

# 13 Simple Linear Regression

## USING STATISTICS @ Sunflowers Apparel

### 13.1 Types of Regression Models

### 13.2 Determining the Simple Linear Regression Equation

The Least-Squares Method  
Predictions in Regression  
Analysis: Interpolation  
Versus Extrapolation  
Computing the Y Intercept,  $b_0$ , and the Slope,  $b_1$

## VISUAL EXPLORATIONS: Exploring Simple Linear Regression Coefficients

### 13.3 Measures of Variation

Computing the Sum of Squares

The Coefficient of Determination  
Standard Error of the Estimate

### 13.4 Assumptions

### 13.5 Residual Analysis

Evaluating the Assumptions

### 13.6 Measuring Autocorrelation: The Durbin-Watson Statistic

Residual Plots to Detect Autocorrelation  
The Durbin-Watson Statistic

### 13.7 Inferences About the Slope and Correlation Coefficient

$t$  Test for the Slope  
 $F$  Test for the Slope

Confidence Interval Estimate for the Slope  
 $t$  Test for the Correlation Coefficient

### 13.8 Estimation of Mean Values and Prediction of Individual Values

The Confidence Interval Estimate  
The Prediction Interval

### 13.9 Pitfalls in Regression

**Think About This: By Any Other Name**

## USING STATISTICS @ Sunflowers Apparel Revisited

## CHAPTER 13 EXCEL GUIDE

## CHAPTER 13 MINITAB GUIDE

## Learning Objectives

In this chapter, you learn:

- How to use regression analysis to predict the value of a dependent variable based on an independent variable
- The meaning of the regression coefficients  $b_0$  and  $b_1$
- How to evaluate the assumptions of regression analysis and know what to do if the assumptions are violated
- How to make inferences about the slope and correlation coefficient
- How to estimate mean values and predict individual values



## USING STATISTICS

### @ Sunflowers Apparel

The sales for Sunflowers Apparel, a chain of upscale clothing stores for women, have increased during the past 12 years as the chain has expanded the number of stores. Until now, Sunflowers managers selected sites based on subjective factors, such as the availability of a good lease or the perception that a location seemed ideal for an apparel store. As the new director of planning, you need to develop a systematic approach that will lead to making better decisions during the site-selection process. As a starting point, you believe that the size of the store significantly contributes to store sales, and you want to use this relationship in the decision-making process. How can you use statistics so that you can forecast the annual sales of a proposed store based on the size of that store?



In this chapter and the next two chapters, you learn how **regression analysis** enables you to develop a model to predict the values of a numerical variable, based on the value of other variables.

In regression analysis, the variable you wish to predict is called the **dependent variable**. The variables used to make the prediction are called **independent variables**. In addition to predicting values of the dependent variable, regression analysis also allows you to identify the type of mathematical relationship that exists between a dependent variable and an independent variable, to quantify the effect that changes in the independent variable have on the dependent variable, and to identify unusual observations. For example, as the director of planning, you might want to predict sales for a Sunflowers store based on the size of the store. Other examples include predicting the monthly rent of an apartment based on its size and predicting the monthly sales of a product in a supermarket based on the amount of shelf space devoted to the product.

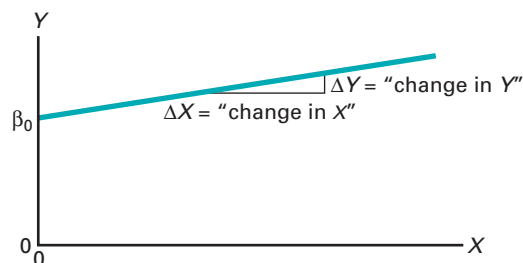
This chapter discusses **simple linear regression**, in which a *single* numerical independent variable,  $X$ , is used to predict the numerical dependent variable  $Y$ , such as using the size of a store to predict the annual sales of the store. Chapters 14 and 15 discuss *multiple regression models*, which use *several* independent variables to predict a numerical dependent variable,  $Y$ . For example, you could use the amount of advertising expenditures, price, and the amount of shelf space devoted to a product to predict its monthly sales.

## 13.1 Types of Regression Models

In Section 2.6, you used a **scatter plot** (also known as a **scatter diagram**) to examine the relationship between an  $X$  variable on the horizontal axis and a  $Y$  variable on the vertical axis. The nature of the relationship between two variables can take many forms, ranging from simple to extremely complicated mathematical functions. The simplest relationship consists of a straight-line relationship, or **linear relationship**. Figure 13.1 illustrates a straight-line relationship.

**FIGURE 13.1**

A straight-line relationship



Equation (13.1) represents the straight-line (linear) model.

### SIMPLE LINEAR REGRESSION MODEL

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad (13.1)$$

where

$\beta_0$  =  $Y$  intercept for the population

$\beta_1$  = slope for the population

$\varepsilon_i$  = random error in  $Y$  for observation  $i$

$Y_i$  = dependent variable (sometimes referred to as the **response variable**) for observation  $i$

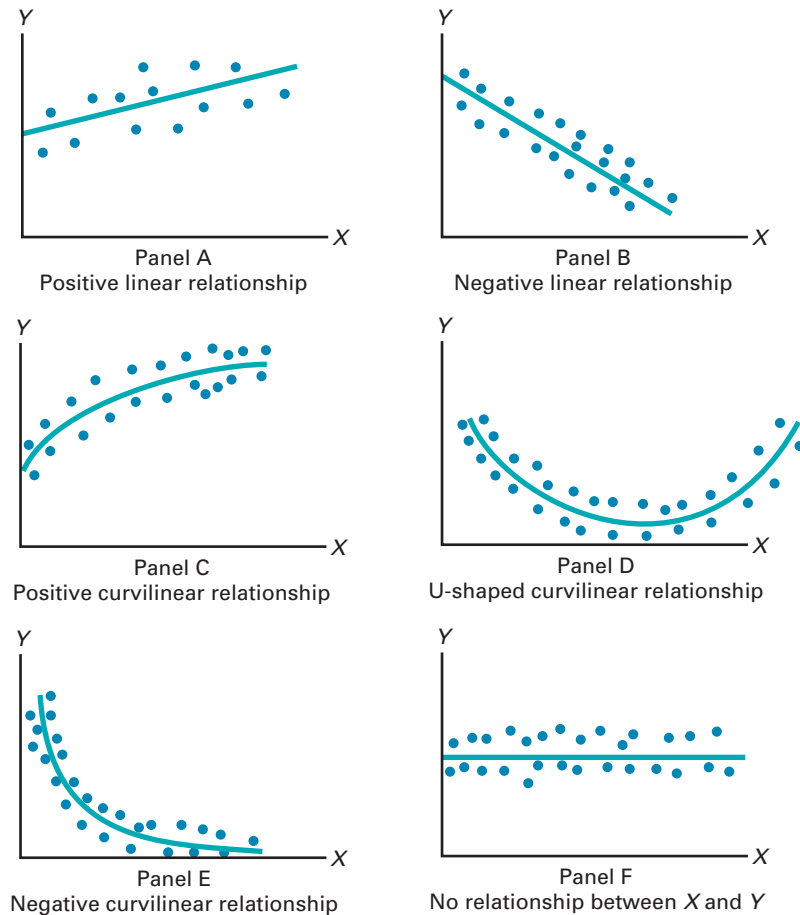
$X_i$  = independent variable (sometimes referred to as the predictor, or **explanatory variable**) for observation  $i$

The  $Y_i = \beta_0 + \beta_1 X_i$  portion of the simple linear regression model expressed in Equation (13.1) is a straight line. The **slope** of the line,  $\beta_1$ , represents the expected change in  $Y$  per unit change in  $X$ . It represents the mean amount that  $Y$  changes (either positively or negatively) for a one-unit change in  $X$ . The  **$Y$  intercept**,  $\beta_0$ , represents the mean value of  $Y$  when  $X$  equals 0. The last component of the model,  $\varepsilon_i$ , represents the random error in  $Y$  for each observation,  $i$ . In other words,  $\varepsilon_i$  is the vertical distance of the actual value of  $Y_i$  above or below the expected value of  $Y_i$  on the line.

