7        25.549

Values of Predictors for New Obs   Predicted Values for New Observations
New Obs  Temp                      New Obs     Fit   SE Fit       95% CI           95% PI
      1  40.0                            1  10.721    0.241  (10.130, 11.312)  (9.015, 12.427)
```

FIGURE 14.11 MINITAB Output of a Simple Linear Regression Analysis of the Starting Salary Data

```
The regression equation is
StartSal = 14.8 + 5.71 GPA

Predictor      Coef  SE Coef      T      P
Constant     14.816    1.235  12.00  0.000
GPA          5.7066   0.3953  14.44  0.000

S = 0.536321   R-Sq = 97.7%   R-Sq(adj) = 97.2%

Analysis of Variance
Source           DF      SS      MS       F      P
Regression        1  59.942  59.942  208.39  0.000
Residual Error    5   1.438   0.288
Total             6  61.380

Values of Predictors for New Obs    Predicted Values for New Observations
New Obs   GPA                       New Obs     Fit   SE Fit       95% CI            95% PI
      1  3.25                             1  33.362    0.213  (32.813, 33.911)  (31.878, 34.846)
```

FIGURE 14.12 Excel Output of a Simple Linear Regression Analysis of the Service Time Data

Regression Statistics	
Multiple R	0.9952
R Square	0.9905
Adjusted R Square	0.9894
Standard Error	4.6152
Observations	11

ANOVA	df	SS	MS	F	Significance F
Regression	1	19918.8438	19918.844	935.149	2.094E-10
Residual	9	191.7017	21.300184		
Total	10	20110.5455			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	11.4641	3.4390	3.3335	0.0087	3.6845	19.2437
Copiers	24.6022	0.8045	30.5802	2.09E-10	22.7823	26.4221

14.22 THE SERVICE TIME CASE DS SrvcTime

The Excel output of a simple linear regression analysis of the data set for this case (see Exercise 14.5 on pages 497 and 498) is given in Figure 14.12. Recall that a labeled Excel regression output is on page 505.

14.23 THE FRESH DETERGENT CASE DS Fresh

The MINITAB output of a simple linear regression analysis of the data set for this case (see Exercise 14.6 on page 498) is given in Figure 14.13. Recall that a labeled MINITAB regression output is on page 505.

14.24 THE DIRECT LABOR COST CASE DS DirLab

The Excel output of a simple linear regression analysis of the data set for this case (see Exercise 14.7 on page 499) is given in Figure 14.14. Recall that a labeled Excel regression output is on page 505.

14.25 THE REAL ESTATE SALES PRICE CASE DS RealEst

The MINITAB output of a simple linear regression analysis of the data set for this case (see Exercise 14.8 on pages 499 and 500) is given in Figure 14.15. Recall that a labeled MINITAB regression output is on page 505.

14.26 Find and interpret a 95 percent confidence interval for the slope β_1 of the simple linear regression model describing the sales volume data in Exercise 14.17 (page 503). DS SalesPlot

FIGURE 14.13 MINITAB Output of a Simple Linear Regression Analysis of the Fresh Detergent Demand Data

```
The regression equation is
Demand = 7.81 + 2.67 PriceDif

Predictor      Coef  SE Coef      T      P
Constant    7.81409  0.07988  97.82  0.000
PriceDif     2.6652   0.2585  10.31  0.000

S = 0.316561   R-Sq = 79.2%   R-Sq(adj) = 78.4%

Analysis of Variance
Source           DF      SS      MS       F      P
Regression        1  10.653  10.653  106.30  0.000
Residual Error   28   2.806   0.100
Total            29  13.459

Values of Predictors for New Obs     Predicted Values for New Observations
New Obs  PriceDif                   New Obs     Fit  SE Fit        95% CI             95% PI
      1     0.100                         1  8.0806  0.0648  (7.9479, 8.2133)  (7.4187, 8.7425)
      2     0.250                         2  8.4804  0.0586  (8.3604, 8.6004)  (7.8209, 9.1398)
```

FIGURE 14.14 Excel Output of a Simple Linear Regression Analysis of the Direct Labor Cost Data

Regression Statistics	
Multiple R	0.9996
R Square	0.9993
Adjusted R Square	0.9992
Standard Error	8.6415
Observations	12

ANOVA	df	SS	MS	F	Significance F
Regression	1	1024592.9043	1024592.9043	13720.4677	5.04E-17
Residual	10	746.7624	74.6762		
Total	11	1025339.6667			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	18.4875	4.6766	3.9532	0.0027	8.0674	28.9076
BatchSize (x)	10.1463	0.0866	117.1344	5.04E-17	9.9533	10.3393

FIGURE 14.15 MINITAB Output of a Simple Linear Regression Analysis of the Real Estate Sales Price Data

```
The regression equation is
SPrice = 48.0 + 5.70 HomeSize

Predictor     Coef  SE Coef     T      P
Constant     48.02    14.41  3.33  0.010
HomeSize    5.7003   0.7457  7.64  0.000

S = 10.5880      R-Sq = 88.0%      R-Sq(adj) = 86.5%

Analysis of Variance
Source           DF      SS      MS      F      P
Regression        1  6550.7  6550.7  58.43  0.000
Residual Error    8   896.8   112.1
Total             9  7447.5

Values of Predictors for New Obs    Predicted Values for New Observations
New Obs  HomeSize                   New Obs     Fit  SE Fit        95% CI             95% PI
      1      20.0                         1  162.03    3.47  (154.04, 170.02)  (136.34, 187.72)
```

FIGURE 14.16 Excel Output of a Simple Linear Regression Analysis of the Fast-Food Restaurant Rating Data

Regression Statistics						
Multiple R	0.9873					
R Square	0.9747					
Adjusted R Square	0.9684					
Standard Error	0.1833					
Observations	6					
ANOVA	**df**	**SS**	**MS**	**F**	**Significance F**	
Regression	1	5.1817	5.1817	154.2792	0.0002	
Residual	4	0.1343	0.0336			
Total	5	5.3160				
	Coefficients	**Standard Error**	**t Stat**	**P-value**	**Lower 95%**	**Upper 95%**
Intercept	−0.1602	0.3029	−0.5289	0.6248	−1.0011	0.6807
MeanTaste (x)	1.2731	0.1025	12.4209	0.0002	0.9885	1.5577

14.27 THE FAST-FOOD RESTAURANT RATING CASE FastFood

In the early 1990s researchers at The Ohio State University studied consumer ratings of six fast-food restaurants: Borden Burger, Hardee's, Burger King, McDonald's, Wendy's, and White Castle. Each of 406 randomly selected individuals gave each restaurant a rating of 1, 2, 3, 4, 5, or 6 on the basis of taste, and then ranked the restaurants from 1 through 6 on the basis of overall preference. In each case, 1 is the best rating and 6 the worst. The mean ratings given by the 406 individuals are given in the following table:

Restaurant	Mean Taste	Mean Preference
Borden Burger	3.5659	4.2552
Hardee's	3.329	4.0911
Burger King	2.4231	3.0052
McDonald's	2.0895	2.2429
Wendy's	1.9661	2.5351
White Castle	3.8061	4.7812

Figure 14.16 gives the Excel output of a simple linear regression analysis of this data. Here, mean preference is the dependent variable and mean taste is the independent variable. Recall that a labeled Excel regression output is given on page 505.

a Find the least squares point estimate b_1 of β_1 on the computer output. Report and interpret this estimate.

b Find the 95 percent confidence interval for β_1 on the output. Report and interpret the interval.

LO14-5 Calculate and interpret a confidence interval for a mean value and a prediction interval for an individual value.

14.4 Confidence and Prediction Intervals

If the regression relationship between y and x is significant, then

$$\hat{y} = b_0 + b_1 x_0$$

is the **point estimate of the mean value of *y*** when the value of the independent variable x is x_0. We have also seen that $\hat{y}$ is the **point prediction of an individual value of *y*** when the value of the independent variable x is x_0. In this section we will assess the accuracy of $\hat{y}$ as both a point estimate and a point prediction. To do this, we will find a **confidence interval for the mean value of *y*** and a **prediction interval for an individual value of *y*.**

Because each possible sample of n values of the dependent variable gives values of b_0 and b_1 that differ from the values given by other samples, different samples give different values of

$\hat{y} = b_0 + b_1 x_0$. If the regression assumptions hold, a confidence interval for the mean value of y is based on the estimated standard deviation of the normally distributed population of all possible values of $\hat{y}$. This estimated standard deviation is called the **standard error of $\hat{y}$,** is denoted $s_{\hat{y}}$, and is given by the formula

$$s_{\hat{y}} = s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}}$$

Here, s is the standard error (see Section 14.2), $\bar{x}$ is the average of the n previously observed values of x, and $SS_{xx} = \Sigma x_i^2 - (\Sigma x_i)^2/n$.

As explained above, a confidence interval for the mean value of y is based on the standard error $s_{\hat{y}}$. A prediction interval for an individual value of y is based on a more complex standard error: the estimated standard deviation of the population of all possible values of $y - \hat{y}$, the prediction error obtained when predicting y by $\hat{y}$. We refer to this estimated standard deviation as the **standard error of $y - \hat{y}$** and denote it as $s_{(y-\hat{y})}$. If the regression assumptions hold, the formula for $s_{(y-\hat{y})}$ is

$$s_{(y-\hat{y})} = s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}}$$

Intuitively, the "extra 1" under the radical in the formula for $s_{(y-\hat{y})}$ accounts for the fact that there is *more uncertainty* in predicting an individual value $y = \beta_0 + \beta_1 x_0 + \varepsilon$ than in estimating the mean value $\beta_0 + \beta_1 x_0$ (because we must predict the error term ε when predicting an individual value). Therefore, as shown in the following summary box, the prediction interval for an individual value of y is longer than the confidence interval for the mean value of y.

A Confidence Interval and a Prediction Interval

If the regression assumptions hold,

1 A **100(1 − α) percent confidence interval for the mean value of y** when x equals x_0 is

$$\left[\hat{y} \pm t_{\alpha/2} s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}}\right]$$

2 A **100(1 − α) percent prediction interval for an individual value of y** when x equals x_0 is

$$\left[\hat{y} \pm t_{\alpha/2} s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}}\right]$$

Here, $t_{\alpha/2}$ is based on $(n - 2)$ degrees of freedom.

The summary box tells us that both the formula for the confidence interval and the formula for the prediction interval use the quantity $1/n + (x_0 - \bar{x})^2/SS_{xx}$. We will call this quantity the **distance value,** because it is a measure of the distance between x_0, the value of x for which we will make a point estimate or a point prediction, and $\bar{x}$, the average of the previously observed values of x. The farther that x_0 is from $\bar{x}$, which represents the center of the experimental region, the larger is the distance value, and thus the longer are both the confidence interval $[\hat{y} \pm t_{\alpha/2} s\sqrt{\text{distance value}}]$ and the prediction interval $[\hat{y} \pm t_{\alpha/2} s\sqrt{1 + \text{distance value}}]$. Said another way, when x_0 is farther from the center of the data, $\hat{y} = b_0 + b_1 x_0$ is likely to be less accurate as both a point estimate and a point prediction.

EXAMPLE 14.6 The Tasty Sub Shop Case: Predicting Revenue and Profit

In the Tasty Sub Shop problem, recall that one of the business entrepreneur's potential sites is near a population of 47,300 residents. Also, recall that

$$\begin{aligned}\hat{y} &= b_0 + b_1 x_0 \\ &= 183.31 + 15.596(47.3) \\ &= 921.0 \text{ (that is, \$921,000)}\end{aligned}$$

is the point estimate of the mean yearly revenue for all Tasty Sub restaurants that could potentially be built near populations of 47,300 residents and is the point prediction of the yearly revenue for a single Tasty Sub restaurant that is built near a population of 47,300 residents. Using the information in Example 14.2 (page 493), we compute

$$\begin{aligned}\text{distance value} &= \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}} \\ &= \frac{1}{10} + \frac{(47.3 - 43.41)^2}{1913.129} \\ &= .1079\end{aligned}$$

Because $s = 61.7052$ (see Example 14.3 on page 502) and because $t_{\alpha/2} = t_{.025}$ based on $n - 2 = 10 - 2 = 8$ degrees of freedom equals 2.306, it follows that a 95 percent confidence interval for the mean yearly revenue when $x = 47.3$ is

$$\begin{aligned}&[\hat{y} \pm t_{\alpha/2} s \sqrt{\text{distance value}}] \\ &= [921.0 \pm 2.306(61.7052)\sqrt{.1079}] \\ &= [921.0 \pm 46.74] \\ &= [874.3, 967.7]\end{aligned}$$

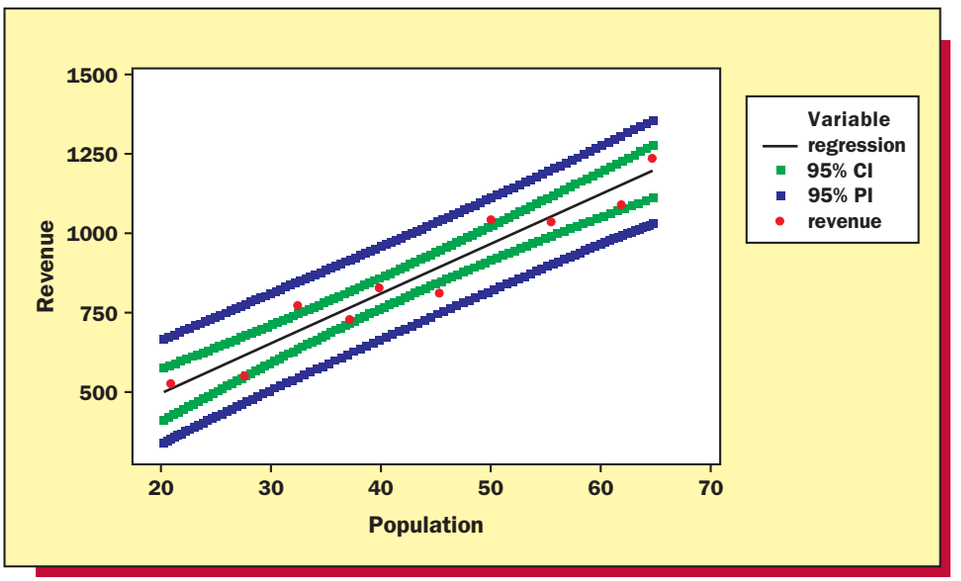
This interval says we are 95 percent confident that the mean yearly revenue for all Tasty Sub restaurants that could potentially be built near populations of 47,300 residents is between \$874,300 and \$967,700.

Because the entrepreneur would be operating a single Tasty Sub restaurant that is built near a population of 47,300 residents, the entrepreneur is interested in obtaining a prediction interval for the yearly revenue of such a restaurant. A 95 percent prediction interval for this revenue is

$$\begin{aligned}&[\hat{y} \pm t_{\alpha/2} s \sqrt{1 + \text{distance value}}] \\ &= [921.0 \pm 2.306(61.7052)\sqrt{1.1079}] \\ &= [921.0 \pm 149.77] \\ &= [771.2, 1070.8]\end{aligned}$$

This interval says that we are 95 percent confident that the yearly revenue for a single Tasty Sub restaurant that is built near a population of 47,300 residents will be between \$771,200 and \$1,070,800. Moreover, recall that the yearly rent and other fixed costs for the entrepreneur's potential restaurant will be \$257,550 and that (according to Tasty Sub corporate headquarters) the yearly food and other variable costs for the restaurant will be 60 percent of the yearly revenue. Using the lower end of the 95 percent prediction interval [771.2, 1070.8], we predict that (1) the restaurant's yearly operating cost will be \$257,550 + .6(\$771,200) = \$720,270 and (2) the restaurant's yearly profit will be \$771,200 − \$720,270 = \$50,930. Using the upper end of the 95 percent prediction interval [771.2, 1070.8], we predict that (1) the restaurant's yearly

FIGURE 14.17 MINITAB Output of 95% Confidence and Prediction Intervals for the Tasty Sub Shop Case



operating cost will be \$257,550 + .6(\$1,070,800) = \$900,030 and (2) the restaurant's yearly profit will be \$1,070,800 − \$900,030 = \$170,770. Combining the two predicted profits, it follows that we are 95 percent confident that the potential restaurant's yearly profit will be between \$50,930 and \$170,770. If the entrepreneur decides that this is an acceptable range of potential yearly profits, then the entrepreneur might decide to purchase a Tasty Sub franchise for the potential restaurant site. In Chapter 15 we will use a *multiple regression model* to reduce the range of the predicted yearly profits for the potential Tasty Sub restaurant.

Below we repeat the bottom of the MINITAB output in Figure 14.8(b) on page 505. This output gives (within rounding) the point estimate and prediction $\hat{y} = 921.0$, the 95 percent confidence interval for the mean value of y when x equals 47.3, and the 95 percent prediction interval for an individual value of y when x equals 47.3.

```
Predicted Values for New Observations
New Obs      Fit  SE Fit      95% CI          95% PI
      1    921.0    20.3  (874.2, 967.7)  (771.2, 1070.7)
```

Although the MINITAB output does not directly give the distance value, it does give $s_{\hat{y}} = s\sqrt{\text{distance value}}$ under the heading "SE Fit." A little algebra shows that this implies that the distance value equals $(s_{\hat{y}}/s)^2$. Specifically, because $s_{\hat{y}} = 20.3$ and $s = 61.7052$, the distance value equals $(20.3/61.7052)^2 = .1082$. Note that, because MINITAB rounds $s_{\hat{y}}$, this calculation of the distance value is slightly less accurate than the previous hand calculation that obtained a distance value of .1079.

To conclude this example, note that Figure 14.17 illustrates the MINITAB output of the 95 percent confidence and prediction intervals corresponding to all values of x in the experimental region. Here $\bar{x} = 43.41$ can be regarded as the center of the experimental region. Notice that the farther x_0 is from $\bar{x} = 43.41$, the larger is the distance value and, therefore, the longer are the 95 percent confidence and prediction intervals. These longer intervals are undesirable because they give us less information about mean and individual values of y.

Exercises for Section 14.4

CONCEPTS

connect

14.28 What is the difference between a confidence interval and a prediction interval?

14.29 What does the distance value measure? How does the distance value affect a confidence or prediction interval?

METHODS AND APPLICATIONS

14.30 THE NATURAL GAS CONSUMPTION CASE GasCon1

The following partial MINITAB regression output for the natural gas consumption data relates to predicting the city's natural gas consumption (in MMcf) in a week that has an average hourly temperature of 40°F.

```
Predicted Values for New Observations
New Obs      Fit    SE Fit        95% CI            95% PI
      1   10.721     0.241  (10.130, 11.312)  (9.015, 12.427)
```

a Report (as shown on the computer output) a point estimate of and a 95 percent confidence interval for the mean natural gas consumption for all weeks having an average hourly temperature of 40°F.

b Report (as shown on the computer output) a point prediction of and a 95 percent prediction interval for the natural gas consumption in a single week that has an average hourly temperature of 40°F.

c Remembering that $s = .6542$; $SS_{xx} = 1{,}404.355$; $\bar{x} = 43.98$; and $n = 8$, hand calculate the distance value when $x_0 = 40$. Remembering that the distance value equals $(s_{\hat{y}}/s)^2$, use s and $s_{\hat{y}}$ from the computer output to calculate (within rounding) the distance value using this formula. Note that, because MINITAB rounds $s_{\hat{y}}$, the first hand calculation is the more accurate calculation of the distance value.

d Remembering that for the natural gas consumption data $b_0 = 15.84$ and $b_1 = -.1279$, calculate (within rounding) the confidence interval of part *a* and the prediction interval of part *b*.

e Suppose that next week the city's average hourly temperature will be 40°F. Also, suppose that the city's natural gas company will use the point prediction $\hat{y} = 10.721$ and order 10.721 MMcf of natural gas to be shipped to the city by a pipeline transmission system. The company will have to pay a fine to the transmission system if the city's actual gas usage y differs from the order of 10.721 MMcf by more than 10.5 percent—that is, is outside of the range $[10.721 \pm .105(10.721)] = [9.595, 11.847]$. Discuss why the 95 percent prediction interval for y, [9.015, 12.427], says that y might be outside of the allowable range and thus does not make the company 95 percent confident that it will avoid paying a fine.
Note: In the exercises of Chapter 15, we will use multiple regression analysis to predict y accurately enough so that the company is likely to avoid paying a fine.

14.31 THE STARTING SALARY CASE StartSal

The following partial MINITAB regression output for the starting salary data relates to predicting the starting salary of a marketing graduate having a grade point average of 3.25.