The selection of the proper mathematical model depends on the distribution of the  $X$  and  $Y$  values on the scatter plot. Figure 13.2 illustrates six different types of relationships.

**FIGURE 13.2**

Six types of relationships found in scatter plots



In Panel A, the values of  $Y$  are generally increasing linearly as  $X$  increases. This panel is similar to Figure 13.3 on page 524, which illustrates the positive relationship between the square footage of the store and the annual sales at branches of the Sunflowers Apparel women's clothing store chain.

Panel B is an example of a negative linear relationship. As  $X$  increases, the values of  $Y$  are generally decreasing. An example of this type of relationship might be the price of a particular product and the amount of sales.

Panel C shows a positive curvilinear relationship between  $X$  and  $Y$ . The values of  $Y$  increase as  $X$  increases, but this increase tapers off beyond certain values of  $X$ . An example of a positive curvilinear relationship might be the age and maintenance cost of a machine. As a machine gets older, the maintenance cost may rise rapidly at first but then level off beyond a certain number of years.

Panel D shows a U-shaped relationship between  $X$  and  $Y$ . As  $X$  increases, at first  $Y$  generally decreases; but as  $X$  continues to increase,  $Y$  not only stops decreasing but actually increases above its minimum value. An example of this type of relationship might be the number of errors per hour at a task and the number of hours worked. The number of errors per hour decreases as the individual becomes more proficient at the task, but then it increases beyond a certain point because of factors such as fatigue and boredom.



Panel E illustrates an exponential relationship between  $X$  and  $Y$ . In this case,  $Y$  decreases very rapidly as  $X$  first increases, but then it decreases much less rapidly as  $X$  increases further. An example of an exponential relationship could be the value of an automobile and its age. The value drops drastically from its original price in the first year, but it decreases much less rapidly in subsequent years.

Finally, Panel F shows a set of data in which there is very little or no relationship between  $X$  and  $Y$ . High and low values of  $Y$  appear at each value of  $X$ .

Although scatter plots are useful in visually displaying the mathematical form of a relationship, more sophisticated statistical procedures are available to determine the most appropriate model for a set of variables. The rest of this chapter discusses the model used when there is a *linear* relationship between variables.

## 13.2 Determining the Simple Linear Regression Equation

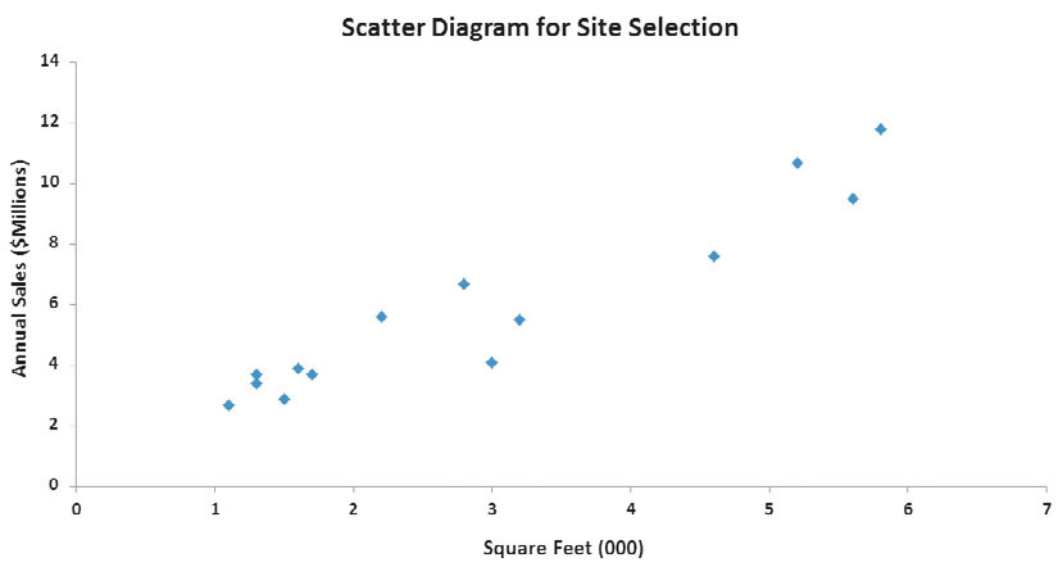
In the Sunflowers Apparel scenario on page 521, the business objective of the director of planning is to forecast annual sales for all new stores, based on store size. To examine the relationship between the store size in square feet and its annual sales, data were collected from a sample of 14 stores. Table 13.1 shows the organized data, which are stored in [Site](#).

Figure 13.3 displays the scatter plot for the data in Table 13.1. Observe the increasing relationship between square feet ( $X$ ) and annual sales ( $Y$ ). As the size of the store increases,

**TABLE 13.1**  
Square Footage (in Thousands of Square Feet) and Annual Sales (in Millions of Dollars) for a Sample of 14 Branches of Sunflowers Apparel

Store	Square Feet (Thousands)	Annual Sales (in Millions of Dollars)	Store	Square Feet (Thousands)	Annual Sales (in Millions of Dollars)
1	1.7	3.7	8	1.1	2.7
2	1.6	3.9	9	3.2	5.5
3	2.8	6.7	10	1.5	2.9
4	5.6	9.5	11	5.2	10.7
5	1.3	3.4	12	4.6	7.6
6	2.2	5.6	13	5.8	11.8
7	1.3	3.7	14	3.0	4.1

**FIGURE 13.3**  
Scatter plot for the Sunflowers Apparel data



annual sales increase approximately as a straight line. Thus, you can assume that a straight line provides a useful mathematical model of this relationship. Now you need to determine the specific straight line that is the *best* fit to these data.

## The Least-Squares Method

In the preceding section, a statistical model is hypothesized to represent the relationship between two variables, square footage and sales, in the entire population of Sunflowers Apparel stores. However, as shown in Table 13.1, the data are collected from a random sample of stores. If certain assumptions are valid (see Section 13.4), you can use the sample  $Y$  intercept,  $b_0$ , and the sample slope,  $b_1$ , as estimates of the respective population parameters,  $\beta_0$  and  $\beta_1$ . Equation (13.2) uses these estimates to form the **simple linear regression equation**. This straight line is often referred to as the **prediction line**.