```
Predicted Values for New Observations
New Obs      Fit    SE Fit        95% CI            95% PI
      1   33.362     0.213  (32.813, 33.911)  (31.878, 34.846)
```

a Report (as shown on the computer output) a point estimate of and a 95 percent confidence interval for the mean starting salary of all marketing graduates having a grade point average of 3.25.

b Report (as shown on the computer output) a point prediction of and a 95 percent prediction interval for the starting salary of an individual marketing graduate having a grade point average of 3.25.

c Remembering that $s = .536321$ and that the distance value equals $(s_{\hat{y}}/s)^2$, use $s_{\hat{y}}$ from the computer output to hand calculate the distance value when $x = 3.25$.

d Remembering that for the starting salary data $n = 7$, $b_0 = 14.816$, and $b_1 = 5.7066$, hand calculate (within rounding) the confidence interval of part *a* and the prediction interval of part *b*.

14.32 THE SERVICE TIME CASE SrvcTime

The following partial Excel add-in (MegaStat) regression output for the service time data relates to predicting service times for 1, 2, 3, 4, 5, 6, and 7 copiers.

Predicted values for: Minutes (y)

		95% Confidence Intervals		95% Prediction Intervals		
Copiers (x)	Predicted	lower	upper	lower	upper	Leverage
1	36.066	29.907	42.226	23.944	48.188	0.348
2	60.669	55.980	65.357	49.224	72.113	0.202
3	85.271	81.715	88.827	74.241	96.300	0.116
4	109.873	106.721	113.025	98.967	120.779	0.091
5	134.475	130.753	138.197	123.391	145.559	0.127
6	159.077	154.139	164.016	147.528	170.627	0.224
7	183.680	177.233	190.126	171.410	195.950	0.381

a Report (as shown on the computer output) a point estimate of and a 95 percent confidence interval for the mean time to service four copiers.

b Report (as shown on the computer output) a point prediction of and a 95 percent prediction interval for the time to service four copiers on a single call.

c For this case: $n = 11$, $b_0 = 11.4641$, $b_1 = 24.6022$, and $s = 4.615$. Using this information and a distance value (called **Leverage** on the add-in output), hand calculate (within rounding) the confidence interval of part *a* and the prediction interval of part *b*.

d If we examine the service time data, we see that there was at least one call on which Accu-Copiers serviced each of 1, 2, 3, 4, 5, 6, and 7 copiers. The 95 percent confidence intervals for the mean service times on these calls might be used to schedule future service calls. To understand this, note that a person making service calls will (in, say, a year or more) make a very large number of service calls. Some of the person's individual service times will be below, and some will be above, the corresponding mean service times. However, because the very large number of individual service times will average out to the mean service times, it seems fair to both the efficiency of the company and to the person making service calls to schedule service calls by using estimates of the mean service times. Therefore, suppose we wish to schedule a call to service five copiers. Examining the computer output, we see that a 95 percent confidence interval for the mean time to service five copiers is [130.753, 138.197]. Because the mean time might be 138.197 minutes, it would seem fair to allow 138 minutes to make the service call. Now suppose we wish to schedule a call to service four copiers. Determine how many minutes to allow for the service call.

14.33 THE FRESH DETERGENT CASE Fresh

The following partial MINITAB regression output for the Fresh detergent data relates to predicting demand for future sales periods in which the price difference will be .10 (see New Obs 1) and .25 (see New Obs 2).

```
Predicted Values for New Observations
New Obs      Fit  SE Fit        95% CI              95% PI
      1   8.0806  0.0648  (7.9479, 8.2133)  (7.4187, 8.7425)
      2   8.4804  0.0586  (8.3604, 8.6004)  (7.8209, 9.1398)
```

a Report (as shown on the computer output) a point estimate of and a 95 percent confidence interval for the mean demand for Fresh in all sales periods when the price difference is .10.

b Report (as shown on the computer output) a point prediction of and a 95 percent prediction interval for the actual demand for Fresh in an individual sales period when the price difference is .10.

c Remembering that $s = .316561$ and that the distance value equals $(s_{\hat{y}}/s)^2$, use $s_{\hat{y}}$ from the computer output to hand calculate the distance value when $x = .10$.

d For this case: $n = 30$, $b_0 = 7.81409$, $b_1 = 2.6652$, and $s = .316561$. Using this information, and your result from part *c*, find 99 percent confidence and prediction intervals for mean and individual demands when $x = .10$.

e Repeat parts *a*, *b*, *c*, and *d* when $x = .25$.

14.34 THE DIRECT LABOR COST CASE DirLab

The following partial Excel add-in (MegaStat) regression output for the direct labor cost data relates to predicting direct labor cost when the batch size is 60.

Predicted values for: LaborCost (y)

		95% Confidence Interval		95% Prediction Interval		
BatchSize (x)	Predicted	lower	upper	lower	upper	Leverage
60	627.263	621.054	633.472	607.032	647.494	0.104

a Report (as shown on the computer output) a point estimate of and a 95 percent confidence interval for the mean direct labor cost of all batches of size 60.

b Report (as shown on the computer output) a point prediction of and a 95 percent prediction interval for the actual direct labor cost of an individual batch of size 60.

c For this case: $n = 12$, $b_0 = 18.4875$, $b_1 = 10.1463$, and $s = 8.6415$. Use this information and the distance value (called **Leverage**) on the computer output to compute 99 percent confidence and prediction intervals for the mean and individual labor costs when $x = 60$.

14.35 THE REAL ESTATE SALES PRICE CASE RealEst

The following partial MINITAB regression output for the real estate sales price data relates to predicting the sales price of a home having 2,000 square feet.

```
Predicted Values for New Observations
New Obs      Fit    SE Fit        95% CI              95% PI
      1   162.03      3.47  (154.04, 170.02)  (136.34, 187.72)
```

a Report (as shown on the MINITAB output) a point estimate of and a 95 percent confidence interval for the mean sales price of all houses having 2,000 square feet.

b Report (as shown on the MINITAB output) a point prediction of and a 95 percent prediction interval for the sales price of an individual house having 2,000 square feet.

c If you were purchasing a home having 2,000 square feet, which of the above intervals would you find to be most useful? Explain.

LO14-6 Calculate and interpret the simple coefficients of determination and correlation.

14.5 Simple Coefficients of Determination and Correlation

The simple coefficient of determination The **simple coefficient of determination** is a measure of the usefulness of a simple linear regression model. To introduce this quantity, which is denoted r^2 (pronounced ***r* squared**), suppose we have observed n values of the dependent variable y. However, we choose to predict y without using a predictor (independent) variable x. In such a case the only reasonable prediction of a specific value of y, say y_i, would be $\bar{y}$, which is simply the average of the n observed values $y_1, y_2, \ldots, y_n$. Here the error of prediction in predicting y_i would be $y_i - \bar{y}$. For example, Figure 14.18(a) illustrates the prediction errors obtained for the Tasty Sub Shop revenue data when we do not use the information provided by the independent variable x, population size.

Next, suppose we decide to employ the predictor variable x and observe the values $x_1, x_2, \ldots, x_n$ corresponding to the observed values of y. In this case the prediction of y_i is

$$\hat{y}_i = b_0 + b_1 x_i$$

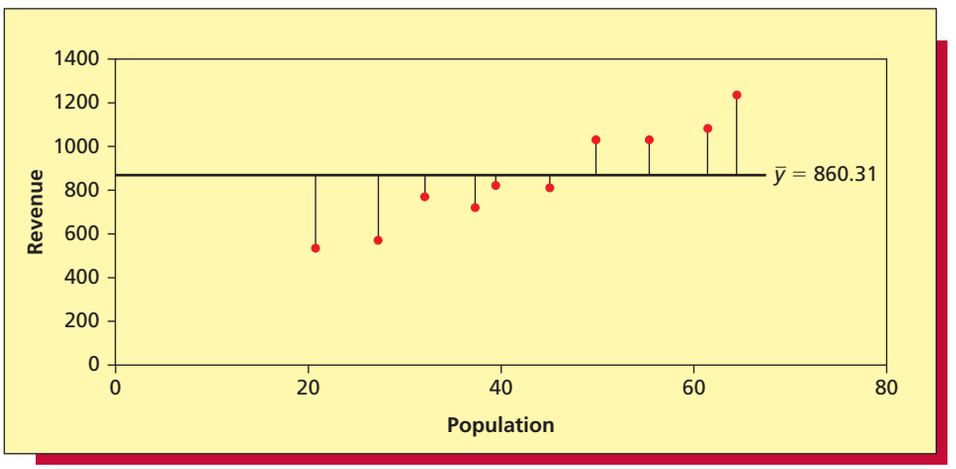
and the error of prediction is $y_i - \hat{y}_i$. For example, Figure 14.18(b) illustrates the prediction errors obtained in the Tasty Sub Shop case when we use the predictor variable x. Together, Figures 14.18(a) and (b) show the reduction in the prediction errors accomplished by employing the predictor variable x (and the least squares line).

Using the predictor variable x decreases the prediction error in predicting y_i from $(y_i - \bar{y})$ to $(y_i - \hat{y}_i)$, or by an amount equal to

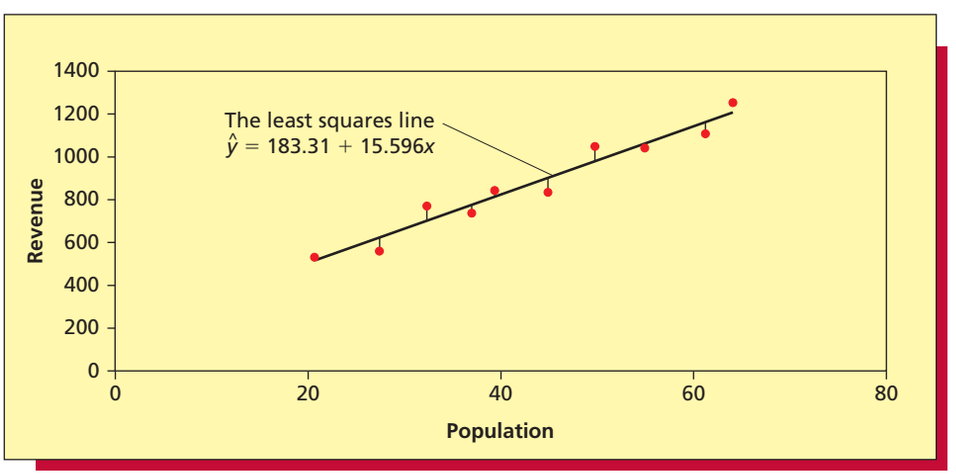
$$(y_i - \bar{y}) - (y_i - \hat{y}_i) = (\hat{y}_i - \bar{y})$$

FIGURE 14.18 The Reduction in the Prediction Errors Accomplished by Employing the Predictor Variable *x*

(a) Prediction errors for the Tasty Sub Shop case when we do not use the information contributed by *x*



(b) Prediction errors for the Tasty Sub Shop case when we use the information contributed by *x* by using the least squares line



It can be shown that in general

$$\sum (y_i - \bar{y})^2 - \sum (y_i - \hat{y}_i)^2 = \sum (\hat{y}_i - \bar{y})^2$$

The sum of squared prediction errors obtained when we do not employ the predictor variable x, $\Sigma(y_i - \bar{y})^2$, is called the **total variation.** Intuitively, this quantity measures the total amount of variation exhibited by the observed values of y. The sum of squared prediction errors obtained when we use the predictor variable x, $\Sigma(y_i - \hat{y}_i)^2$, is called the **unexplained variation** (**this is another name for *SSE***). Intuitively, this quantity measures the amount of variation in the values of y that is not explained by the predictor variable. The quantity $\Sigma(\hat{y}_i - \bar{y})^2$ is called the **explained variation.** Using these definitions and the above equation involving these summations, we see that

Total variation − Unexplained variation = Explained variation

It follows that the explained variation is the reduction in the sum of squared prediction errors that has been accomplished by using the predictor variable x to predict y. It also follows that

Total variation = Explained variation + Unexplained variation

Intuitively, this equation implies that the explained variation represents the amount of the total variation in the observed values of y that is explained by the predictor variable x (and the simple linear regression model).

We now define the **simple coefficient of determination** to be

$$r^2 = \frac{\text{Explained variation}}{\text{Total variation}}$$

That is, r^2 is the proportion of the total variation in the n observed values of y that is explained by the simple linear regression model. Neither the explained variation nor the total variation can be negative (both quantities are sums of squares). Therefore, r^2 is greater than or equal to 0. Because the explained variation must be less than or equal to the total variation, r^2 cannot be greater than 1. The nearer r^2 is to 1, the larger is the proportion of the total variation that is explained by the model, and the greater is the utility of the model in predicting y. If the value of r^2 is not reasonably close to 1, the independent variable in the model does not provide accurate predictions of y. In such a case, a different predictor variable must be found in order to accurately predict y. It is also possible that no regression model employing a single predictor variable will accurately predict y. In this case the model must be improved by including more than one independent variable. We show how to do this in Chapter 15.

The Simple Coefficient of Determination, r^2

For the simple linear regression model

1. **Total variation** $= \sum (y_i - \bar{y})^2$
2. **Explained variation** $= \sum (\hat{y}_i - \bar{y})^2$
3. **Unexplained variation** $= \sum (y_i - \hat{y}_i)^2$
4. **Total variation = Explained variation + Unexplained variation**
5. **The simple coefficient of determination is**
$$r^2 = \frac{\text{Explained variation}}{\text{Total variation}}$$
6. r^2 is the proportion of the total variation in the n observed values of the dependent variable that is explained by the simple linear regression model.

EXAMPLE 14.7 The Tasty Sub Shop Case: Calculating and Interpreting r^2 C

For the Tasty Sub data we have seen that $\bar{y} = 860.31$ (see Example 14.2 on page 493). It follows that the total variation is

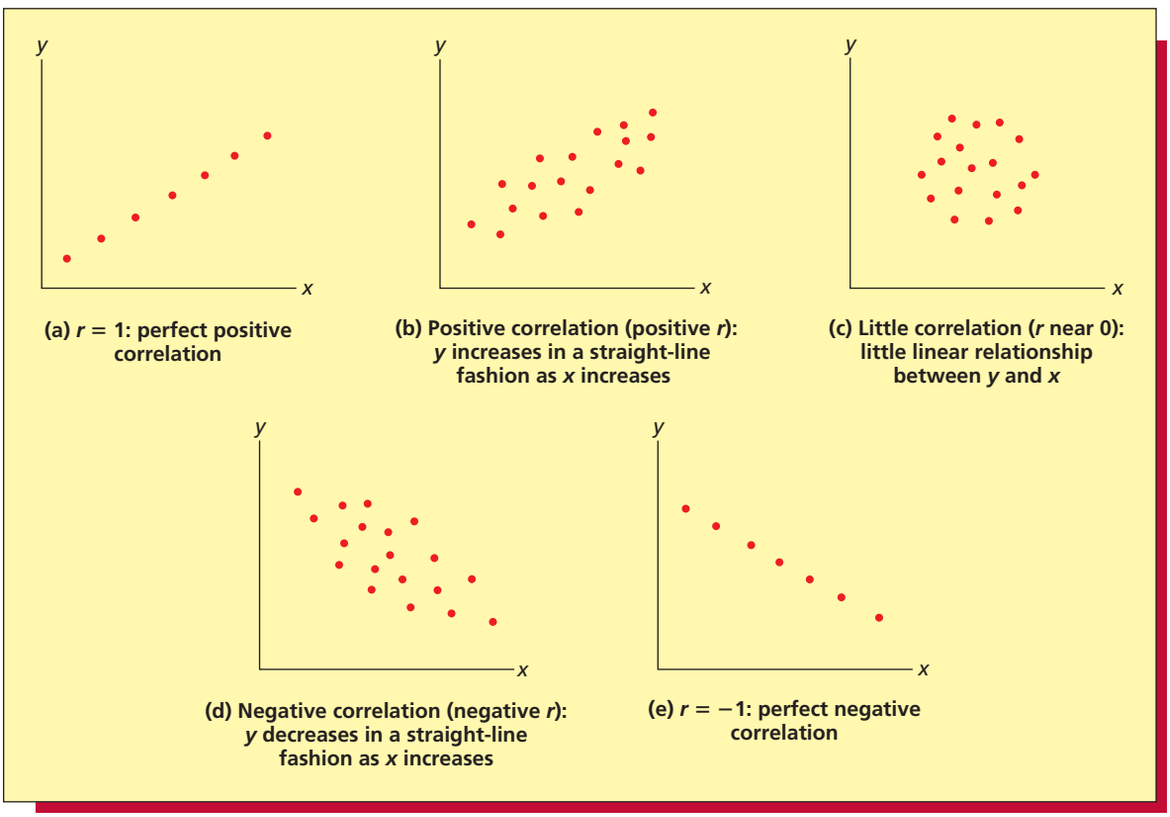
$$\sum (y_i - \bar{y})^2 = (527.1 - 860.31)^2 + (548.7 - 860.31)^2 + \cdots + (1235.8 - 860.31)^2$$
$$= 495{,}776.51$$

Furthermore, we found in Table 14.2 (page 494) that the unexplained variation is $SSE = 30{,}460.21$. Therefore, we can compute the explained variation and r^2 as follows:

$$\text{Explained variation} = \text{Total variation} - \text{Unexplained variation}$$
$$= 495{,}776.51 - 30{,}460.21 = 465{,}316.30$$
$$r^2 = \frac{\text{Explained variation}}{\text{Total variation}} = \frac{465{,}316.30}{495{,}776.51} = .939$$

This value of r^2 says that the regression model explains 93.9 percent of the total variation in the 10 observed yearly revenues.

FIGURE 14.19 An Illustration of Different Values of the Simple Correlation Coefficient



The simple correlation coefficient, *r* People often claim that two variables are correlated. For example, a college admissions officer might feel that the academic performance of college students (measured by grade point average) is correlated with the students' scores on a standardized college entrance examination. This means that college students' grade point averages are related to their college entrance exam scores. One measure of the relationship between two variables y and x is the **simple correlation coefficient.** We define this quantity as follows:

The Simple Correlation Coefficient

The **simple correlation coefficient between *y* and *x*,** denoted by ***r***, is

$$r = +\sqrt{r^2} \quad \text{if } b_1 \text{ is positive} \qquad \text{and} \qquad r = -\sqrt{r^2} \quad \text{if } b_1 \text{ is negative}$$

where b_1 is the slope of the least squares line relating y to x. This correlation coefficient **measures the strength of the linear relationship between *y* and *x*.**

Because r^2 is always between 0 and 1, the correlation coefficient r is between -1 and 1. A value of r near 0 implies little linear relationship between y and x. A value of r close to 1 says that y and x have a strong tendency to move together in a straight-line fashion with a positive slope and, therefore, that y and x are highly related and **positively correlated.** A value of r close to -1 says that y and x have a strong tendency to move together in a straight-line fashion with a negative slope and, therefore, that y and x are highly related and **negatively correlated.** Figure 14.19 illustrates these relationships. Notice that when $r = 1$, y and x have a perfect linear relationship with a positive slope, whereas when $r = -1$, y and x have a perfect linear relationship with a negative slope.

EXAMPLE 14.8 The Tasty Sub Shop Case: Simple Correlation Coefficient

In the Tasty Sub Shop case, we found that $b_1 = 15.596$ and $r^2 = .939$. It follows that the simple correlation coefficient between y (yearly revenue) and x (population size) is

$$r = +\sqrt{r^2} = +\sqrt{.939} = .969$$

This simple correlation coefficient says that x and y have a strong tendency to move together in a linear fashion with a positive slope. We have seen this tendency in Figure 14.1 (page 488), which indicates that y and x are positively correlated.

If we have computed the least squares slope b_1 and r^2, the method given in the previous box provides the easiest way to calculate r. The simple correlation coefficient can also be calculated using the formula

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}$$

Here SS_{xy} and SS_{xx} have been defined in Section 14.1 on page 492, and SS_{yy} denotes the total variation, which has been defined in this section. Furthermore, this formula for r automatically gives r the correct ($+$ or $-$) sign. For instance, in the Tasty Sub Shop case, $SS_{xy} = 29{,}836.389$, $SS_{xx} = 1913.129$, and $SS_{yy} = 495{,}776.51$ (see Examples 14.2 on page 493 and 14.7 on page 518). Therefore

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}$$

$$= \frac{29{,}836.389}{\sqrt{(1{,}913.129)(495{,}776.51)}} = .969$$

It is important to make two points. First, **the value of the simple correlation coefficient is not the slope of the least squares line.** If we wish to find this slope, we should use the previously given formula for b_1.[2] Second, **high correlation does not imply that a cause-and-effect relationship exists.** When r indicates that y and x are highly correlated, this says that y and x have a strong tendency to move together in a straight-line fashion. The correlation does not mean that changes in x cause changes in y. Instead, some other variable (or variables) could be causing the apparent relationship between y and x. For example, suppose that college students' grade point averages and college entrance exam scores are highly positively correlated. This does not mean that earning a high score on a college entrance exam causes students to receive a high grade point average. Rather, other factors such as intellectual ability, study habits, and attitude probably determine both a student's score on a college entrance exam and a student's college grade point average. In general, while the simple correlation coefficient can show that variables tend to move together in a straight-line fashion, scientific theory must be used to establish cause-and-effect relationships.

A technical note In optional Section 14.10 we present some shortcut formulas for calculating the total, explained, and unexplained variations. Also, for those who have already read Section 14.3, r^2, the explained variation, the unexplained variation, and the total variation are calculated by Excel and MINITAB. These quantities are identified on the Excel and MINITAB outputs of Figure 14.8 (page 505) by, respectively, the labels 9, 10, 11, and 12. These outputs also give an "adjusted r^2." We will explain the meaning of this quantity in Chapter 15.

[2] Essentially, the difference between r and b_1 is a change of scale. It can be shown that b_1 and r are related by the equation $b_1 = (SS_{yy}/SS_{xx})^{1/2} r$.

Exercises for Section 14.5

CONCEPTS

connect

14.36 Discuss the meanings of the total variation, the unexplained variation, and the explained variation.

14.37 What does the simple coefficient of determination measure?

METHODS AND APPLICATIONS

In Exercises 14.38 through 14.43, we give the total variation, the unexplained variation (*SSE*), and the least squares point estimate b_1 that are obtained when simple linear regression is used to analyze the data set related to each of six previously discussed case studies. Using the information given in each exercise, find the explained variation, the simple coefficient of determination (r^2), and the simple correlation coefficient (r). Interpret r^2.

14.38 **THE NATURAL GAS CONSUMPTION CASE** DS GasCon1

Total variation = 25.549; $SSE = 2.568$; $b_1 = -.12792$

14.39 **THE STARTING SALARY CASE** DS StartSal

Total variation = 61.380; $SSE = 1.438$; $b_1 = 5.7066$

14.40 **THE SERVICE TIME CASE** DS SrvcTime

Total variation = 20,110.5455; $SSE = 191.7017$; $b_1 = 24.6022$

14.41 **THE FRESH DETERGENT CASE** DS Fresh

Total variation = 13.459; $SSE = 2.806$; $b_1 = 2.6652$

14.42 **THE DIRECT LABOR COST CASE** DS DirLab

Total variation = 1,025,339.6667; $SSE = 746.7624$; $b_1 = 10.1463$

14.43 **THE REAL ESTATE SALES PRICE CASE** DS RealEst

Total variation = 7447.5; $SSE = 896.8$; $b_1 = 5.7003$

14.6 Testing the Significance of the Population Correlation Coefficient (Optional)

LO14-7 Test hypotheses about the population correlation coefficient.

We have seen that the simple correlation coefficient measures the linear relationship between the observed values of x and the observed values of y that make up the sample. A similar coefficient of linear correlation can be defined for the population of *all possible combinations of observed values of x and y.* We call this coefficient the **population correlation coefficient** and denote it by the symbol $\boldsymbol{\rho}$ (pronounced **rho**). We use r as the point estimate of ρ. In addition, we can carry out a hypothesis test. Here we test the null hypothesis $\boldsymbol{H_0: \rho = 0}$**, which says there is no linear relationship between** $\boldsymbol{x}$ **and** $\boldsymbol{y}$**,** against the alternative $\boldsymbol{H_a: \rho \neq 0}$**, which says there is a positive or negative linear relationship between** $\boldsymbol{x}$ **and** $\boldsymbol{y}$**.** This test employs the test statistic

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

and is based on the assumption that the population of all possible observed combinations of values of x and y has a **bivariate normal probability distribution.** See Wonnacott and Wonnacott (1981) for a discussion of this distribution. It can be shown that the preceding test statistic t and the p-value used to test H_0: $\rho = 0$ versus H_a: $\rho \neq 0$ are equal to, respectively, the test statistic $t = b_1/s_{b_1}$ and the p-value used to test H_0: $\beta_1 = 0$ versus H_a: $\beta_1 \neq 0$, where β_1 is the slope in the simple linear regression model. Keep in mind, however, that although the mechanics involved in these hypothesis tests are the same, these tests are based on different assumptions (remember that the test for significance of the slope is based on the regression assumptions). If the bivariate normal distribution assumption for the test concerning ρ is badly violated, we can use a nonparametric approach to correlation. One such approach is **Spearman's rank correlation coefficient.** This approach is discussed in Chapter 18.

EXAMPLE 14.9 The Tasty Sub Shop Case: The Correlation Between x and y C

Again consider testing the significance of the slope in the Tasty Sub Shop problem. Recall that in Example 14.4 (page 504) we found that $t = 11.05$ and that the p-value related to this t statistic is less than .001. We therefore (if the regression assumptions hold) can reject H_0: $\beta_1 = 0$ at level of significance .05, .01, or .001, and we have extremely strong evidence that x is significantly related to y. This also implies (if the population of all possible observed combinations of x and y has a bivariate normal probability distribution) that we can reject H_0: $\rho = 0$ in favor of H_a: $\rho \neq 0$ at level of significance .05, .01, or .001. It follows that we have extremely strong evidence of a linear relationship, or correlation, between x and y. Furthermore, because we have previously calculated r to be .969, we estimate that x and y are positively correlated.

Exercises for Section 14.6

connect

CONCEPTS

14.44 Explain what is meant by the population correlation coefficient ρ.

14.45 Explain how we test H_0: $\rho = 0$ versus H_a: $\rho \neq 0$. What do we conclude if we reject H_0: $\rho = 0$?

METHODS AND APPLICATIONS

14.46 THE STARTING SALARY CASE DS StartSal

Consider testing H_0: $\beta_1 = 0$ versus H_a: $\beta_1 \neq 0$. Figure 14.11 (page 508) tells us that $t = 14.44$ and that the related p-value is less than .001. Assuming that the bivariate normal probability distribution assumption holds, test H_0: $\rho = 0$ versus H_a: $\rho \neq 0$ by setting α equal to .05, .01, and .001. What do you conclude about how x and y are related?

14.47 THE SERVICE TIME CASE DS SrvcTime

Consider testing H_0: $\beta_1 = 0$ versus H_a: $\beta_1 \neq 0$. Figure 14.12 (page 508) tells us that $t = 30.5802$ and that the related p-value is less than .001. Assuming that the bivariate normal probability distribution assumption holds, test H_0: $\rho = 0$ versus H_a: $\rho \neq 0$ by setting α equal to .05, .01, and .001. What do you conclude about how x and y are related?

LO14-8 Test the significance of a simple linear regression model by using an F-test.

14.7 An F-Test for the Model

In this section we discuss an F-test that can be used to test the significance of the regression relationship between x and y. Sometimes people refer to this as testing the significance of the simple linear regression model. For simple linear regression, this test is another way to test the null hypothesis H_0: $\beta_1 = 0$ (the relationship between x and y is not significant) versus H_a: $\beta_1 \neq 0$ (the relationship between x and y is significant). If we can reject H_0 at level of significance α, we often say that **the simple linear regression model is significant at level of significance α.**

An F Test for the Simple Linear Regression Model

Suppose that the regression assumptions hold, and define the **overall F statistic** to be

$$F(\text{model}) = \frac{\text{Explained variation}}{(\text{Unexplained variation})/(n-2)}$$

Also define the p-value related to F(model) to be the area under the curve of the F distribution (having 1 numerator and $n - 2$ denominator degrees of freedom) to the right of F(model)—see Figure 14.20(b).

We can reject H_0: $\beta_1 = 0$ in favor of H_a: $\beta_1 \neq 0$ at level of significance α if either of the following equivalent conditions holds:

1. $F(\text{model}) > F_\alpha$
2. $p\text{-value} < \alpha$

Here the point F_α is based on 1 numerator and $n - 2$ denominator degrees of freedom.

FIGURE 14.20(a) The *F*-Test Critical Value

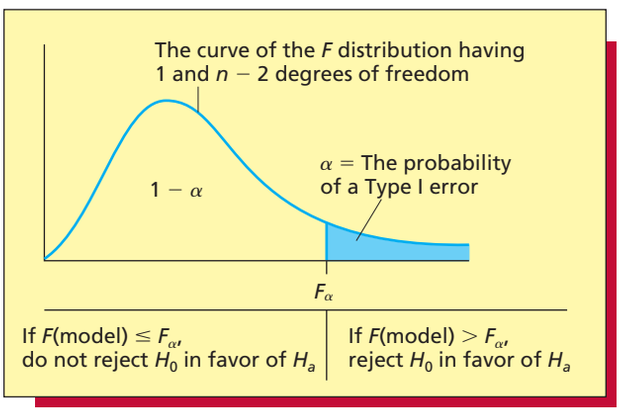
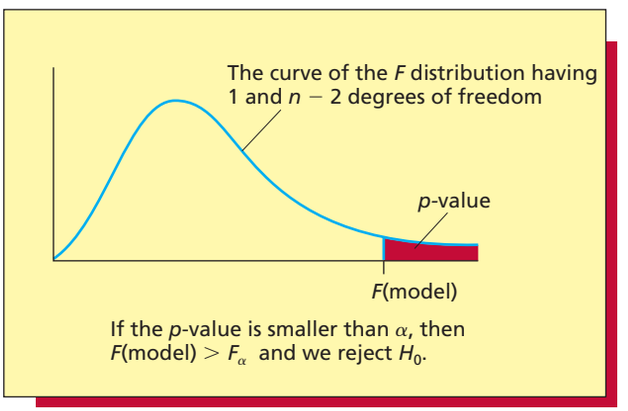


FIGURE 14.20(b) The *F*-Test *p*-Value



The first condition in the box says we should reject H_0: $\beta_1 = 0$ (and conclude that the relationship between x and y is significant) when F(model) is large. This is intuitive because a large overall F statistic would be obtained when the explained variation is large compared to the unexplained variation. This would occur if x is significantly related to y, which would imply that the slope β_1 is not equal to 0. Figure 14.20(a) illustrates that we reject H_0 when F(model) is greater than F_α. As can be seen in Figure 14.20(b), when F(model) is large, the related p-value is small. When the p-value is small enough [resulting from an F(model) statistic that is large enough], we reject H_0. Figure 14.20(b) illustrates that the second condition in the box (p-value $< \alpha$) is an equivalent way to carry out this test.

EXAMPLE 14.10 The Tasty Sub Shop Case: An *F*-Test for the Model

Consider the Tasty Sub Shop problem and the following partial MINITAB output of the simple linear regression analysis relating yearly revenue y to population size x:

```
Analysis of Variance
Source              DF          SS          MS          F      P-value
Regression           1      465316      465316     122.21        0.000
Residual Error       8       30460        3808
Total                9      495777
```

Looking at this output, we see that the explained variation is 465,316 and the unexplained variation is 30,460. It follows that

$$F(\text{model}) = \frac{\text{Explained variation}}{(\text{Unexplained variation})/(n-2)} = \frac{465{,}316}{30{,}460/(10-2)} = \frac{465{,}316}{3808} = 122.21$$

Note that this overall F statistic is given on the MINITAB output and is also given on the following partial Excel output:

ANOVA	df	SS	MS	F	Significance F
Regression	1	465316.3004	465316.3004	122.2096	0.0000
Residual	8	30460.2086	3807.5261		
Total	9	495776.5090			

The p-value related to F(model) is the area to the right of 122.21 under the curve of the F distribution having 1 numerator and 8 denominator degrees of freedom. This p-value is given on both the MINITAB output (labeled "p") and the Excel output (labeled "Significance F") and is less than .001. If we wish to test the significance of the regression relationship with level of significance $\alpha = .05$, we use the critical value $F_{.05}$ based on 1 numerator and 8 denominator degrees of freedom. Using Table A.7 (page 796), we find that $F_{.05} = 5.32$. Because $F(\text{model}) = 122.21 > F_{.05} = 5.32$, we can reject H_0: $\beta_1 = 0$ in favor of H_a: $\beta_1 \neq 0$ at level of significance .05. Alternatively, because the p-value is smaller than .05, .01, and .001, we can reject H_0 at level of significance .05, .01, or .001. Therefore, we have extremely strong evidence that H_0: $\beta_1 = 0$ should be rejected and that the regression relationship between x and y is significant. That is, we might say that we have extremely strong evidence that the simple linear model relating y to x is significant.

Testing the significance of the regression relationship between y and x by using the overall F statistic and its related p-value is equivalent to doing this test by using the t statistic and its related p-value. Specifically, it can be shown that $(t)^2 = F(\text{model})$ and that $(t_{\alpha/2})^2$ based on $n - 2$ degrees of freedom equals F_α based on 1 numerator and $n - 2$ denominator degrees of freedom. It follows that the critical value conditions

$$|t| > t_{\alpha/2} \quad \text{and} \quad F(\text{model}) > F_\alpha$$

are equivalent. Furthermore, the p-values related to t and F(model) can be shown to be equal. Because these tests are equivalent, it would be logical to ask why we have presented the F-test. There are two reasons. First, most standard regression computer packages include the results of the F-test as a part of the regression output. Second, the F-test has a useful generalization in multiple regression analysis (where we employ more than one predictor variable). The F-test in multiple regression is not equivalent to a t test. This is further explained in Chapter 15.

Exercises for Section 14.7

CONCEPTS

connect

14.48 What are the null and alternative hypotheses for the F-test in simple linear regression?

14.49 The F-test in simple linear regression is equivalent to what other test?

METHODS AND APPLICATIONS

In Exercises 14.50 through 14.55, we give MINITAB and Excel outputs of simple linear regression analyses of the data sets related to six previously discussed case studies. Using the appropriate computer output,

a Use the explained variation and the unexplained variation as given on the computer output to calculate (within rounding) the F(model) statistic.

b Utilize the F(model) statistic and the appropriate critical value to test H_0: $\beta_1 = 0$ versus H_a: $\beta_1 \neq 0$ by setting α equal to .05. What do you conclude about the regression relationship between y and x?

c Utilize the F(model) statistic and the appropriate critical value to test H_0: $\beta_1 = 0$ versus H_a: $\beta_1 \neq 0$ by setting α equal to .01. What do you conclude about the regression relationship between y and x?

d Find the p-value related to F(model) on the computer output and report its value. Using the p-value, test the significance of the regression model at the .10, .05, .01, and .001 levels of significance. What do you conclude?

e Show that the F(model) statistic is (within rounding) the square of the t statistic for testing H_0: $\beta_1 = 0$ versus H_a: $\beta_1 \neq 0$. Also, show that the $F_{.05}$ critical value is the square of the $t_{.025}$ critical value.

Note that in the lower right hand corner of each output we give (in parentheses) the number of observations, n, used to perform the regression analysis and the t statistic for testing H_0: $\beta_1 = 0$ versus H_a: $\beta_1 \neq 0$.

14.50 THE NATURAL GAS CONSUMPTION CASE DS GasCon1

ANOVA	df	SS	MS	F	Significance F
Regression	1	22.9808	22.9808	53.6949	0.0003
Residual	6	2.5679	0.4280		
Total	7	25.5488			(n=8; t=-7.33)

14.51 THE STARTING SALARY CASE StartSal

```
Analysis of Variance
Source           DF      SS      MS        F       P
Regression        1  59.942  59.942   208.39   0.000
Residual Error    5   1.438   0.288
Total             6  61.380          (n=7; t=14.44)
```

14.52 THE SERVICE TIME CASE SrvcTime

ANOVA	SS	df	MS	F	p-value
Regression	19,918.8438	1	19,918.8438	935.15	2.09E-10
Residual	191.7017	9	21.3002		
Total	20,110.5455	10		(n=11; t=30.580)	

14.53 THE FRESH DETERGENT CASE Fresh

```
Analysis of Variance
Source           DF      SS      MS        F       P
Regression        1  10.653  10.653   106.30   0.000
Residual Error   28   2.806   0.100
Total            29  13.459          (n=30; t=10.31)
```

14.54 THE DIRECT LABOR COST CASE DirLab

ANOVA	df	SS	MS	F	Significance F
Regression	1	1024592.9043	1024592.9043	13720.4677	5.04E-17
Residual	10	746.7624	74.6762		
Total	11	1025339.6667		(n=12; t=117.1344)	

14.55 THE REAL ESTATE SALES PRICE CASE RealEst