### SIMPLE LINEAR REGRESSION EQUATION: THE PREDICTION LINE

The predicted value of  $Y$  equals the  $Y$  intercept plus the slope multiplied by the value of  $X$ .

$$\hat{Y}_i = b_0 + b_1X_i \quad (13.2)$$

where

$\hat{Y}_i$  = predicted value of  $Y$  for observation  $i$

$X_i$  = value of  $X$  for observation  $i$

$b_0$  = sample  $Y$  intercept

$b_1$  = sample slope

Equation (13.2) requires you to determine two **regression coefficients**— $b_0$  (the sample  $Y$  intercept) and  $b_1$  (the sample slope). The most common approach to finding  $b_0$  and  $b_1$  is using the least-squares method. This method minimizes the sum of the squared differences between the actual values ( $Y_i$ ) and the predicted values ( $\hat{Y}_i$ ) using the simple linear regression equation [i.e., the prediction line; see Equation (13.2)]. This sum of squared differences is equal to

$$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

Because  $\hat{Y}_i = b_0 + b_1X_i$ ,

$$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n [Y_i - (b_0 + b_1X_i)]^2$$

Because this equation has two unknowns,  $b_0$  and  $b_1$ , the sum of squared differences depends on the sample  $Y$  intercept,  $b_0$ , and the sample slope,  $b_1$ . The **least-squares method** determines the values of  $b_0$  and  $b_1$  that minimize the sum of squared differences around the prediction line. Any values for  $b_0$  and  $b_1$  other than those determined by the least-squares method result in a greater sum of squared differences between the actual values ( $Y_i$ ) and the predicted values ( $\hat{Y}_i$ ). Figure 13.4<sup>1</sup> presents the simple linear regression model for the Table 13.1 Sunflowers Apparel data.

<sup>1</sup>The equations used to compute these results are shown in Examples 13.3 and 13.4 on pages 528–530 and 535–536. You should use software to do these computations for large data sets, given the complex nature of the computations.

**FIGURE 13.4**

Excel and Minitab simple linear regression models for the Sunflowers Apparel data

	A	B	C	D	E	F	G	H	I
1	Simple Linear Regression								
2									
3	Regression Statistics								
4	Multiple R	0.9509							
5	R Square	0.9042							
6	Adjusted R Square	0.8962							
7	Standard Error	0.9664							
8	Observations	14							
9									
10	ANOVA								
11		df	SS	MS	F	Significance F			
12	Regression	1	105.7476	105.7476	113.2335	0.0000			
13	Residual	12	11.2067	0.9339					
14	Total	13	116.9543						
15									
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
17	Intercept	0.9645	0.5262	1.8329	0.0917	-0.1820	2.1110	-0.1820	2.11095
18	Square Feet	1.6699	0.1569	10.6411	0.0000	1.3280	2.0118	1.3280	2.01177

**Regression Analysis: Annual Sales versus Square Feet**

The regression equation is

$$\text{Annual Sales} = 0.964 + 1.67 \text{ Square Feet}$$

Predictor	Coef	SE Coef	T	P
Constant	0.9645	0.5262	1.83	0.092
Square Feet	1.6699	0.1569	10.64	0.000

$$S = 0.966380 \quad R\text{-Sq} = 90.4\% \quad R\text{-Sq(adj)} = 89.6\%$$

**Analysis of Variance**

Source	DF	SS	MS	F	P
Regression	1	105.75	105.75	113.23	0.000
Residual Error	12	11.21	0.93		
Total	13	116.95			

**Predicted Values for New Observations**

New Obs	Fit	SE Fit	95% CI	95% PI
1	7.644	0.309	(6.971, 8.317)	(5.433, 9.854)

**Values of Predictors for New Observations**

New Obs	Square Feet
1	4.00

In Figure 13.4, observe that  $b_0 = 0.9645$  and  $b_1 = 1.6699$ . Using Equation (13.2) on page 525, the prediction line for these data is

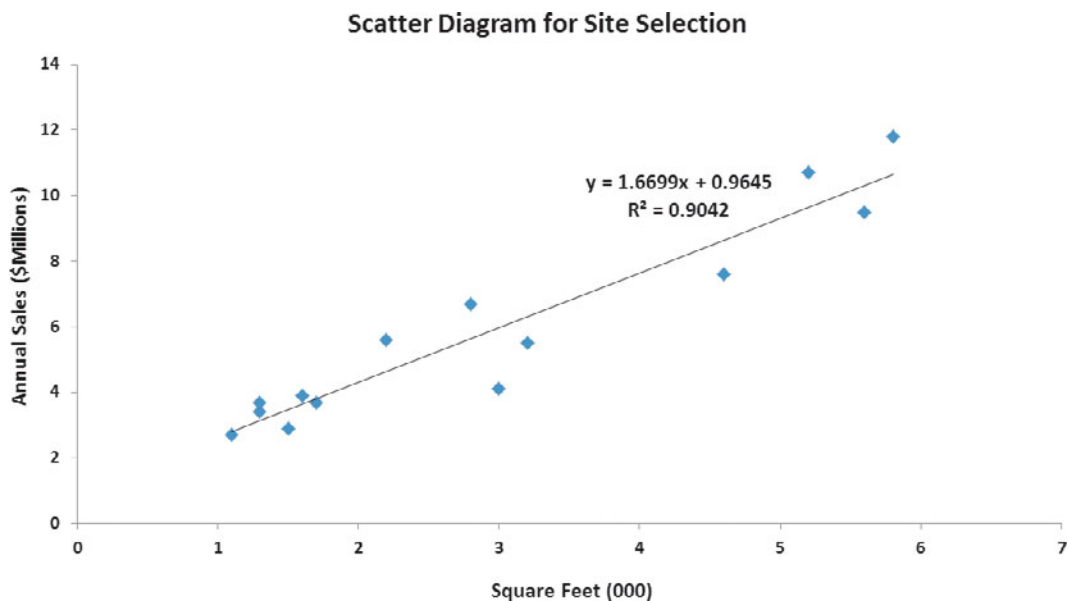
$$\hat{Y}_i = 0.9645 + 1.6699X_i$$

The slope,  $b_1$ , is +1.6699. This means that for each increase of 1 unit in  $X$ , the predicted value of  $Y$  is estimated to increase by 1.6699 units. In other words, for each increase of 1.0 thousand square feet in the size of the store, the predicted annual sales are estimated to increase by 1.6699 millions of dollars. Thus, the slope represents the portion of the annual sales that are estimated to vary according to the size of the store.

The  $Y$  intercept,  $b_0$ , is +0.9645. The  $Y$  intercept represents the predicted value of  $Y$  when  $X$  equals 0. Because the square footage of the store cannot be 0, this  $Y$  intercept has little or no practical interpretation. Also, the  $Y$  intercept for this example is outside the range of the observed values of the  $X$  variable, and therefore interpretations of the value of  $b_0$  should be made cautiously. Figure 13.5 displays the actual values and the prediction line. To illustrate a situation in which there is a direct interpretation for the  $Y$  intercept,  $b_0$ , see Example 13.1.