```
Analysis of Variance
Source           DF      SS      MS        F       P
Regression        1  6550.7  6550.7   58.43   0.000
Residual Error    8   896.8   112.1
Total             9  7447.5          (n=10; t=7.64)
```

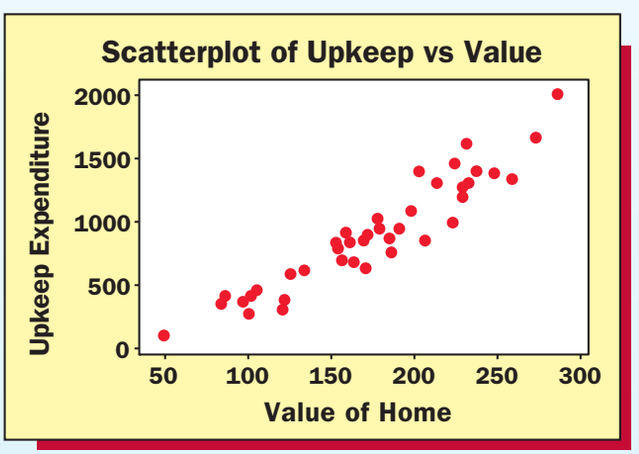
14.8 The QHIC Case: Developing An Advertising Strategy

Quality Home Improvement Center (QHIC) operates five stores in a large metropolitan area. The marketing department at QHIC wishes to study the relationship between x, home value (in thousands of dollars), and y, yearly expenditure on home upkeep (in dollars). A random sample of 40 homeowners is taken and survey participants are asked to estimate their expenditures during the previous year on the types of home upkeep products and services offered by QHIC. Public records of the county auditor are used to obtain the previous year's assessed values of the homeowner's homes. The resulting x and y values, as well as a scatter plot of these values, are given in Figure 14.21. The Excel output included in this figure tells us that the least squares point estimates of the y-intercept β_0 and the slope β_1 are $b_0 = -348.3921$ and $b_1 = 7.2583$. The p-value associated with b_1 implies there is a significant linear relationship between x and y. In addition, because $b_1 = 7.2583$, we estimate that mean yearly upkeep expenditure increases by \$7.26 for each additional \$1,000 increase in home value. Consider a home worth \$220,000, and note that $x_0 = 220$ is in the range of previously observed values of x: 48.9 to 286.18. It follows that

$$\begin{aligned}\hat{y} &= b_0 + b_1 x_0 \\ &= -348.3921 + 7.2583(220) \\ &= 1{,}248.43 \text{ (or \$1,248.43)}\end{aligned}$$

FIGURE 14.21 The QHIC Upkeep Expenditure Data, Scatterplot, and Excel Output QHIC

Home	Value of Home, x (Thousands of Dollars)	Upkeep Expenditure, y (Dollars)	Home	Value of Home, x (Thousands of Dollars)	Upkeep Expenditure, y (Dollars)
1	237.00	1,412.08	21	153.04	849.14
2	153.08	797.20	22	232.18	1,313.84
3	184.86	872.48	23	125.44	602.06
4	222.06	1,003.42	24	169.82	642.14
5	160.68	852.90	25	177.28	1,038.80
6	99.68	288.48	26	162.82	697.00
7	229.04	1,288.46	27	120.44	324.34
8	101.78	423.08	28	191.10	965.10
9	257.86	1,351.74	29	158.78	920.14
10	96.28	378.04	30	178.50	950.90
11	171.00	918.08	31	272.20	1,670.32
12	231.02	1,627.24	32	48.90	125.40
13	228.32	1,204.76	33	104.56	479.78
14	205.90	857.04	34	286.18	2,010.64
15	185.72	775.00	35	83.72	368.36
16	168.78	869.26	36	86.20	425.60
17	247.06	1,396.00	37	133.58	626.90
18	155.54	711.50	38	212.86	1,316.94
19	224.20	1,475.18	39	122.02	390.16
20	202.04	1,413.32	40	198.02	1,090.84



Regression Statistics	
Multiple R	0.9430
R Square	0.8892
Adjusted R Square	0.8863
Standard Error	146.8973
Observations	40

ANOVA	df	SS	MS	F	Significance F
Regression	1	6582759.6972	6582759.6972	305.0564	0.0000
Residual	38	819995.5427	21578.8301		
Total	39	7402755.2399			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	−348.3921	76.1410	−4.5756	0.0000	−502.5314	−194.2527
Value	7.2583	0.4156	17.4659	0.0000	6.4170	8.0995

is the point estimate of the mean yearly upkeep expenditure for all homes worth $220,000 and is the point prediction of a yearly upkeep expenditure for an individual home worth $220,000.

The marketing department at QHIC wishes to determine which homes should be sent advertising brochures promoting QHIC's products and services. If the marketing department has decided to send an advertising brochure to any home that has a predicted yearly upkeep expenditure of at least $500, then a home worth $220,000 would be sent an advertising brochure. This is because the predicted yearly upkeep expenditure for such a home is (as calculated above) $1,248.43. Other homes can be evaluated in a similar fashion.

BI

14.9 Residual Analysis

LO14-9 Use residual analysis to check the assumptions of simple linear regression.

In this section we explain how to check the validity of the regression assumptions. The required checks are carried out by analyzing the **regression residuals.** The residuals are defined as follows:

For any particular observed value of y, the corresponding **residual** is

$$e = y - \hat{y} = (\text{observed value of } y - \text{predicted value of } y)$$

where the predicted value of y is calculated using the **least squares prediction equation**

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

The linear regression model $y = \beta_0 + \beta_1 x + \varepsilon$ implies that the error term ε is given by the equation $\varepsilon = y - (\beta_0 + \beta_1 x)$. Because $\hat{y}$ in the previous box is clearly the point estimate of $\beta_0 + \beta_1 x$, we see that the residual $e = y - \hat{y}$ is the point estimate of the error term ε. If the regression assumptions are valid, then, for any given value of the independent variable, the population of potential error term values will be normally distributed with mean 0 and variance σ^2 (see the regression assumptions in Section 14.2 on page 500). Furthermore, the different error terms will be statistically independent. Because the residuals provide point estimates of the error terms, it follows that

If the regression assumptions hold, the residuals should look like they have been randomly and independently selected from normally distributed populations having mean 0 and variance σ^2.

In any real regression problem, the regression assumptions will not hold exactly. In fact, it is important to point out that mild departures from the regression assumptions do not seriously hinder our ability to use a regression model to make statistical inferences. Therefore, we are looking for pronounced, rather than subtle, departures from the regression assumptions. Because of this, we will require that the residuals only approximately fit the description just given.

Residual plots One useful way to analyze residuals is to plot them versus various criteria. The resulting plots are called **residual plots.** To construct a residual plot, we compute the residual for each observed y value. The calculated residuals are then plotted versus some criterion. To validate the regression assumptions, we make residual plots against (1) values of the independent variable x; (2) values of $\hat{y}$, the predicted value of the dependent variable; and (3) the time order in which the data have been observed (if the regression data are time series data).

We next look at an example of constructing residual plots. Then we explain how to use these plots to check the regression assumptions.

EXAMPLE 14.11 The QHIC Case: Constructing Residual Plots

C

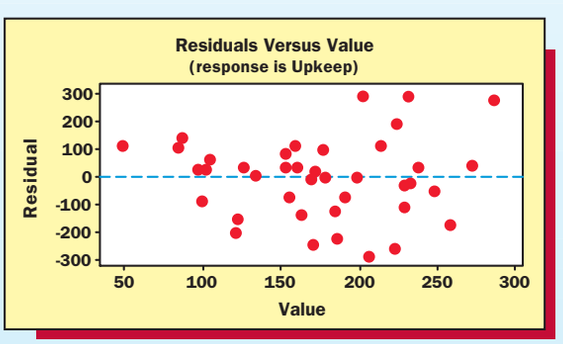
Figure 14.21 gives the QHIC upkeep expenditure data and a scatterplot of the data. If we use a simple linear regression model to describe the QHIC data, we find that the least squares point estimates of β_0 and β_1 are $b_0 = -348.3921$ and $b_1 = 7.2583$. The Excel add-in (MegaStat) output in

FIGURE 14.22 Residuals and Residual Plots for the QHIC Simple Linear Regression Model

(a) Excel add-in (MegaStat) output of the residuals

Observation	Upkeep	Predicted	Residual	Observation	Upkeep	Predicted	Residual
1	1,412.080	1,371.816	40.264	21	849.140	762.413	86.727
2	797.200	762.703	34.497	22	1,313.840	1,336.832	−22.992
3	872.480	993.371	−120.891	23	602.060	562.085	39.975
4	1,003.420	1,263.378	−259.958	24	642.140	884.206	−242.066
5	852.900	817.866	35.034	25	1,038.800	938.353	100.447
6	288.480	375.112	−86.632	26	697.000	833.398	−136.398
7	1,288.460	1,314.041	−25.581	27	324.340	525.793	−201.453
8	423.080	390.354	32.726	28	965.100	1,038.662	−73.562
9	1,351.740	1,523.224	−171.484	29	920.140	804.075	116.065
10	378.040	350.434	27.606	30	950.900	947.208	3.692
11	918.080	892.771	25.309	31	1,670.320	1,627.307	43.013
12	1,627.240	1,328.412	298.828	32	125.400	6.537	118.863
13	1,204.760	1,308.815	−104.055	33	479.780	410.532	69.248
14	857.040	1,146.084	−289.044	34	2,010.640	1,728.778	281.862
15	775.000	999.613	−224.613	35	368.360	259.270	109.090
16	869.260	876.658	−7.398	36	425.600	277.270	148.330
17	1,396.000	1,444.835	−48.835	37	626.900	621.167	5.733
18	711.500	780.558	−69.058	38	1,316.940	1,196.602	120.338
19	1,475.180	1,278.911	196.269	39	390.160	537.261	−147.101
20	1,413.320	1,118.068	295.252	40	1,090.840	1,088.889	1.951

(b) MINITAB output of a residual plot versus *x*



(c) MINITAB output of a residual plot versus $\hat{y}$

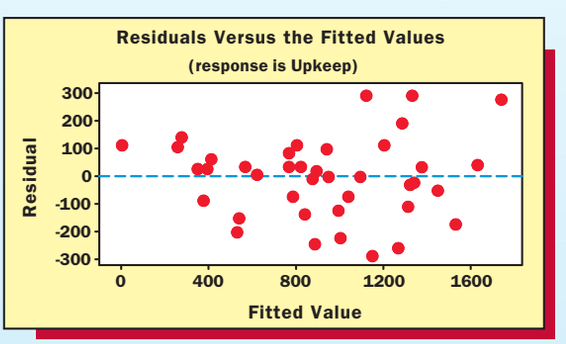


Figure 14.22(a) presents the predicted home upkeep expenditures and residuals that are given by the simple linear regression model. Here each residual is computed as

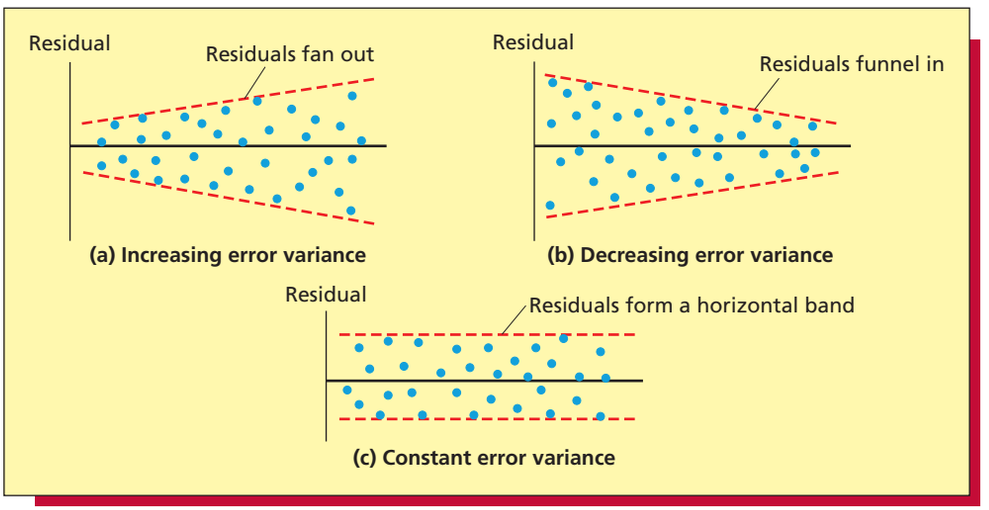
$$e = y - \hat{y} = y - (b_0 + b_1 x) = y - (-348.3921 + 7.2583x)$$

For instance, for the first observation (home) when $y = 1{,}412.08$ and $x = 237.00$ (see Figure 14.21 on page 526), the residual is

$$\begin{aligned} e &= 1{,}412.08 - (-348.3921 + 7.2583(237)) \\ &= 1{,}412.08 - 1{,}371.816 = 40.264 \end{aligned}$$

The MINITAB output in Figure 14.22(b) and (c) gives plots of the residuals for the QHIC simple linear regression model against values of x and $\hat{y}$. To understand how these plots are constructed, recall that for the first observation (home) $y = 1{,}412.08$, $x = 237.00$, $\hat{y} = 1{,}371.816$, and the residual is 40.264. It follows that the point plotted in Figure 14.22(b) corresponding to the first observation has a horizontal axis coordinate of the x value 237.00 and a vertical axis coordinate of the residual 40.264. It also follows that the point plotted in Figure 14.22(c) corresponding to the first observation has a horizontal axis coordinate of the $\hat{y}$ value 1,371.816, and a vertical axis coordinate of the residual 40.264. Finally, note that the QHIC data are cross-sectional data, not time series data. Therefore, we cannot make a residual plot versus time.

FIGURE 14.23 Residual Plots and the Constant Variance Assumption



The constant variance assumption To check the validity of the constant variance assumption, we examine plots of the residuals against values of x, $\hat{y}$, and time (if the regression data are time series data). When we look at these plots, the pattern of the residuals' fluctuation around 0 tells us about the validity of the constant variance assumption. A residual plot that "fans out" [as in Figure 14.23(a)] suggests that the error terms are becoming more spread out as the horizontal plot value increases and that the constant variance assumption is violated. Here we would say that an **increasing error variance** exists. A residual plot that "funnels in" [as in Figure 14.23(b)] suggests that the spread of the error terms is decreasing as the horizontal plot value increases and that again the constant variance assumption is violated. In this case we would say that a **decreasing error variance** exists. A residual plot with a "horizontal band appearance" [as in Figure 14.23(c)] suggests that the spread of the error terms around 0 is not changing much as the horizontal plot value increases. Such a plot tells us that the constant variance assumption (approximately) holds.

As an example, consider the QHIC case and the residual plot in Figure 14.22(b). This plot appears to fan out as x increases, indicating that the spread of the error terms is increasing as x increases. That is, an increasing error variance exists. This is equivalent to saying that the variance of the population of potential yearly upkeep expenditures for houses worth x (thousand dollars) appears to increase as x increases. The reason is that the model $y = \beta_0 + \beta_1 x + \varepsilon$ says that the variation of y is the same as the variation of ε. For example, the variance of the population of potential yearly upkeep expenditures for houses worth \$200,000 would be larger than the variance of the population of potential yearly upkeep expenditures for houses worth \$100,000. Increasing variance makes some intuitive sense because people with more expensive homes generally have more discretionary income. These people can choose to spend either a substantial amount or a much smaller amount on home upkeep, thus causing a relatively large variation in upkeep expenditures.

Another residual plot showing the increasing error variance in the QHIC Case is Figure 14.22(c). This plot tells us that the residuals appear to fan out as $\hat{y}$ (predicted y) increases, which is logical because $\hat{y}$ is an increasing function of x. Also, note that the scatter plot of y versus x in Figure 14.21 (page 526) shows the increasing error variance—the y values appear to fan out as x increases. In fact, one might ask why we need to consider residual plots when we can simply look at scatter plots of y versus x. One answer is that, in general, because of possible differences in scaling between residual plots and scatter plots of y versus x, one of these types of plots might be more informative in a particular situation. Therefore, we should always consider both types of plots.

When the constant variance assumption is violated, we cannot use the formulas of this chapter to make statistical inferences. Later in this section we discuss how we can make statistical inferences when a nonconstant error variance exists.

The assumption of correct functional form If the functional form of a regression model is incorrect, the residual plots constructed by using the model often display a pattern suggesting the form of a more appropriate model. For instance, if we use a simple linear regression model when the true relationship between y and x is curved, the residual plot will have a curved appearance. For example, the scatter plot of upkeep expenditure, y, versus home value, x, in Figure 14.21 (page 526) has either a straight-line or slightly curved appearance. We used a simple linear regression model to describe the relationship between y and x, but note that there is a "dip," or slightly curved appearance, in the upper left portion of each residual plot in Figure 14.22 (page 528). Therefore, both the scatter plot and residual plots indicate that there might be a slightly curved relationship between y and x. Later in this section we discuss one way to model curved relationships.

The normality assumption If the normality assumption holds, a histogram and/or stem-and-leaf display of the residuals should look reasonably bell-shaped and reasonably symmetric about 0. Figure 14.24(a) gives the MINITAB output of a stem-and-leaf display of the residuals from the simple linear regression model describing the QHIC data. The stem-and-leaf display looks fairly bell-shaped and symmetric about 0. However, the tails of the display look somewhat long and "heavy" or "thick," indicating a possible violation of the normality assumption.

Another way to check the normality assumption is to construct a **normal plot** of the residuals. To make a normal plot, we first arrange the residuals in order from smallest to largest. Letting the ordered residuals be denoted as $e_{(1)}, e_{(2)}, \ldots, e_{(n)}$ we denote the ith residual in the ordered listing as $e_{(i)}$. We plot $e_{(i)}$ on the vertical axis against a point called $z_{(i)}$ on the horizontal axis. Here $z_{(i)}$ is defined to be the point on the horizontal axis under the standard normal curve so that the area under this curve to the left of $z_{(i)}$ is $(3i - 1)/(3n + 1)$. For example, recall in the QHIC case that there are $n = 40$ residuals given in Figure 14.22(a). It follows that, when $i = 1$, then

$$\frac{3i - 1}{3n + 1} = \frac{3(1) - 1}{3(40) + 1} = \frac{2}{121} = .0165$$

Therefore, $z_{(1)}$ is the normal point having an area of .0165 under the standard normal curve to its left. Thus, as illustrated in Figure 14.24(b), $z_{(1)}$ equals -2.13. Because the smallest residual in Figure 14.22(a) is -289.044, the first point plotted is $e_{(1)} = -289.044$ on the vertical scale versus $z_{(1)} = -2.13$ on the horizontal scale. When $i = 2$, it can be verified that $(3i - 1)/(3n + 1)$ equals .0413 and thus that $z_{(2)} = -1.74$. Therefore, because the second-smallest residual in Figure 14.24(a) is -259.958, the second point plotted is $e_{(2)} = -259.958$ on the vertical scale versus $z_{(2)} = -1.74$ on the horizontal scale. This process is continued until the entire normal plot is constructed. The Excel add-in (MegaStat) output of this plot is given in Figure 14.24(c).

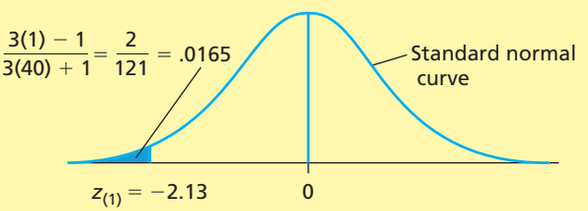
An equivalent plot is shown in Figure 14.24(d), which is a MINITAB output. In this figure, we plot the percentage $p_{(i)}$ of the area under the standard normal curve to the left of $z_{(i)}$ on the vertical axis. Thus, the first point plotted in this normal plot is $e_{(1)} = -289.044$ on the horizontal scale versus $p_{(1)} = (.0165)(100) = 1.65$ on the vertical scale, and the second point plotted is $e_{(2)} = -259.958$ on the horizontal scale versus $p_{(2)} = (.0413)(100) = 4.13$ on the vertical scale. It is important to note that the scale on the vertical axis does not have the usual spacing between the percentages. The spacing reflects the distance between the z-scores that correspond to the percentages in the standard normal distribution. Hence, if we wished to create the plot in Figure 14.24(d) by hand, we would need special graphing paper with this vertical scale.

It can be proven that, if the normality assumption holds, then the expected value of the ith ordered residual $e_{(i)}$ is proportional to $z_{(i)}$. Therefore, a plot of the $e_{(i)}$ values on the horizontal scale versus the $z_{(i)}$ values on the vertical scale (or equivalently, the $e_{(i)}$ values on the horizontal scale versus the $p_{(i)}$ values on the vertical scale) should have a straight-line appearance. That is, if the normality assumption holds, then the normal plot should have a straight-line appearance. A normal plot that does not look like a straight line (admittedly, a subjective decision) indicates that the normality assumption is violated. Because the normal plots in Figure 14.24 have some curvature (particularly in the upper right portion), there is a possible violation of the normality assumption.

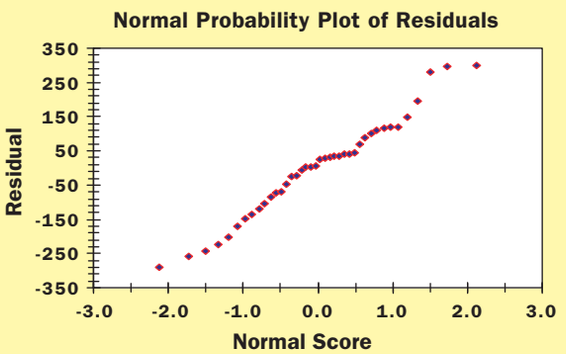
FIGURE 14.24 Stem-and-Leaf Display and Normal Plots of the Residuals from the Simple Linear Regression Model Describing the QHIC Data