**FIGURE 13.5**

Scatter plot and prediction line for Sunflowers Apparel data



**EXAMPLE 13.1**

Interpreting the  $Y$  Intercept,  $b_0$ , and the Slope,  $b_1$

A statistics professor wants to use the number of hours a student studies for a statistics final exam ( $X$ ) to predict the final exam score ( $Y$ ). A regression model was fit based on data collected from a class during the previous semester, with the following results:

$$\hat{Y}_i = 35.0 + 3X_i$$

What is the interpretation of the  $Y$  intercept,  $b_0$ , and the slope,  $b_1$ ?

**SOLUTION** The  $Y$  intercept  $b_0 = 35.0$  indicates that when the student does not study for the final exam, the predicted final exam score is 35.0. The slope  $b_1 = 3$  indicates that for each increase of one hour in studying time, the predicted change in the final exam score is +3.0. In other words, the final exam score is predicted to increase by a mean of 3 points for each one-hour increase in studying time.

Return to the Sunflowers Apparel scenario on page 521. Example 13.2 illustrates how you use the prediction line to predict the annual sales.

**EXAMPLE 13.2**

Predicting Annual Sales Based on Square Footage

Use the prediction line to predict the annual sales for a store with 4,000 square feet.

**SOLUTION** You can determine the predicted value by substituting  $X = 4$  (thousands of square feet) into the simple linear regression equation:

$$\begin{aligned}\hat{Y}_i &= 0.9645 + 1.6699X_i \\ \hat{Y}_i &= 0.9645 + 1.6699(4) = 7.644 \text{ or } \$7,644,000\end{aligned}$$

Thus, a store with 4,000 square feet has predicted annual sales of \$7,644,000.

### Predictions in Regression Analysis: Interpolation Versus Extrapolation

When using a regression model for prediction purposes, you should consider only the **relevant range** of the independent variable in making predictions. This relevant range includes all values from the smallest to the largest  $X$  used in developing the regression model. Hence, when predicting  $Y$  for a given value of  $X$ , you can interpolate within this relevant range of the  $X$  values, but you should not extrapolate beyond the range of  $X$  values. When you use the square footage to predict annual sales, the square footage (in thousands of square feet) varies from 1.1 to 5.8 (see Table 13.1 on page 524). Therefore, you should predict annual sales *only* for stores whose size is between 1.1 and 5.8 thousands of square feet. Any prediction of annual sales for stores outside this range assumes that the observed relationship between sales and store size for store sizes from 1.1 to 5.8 thousand square feet is the same as for stores outside this range. For example, you cannot extrapolate the linear relationship beyond 5,800 square feet in Example 13.2. It would be improper to use the prediction line to forecast the sales for a new store containing 8,000 square feet because the relationship between sales and store size may have a point of diminishing returns. If that is true, as square footage increases beyond 5,800 square feet, the effect on sales may become smaller and smaller.



### Computing the Y Intercept, $b_0$ , and the Slope, $b_1$

For small data sets, you can use a hand calculator to compute the least-squares regression coefficients. Equations (13.3) and (13.4) give the values of  $b_0$  and  $b_1$ , which minimize

$$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n [Y_i - (b_0 + b_1 X_i)]^2$$

COMPUTATIONAL FORMULA FOR THE SLOPE,  $b_1$

$$b_1 = \frac{SSXY}{SSX} \quad (13.3)$$

where

$$SSXY = \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^n X_i Y_i - \frac{\left(\sum_{i=1}^n X_i\right)\left(\sum_{i=1}^n Y_i\right)}{n}$$

$$SSX = \sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n X_i^2 - \frac{\left(\sum_{i=1}^n X_i\right)^2}{n}$$

COMPUTATIONAL FORMULA FOR THE Y INTERCEPT,  $b_0$

$$b_0 = \bar{Y} - b_1 \bar{X} \quad (13.4)$$

where

$$\bar{Y} = \frac{\sum_{i=1}^n Y_i}{n}$$

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

### EXAMPLE 13.3

Computing the Y Intercept,  $b_0$ , and the Slope,  $b_1$

Compute the Y intercept,  $b_0$ , and the slope,  $b_1$ , for the Sunflowers Apparel data.

**SOLUTION** In Equations (13.3) and (13.4), five quantities need to be computed to determine  $b_1$  and  $b_0$ . These are  $n$ , the sample size;  $\sum_{i=1}^n X_i$ , the sum of the  $X$  values;  $\sum_{i=1}^n Y_i$ , the sum of the  $Y$  values;  $\sum_{i=1}^n X_i^2$ , the sum of the squared  $X$  values; and  $\sum_{i=1}^n X_i Y_i$ , the sum of the product of  $X$  and  $Y$ . For the Sunflowers Apparel data, the number of square feet ( $X$ ) is used to predict the annual sales ( $Y$ ) in a store. Table 13.2 presents the computations of the sums needed for the site selection problem. The table also includes  $\sum_{i=1}^n Y_i^2$ , the sum of the squared  $Y$  values that will be used to compute  $SST$  in Section 13.3.

**TABLE 13.2**

Computations for  
the Sunflowers  
Apparel Data

Store	Square Feet (X)	Annual Sales (Y)	$X^2$	$Y^2$	$XY$
1	1.7	3.7	2.89	13.69	6.29
2	1.6	3.9	2.56	15.21	6.24
3	2.8	6.7	7.84	44.89	18.76
4	5.6	9.5	31.36	90.25	53.20
5	1.3	3.4	1.69	11.56	4.42
6	2.2	5.6	4.84	31.36	12.32
7	1.3	3.7	1.69	13.69	4.81
8	1.1	2.7	1.21	7.29	2.97
9	3.2	5.5	10.24	30.25	17.60
10	1.5	2.9	2.25	8.41	4.35
11	5.2	10.7	27.04	114.49	55.64
12	4.6	7.6	21.16	57.76	34.96
13	5.8	11.8	33.64	139.24	68.44
14	3.0	4.1	9.00	16.81	12.30
Totals	40.9	81.8	157.41	594.90	302.30

Using Equations (13.3) and (13.4), you can compute  $b_0$  and  $b_1$ :

$$SSXY = \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^n X_i Y_i - \frac{\left(\sum_{i=1}^n X_i\right)\left(\sum_{i=1}^n Y_i\right)}{n}$$

$$\begin{aligned} SSXY &= 302.3 - \frac{(40.9)(81.8)}{14} \\ &= 302.3 - 238.97285 \\ &= 63.32715 \end{aligned}$$

$$\begin{aligned} SSX &= \sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n X_i^2 - \frac{\left(\sum_{i=1}^n X_i\right)^2}{n} \\ &= 157.41 - \frac{(40.9)^2}{14} \\ &= 157.41 - 119.48642 \\ &= 37.92358 \end{aligned}$$

Therefore,

$$\begin{aligned} b_1 &= \frac{SSXY}{SSX} \\ &= \frac{63.32715}{37.92358} \\ &= 1.6699 \end{aligned}$$

And,

$$\begin{aligned} \bar{Y} &= \frac{\sum_{i=1}^n Y_i}{n} = \frac{81.8}{14} = 5.842857 \\ \bar{X} &= \frac{\sum_{i=1}^n X_i}{n} = \frac{40.9}{14} = 2.92143 \end{aligned}$$

Therefore,

$$\begin{aligned} b_0 &= \bar{Y} - b_1 \bar{X} \\ &= 5.842857 - (1.6699)(2.92143) \\ &= 0.9645 \end{aligned}$$

## VISUAL EXPLORATIONS

### Exploring Simple Linear Regression Coefficients

Use the Visual Explorations Simple Linear Regression procedure to create a prediction line that is as close as possible to the prediction line defined by the least-squares solution. Open the **Visual Explorations** add-in workbook (see Appendix Section D.4) and select **Add-ins** → **VisualExplorations** → **Simple Linear Regression**.