```
Stem-and-leaf of RESI1 N = 40
Leaf Unit = 10
 2     -2  85
 5     -2  420
 6     -1  7
 10    -1  4320
 13    -0  876
 17    -0  4220
(11)    0  00022333344
 12     0  68
 10     1  001124
 4      1  9
 3      2
 3      2  899
```

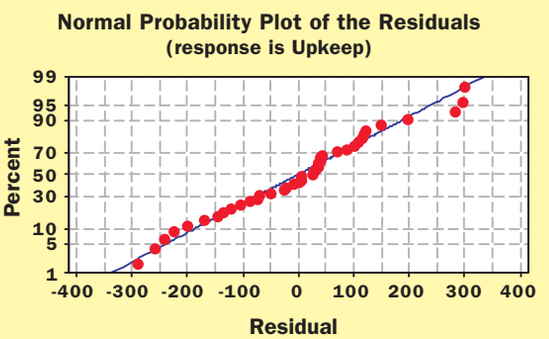
(a) MINITAB output of the stem-and-leaf display



(b) Calculating $z_{(1)}$ for a normal plot



(c) Excel add-in (MegaStat) normal plot



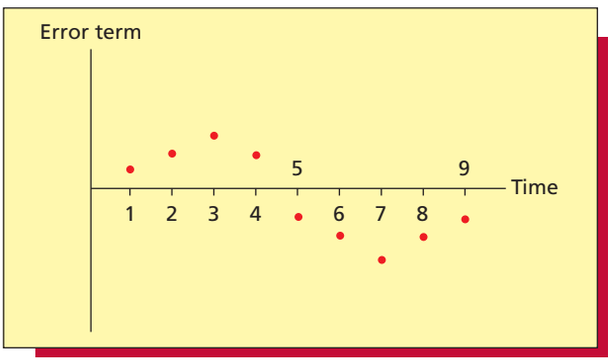
(d) MINITAB normal plot

It is important to realize that violations of the constant variance and correct functional form assumptions can often cause a histogram and/or stem-and-leaf display of the residuals to look nonnormal and can cause the normal plot to have a curved appearance. Because of this, it is usually a good idea to use residual plots to check for nonconstant variance and incorrect functional form before making any final conclusions about the normality assumption. Later in this section we discuss a procedure that sometimes remedies simultaneous violations of the constant variance, correct functional form, and normality assumptions.

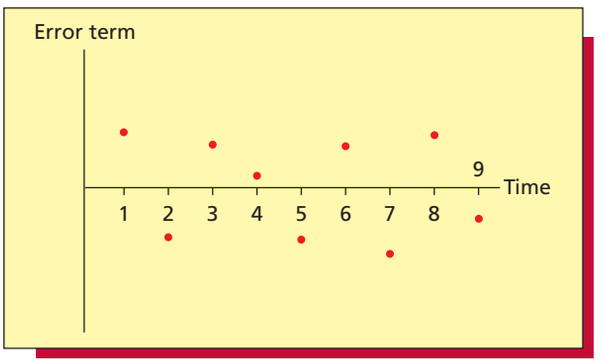
The independence assumption The independence assumption is most likely to be violated when the regression data are **time series data**—that is, data that have been collected in a time sequence. For such data the time-ordered error terms can be **autocorrelated.** Intuitively, we say that error terms occurring over time have **positive autocorrelation** if a positive error term in time period i tends to produce, or be followed by, another positive error term in time period $i + k$ (some later time period) and if a negative error term in time period i tends to produce, or be followed by, another negative error term in time period $i + k$. In other words, positive autocorrelation exists when positive error terms tend to be followed over time by positive error terms and when negative error terms tend to be followed over time by negative error terms. Positive autocorrelation in the error terms is depicted in Figure 14.25(a), which illustrates that **positive autocorrelation can produce a cyclical error term pattern over time.** The simple linear regression model implies that a positive error term produces a greater-than-average value of y and a negative error term produces a smaller-than-average value of y. It follows that positive autocorrelation in the error terms means that greater-than-average values of y tend to be followed by greater-than-average values of y, and smaller-than-average values of y tend to be followed by smaller-than-average values of y. An example of positive autocorrelation could hypothetically be

FIGURE 14.25 Positive and Negative Autocorrelation

(a) Positive Autocorrelation in the Error Terms: Cyclical Pattern



(b) Negative Autocorrelation in the Error Terms: Alternating Pattern



provided by a simple linear regression model relating demand for a product to advertising expenditure. Here we assume that the data are time series data observed over a number of consecutive sales periods. One of the factors included in the error term of the simple linear regression model is competitors' advertising expenditure for their similar products. If, for the moment, we assume that competitors' advertising expenditure significantly affects the demand for the product, then a higher-than-average competitors' advertising expenditure probably causes demand for the product to be lower than average and hence probably causes a negative error term. On the other hand, a lower-than-average competitors' advertising expenditure probably causes the demand for the product to be higher than average and hence probably causes a positive error term. If, then, competitors tend to spend money on advertising in a cyclical fashion—spending large amounts for several consecutive sales periods (during an advertising campaign) and then spending lesser amounts for several consecutive sales periods—a negative error term in one sales period will tend to be followed by a negative error term in the next sales period, and a positive error term in one sales period will tend to be followed by a positive error term in the next sales period. In this case the error terms would display positive autocorrelation, and thus these error terms would not be statistically independent.

Intuitively, error terms occurring over time have **negative autocorrelation** if a positive error term in time period i tends to produce, or be followed by, a negative error term in time period $i + k$ and if a negative error term in time period i tends to produce, or be followed by, a positive error term in time period $i + k$. In other words, negative autocorrelation exists when positive error terms tend to be followed over time by negative error terms and negative error terms tend to be followed over time by positive error terms. An example of negative autocorrelation in the error terms is depicted in Figure 14.25(b), which illustrates that **negative autocorrelation in the error terms can produce an alternating pattern over time.** It follows that negative autocorrelation in the error terms means that greater-than-average values of y tend to be followed by smaller-than-average values of y and smaller-than-average values of y tend to be followed by greater-than-average values of y. An example of negative autocorrelation might be provided by a retailer's weekly stock orders. Here a larger-than-average stock order one week might result in an oversupply and hence a smaller-than-average order the next week.

The **independence assumption** basically says that the time-ordered error terms display no positive or negative autocorrelation. This says that **the error terms occur in a random pattern over time.** Such a random pattern would imply that the error terms (and their corresponding y values) are statistically independent.

Because the residuals are point estimates of the error terms, a residual plot versus time is used to check the independence assumption. If a residual plot versus the data's time sequence has a cyclical appearance, the error terms are positively autocorrelated, and the independence assumption is violated. If a plot of the time-ordered residuals has an alternating pattern, the error terms are negatively autocorrelated, and again the independence assumption is violated.

However, if a plot of the time-ordered residuals displays a random pattern, the error terms have little or no autocorrelation. In such a case, it is reasonable to conclude that the independence assumption holds.

EXAMPLE 14.12 Pages Bookstore: Positive Autocorrelation

Figure 14.26(a) presents data concerning weekly sales at Pages Bookstore (Sales), Pages weekly advertising expenditure (Adver), and the weekly advertising expenditure of Pages main competitor (Compadv). Here the sales values are expressed in thousands of dollars, and the advertising expenditure values are expressed in hundreds of dollars. Figure 14.26(a) also gives the MINITAB output of the residuals that are obtained when a simple linear regression analysis is performed relating Pages sales to Pages advertising expenditure. These residuals are plotted versus time in Figure 14.26(b). We see that the residual plot has a cyclical pattern. This tells us that the error terms for the model are positively autocorrelated and the independence assumption is violated. Furthermore, there tend to be positive residuals when the competitor's advertising expenditure is lower (in weeks 1 through 8 and weeks 14, 15, and 16) and negative residuals when the competitor's advertising expenditure is higher (in weeks 9 through 13). Therefore, the competitor's advertising expenditure seems to be causing the positive autocorrelation.

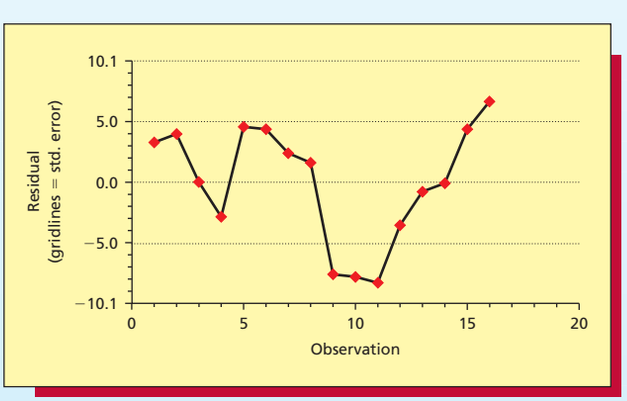
FIGURE 14.26 Pages Bookstore Sales and Advertising Data, and Residual Analysis

(a) The data and the MINITAB output of the residuals from a simple linear regression relating Pages sales to Pages advertising expenditure BookSales

Observation	Adver	Compadv	Sales	Predicted	Residual
1	18	10	22	18.7	3.3
2	20	10	27	23.0	4.0
3	20	15	23	23.0	−0.0
4	25	15	31	33.9	−2.9
5	28	15	45	40.4	4.6
6	29	20	47	42.6	4.4
7	29	20	45	42.6	2.4
8	28	25	42	40.4	1.6
9	30	35	37	44.7	−7.7
10	31	35	39	46.9	−7.9
11	34	35	45	53.4	−8.4
12	35	30	52	55.6	−3.6
13	36	30	57	57.8	−0.8
14	38	25	62	62.1	−0.1
15	41	20	73	68.6	4.4
16	45	20	84	77.3	6.7

Durbin-Watson = 0.65

(b) A plot of the residuals in Figure 14.26(a) versus time



To conclude this example, note that the simple linear regression model relating Pages sales to Pages advertising expenditure has a standard error, s, of 5.038. The residual plot in Figure 14.26(b) includes grid lines that are placed one and two standard errors above and below the residual mean of 0. Such grid lines help us to better diagnose potential violations of the regression assumptions.

When the independence assumption is violated, various remedies can be employed. One approach is to identify which independent variable left in the error term (for example, competitors' advertising expenditure) is causing the error terms to be autocorrelated. We can then remove this independent variable from the error term and insert it directly into the regression model, forming a **multiple regression model.** (Multiple regression models are discussed in Chapter 15.)

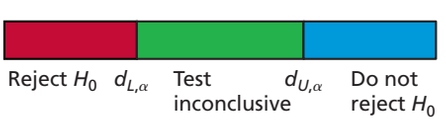
The Durbin–Watson test One type of positive or negative autocorrelation is called **first-order autocorrelation.** It says that ε_t, the error term in time period t, is related to ε_{t-1}, the error term in time period $t-1$. To check for first-order autocorrelation, we can use the **Durbin–Watson statistic**

$$d = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

where $e_1, e_2, \ldots, e_n$ are the time-ordered residuals.

Intuitively, small values of d lead us to conclude that there is positive autocorrelation. This is because, if d is small, the differences $(e_t - e_{t-1})$ are small. This indicates that the adjacent residuals e_t and e_{t-1} are of the same magnitude, which in turn says that the adjacent error terms ε_t and ε_{t-1} are positively correlated. Consider testing the null hypothesis $\boldsymbol{H_0}$ **that the error terms are not autocorrelated** versus the alternative hypothesis $\boldsymbol{H_a}$ **that the error terms are positively autocorrelated.** Durbin and Watson have shown that there are points (denoted $d_{L,\alpha}$ and $d_{U,\alpha}$) such that, if α is the probability of a Type I error, then

1. If $d < d_{L,\alpha}$, we reject H_0.
2. If $d > d_{U,\alpha}$, we do not reject H_0.
3. If $d_{L,\alpha} \le d \le d_{U,\alpha}$, the test is inconclusive.



So that the Durbin–Watson test may be easily done, tables containing the points $d_{L,\alpha}$ and $d_{U,\alpha}$ have been constructed. These tables give the appropriate $d_{L,\alpha}$ and $d_{U,\alpha}$ points for various values of α; k, the number of independent variables used by the regression model; and n, the number of observations. Tables A.11, A.12, and A.13 (pages 801–802) give these points for $\alpha = .05$, $\alpha = .025$, and $\alpha = .01$. A portion of Table A.11 is given in Table 14.3. Note that when we are considering a simple linear regression model, which uses *one* independent variable, we look up the points $d_{L,\alpha}$ and $d_{U,\alpha}$ under the heading "$k = 1$." Other values of k are used when we study multiple regression models in Chapter 15. Using the residuals in Figure 14.26(a), the Durbin–Watson statistic for the simple linear regression model relating Pages sales to Pages advertising expenditure is calculated to be

$$\begin{aligned} d &= \frac{\sum_{t=2}^{16} (e_t - e_{t-1})^2}{\sum_{t=1}^{16} e_t^2} \\ &= \frac{(4.0 - 3.3)^2 + (0.0 - 4.0)^2 + \cdots + (6.7 - 4.4)^2}{(3.3)^2 + (4.0)^2 + \cdots + (6.7)^2} \\ &= .65 \end{aligned}$$

A MINITAB output of the Durbin–Watson statistic is given at the bottom of Figure 14.26(a). To test for positive autocorrelation, we note that there are $n = 16$ observations and the regression

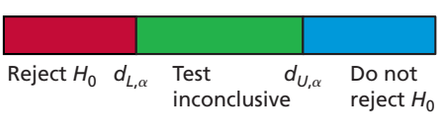
TABLE 14.3 Critical Values for the Durbin–Watson d Statistic ($\alpha = .05$)

	$k = 1$		$k = 2$		$k = 3$		$k = 4$	
n	$d_{L,.05}$	$d_{U,.05}$	$d_{L,.05}$	$d_{U,.05}$	$d_{L,.05}$	$d_{U,.05}$	$d_{L,.05}$	$d_{U,.05}$
15	1.08	1.36	0.95	1.54	0.82	1.75	0.69	1.97
16	1.10	1.37	0.98	1.54	0.86	1.73	0.74	1.93
17	1.13	1.38	1.02	1.54	0.90	1.71	0.78	1.90
18	1.16	1.39	1.05	1.53	0.93	1.69	0.82	1.87
19	1.18	1.40	1.08	1.53	0.97	1.68	0.86	1.85
20	1.20	1.41	1.10	1.54	1.00	1.68	0.90	1.83

model uses $k = 1$ independent variable. Therefore, if we set $\alpha = .05$, Table 14.3 tells us that $d_{L,.05} = 1.10$ and $d_{U,.05} = 1.37$. Because $d = .65$ is less than $d_{L,.05} = 1.10$, we reject the null hypothesis of no autocorrelation. That is, we conclude (at an α of .05) that there is positive (first-order) autocorrelation.

It can be shown that the Durbin–Watson statistic d is always between 0 and 4. Large values of d (and hence small values of $4 - d$) lead us to conclude that there is negative autocorrelation because if d is large, this indicates that the differences $(e_t - e_{t-1})$ are large. This says that the adjacent error terms ε_t and ε_{t-1} are negatively autocorrelated. Consider testing the null hypothesis $\boldsymbol{H_0}$ **that the error terms are not autocorrelated** versus the alternative hypothesis $\boldsymbol{H_a}$ **that the error terms are negatively autocorrelated.** Durbin and Watson have shown that based on setting the probability of a Type I error equal to α, the points $d_{L,\alpha}$ and $d_{U,\alpha}$ are such that

1. If $(4 - d) < d_{L,\alpha}$, we reject H_0.
2. If $(4 - d) > d_{U,\alpha}$, we do not reject H_0.
3. If $d_{L,\alpha} \leq (4 - d) \leq d_{U,\alpha}$, the test is inconclusive.



As an example, for the Pages sales simple linear regression model, we see that

$$(4 - d) = (4 - .65) = 3.35 > d_{U,.05} = 1.37$$

Therefore, on the basis of setting α equal to .05, we do not reject the null hypothesis of no autocorrelation. That is, there is no evidence of negative (first-order) autocorrelation.

We can also use the Durbin–Watson statistic to test for positive or negative autocorrelation. Specifically, consider testing the null hypothesis $\boldsymbol{H_0}$ **that the error terms are not autocorrelated** versus the alternative hypothesis $\boldsymbol{H_a}$ **that the error terms are positively or negatively autocorrelated.** Durbin and Watson have shown that, based on setting the probability of a Type I error equal to α, we perform both the above described test for positive autocorrelation and the above described test for negative autocorrelation by using the critical values $d_{L,\alpha/2}$ and $d_{U,\alpha/2}$ for each test. If **either test** says to reject H_0, then we reject H_0. If **both tests** say to not reject H_0, then we do not reject H_0. Finally, if **either test** is inconclusive, then the overall test is inconclusive. For example, consider testing for positive or negative autocorrelation in the Pages sales model. If we set α equal to .05, then $\alpha/2 = .025$, and we need to find the points $d_{L,.025}$ and $d_{U,.025}$ when $n = 16$ and $k = 1$. Looking up these points in Table A.12 (page 801), we find that $d_{L,.025} = .98$ and $d_{U,.025} = 1.24$. Because $d = .65$ is less than $d_{L,.025} = .98$, we reject the null hypothesis of no autocorrelation. That is, we conclude (at an α of .05) that there is first-order autocorrelation.

Although we have used the Pages sales model in these examples to demonstrate the Durbin–Watson tests for (1) positive autocorrelation, (2) negative autocorrelation, and (3) positive or negative autocorrelation, we must in practice choose one of these Durbin–Watson tests in a particular situation. Because positive autocorrelation is more common in real time series data than negative autocorrelation, the Durbin–Watson test for positive autocorrelation is used more often than the other two tests. Also, note that each Durbin–Watson test assumes that the population of all possible residuals at any time t has a normal distribution.

Transforming the dependent variable: A possible remedy for violations of the constant variance, correct functional form, and normality assumptions In general, if a data or residual plot indicates that the error variance of a regression model increases as an

independent variable or the predicted value of the dependent variable increases, then we can sometimes remedy the situation by transforming the dependent variable. One transformation that works well is to take each y value to a fractional power. As an example, we might use a transformation in which we take the square root (or one-half power) of each y value. Letting y^* denote the value obtained when the transformation is applied to y, we would write the **square root transformation** as

$$y^* = \sqrt{y} = y^{.5}$$

Another commonly used transformation is the **quartic root transformation.** Here we take each y value to the one-fourth power. That is,

$$y^* = y^{.25}$$

If we consider a transformation that takes each y value to a fractional power (such as .5, .25, or the like), as the power approaches 0, the transformed value y^* approaches the natural logarithm of y (commonly written ln y). In fact, we sometimes use the **logarithmic transformation**

$$y^* = \ln y$$

which takes the natural logarithm of each y value. In general, when we take a fractional power (including the natural logarithm) of the dependent variable, the transformation not only tends to equalize the error variance but also tends to "straighten out" certain types of nonlinear data plots. Specifically, if a data plot indicates that the dependent variable is increasing at an increasing rate, then a fractional power transformation tends to straighten out the data plot. A fractional power transformation can also help to remedy a violation of the normality assumption. Because we cannot know which fractional power to use before we actually take the transformation, we recommend taking all of the square root, quartic root, and natural logarithm transformations and seeing which one best equalizes the error variance and (possibly) straightens out a nonlinear data plot.

EXAMPLE 14.13 The QHIC Case: Using a Data Transformation

Consider the QHIC upkeep expenditures in Figure 14.21. In Figures 14.27, 14.28, and 14.29 we show the plots that result when we take the square root, quartic root, and natural logarithmic transformations of the upkeep expenditures and plot the transformed values versus the home values. The square root transformation seems to best equalize the error variance and straighten out the curved data plot in Figure 14.21. Note that the natural logarithm transformation seems to

FIGURE 14.27 MINITAB Plot of the Square Roots of the Upkeep Expenditures versus the Home Values

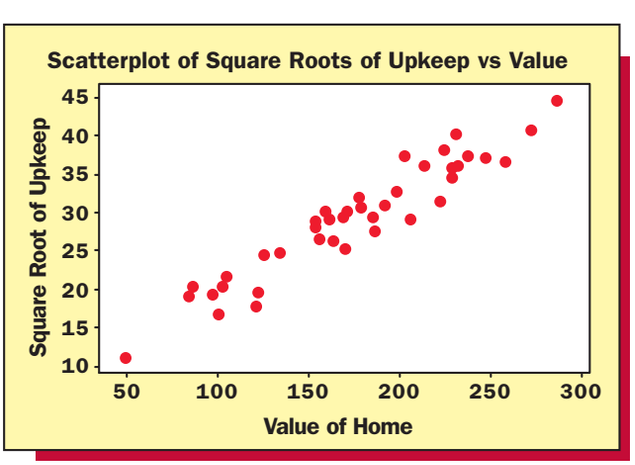


FIGURE 14.28 MINITAB Plot of the Quartic Roots of the Upkeep Expenditures versus the Home Values

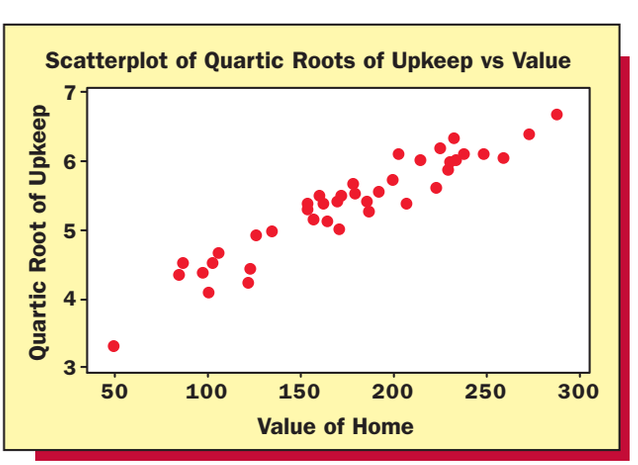


FIGURE 14.29 MINITAB Plot of the Natural Logarithms of the Upkeep Expenditures versus the Home Values

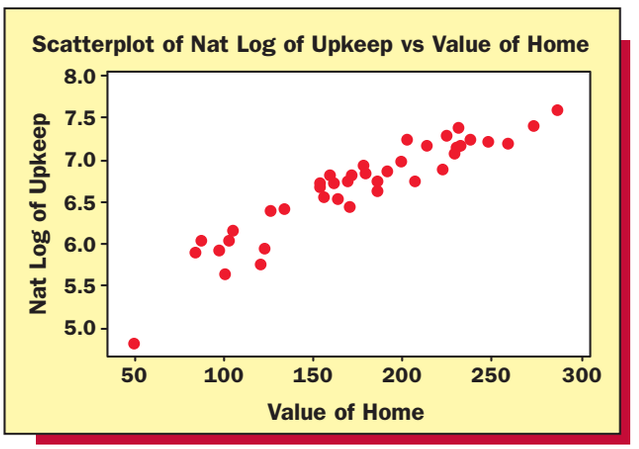


FIGURE 14.30 MINITAB Output of a Regression Analysis of the Upkeep Expenditure Data by Using the Model $y^* = \beta_0 + \beta_1 x + \varepsilon$ where $y^* = y^{.5}$, and a Residual Plot versus x

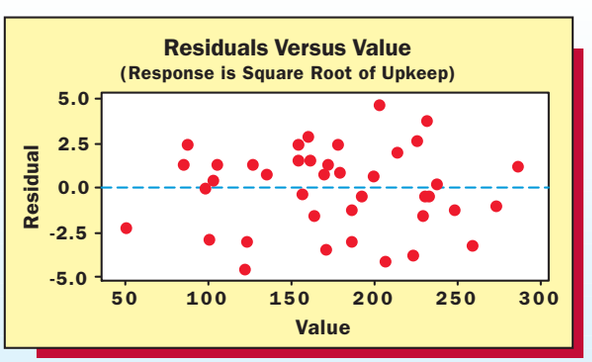
```
The regression equation is
SqRtUpkeep = 7.20 + 0.127 Value

Predictor       Coef    SE Coef       T      P
Constant       7.201      1.205    5.98  0.000
Value       0.127047   0.006577   19.32  0.000

S = 2.32479    R-Sq = 90.8%    R-Sq(adj) = 90.5%

Analysis of Variance
Source           DF      SS      MS       F      P
Regression        1  2016.8  2016.8  373.17  0.000
Residual Error   38   205.4     5.4
Total            39  2222.2