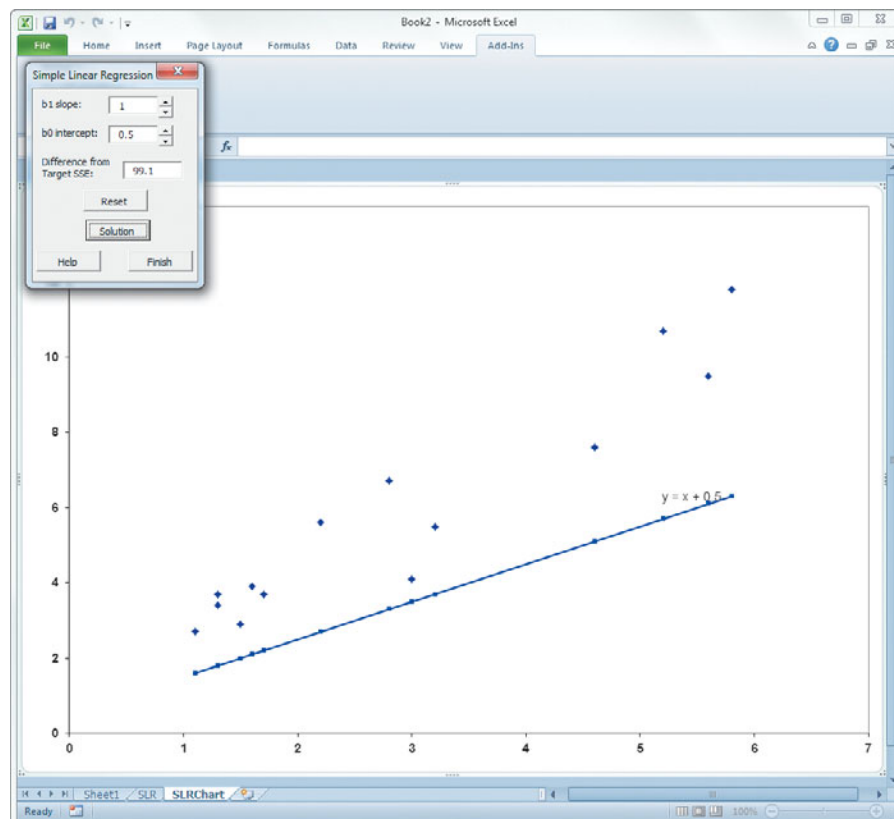
In the Simple Linear Regression dialog box (shown below):

1. Click for the spinner buttons for **b1 slope** (the slope of the prediction line), and **b0 intercept** (the Y intercept of the prediction line) to change the prediction line.
2. Using the visual feedback of the chart, try to create a prediction line that is as close as possible to the prediction line defined by the least-squares estimates. In other words, try to make the **Difference from Target SSE** value as small as possible. (See page 533 for an explanation of SSE.)

At any time, click **Reset** to reset the  $b_1$  and  $b_0$  values or **Solution** to reveal the prediction line defined by the least-squares method. Click **Finish** when you are finished with this exercise.

#### Using Your Own Regression Data

Select **Simple Linear Regression with your worksheet data** from the **VisualExplorations** menu to explore the simple linear regression coefficients using data you supply from a worksheet. In the procedure's dialog box, enter the cell range of your Y variable as the **Y Variable Cell Range** and the cell range of your X variable as the **X Variable Cell Range**. Click **First cells in both ranges contain a label**, enter a **Title**, and click **OK**. After the scatter plot appears onscreen, continue with the step 1 and step 2 instructions.



## Problems for Section 13.2

### LEARNING THE BASICS

**13.1** Fitting a straight line to a set of data yields the following prediction line:

$$\hat{Y}_i = 2 + 5X_i$$

- Interpret the meaning of the  $Y$  intercept,  $b_0$ .
- Interpret the meaning of the slope,  $b_1$ .
- Predict the value of  $Y$  for  $X = 3$ .

**13.2** If the values of  $X$  in Problem 13.1 range from 2 to 25, should you use this model to predict the mean value of  $Y$  when  $X$  equals

- 3?
- 3?
- 0?
- 24?

**13.3** Fitting a straight line to a set of data yields the following prediction line:

$$\hat{Y}_i = 16 - 0.5X_i$$

- Interpret the meaning of the  $Y$  intercept,  $b_0$ .
- Interpret the meaning of the slope,  $b_1$ .
- Predict the value of  $Y$  for  $X = 6$ .

### APPLYING THE CONCEPTS

**SELF Test** **13.4** The marketing manager of a large supermarket chain would like to use shelf space to predict the sales of pet food. A random sample of 12 equal-sized stores is selected, with the following results (stored in **Petfood**):

Store	Shelf Space (X) (Feet)	Weekly Sales (Y) (\$)
1	5	160
2	5	220
3	5	140
4	10	190
5	10	240
6	10	260
7	15	230
8	15	270
9	15	280
10	20	260
11	20	290
12	20	310

- Construct a scatter plot.  
For these data,  $b_0 = 145$  and  $b_1 = 7.4$ .
- Interpret the meaning of the slope,  $b_1$ , in this problem.
- Predict the weekly sales of pet food for stores with 8 feet of shelf space for pet food.

**13.5** Zagat's publishes restaurant ratings for various locations in the United States. The file **Restaurants** contains the Zagat rating for food, décor, service, and the cost per person for a sample of 100 restaurants located in New York City and in a suburb of New York City. Develop a regression model to predict the price per person, based on a variable that represents the sum of the ratings for food, décor, and service.

Sources: Extracted from *Zagat Survey 2010, New York City Restaurants*; and *Zagat Survey 2009–2010, Long Island Restaurants*.

- Construct a scatter plot.  
For these data,  $b_0 = -28.1975$  and  $b_1 = 1.2409$ .
- Assuming a linear cost relationship, use the least-squares method to compute the regression coefficients  $b_0$  and  $b_1$ .
- Interpret the meaning of the  $Y$  intercept,  $b_0$ , and the slope,  $b_1$ , in this problem.
- Predict the cost per person for a restaurant with a summated rating of 50.

**13.6** The owner of a moving company typically has his most experienced manager predict the total number of labor hours that will be required to complete an upcoming move. This approach has proved useful in the past, but the owner has the business objective of developing a more accurate method of predicting labor hours. In a preliminary effort to provide a more accurate method, the owner has decided to use the number of cubic feet moved as the independent variable and has collected data for 36 moves in which the origin and destination were within the borough of Manhattan in New York City and in which the travel time was an insignificant portion of the hours worked. The data are stored in **Moving**.

- Construct a scatter plot.
- Assuming a linear relationship, use the least-squares method to determine the regression coefficients  $b_0$  and  $b_1$ .
- Interpret the meaning of the slope,  $b_1$ , in this problem.
- Predict the labor hours for moving 500 cubic feet.