Values of Predictors for New Obs    Predicted Values for New Observations
New Obs  Value                      New Obs     Fit  SE Fit       95% CI            95% PI
      1    220                            1  35.151   0.474  (34.191, 36.111)  (30.348, 39.954)
```



"overtransform" the data—the error variance tends to decrease as the home value increases and the data plot seems to "bend down." The plot of the quartic roots indicates that the quartic root transformation also seems to overtransform the data (but not by as much as the logarithmic transformation). In general, as the fractional power gets smaller, the transformation gets stronger. Different fractional powers are best in different situations.

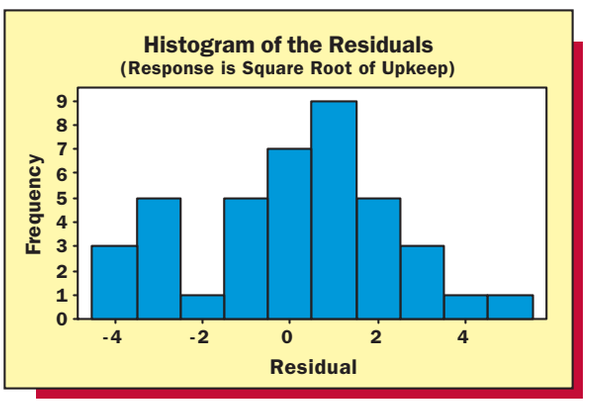
Because the plot in Figure 14.27 of the square roots of the upkeep expenditures versus the home values has a straight-line appearance, we consider the model

$$y^* = \beta_0 + \beta_1 x + \varepsilon \quad \text{where } y^* = y^{.5}$$

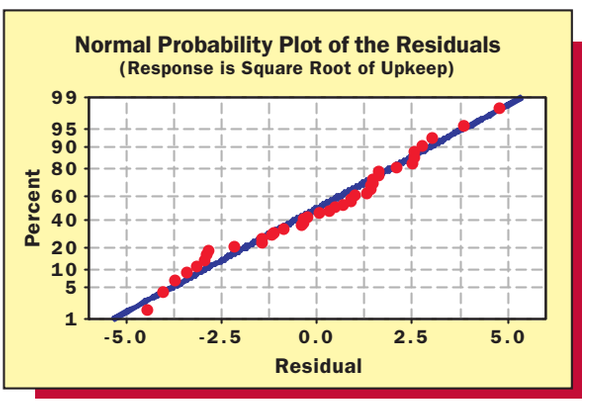
The MINITAB output of a regression analysis using this transformed model is given in Figure 14.30, along with a plot of the transformed model's residuals versus x. Note that this residual plot has a horizontal band appearance. It can also be verified that the transformed model's residual plot versus $\hat{y}$, which we do not give here, has a similar horizontal band appearance.

FIGURE 14.31 MINITAB Output of Normality Assumption Analysis for the Upkeep Expenditure Model $y^* = \beta_0 + \beta_1 x + \varepsilon$ where $y^* = y^{.5}$

(a) Histogram of the residuals



(b) Normal plot of the residuals



Therefore, we conclude that the constant variance and the correct functional form assumptions approximately hold for the transformed model. Next, note that the histogram of the transformed model's residuals in Figure 14.31(a) looks reasonably bell-shaped and symmetric, and note that the normal plot of these residuals in Figure 14.31(b) looks straighter than the normal plots for the untransformed model (see Figure 14.24 on page 531). Therefore, we also conclude that the normality assumption approximately holds for the transformed model.

Because the regression assumptions approximately hold for the transformed regression model, we can use this model to make statistical inferences. Consider a home worth \$220,000. Using the least squares point estimates on the MINITAB output in Figure 14.30 on the previous page, it follows that a point prediction of y^* for such a home is

$$\hat{y}^* = 7.201 + .127047(220)$$
$$= 35.151$$

This point prediction is given at the bottom of the MINITAB output, as is the 95 percent prediction interval for y^*, which is [30.348, 39.954]. It follows that a point prediction of the upkeep expenditure for a home worth \$220,000 is $(35.151)^2 = \$1{,}235.59$ and that a 95 percent prediction interval for this upkeep expenditure is $[(30.348)^2, (39.954)^2] = [\$921.00, \$1596.32]$. Suppose that QHIC wishes to send an advertising brochure to any home that has a predicted upkeep expenditure of at least \$500. It follows that a home worth \$220,000 would be sent an advertising brocedure. This is because the predicted yearly upkeep expenditure for such a home is (as just calculated) \$1,235.59. Other homes can be evaluated in a similar fashion.

BI

Exercises for Section 14.9

connect

CONCEPTS

14.56 In regression analysis, what variables should the residuals be plotted against?

14.57 What patterns in residual plots indicate violations of the regression assumptions?

14.58 In regression analysis, how do we check the normality assumption?

METHODS AND APPLICATIONS

14.59 **THE SERVICE TIME CASE** DS SrvcTime

The residuals given by the service time prediction equation $\hat{y} = 11.4641 + 24.6022x$ are listed in Table 14.4(a), and residual plots versus x and $\hat{y}$ are given in Figures 14.32(a) and (b). Do the plots indicate any violations of the regression assumptions? Explain.

TABLE 14.4 Service Time Model Residuals and Normal Plot Calculations

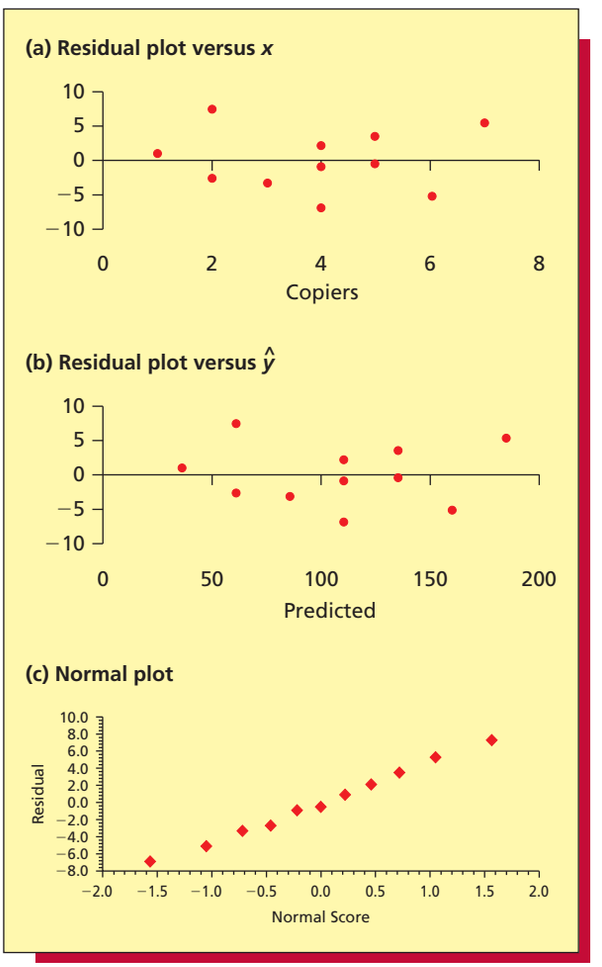
(a) Predicted values and residuals using $\hat{y} = 11.4641 + 24.6022x$

Observation	Copiers	Minutes	Predicted	Residual
1	4	109.0	109.9	−0.9
2	2	58.0	60.7	−2.7
3	5	138.0	134.5	3.5
4	7	189.0	183.7	5.3
5	1	37.0	36.1	0.9
6	3	82.0	85.3	−3.3
7	4	103.0	109.9	−6.9
8	5	134.0	134.5	−0.5
9	2	68.0	60.7	7.3
10	6	112.0	109.9	2.1
11	4	154.0	159.1	−5.1

(b) Ordered residuals and normal plot calculations

i	Ordered Residual, $e_{(i)}$	$\frac{3i-1}{3n+1}$	$z_{(i)}$
1	−6.9	.0588	−1.565
2	−5.1	.1470	−1.05
3	−3.3	.2353	−.72
4	−2.7	.3235	−.46
5	−0.9	.4118	−.22
6	−0.5	.5000	0
7	0.9	.5882	.22
8	2.1	.6765	.46
9	3.5	.7647	.72
10	5.3	.8529	1.05
11	7.3	.9412	1.565

FIGURE 14.32 Service Time Model Residual and Normal Plots



14.60 THE SERVICE TIME CASE SrvcTime

The residuals given by the service time prediction equation $\hat{y} = 11.4641 + 24.6022x$ are listed in Table 14.4(a).

a In this exercise we construct a normal plot of these residuals. To construct this plot, we must first arrange the residuals in order from smallest to largest. These ordered residuals are given in Table 14.4(b). Denoting the ith ordered residual as $e_{(i)}$ ($i = 1, 2, \ldots, 11$), we next compute for each value of i the point $z_{(i)}$. These computations are summarized in Table 14.4(b). Show how $z_{(4)} = -.46$ and $z_{(10)} = 1.05$ have been obtained.

b The ordered residuals (the $e_{(i)}$'s) are plotted against the $z_{(i)}$'s in Figure 14.32(c). Does this figure indicate a violation of the normality assumption? Explain.

14.61 A simple linear regression model is employed to analyze the 24 monthly observations given in Table 14.5 on the next page. Residuals are computed and are plotted versus time. The resulting residual plot is shown in Figure 14.33 on the next page. **(1)** Discuss why the residual plot suggests the existence of positive autocorrelation. **(2)** The Durbin–Watson statistic d can be calculated to be .473. Test for positive (first-order) autocorrelation at $\alpha = .05$, and test for negative (first-order) autocorrelation at $\alpha = .05$ SalesAdv

14.62 THE LAPTOP SERVICE TIME CASE SrvcTime2

The page margin shows data concerning the time, y, required to perform service (in minutes) and the number of laptop computers serviced, x, for 15 service calls. Figure 14.34 on the next page gives a plot of y versus x, and Figure 14.35(a) gives the Excel output of a plot of the residuals versus x for a simple linear regression model. What regression assumption appears to be violated? Explain.

The Laptop Service Time Data

Service Time, y	Laptops Serviced, x
92	3
63	2
126	6
247	8
49	2
90	4
119	5
114	6
67	2
115	4
188	6
298	11
77	3
151	10
27	1

TABLE 14.5 Sales and Advertising Data for Exercise 14.61 DS SalesAdv

Month	Monthly Total Sales, y	Advertising Expenditures, x
1	202.66	116.44
2	232.91	119.58
3	272.07	125.74
4	290.97	124.55
5	299.09	122.35
6	296.95	120.44
7	279.49	123.24
8	255.75	127.55
9	242.78	121.19
10	255.34	118.00
11	271.58	121.81
12	268.27	126.54
13	260.51	129.85
14	266.34	122.65
15	281.24	121.64
16	286.19	127.24
17	271.97	132.35
18	265.01	130.86
19	274.44	122.90
20	291.81	117.15
21	290.91	109.47
22	264.95	114.34
23	228.40	123.72
24	209.33	130.33

Source: *Forecasting Methods and Applications,* "Sales and Advertising Data," by S. Makridakis, S. C. Wheelwright, and V. E. McGee, *Forecasting: Methods and Applications* (Copyright © 1983 John Wiley & Sons, Inc.). Reprinted by permission of John Wiley & Sons, Inc.

FIGURE 14.33 Residual Plot for Exercise 14.61

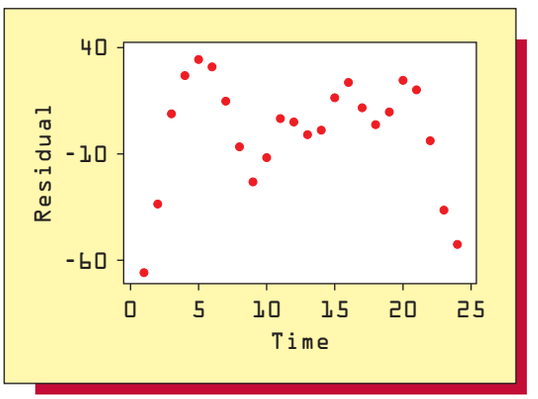


FIGURE 14.34 The Laptop Service Time Scatter Plot (for Exercise 14.62) DS SrvcTime2

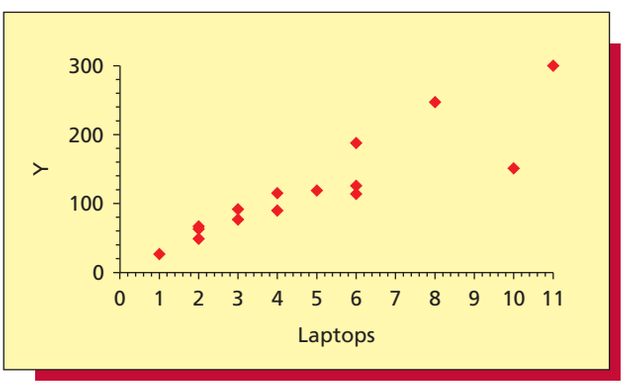
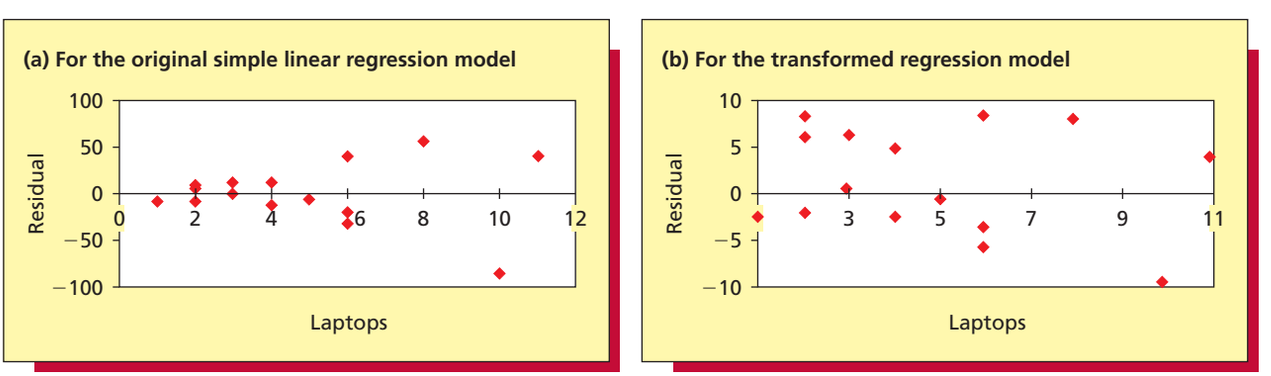


FIGURE 14.35 Residual Plots in the Laptop Service Time Case (for Exercises 14.62 and 14.63).



14.63 THE LAPTOP SERVICE TIME CASE DS SrvcTime2

Figure 14.35(a) shows the residual plot versus x for the simple linear regression of the laptop service time data. This plot fans out, indicating that the error term ε tends to become larger in magnitude as x increases. To remedy this violation of the constant variance assumption, we divide all terms in the simple linear regression model by x. This gives the transformed model

$$\frac{y}{x} = \beta_0\left(\frac{1}{x}\right) + \beta_1 + \frac{\varepsilon}{x} \quad \text{or, equivalently,} \quad \frac{y}{x} = \beta_0 + \beta_1\left(\frac{1}{x}\right) + \frac{\varepsilon}{x}$$

FIGURE 14.36 MINITAB Output of a Regression Analysis of the Laptop Service Time Data Using the Model $y/x = \beta_0 + \beta_1(1/x) + \varepsilon/x$

```
The regression equation is
Y/X = 24.0 + 6.76 1/X

Predictor       Coef     SE Coef         T         P
Constant      24.041       2.246     10.70     0.000
1/X            6.764       5.794      1.17     0.264

          S = 5.15816     R-Sq = 9.50%     R-Sq(adj) = 2.5%

Analysis of Variance
Source            DF         SS         MS        F       P
Regression         1      36.27      36.27     1.36   0.264
Residual Error    13     345.89      26.61
Total             14     382.15

Values of Predictors for New Obs  Predicted Values for New Observations
New Obs           1/X              Obs       Fit     SE Fit            95% CI            95% PI
      1         0.143                1     25.01       1.65     (21.43, 28.58)    (13.30, 36.71)
```

Figure 14.36 and Figure 14.35(b) give a regression output and a residual plot versus x for this model. Does the residual plot indicate that the constant variance assumption holds for the transformed model? Explain.

14.64 THE LAPTOP SERVICE TIME CASE SrvcTime2

Consider a future service call on which seven laptops will be serviced. Let μ_0 represent the mean service time for all service calls on which seven laptops will be serviced, and let y_0 represent the actual service time for an individual service call on which seven laptops will be serviced. The bottom of the MINITAB output in Figure 14.36 tells us that $\hat{y}/7 = 24.041 + 6.764(1/7) = 25.01$ is a point estimate of $\mu_0/7$ and a point prediction of $y_0/7$. **(1)** Multiply this result by 7 to obtain $\hat{y}$. **(2)** Multiply the ends of the confidence interval and prediction interval shown on the MINITAB output by 7. This will give a 95 percent confidence interval for μ_0 and a 95 percent prediction interval for y_0. **(3)** If the number of minutes we will allow for the future service call is the upper limit of the 95 percent confidence interval for μ_0, how many minutes will we allow?

14.65 USING A NATURAL LOGARITHM TRANSFORMATION WestStk

Western Steakhouses, a fast-food chain, opened 15 years ago. Each year since then the number of steakhouses in operation, y, was recorded. An analyst for the firm wishes to use these data to predict the number of steakhouses that will be in operation next year. The data are given in the page margin, and a plot of the data is given in Figure 14.37(a) on the next page. Examining the data plot, we see that the number of steakhouse openings has increased over time at an increasing rate and with increasing variation. A plot of the natural logarithms of the steakhouse values versus time (see Figure 14.37(b) on the next page) has a straight-line appearance with constant variation. Therefore, we consider the model $\ln y_t = \beta_0 + \beta_1 t + \varepsilon_t$. If we use MINITAB, we find that the least squares point estimates of β_0 and β_1 are $b_0 = 2.07012$ and $b_1 = .256880$. We also find that a point prediction of and a 95 percent prediction interval for the natural logarithm of the number of steakhouses in operation next year (year 16) are 6.1802 and [5.9945, 6.3659]. See the MINITAB output in Figure 14.38 on the next page.

The Steakhouse Data

Year, t	Steakhouses, y
1	11
2	14
3	16
4	22
5	28
6	36
7	46
8	67
9	82
10	99
11	119
12	156
13	257
14	284
15	403

a Use the least squares point estimates to calculate the point prediction.

b By exponentiating the point prediction and prediction interval—that is, by calculating $e^{6.1802}$ and $[e^{5.9945}, e^{6.3659}]$—find a point prediction of and a 95 percent prediction interval for the number of steakhouses in operation next year.

c The model $\ln y_t = \beta_0 + \beta_1 t + \varepsilon_t$ is called a **growth curve model** because it implies that

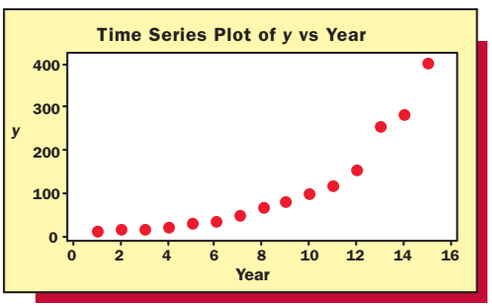
$$y_t = e^{(\beta_0 + \beta_1 t + \varepsilon_t)} = (e^{\beta_0})(e^{\beta_1 t})(e^{\varepsilon_t}) = \alpha_0 \alpha_1^t \eta_t$$

where $\alpha_0 = e^{\beta_0}$, $\alpha_1 = e^{\beta_1}$, and $\eta_t = e^{\varepsilon_t}$. Here $\alpha_1 = e^{\beta_1}$ is called the **growth rate** of the y values. Noting that the least squares point estimate of β_1 is $b_1 = .256880$, estimate the growth rate α_1.

d We see that $y_t = \alpha_0 \alpha_1^t \eta_t = (\alpha_0 \alpha_1^{t-1})\alpha_1 \eta_t \approx (y_{t-1})\alpha_1 \eta_t$. This says that y_t is expected to be approximately α_1 times y_{t-1}. Noting this, interpret the growth rate of part (c).

FIGURE 14.37 Data Plots for Exercise 14.65 WestStk

(a) Time Series Plot of *y* versus *t*



(b) Time Series Plot of Natural Logarithm of *y* versus *t*

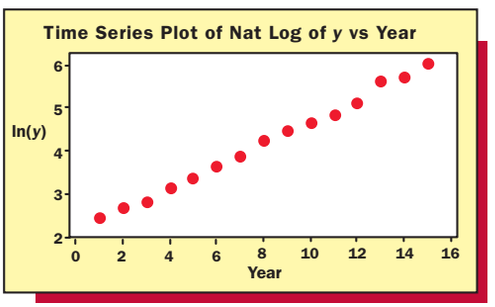


FIGURE 14.38 MINITAB Output of a Regression Analysis of the Steakhouse Data Using the Model $\ln y_t = \beta_0 + \beta_1 t + \varepsilon_t$ (for Exercise 14.65)

```
The regression equation is
ln(y) = 2.07 + 0.257 Year

Predictor        Coef     SE Coef        T        P
Constant      2.07012     0.04103    50.45    0.000
Year         0.256880    0.004513    56.92    0.000

S = 0.0755161   R-Sq = 99.6%   R-Sq(adj) = 99.6%   Durbin-Watson statistic = 1.87643

Analysis of Variance
Source            DF        SS        MS          F        P
Regression         1    18.477    18.477    3239.97    0.000
Residual Error    13     0.074     0.006
Total             14    18.551

Values of Predictors for New Obs   Predicted Values for New Observations
New Obs  Year                      Obs     Fit  SE Fit       95% CI            95% PI
      1    16                        1  6.1802  0.0410  (6.0916, 6.2689)  (5.9945, 6.3659)
```

14.10 Some Shortcut Formulas (Optional)

Calculating the sum of squared residuals A shortcut formula for the sum of squared residuals is

$$SSE = SS_{yy} - \frac{SS_{xy}^2}{SS_{xx}}$$

where

$$SS_{yy} = \sum (y_i - \bar{y})^2 = \sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}$$

For example, consider the Tasty Sub Shop case. If we square each of the ten observed yearly revenues in Table 14.1 (page 488) and add up the resulting squared values, we find that $\Sigma y_i^2 = 7{,}897{,}109.47$. We have also found in Example 14.2 (page 493) that $\Sigma y_i = 8603.1$, $SS_{xy} = 29{,}836.389$ and $SS_{xx} = 1{,}913.129$. It follows that

$$SS_{yy} = \sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n} = 7{,}897{,}109.47 - \frac{(8603.1)^2}{10} = 495{,}776.51$$

and

$$SSE = SS_{yy} - \frac{SS_{xy}^2}{SS_{xx}} = 495{,}776.51 - \frac{(29{,}836.389)^2}{1913.129}$$

$$= 495{,}776.51 - 465{,}316.30 = 30{,}460.21$$

Finally, note that SS_{xy}^2/SS_{xx} equals $b_1 SS_{xy}$. However, we recommend using the first of these expressions, because doing so usually gives less round-off error.

Calculating the total, explained, and unexplained variations The **unexplained variation** is *SSE*, and thus the shortcut formula for *SSE* is a shortcut formula for the unexplained variation. The quantity SS_{yy} defined on page 542 is the **total variation,** and thus the shortcut formula for SS_{yy} is a shortcut formula for the total variation. Lastly, it can be shown that the expression SS_{xy}^2/SS_{xx} equals the **explained variation** and thus is a shortcut formula for this quantity.

Chapter Summary

This chapter has discussed **simple linear regression analysis,** which relates a **dependent variable** to a single **independent** (predictor) **variable.** We began by considering the **simple linear regression model,** which employs two parameters: the **slope** and ***y* intercept.** We next discussed how to compute the **least squares point estimates** of these parameters and how to use these estimates to calculate a **point estimate of the mean value of the dependent variable** and a **point prediction of an individual value** of the dependent variable. Then, after considering the assumptions behind the simple linear regression model, we discussed **testing the significance of the regression relationship (slope),** calculating a **confidence interval** for the mean value of the dependent variable, and calculating a **prediction interval** for an individual value of the dependent variable. We next explained several measures of the utility of the simple linear regression model. These include the **simple coefficient of determination** and an ***F*-test for the simple linear model.** We concluded this chapter by giving a discussion of using **residual analysis** to detect violations of the regression assumptions. We learned that we can sometimes remedy violations of these assumptions by **transforming** the dependent variable.

Glossary of Terms

dependent variable: The variable that is being described, predicted, or controlled. (page 487)

distance value: A measure of the distance between a particular value x_0 of the independent variable x and $\bar{x}$, the average of the previously observed values of x (the center of the experimental region). (page 511)

error term: The difference between an individual value of the dependent variable and the corresponding mean value of the dependent variable. (page 490)

experimental region: The range of the previously observed values of the independent variable. (page 494)

explained variation: A quantity that measures the amount of the total variation in the observed values of y that is explained by the predictor variable x. (pages 517 and 518)

independent variable: A variable used to describe, predict, and control the dependent variable. (page 487)

least squares point estimates: The point estimates of the slope and y intercept of the simple linear regression model that minimize the sum of squared residuals. (pages 492–493)

negative autocorrelation: The situation in which positive error terms tend to be followed over time by negative error terms and negative error terms tend to be followed over time by positive error terms. (page 532)

normal plot: A residual plot that is used to check the normality assumption. (page 530)

positive autocorrelation: The situation in which positive error terms tend to be followed over time by positive error terms and negative error terms tend to be followed over time by negative error terms. (page 531)

residual: The difference between the observed value of the dependent variable and the corresponding predicted value of the dependent variable. (pages 492, 527)

residual plot: A plot of the residuals against some criterion. The plot is used to check the validity of one or more regression assumptions. (page 527)

simple coefficient of determination r^2: The proportion of the total variation in the observed values of the dependent variable that is explained by the simple linear regression model. (page 518)

simple correlation coefficient: A measure of the linear association between two variables. (page 519)

simple linear regression model: An equation that describes the straight-line relationship between a dependent variable and an independent variable. (page 490)

slope (of the simple linear regression model): The change in the mean value of the dependent variable that is associated with a one-unit increase in the value of the independent variable. (page 490)

total variation: A quantity that measures the total amount of variation exhibited by the observed values of the dependent variable y. (pages 517 and 518)

unexplained variation: A quantity that measures the amount of of the total variation in the observed values of y that is not explained by the predictor variable x. (pages 517 and 518)

***y*-intercept (of the simple linear regression model):** The mean value of the dependent variable when the value of the independent variable is 0. (page 490)

Important Formulas and Tests

Simple linear regression model: page 490
Least squares point estimates of β_0 and β_1: pages 492 and 493
Least squares line (prediction equation): page 492
The predicted value of y: page 492
The residual: pages 492 and 527
Sum of squared residuals: pages 492 and 502
Point estimate of a mean value of y: pages 496 and 510
Point prediction of an individual value of y: pages 496 and 510
Mean square error: page 502
Standard error: page 502
Standard error of the estimate b_1: page 503
Testing the significance of the slope: page 504
Testing the significance of the y-intercept: page 506
Confidence interval for the slope: page 506
Standard error of $\hat{y}$: page 511
Confidence interval for a mean value of y: page 511
Prediction interval for an individual value of y: page 511
Explained variation: page 518
Unexplained variation: page 518
Total variation: page 518
Simple coefficient of determination r^2: page 518
Simple correlation coefficient: page 519
Testing the significance of the population correlation coefficient: page 521
An F-test for the simple linear regression model: page 522
Normal plot calculations: page 530
Durbin–Watson test: pages 534–535
Shortcut formulas: pages 542–543

Supplementary Exercises

connect

14.66 Consider the following data concerning the demand (y) and price (x) of a consumer product. Demand

Demand, *y*	252	244	241	234	230	223
Price, *x*	$2.00	$2.20	$2.40	$2.60	$2.80	$3.00

a Plot y versus x. Does it seem reasonable to use the simple linear regression model to relate y to x?
b Calculate the least squares point estimates of the parameters in the simple linear regression model.
c Write the least squares prediction equation. Graph this equation on the plot of y versus x.
d Test the significance of the regression relationship between y and x.
e Find a point prediction of and a 95 percent prediction interval for the demand corresponding to each of the prices $2.10, $2.75, and $3.10.

14.67 In an article in *Public Roads* (1983), Bissell, Pilkington, Mason, and Woods study bridge safety (measured in accident rates per 100 million vehicles) and the **difference** between the width of the bridge and the width of the roadway approach (road plus shoulder):[3] AutoAcc

WidthDiff.	−6	−4	−2	0	2	4	6	8	10	12
Accident	120	103	87	72	58	44	31	20	12	7

The MINITAB output of a simple linear regression analysis relating accident to width difference is as follows:

```
The regression equation is
Accident Rate = 74.7 - 6.44 WidthDif

Predictor        Coef     SE Coef          T          P
Constant       74.727       1.904      39.25      0.000
WidthDif      -6.4424      0.2938     -21.93      0.000

S = 5.33627   R-Sq = 98.4%   R-Sq(adj) = 98.2%

Analysis of Variance
Source             DF        SS        MS          F          P
Regression          1     13697     13697     480.99      0.000
Residual Error      8       228        28
Total               9     13924
```

[3] Source: H. H. Bissell, G. B. Pilkington II, J. M. Mason, and D. L. Woods, "Roadway Cross Section and Alignment," *Public Roads* 46 (March 1983), pp. 132–41.

Using the MINITAB output

a Identify and interpret the least squares point estimate of the slope of the simple linear regression model.

b Identify and interpret the p-value for testing H_0: $\beta_1 = 0$ versus H_a: $\beta_1 \neq 0$.

c Identify and interpret r^2.

14.68 The data in Table 14.6 concerning the relationship between smoking and lung cancer death are presented in a course of The Open University, *Statistics in Society,* Unit C4, The Open University Press, Milton Keynes, England, 1983. The original source of the data is *Occupational Mortality: The Registrar General's Decennial Supplement for England and Wales, 1970–1972,* Her Majesty's Stationery Office, London, 1978. In the table, a smoking index greater (less) than 100 indicates that men in the occupational group smoke more (less) than average when compared to all men of the same age. Similarly, a lung cancer death index greater (less) than 100 indicates that men in the occupational group have a greater (less) than average lung cancer death rate when compared to all men of the same age. In Figure 14.39 we present a portion of a MINITAB output of a simple linear regression analysis relating the lung cancer death index to the smoking index. In Figure 14.40 we present a plot of the lung cancer death index versus the smoking index. DS Smoking

TABLE 14.6 The Smoking and Lung Cancer Death Data DS Smoking

Occupational Group	Smoking Index	Lung Cancer Death Index
Farmers, foresters, and fisherman	77	84
Miners and quarrymen	137	116
Gas, coke, and chemical makers	117	123
Glass and ceramics makers	94	128
Furnace, forge, foundry, and rolling mill workers	116	155
Electrical and electronics workers	102	101
Engineering and allied trades	111	118
Woodworkers	93	113
Leather workers	88	104
Textile workers	102	88
Clothing workers	91	104
Food, drink, and tobacco workers	104	129
Paper and printing workers	107	86
Makers of other products	112	96
Construction workers	113	144
Painters and decorators	110	139
Drivers of stationary engines, cranes, etc.	125	113
Laborers not included elsewhere	133	146
Transport and communications workers	115	128
Warehousemen, storekeepers, packers, and bottlers	105	115
Clerical workers	87	79
Sales workers	91	85
Service, sport, and recreation workers	100	120
Administrators and managers	76	60
Professionals, technical workers, and artists	66	51

FIGURE 14.39 MINITAB Output of a Simple Linear Regression Analysis of the Data in Table 14.6

```
The regression equation is
Death Index = - 2.9 + 1.09 Smoking Index

Predictor          Coef       SE Coef           T          P
Constant          -2.89         23.03       -0.13      0.901
Smoking Index    1.0875        0.2209        4.92       0.00