**13.7** A critically important aspect of customer service in a supermarket is the waiting time at the checkout (defined as the time the customer enters the line until he or she is served). Data were collected during time periods in which a constant number of checkout counters were open. The total number of customers in the store and the waiting times (in minutes) were recorded. The results are stored in **Supermarket**.

- Construct a scatter plot.
- Assuming a linear relationship, use the least-squares method to determine the regression coefficients  $b_0$  and  $b_1$ .
- Interpret the meaning of the slope,  $b_1$ , in this problem.
- Predict the waiting time when there are 20 customers in the store.

**13.8** The value of a sports franchise is directly related to the amount of revenue that a franchise can generate. The file **BBRevenue** represents the value in 2010 (in millions

of dollars) and the annual revenue (in millions of dollars) for the 30 major league baseball franchises. (Data extracted from [www.forbes.com/2010/04/07/most-valuable-baseball-teams-business-sportsmoney-baseball-valuations-10\\_values.html](http://www.forbes.com/2010/04/07/most-valuable-baseball-teams-business-sportsmoney-baseball-valuations-10_values.html).) Suppose you want to develop a simple linear regression model to predict franchise value based on annual revenue generated.

- Construct a scatter plot.
- Use the least-squares method to determine the regression coefficients  $b_0$  and  $b_1$ .
- Interpret the meaning of  $b_0$  and  $b_1$  in this problem.
- Predict the value of a baseball franchise that generates \$150 million of annual revenue.

**13.9** An agent for a residential real estate company in a large city would like to be able to predict the monthly rental cost for apartments, based on the size of an apartment, as defined by square footage. The agent selects a sample of 25 apartments in a particular residential neighborhood and gathers the following data (stored in [Rent](#)).

Rent (\$)	Size (Square Feet)
950	850
1,600	1,450
1,200	1,085
1,500	1,232
950	718
1,700	1,485
1,650	1,136
935	726
875	700
1,150	956
1,400	1,100
1,650	1,285
2,300	1,985
1,800	1,369
1,400	1,175
1,450	1,225
1,100	1,245
1,700	1,259
1,200	1,150
1,150	896
1,600	1,361
1,650	1,040
1,200	755
800	1,000
1,750	1,200

- Construct a scatter plot.
- Use the least-squares method to determine the regression coefficients  $b_0$  and  $b_1$ .
- Interpret the meaning of  $b_0$  and  $b_1$  in this problem.
- Predict the monthly rent for an apartment that has 1,000 square feet.

- Why would it not be appropriate to use the model to predict the monthly rent for apartments that have 500 square feet?
- Your friends Jim and Jennifer are considering signing a lease for an apartment in this residential neighborhood. They are trying to decide between two apartments, one with 1,000 square feet for a monthly rent of \$1,275 and the other with 1,200 square feet for a monthly rent of \$1,425. Based on (a) through (d), which apartment do you think is a better deal?

**13.10** A company that holds the DVD distribution rights to movies previously released only in theaters wants to estimate sales revenue of DVDs based on box office success. The box office gross (in \$millions) for each of 22 movies in the year that they were released and the DVD revenue (in \$millions) in the following year are shown below and stored in [Movie](#).

Title	Gross	DVD Revenue
<i>Bolt</i>	109.92	81.60
<i>Madagascar: Escape 2 Africa</i>	177.02	107.54
<i>Quantum of Solace</i>	166.82	44.41
<i>Beverly Hills Chihuahua</i>	93.78	60.21
<i>Marley and Me</i>	106.66	62.82
<i>High School Musical 3 Senior Year</i>	90.22	58.81
<i>Bedtime Stories</i>	85.54	48.79
<i>Role Models</i>	66.70	38.78
<i>Pineapple Express</i>	87.34	44.67
<i>Eagle Eye</i>	101.40	34.88
<i>Fireproof</i>	33.26	31.05
<i>Mamma Mia!</i>	144.13	33.14
<i>Seven Pounds</i>	60.15	27.12
<i>Australia</i>	46.69	28.16
<i>Valkyrie</i>	60.73	26.43
<i>Saw V</i>	56.75	26.10
<i>The Curious Case of Benjamin Button</i>	79.30	42.04
<i>Max Payne</i>	40.68	25.03
<i>Body of Lies</i>	39.32	21.45
<i>Nights in Rodanthe</i>	41.80	17.51
<i>Lakeview Terrace</i>	39.26	21.08
<i>The Spirit</i>	17.74	18.78

Sources: Data extracted from [www.the-numbers.com/market/movies2008.php](http://www.the-numbers.com/market/movies2008.php); and [www.the-numbers.com/dvd/charts/annual/2009.php](http://www.the-numbers.com/dvd/charts/annual/2009.php).

For these data,

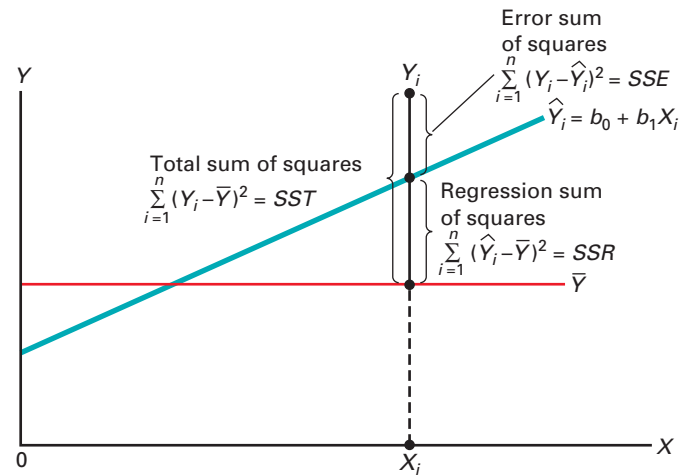
- construct a scatter plot.
- assuming a linear relationship, use the least-squares method to determine the regression coefficients  $b_0$  and  $b_1$ .
- interpret the meaning of the slope,  $b_1$ , in this problem.
- predict the sales revenue for a movie DVD that had a box office gross of \$75 million.



## 13.3 Measures of Variation

When using the least-squares method to determine the regression coefficients for a set of data, you need to compute three measures of variation. The first measure, the **total sum of squares (SST)**, is a measure of variation of the  $Y_i$  values around their mean,  $\bar{Y}$ . The **total variation**, or total sum of squares, is subdivided into **explained variation** and **unexplained variation**. The explained variation, or **regression sum of squares (SSR)**, represents variation that is explained by the relationship between  $X$  and  $Y$ , and the unexplained variation, or **error sum of squares (SSE)**, represents variation due to factors other than the relationship between  $X$  and  $Y$ . Figure 13.6 shows these different measures of variation.

**FIGURE 13.6**  
Measures of variation



### Computing the Sum of Squares

The regression sum of squares ( $SSR$ ) is based on the difference between  $\hat{Y}_i$  (the predicted value of  $Y$  from the prediction line) and  $\bar{Y}$  (the mean value of  $Y$ ). The error sum of squares ( $SSE$ ) represents the part of the variation in  $Y$  that is not explained by the regression. It is based on the difference between  $Y_i$  and  $\hat{Y}_i$ . Equations (13.5), (13.6), (13.7), and (13.8) define these measures of variation and the total sum of squares ( $SST$ ).

#### MEASURES OF VARIATION IN REGRESSION

The total sum of squares is equal to the regression sum of squares ( $SSR$ ) plus the error sum of squares ( $SSE$ ).

$$SST = SSR + SSE \quad (13.5)$$

#### TOTAL SUM OF SQUARES ( $SST$ )

The total sum of squares ( $SST$ ) is equal to the sum of the squared differences between each observed value of  $Y$  and the mean value of  $Y$ .

$$\begin{aligned} SST &= \text{Total sum of squares} \\ &= \sum_{i=1}^n (Y_i - \bar{Y})^2 \end{aligned} \quad (13.6)$$

REGRESSION SUM OF SQUARES (*SSR*)

The regression sum of squares (*SSR*) is equal to the sum of the squared differences between each predicted value of  $Y$  and the mean value of  $Y$ .

$$\begin{aligned} SSR &= \text{Explained variation or regression sum of squares} \\ &= \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 \end{aligned} \quad (13.7)$$

ERROR SUM OF SQUARES (*SSE*)

The error sum of squares (*SSE*) is equal to the sum of the squared differences between each observed value of  $Y$  and the predicted value of  $Y$ .

$$\begin{aligned} SSE &= \text{Unexplained variation or error sum of squares} \\ &= \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \end{aligned} \quad (13.8)$$

Figure 13.7 shows the sum of squares portion of the Figure 13.4 results for the Sunflowers Apparel data. The total variation, *SST*, is equal to 116.9543. This amount is subdivided into the sum of squares explained by the regression (*SSR*), equal to 105.7476, and the sum of squares unexplained by the regression (*SSE*), equal to 11.2067. From Equation (13.5) on page 533:

$$\begin{aligned} SST &= SSR + SSE \\ 116.9543 &= 105.7476 + 11.2067 \end{aligned}$$

FIGURE 13.7

Excel and Minitab sum of squares portion for the Sunflowers Apparel data

	A	B	C	D	E	F
10	ANOVA					
11		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
12	Regression	1	105.7476	105.7476	113.2335	0.0000
13	Residual	12	11.2067	0.9339		
14	Total	13	116.9543			

## Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	105.75	105.75	113.23	0.000
Residual Error	12	11.21	0.93		
Total	13	116.95			

## The Coefficient of Determination

By themselves, *SSR*, *SSE*, and *SST* provide little information. However, the ratio of the regression sum of squares (*SSR*) to the total sum of squares (*SST*) measures the proportion of variation in  $Y$  that is explained by the independent variable  $X$  in the regression model. This ratio, called the coefficient of determination,  $r^2$ , is defined in Equation (13.9).

## COEFFICIENT OF DETERMINATION

The coefficient of determination is equal to the regression sum of squares (i.e., explained variation) divided by the total sum of squares (i.e., total variation).

$$r^2 = \frac{\text{Regression sum of squares}}{\text{Total sum of squares}} = \frac{SSR}{SST} \quad (13.9)$$

The **coefficient of determination** measures the proportion of variation in  $Y$  that is explained by the variation in the independent variable  $X$  in the regression model.

For the Sunflowers Apparel data, with  $SSR = 105.7476$ ,  $SSE = 11.2067$ , and  $SST = 116.9543$ ,

$$r^2 = \frac{105.7476}{116.9543} = 0.9042$$

Therefore, 90.42% of the variation in annual sales is explained by the variability in the size of the store as measured by the square footage. This large  $r^2$  indicates a strong linear relationship between these two variables because the regression model has explained 90.42% of the variability in predicting annual sales. Only 9.58% of the sample variability in annual sales is due to factors other than what is accounted for by the linear regression model that uses square footage.

Figure 13.8 presents the regression statistics table portion of the Figure 13.4 results for the Sunflowers Apparel data. This table contains the coefficient of determination (labeled R Square in Excel and R-Sq in Minitab).

**FIGURE 13.8** Excel and Minitab regression statistics for the Sunflowers Apparel data

	A	B
3	<b>Regression Statistics</b>	
4	Multiple R	0.9509
5	R Square	0.9042
6	Adjusted R Square	0.8962
7	Standard Error	0.9664
8	Observations	14

**Predictor**      **Coef**    **SE Coef**      **T**      **P**  
**Constant**      **0.9645**    **0.5262**      **1.83**    **0.092**  
**Square Feet**    **1.6699**    **0.1569**      **10.64**   **0.000**  
**S = 0.966380    R-Sq = 90.4%    R-Sq(adj) = 89.6%**

### EXAMPLE 13.4

#### Computing the Coefficient of Determination

Compute the coefficient of determination,  $r^2$ , for the Sunflowers Apparel data.

**SOLUTION** You can compute  $SST$ ,  $SSR$ , and  $SSE$ , which are defined in Equations (13.6), (13.7), and (13.8) on pages 533 and 534, by using Equations (13.10), (13.11), and (13.12).

#### COMPUTATIONAL FORMULA FOR $SST$

$$SST = \sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n Y_i^2 - \frac{\left(\sum_{i=1}^n Y_i\right)^2}{n} \quad (13.10)$$

#### COMPUTATIONAL FORMULA FOR $SSR$

$$\begin{aligned} SSR &= \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 \\ &= b_0 \sum_{i=1}^n Y_i + b_1 \sum_{i=1}^n X_i Y_i - \frac{\left(\sum_{i=1}^n Y_i\right)^2}{n} \end{aligned} \quad (13.11)$$

#### COMPUTATIONAL FORMULA FOR $SSE$

$$SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n Y_i^2 - b_0 \sum_{i=1}^n Y_i - b_1 \sum_{i=1}^n X_i Y_i \quad (13.12)$$

Using the summary results from Table 13.2 on page 529,

$$\begin{aligned}
 SST &= \sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n Y_i^2 - \frac{\left(\sum_{i=1}^n Y_i\right)^2}{n} \\
 &= 594.9 - \frac{(81.8)^2}{14} \\
 &= 594.9 - 477.94571 \\
 &= 116.95429 \\
 SSR &= \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 \\
 &= b_0 \sum_{i=1}^n Y_i + b_1 \sum_{i=1}^n X_i Y_i - \frac{\left(\sum_{i=1}^n Y_i\right)^2}{n} \\
 &= (0.9645)(81.8) + (1.6699)(302.3) - \frac{(81.8)^2}{14} \\
 &= 105.74726 \\
 SSE &= \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \\
 &= \sum_{i=1}^n Y_i^2 - b_0 \sum_{i=1}^n Y_i - b_1 \sum_{i=1}^n X_i Y_i \\
 &= 594.9 - (0.9645)(81.8) - (1.6699)(302.3) \\
 &= 11.2067
 \end{aligned}$$

Therefore,

$$r^2 = \frac{105.74726}{116.95429} = 0.9042$$

## Standard Error of the Estimate

Although the least-squares method produces the line that fits the data with the minimum amount of prediction error, unless all the observed data points fall on a straight line, the prediction line is not a perfect predictor. Just as all data values cannot be expected to be exactly equal to their mean, neither can all the values in a regression analysis be expected to fall exactly on the prediction line. Figure 13.5 on page 526 illustrates the variability around the prediction line for the Sunflowers Apparel data. Observe that many of the actual values of  $Y$  fall near the prediction line, but none of the values are exactly on the line.