S = 18.6154   R-Sq = 51.3%   R-Sq(adj) = 49.2%
```

FIGURE 14.40 A Plot of the Lung Cancer Death Index versus the Smoking Index

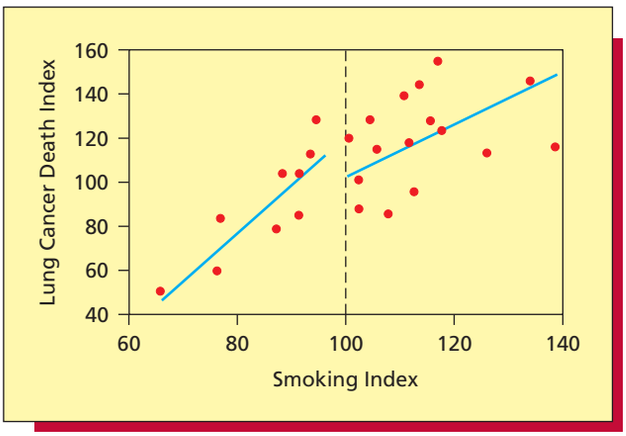


TABLE 14.7 Return Length Estimates of β_1

Average Estimate of β_1 Beta

Return Length Time	Aggressive Stocks	Defensive Stocks	Neutral Stocks
1	1.37	.50	.98
3	1.42	.44	.95
6	1.53	.41	.94
9	1.69	.39	1.00
12	1.83	.40	.98
15	1.67	.38	1.00
18	1.78	.39	1.02
24	1.86	.35	1.14
30	1.83	.33	1.22

Source: Reprinted by permission from H. Levy, "Measuring Risk and Performance over Alternative Investment Horizons," *Financial Analysts Journal* (March–April 1984), pp. 61–68. Copyright © 1984, CFA Institute. Reproduced and modified from Financial Analysts Journal with permission of CFA Institute.

a Although the data do not prove that smoking increases your chance of getting lung cancer, can you think of a third factor that would cause the two indexes to move together?

b Does the slope of the hypothetical line relating the two indexes when the smoking index is less than 100 seem to equal the slope of the hypothetical line relating the two indexes when the smoking index is greater than 100? If you wish, use simple linear regression to make a more precise determination. What practical conclusion might you make?

14.69 In analyzing the stock market, we sometimes use the model $y = \beta_0 + \beta_1 x + \varepsilon$ to relate y, the rate of return on a particular stock, to x, the rate of return on the overall stock market. When using the preceding model, we can interpret β_1 to be the percentage point change in the mean (or expected) rate of return on the particular stock that is associated with an increase of one percentage point in the rate of return on the overall stock market.

If regression analysis can be used to conclude (at a high level of confidence) that β_1 is greater than 1 (for example, if the 95 percent confidence interval for β_1 were [1.1826, 1.4723]), this indicates that the mean rate of return on the particular stock changes more quickly than the rate of return on the overall stock market. Such a stock is called an *aggressive stock* because gains for such a stock tend to be greater than overall market gains (which occur when the market is bullish). However, losses for such a stock tend to be greater than overall market losses (which occur when the market is bearish). Aggressive stocks should be purchased if you expect the market to rise and avoided if you expect the market to fall.

If regression analysis can be used to conclude (at a high level of confidence) that β_1 is less than 1 (for example, if the 95 percent confidence interval for β_1 were [.4729, .7861]), this indicates that the mean rate of return on the particular stock changes more slowly than the rate of return on the overall stock market. Such a stock is called a *defensive stock.* Losses for such a stock tend to be less than overall market losses, whereas gains for such a stock tend to be less than overall market gains. Defensive stocks should be held if you expect the market to fall and sold off if you expect the market to rise.

If the least squares point estimate b_1 of β_1 is nearly equal to 1, and if the 95 percent confidence interval for β_1 contains 1, this might indicate that the mean rate of return on the particular stock changes at roughly the same rate as the rate of return on the overall stock market. Such a stock is called a *neutral stock.*

In an article in *Financial Analysts Journal,* Haim Levy considers how a stock's value of β_1 depends on the length of time for which the rate of return is calculated. Levy calculated estimated values of β_1 for return length times varying from 1 to 30 months for each of 38 aggressive stocks, 38 defensive stocks, and 68 neutral stocks. In Table 14.7 we present the average estimate of β_1 for each stock type for different return length times.

a Let y = average estimate of β_1 and x = return length time, and consider relating y to x for each stock type by using the simple linear regression model $y = \beta_0^* + \beta_1^* x + \varepsilon$. Here β_0^* and β_1^* are regression parameters relating y to x. We use the asterisks to indicate that these regression parameters are different from β_0 and β_1. Calculate a 95 percent confidence interval for β_1^* for each stock type.

b Carefully interpret the meaning of each interval that you found in part(a).

14.70 On January 28, 1986, the space shuttle *Challenger* exploded soon after takeoff, killing all eight astronauts aboard. The temperature at the Kennedy Space Center at liftoff was 31°F. Before the launch, several scientists argued that the launch should be delayed because the shuttle's O-rings might harden in the cold and leak. Other scientists used the data plot in Figure 14.41 to argue that there was no relationship between temperature and O-ring failure. On the basis of this figure and other considerations, *Challenger* was launched to its disastrous, last flight.

Scientists using the data plot in Figure 14.41 made a horrible mistake. They relied on a data plot that was created by using only the seven previous launches where there was at least one O-ring failure. A plot based on all 24 previous launches—17 of which had no O-ring failures—is given in Figure 14.42.

a Intuitively, do you think that Figure 14.42 indicates that there is a relationship between temperature and O-ring failure? Use simple linear regression to justify your answer.

b Even though the figure using only seven launches is incomplete, what about it should have cautioned the scientists not to make the launch?

14.71 In New Jersey, banks have been charged with withdrawing from counties having a high percentage of minorities. To substantiate this charge, P. D'Ambrosio and S. Chambers (1995) present the data in Table 14.8 concerning the percentage, x, of minority population and the number of county residents, y, per bank branch in each of New Jersey's 21 counties. If we use Excel to perform a simple linear regression analysis of this data, we obtain the output given in Figure 14.43 on the next page. DS NJBank

a Determine if there is a significant relationship between x and y.

b Describe the exact nature of any relationship that exists between x and y. (Hint: Estimate β_1 by a point estimate and a confidence interval.)

Note: The Internet exercise for this chapter is on page 553.

FIGURE 14.41 A Data Plot Based on Seven Launches

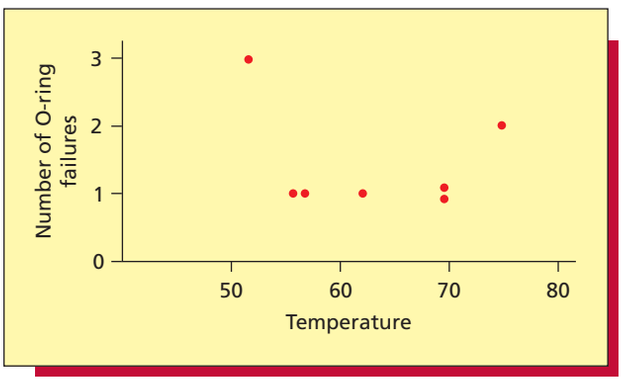


FIGURE 14.42 A Data Plot Based on All 24 Launches

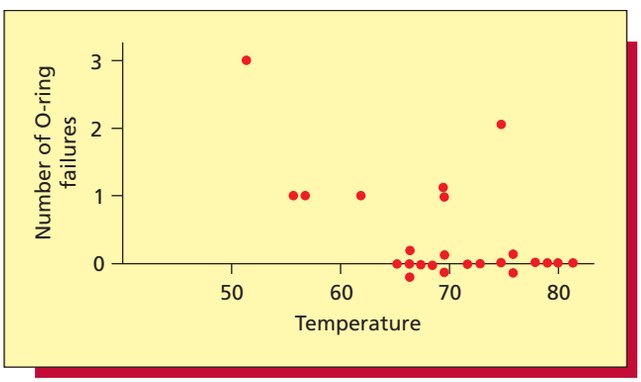


TABLE 14.8 The New Jersey Bank Data
DS NJBank

County	Percentage of Minority Population, x	Number of Residents Per Bank Branch, y
Atlantic	23.3	3,073
Bergen	13.0	2,095
Burlington	17.8	2,905
Camden	23.4	3,330
Cape May	7.3	1,321
Cumberland	26.5	2,557
Essex	48.8	3,474
Gloucester	10.7	3,068
Hudson	33.2	3,683
Hunterdon	3.7	1,998
Mercer	24.9	2,607
Middlesex	18.1	3,154
Monmouth	12.6	2,609
Morris	8.2	2,253
Ocean	4.7	2,317
Passaic	28.1	3,307
Salem	16.7	2,511
Somerset	12.0	2,333
Sussex	2.4	2,568
Union	25.6	3,048
Warren	2.8	2,349

Source: P. D'Ambrosio and S. Chambers, "No Checks and Balances," *Asbury Park Press,* September 10, 1995. Copyright © 1995 Asbury Park Press. Used with permission.

FIGURE 14.43 Excel Output of a Simple Linear Regression Analysis of the New Jersey Bank Data

Regression Statistics	
Multiple R	0.7256
R Square	0.5265
Adjusted R Square	0.5016
Standard Error	400.2546
Observations	21

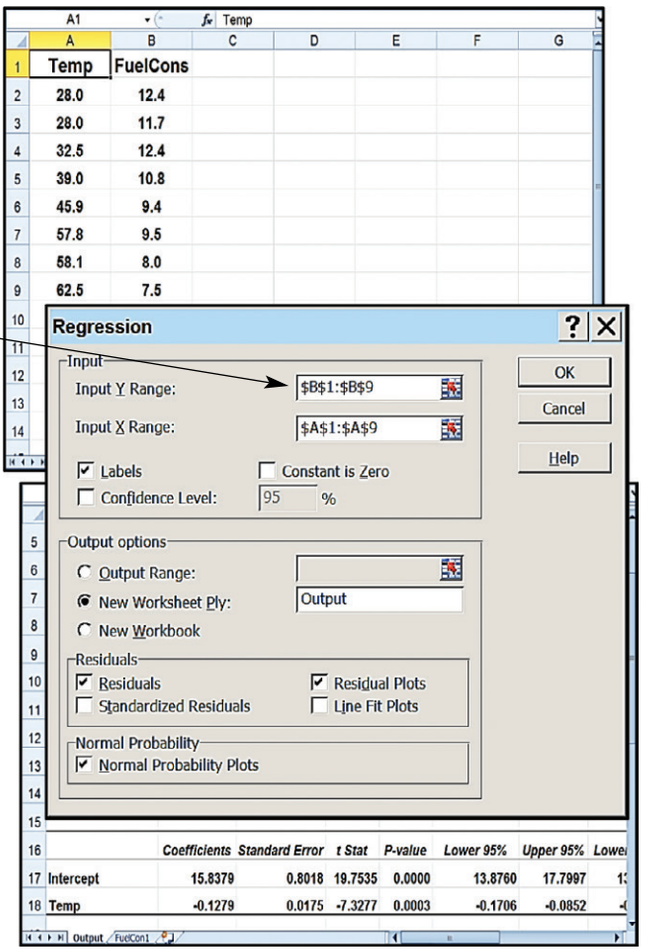
ANOVA	df	SS	MS	F	Significance F
Regression	1	3385090.234	3385090	21.1299	0.0002
Residual	19	3043870.432	160203.7		
Total	20	6428960.667			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	2082.0153	159.1070	13.0856	5.92E-11	1749.0005	2415.0301
% Minority Pop (x)	35.2877	7.6767	4.5967	0.0002	19.2202	51.3553

Appendix 14.1 ■ Simple Linear Regression Analysis Using Excel

Simple linear regression in Figure 14.9 on page 507 (data file: GasCon1.xlsx):

- Enter the natural gas consumption data (page 497)—the temperatures in column A with label Temp and the gas consumptions in column B with label FuelCons.
- Select **Data : Data Analysis : Regression** and click OK in the Data Analysis dialog box.
- In the Regression dialog box:
  Enter B1:B9 into the "Input Y Range" box.
  Enter A1:A9 into the "Input X Range" box.
- Place a checkmark in the Labels checkbox.
- Be sure that the "Constant is Zero" checkbox is NOT checked.
- Select the "New Worksheet Ply" option and enter the name Output into the New Worksheet window.
- Click OK in the Regression dialog box to obtain the regression results in a new worksheet.



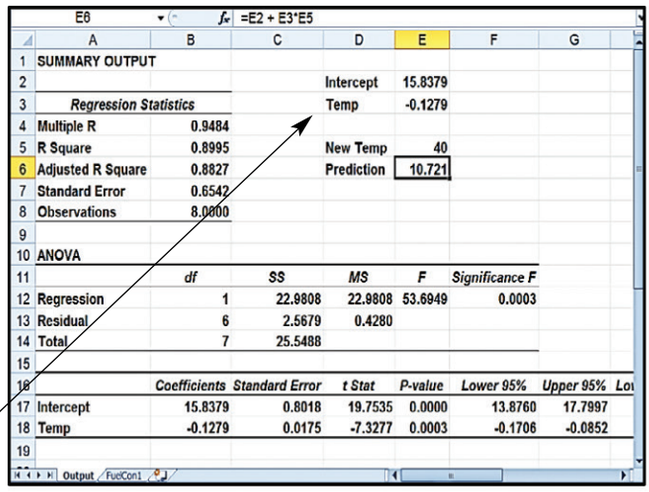
To produce residual plots similar to Figures 14.22(b) and 14.22(c) on page 528:

- In the Regression dialog box, place a checkmark in the Residuals checkbox to request predicted values and residuals.
- Place a checkmark in the Residual Plots checkbox.
- Place a checkmark in the Normal Probability Plots checkbox.
- Click OK in the Regression dialog box.

- Move the plots to chart sheets to format them for effective viewing. Additional residual plots—residuals versus predicted values and residuals versus time—can be produced using the Excel charting features.

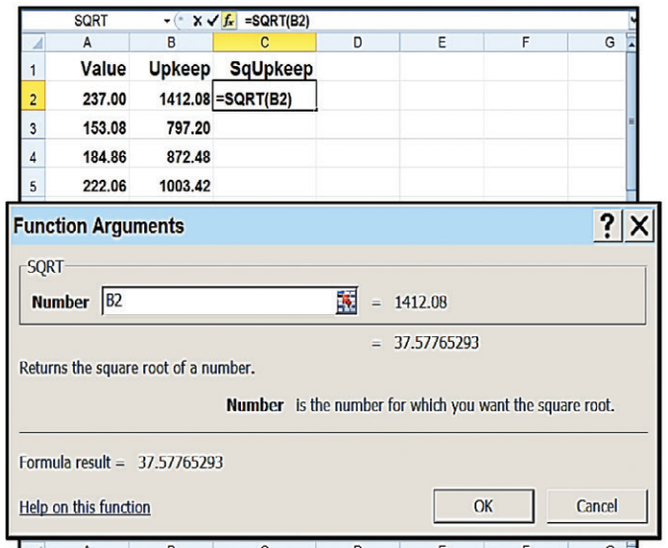
To compute a point prediction of gas consumption when temperature is 40°F (data file: GasCon1.xlsx):

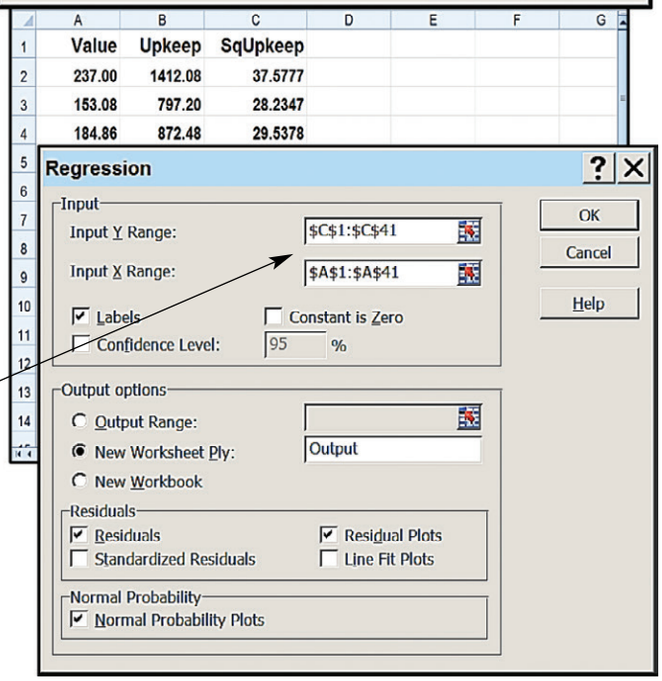
- The Excel Analysis ToolPak does not provide an option for computing point or interval predictions. A point prediction can be computed from the regression results using Excel cell formulas.
- In the regression output, the estimated intercept and slope parameters from cells A17:B18 have been copied to cells D2:E3 and the predictor value 40 has been placed in cell E5.
- In cell E6, enter the Excel formula = E2 + E3*E5 (= 10.721) to compute the prediction.



Simple linear regression with a transformed response similar to Figure 14.30 on page 537 (data file: QHIC.xlsx):

- Enter the QHIC upkeep expenditure data from Figure 14.21 (page 526). Enter the label Value in cell A1 with the home values in cells A2 to A41 and enter the label Upkeep in cell B1 with the upkeep expenditures in cells B2 to B41.
- Enter the label SqUpkeep in cell C1.
- Click on cell C2 and then select the Insert Function button f_x on the Excel ribbon.
- Select **Math & Trig** from the "Or select a category:" menu, select **SQRT** from the "Select a function:" menu, and click OK in the Insert Function dialog box.
- In the "SQRT Function Arguments" dialog box, enter B2 in the Number box and click OK to compute the square root of the value in cell B2.
- Copy the cell formula of C2 through cell C41 by double-clicking the drag handle (in the lower right corner) of cell C2 to compute the square roots of the remaining upkeep values.
- Follow the steps for **simple linear regression** (on page 548) using cells C1:C41 as the response (Input Y Range) and cells A1:A41 as the predictor (Input X Range).





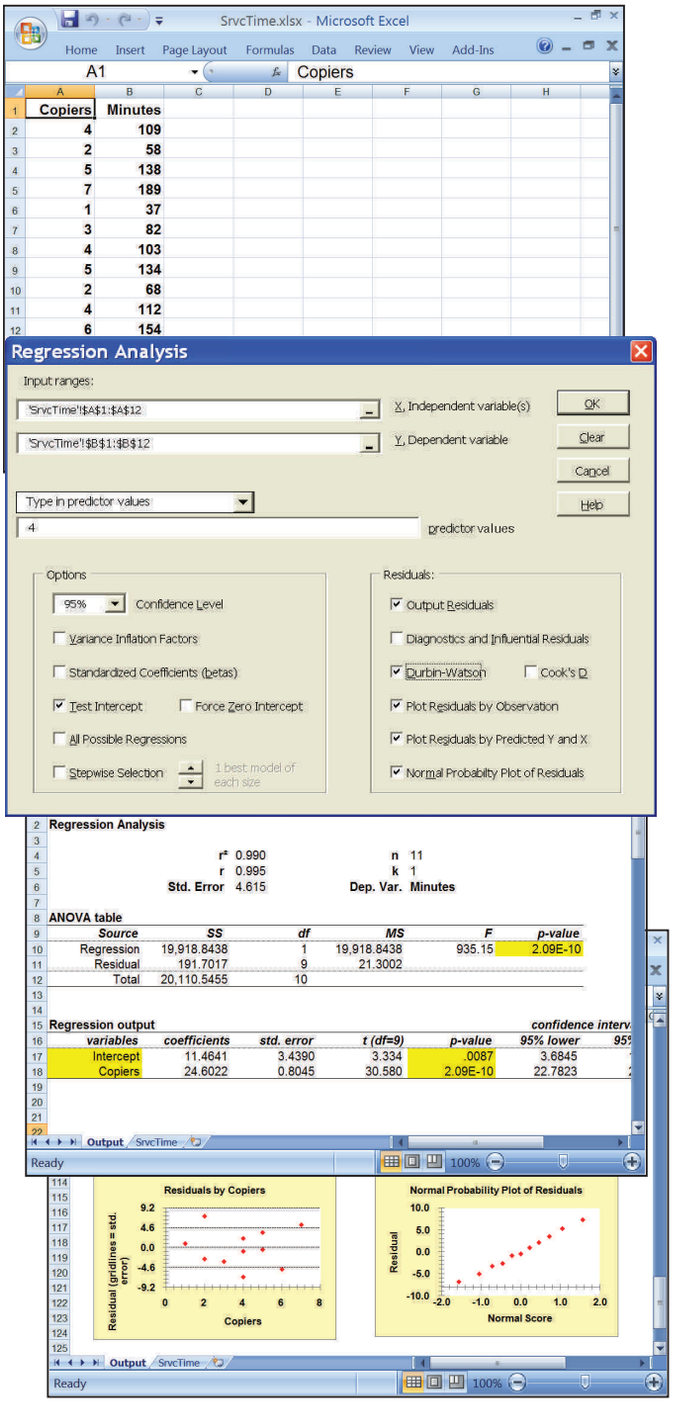
Appendix 14.2 ■ Simple Linear Regression Analysis Using MegaStat

Simple linear regression for the service time data in Exercise 14.5 on page 497 (data file: SrvcTime.xlsx):

- Enter the service time data (page 498) with the numbers of copiers serviced in column A with label Copiers and with the service times in column B with label Minutes.
- Select **Add-Ins : MegaStat : Correlation/ Regression : Regression Analysis.**
- In the Regression Analysis dialog box, click in the Independent Variables window and use the autoexpand feature to enter the range A1:A12.
- Click in the Dependent Variable window and use the AutoExpand feature to enter the range B1:B12.
- Check the appropriate Options and Residuals checkboxes as follows:
  1. Check "Test Intercept" to include a *y*-intercept and to test its significance.
  2. Check "Output Residuals" to obtain a list of the model residuals.
  3. Check "Plot Residuals by Observation," and "Plot Residuals by Predicted Y and X" to obtain residual plots versus time, versus the predicted values of *y*, and versus the values of the independent variable.
  4. Check "Normal Probability Plot of Residuals" to obtain a normal plot.
  5. Check "Durbin-Watson" for the Durbin–Watson statistic.

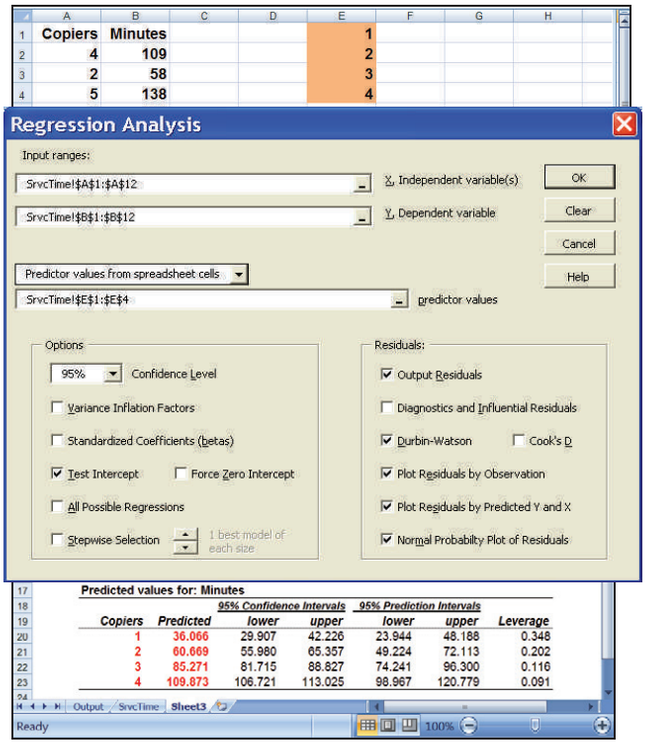
To obtain a **point prediction** of *y* when four computers will be serviced (as well as a confidence interval and prediction interval):

- Click on the drop-down menu above the Predictor Values window and select "Type in predictor values."
- Type the value of the independent variable for which a prediction is desired (here equal to 4) into the Predictor Values window.
- Select a desired level of confidence (here 95%) from the Confidence Level drop-down menu or type in a value.
- Click OK in the Regression Analysis dialog box.



To compute several point predictions of *y*—say, when 1, 2, 3, and 4 computers will be serviced—(and to compute corresponding confidence and prediction intervals):

- Enter the values of *x* for which predictions are desired into a column in the spreadsheet—these values can be in any column. Here we have entered the values 1, 2, 3, and 4 into cells E1 through E4.
- Click on the drop-down menu above the predictor values window and select "Predictor values from spreadsheet cells."
- Enter the range E1:E4 into the predictor values window.
- Click OK in the Regression Analysis dialog box.

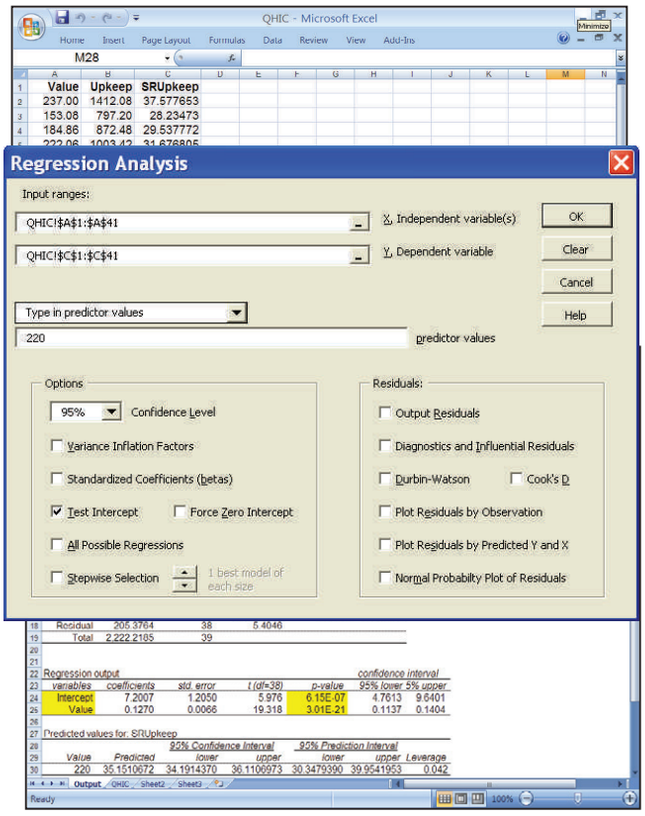


Simple linear regression with a transformed response similar to Figure 14.30 on page 537 (data file: QHIC.xlsx):

- Enter the QHIC data from Figure 14.21 (page 526)—the home values in column A (with label Value) and the upkeep expenditures in column B (with label Upkeep).
- Follow the instructions on page 549 in Appendix 14.1 to calculate the square roots of the upkeep expenditures in column C (with label SRUpkeep).
- Select **Add-Ins : MegaStat : Correlation/ Regression : Regression Analysis.**
- In the Regression Analysis dialog box, click in the Independent variables window, and use the AutoExpand feature to enter the range A1:A41.
- Click in the Dependent variable window and use the AutoExpand feature to enter the range C1:C41.
- Check the "Test Intercept" checkbox to include a *y*-intercept and test its significance.

To compute a **point prediction of the square root of *y*** (as well as a confidence interval and prediction interval) for a house having a value of $220,000:

- Select "Type in predictor values" from the drop-down menu above the predictor values window.
- Type 220 into the predictor values window.
- Select a desired level of confidence (here 95%) from the drop-down menu in the Confidence Level box or type in a value.
- Click OK in the Regression Analysis dialog box.



Appendix 14.3 ■ Simple Linear Regression Analysis Using MINITAB

Simple linear regression of the natural gas consumption data in Figure 14.10 on page 507 (data file: GasCon1.MTW):

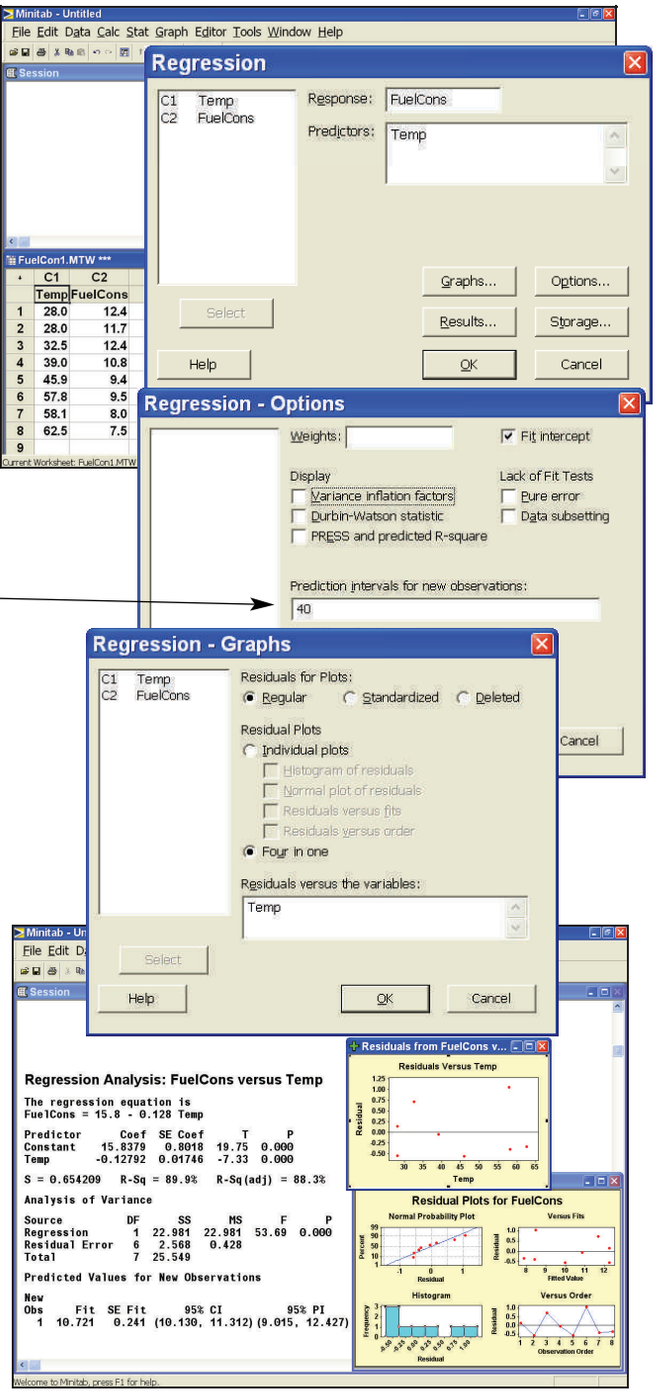
- In the Data window, enter the gas consumption data from Exercise 14.3 (page 497)—average hourly temperatures in column C1 with variable name Temp and weekly gas consumptions in column C2 with variable name FuelCons.
- Select **Stat : Regression : Regression.**
- In the Regression dialog box, select FuelCons into the Response window.
- Select Temp into the Predictors window.

To compute a **prediction** of natural gas consumption when temperature is 40°F:

- In the Regression dialog box, click on the Options... button.
- In the "Regression—Options" dialog box, type 40 in the "Prediction intervals for new observations" window.
- Click OK in the "Regression—Options" dialog box.

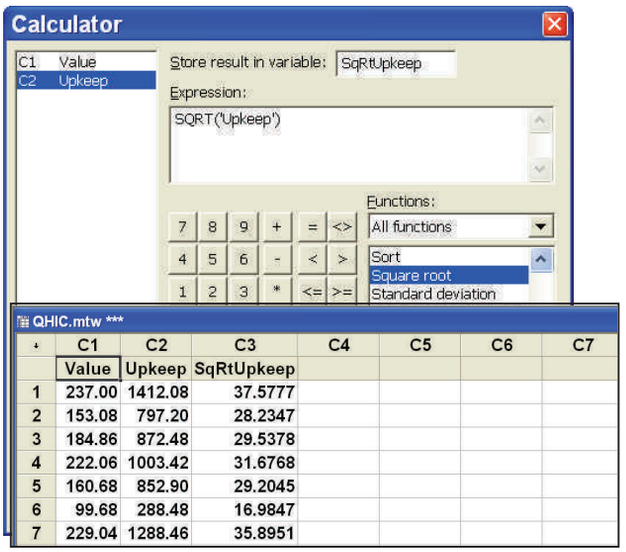
To produce **residual analysis** similar to Figure 14.32 on page 539:

- In the Regression dialog box, click on the Graphs... button.
- In the "Regression—Graphs" dialog box, select the "Residuals for Plots: Regular" option.
- To obtain a histogram and normal plot of the residuals, a plot of the residuals versus the fitted values, and a plot of the residuals versus time order, select "Four in one" in the list of options under Residual Plots. (Note that the plot versus time order is generally informative only if the data are in time sequence order.)
- Enter Temp in the "Residuals versus the variables" window to obtain a plot of the residuals versus the values of average hourly temperature.
- Click OK in the "Regression—Graphs" dialog box.
- To see the regression results in the Session window and high-resolution graphs in two graphics windows, click OK in the Regression dialog box.



Simple linear regression with a transformed response in Figure 14.30 on page 537 (data file: QHIC.MTW):

- In the Data window, enter the QHIC upkeep expenditure data from Figure 14.21 (page 526)—home values in column C1 with variable name Value and upkeep expenditures in column C2 with variable name Upkeep.
- Select **Calc : Calculator.**
- In the Calculator dialog box, enter SqRtUpkeep in the "Store result in variable" window.
- From the Functions menu list, double-click on "Square root," giving SQRT(number) in the Expression window.
- Replace "number" in the Expression window with Upkeep by double-clicking Upkeep in the variables list.
- Click OK in the Calculator dialog box to obtain a new column, SqRtUpkeep, containing the square roots of the Upkeep values.
- Follow the steps for **simple linear regression** on page 552 using SqRtUpkeep as the response and Value as the predictor.



	C1	C2	C3	C4	C5	C6	C7
	Value	Upkeep	SqRtUpkeep				
1	237.00	1412.08	37.5777				
2	153.08	797.20	28.2347				
3	184.86	872.48	29.5378				
4	222.06	1003.42	31.6768				
5	160.68	852.90	29.2045				
6	99.68	288.48	16.9847				
7	229.04	1288.46	35.8951				

14.72 Internet Exercise

The *U.S. News & World Report* website provides rankings of the best colleges and universities in the United States. The free version of "Best Colleges" gives information such as number of students enrolled, tuition rates, and so forth. Among the data provided are the percentage acceptance rate and the average freshman retention rate.

One might wonder if there is a statistically significant relationship between average freshman percentage retention rate and the percentage acceptance rate at colleges and universities in the United States. To investigate this possible relationship, go to the *U.S. News & World Report* website (www.usnews.com). Then make selections as follows: Education; Best Colleges; National Universities; View National Universities Rankings. From the rankings, randomly select a list of the ranked universities and their most recent acceptance rates. Note that following the list of ranked universities is a list of unranked schools. Omit these unranked schools from your analysis. Next, return to the National Universities page and select "Freshman Retention Rate" from the "National Universities Quick Comparison" list. Compile a list of average retention rates for the schools in your list of ranked universities. Finally, enter the ranked universities and their corresponding acceptance rates and average freshman retention rates into a spreadsheet.

Using Excel or MINITAB, construct a scatter plot of average freshman retention rate versus acceptance rate. Describe any apparent relationship between these variables. Develop a simple linear regression model expressing average freshman retention rate as a linear function of acceptance rate. Then use Excel or MINITAB to fit the model. Using the computer output, identify the key summary measures—r^2, the standard error, and the F-statistic from the ANOVA table. Identify and interpret the estimated regression coefficients. Suppose that a university has an acceptance rate of 90 percent. Use your regression model to predict the average freshman retention rate for this school. Prepare a brief report summarizing your analysis and conclusions.