The **standard error of the estimate** measures the variability of the actual  $Y$  values from the predicted  $Y$  values in the same way that the standard deviation in Chapter 3 measures the variability of each value around the sample mean. In other words, the standard error of the estimate is the standard deviation *around* the prediction line, whereas the standard deviation in Chapter 3 is the standard deviation *around* the sample mean. Equation (13.13) defines the standard error of the estimate, represented by the symbol  $S_{YX}$ .

## STANDARD ERROR OF THE ESTIMATE

$$S_{YX} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2}} \quad (13.13)$$

where

$Y_i$  = actual value of  $Y$  for a given  $X_i$

$\hat{Y}_i$  = predicted value of  $Y$  for a given  $X_i$

$SSE$  = error sum of squares

From Equation (13.8) and Figure 13.4 or Figure 13.7 on pages 526 or 534,  $SSE = 11.2067$ . Thus,

$$S_{YX} = \sqrt{\frac{11.2067}{14-2}} = 0.9664$$

This standard error of the estimate, equal to 0.9664 millions of dollars (i.e., \$966,400), is labeled Standard Error in the Figure 13.8 Excel results and S in the Minitab results. The standard error of the estimate represents a measure of the variation around the prediction line. It is measured in the same units as the dependent variable  $Y$ . The interpretation of the standard error of the estimate is similar to that of the standard deviation. Just as the standard deviation measures variability around the mean, the standard error of the estimate measures variability around the prediction line. For Sunflowers Apparel, the typical difference between actual annual sales at a store and the predicted annual sales using the regression equation is approximately \$966,400.

## Problems for Section 13.3

### LEARNING THE BASICS

**13.11** How do you interpret a coefficient of determination,  $r^2$ , equal to 0.80?

**13.12** If  $SSR = 36$  and  $SSE = 4$ , determine  $SST$ , then compute the coefficient of determination,  $r^2$ , and interpret its meaning.

**13.13** If  $SSR = 66$  and  $SST = 88$ , compute the coefficient of determination,  $r^2$ , and interpret its meaning.

**13.14** If  $SSE = 10$  and  $SSR = 30$ , compute the coefficient of determination,  $r^2$ , and interpret its meaning.

**13.15** If  $SSR = 120$ , why is it impossible for  $SST$  to equal 110?

### APPLYING THE CONCEPTS



**13.16** In Problem 13.4 on page 531, the marketing manager used shelf space for pet food to predict

weekly sales (stored in **Petfood**). For those data,  $SSR = 20,535$  and  $SST = 30,025$ .

- Determine the coefficient of determination,  $r^2$ , and interpret its meaning.
- Determine the standard error of the estimate.
- How useful do you think this regression model is for predicting sales?

**13.17** In Problem 13.5 on page 531, you used the summarized rating to predict the cost of a restaurant meal (stored in **Restaurants**). For those data,  $SSR = 6,951.3963$  and  $SST = 15,890.11$

- Determine the coefficient of determination,  $r^2$ , and interpret its meaning.
- Determine the standard error of the estimate.
- How useful do you think this regression model is for predicting audited sales?

**13.18** In Problem 13.6 on page 531, an owner of a moving company wanted to predict labor hours, based on the



cubic feet moved (stored in **Moving**). Using the results of that problem,

- determine the coefficient of determination,  $r^2$ , and interpret its meaning.
- determine the standard error of the estimate.
- How useful do you think this regression model is for predicting labor hours?

**13.19** In Problem 13.7 on page 531, you used the number of customers to predict the waiting time at the checkout line in a supermarket (stored in **Supermarket**). Using the results of that problem,

- determine the coefficient of determination,  $r^2$ , and interpret its meaning.
- determine the standard error of the estimate.
- How useful do you think this regression model is for predicting the waiting time at the checkout line in a supermarket?

**13.20** In Problem 13.8 on page 531, you used annual revenues to predict the value of a baseball franchise (stored in **BBRevenue**). Using the results of that problem,

- determine the coefficient of determination,  $r^2$ , and interpret its meaning.
- determine the standard error of the estimate.

- How useful do you think this regression model is for predicting the value of a baseball franchise?

**13.21** In Problem 13.9 on page 532, an agent for a real estate company wanted to predict the monthly rent for apartments, based on the size of the apartment (stored in **Rent**). Using the results of that problem,

- determine the coefficient of determination,  $r^2$ , and interpret its meaning.
- determine the standard error of the estimate.
- How useful do you think this regression model is for predicting the monthly rent?
- Can you think of other variables that might explain the variation in monthly rent?

**13.22** In Problem 13.10 on page 532, you used box office gross to predict DVD revenue (stored in **Movie**). Using the results of that problem,

- determine the coefficient of determination,  $r^2$ , and interpret its meaning.
- determine the standard error of the estimate.
- How useful do you think this regression model is for predicting DVD revenue?
- Can you think of other variables that might explain the variation in DVD revenue?

## 13.4 Assumptions

When hypothesis testing and the analysis of variance were discussed in Chapters 9 through 12, the importance of the assumptions to the validity of any conclusions reached was emphasized. The assumptions necessary for regression are similar to those of the analysis of variance because both are part of the general category of *linear models* (reference 4).

The four **assumptions of regression** (known by the acronym LINE) are as follows:

- **Linearity**
- **Independence of errors**
- **Normality of error**
- **Equal variance**

The first assumption, **linearity**, states that the relationship between variables is linear. Relationships between variables that are not linear are discussed in Chapter 15.

The second assumption, **independence of errors**, requires that the errors ( $\varepsilon_i$ ) are independent of one another. This assumption is particularly important when data are collected over a period of time. In such situations, the errors in a specific time period are sometimes correlated with those of the previous time period.

The third assumption, **normality**, requires that the errors ( $\varepsilon_i$ ) are normally distributed at each value of  $X$ . Like the  $t$  test and the ANOVA  $F$  test, regression analysis is fairly robust against departures from the normality assumption. As long as the distribution of the errors at each level of  $X$  is not extremely different from a normal distribution, inferences about  $\beta_0$  and  $\beta_1$  are not seriously affected.

The fourth assumption, **equal variance**, or **homoscedasticity**, requires that the variance of the errors ( $\varepsilon_i$ ) be constant for all values of  $X$ . In other words, the variability of  $Y$  values is the same when  $X$  is a low value as when  $X$  is a high value. The equal-variance assumption is important when making inferences about  $\beta_0$  and  $\beta_1$ . If there are serious departures from this assumption, you can use either data transformations or weighted least-squares methods (see reference 4).

## 13.5 Residual Analysis

Sections 13.2 and 13.3 developed a regression model using the least-squares method for the Sunflowers Apparel data. Is this the correct model for these data? Are the assumptions presented in Section 13.4 valid? **Residual analysis** visually evaluates these assumptions and helps you to determine whether the regression model that has been selected is appropriate.

The **residual**, or estimated error value,  $e_i$ , is the difference between the observed ( $Y_i$ ) and predicted ( $\hat{Y}_i$ ) values of the dependent variable for a given value of  $X_i$ . A residual appears on a scatter plot as the vertical distance between an observed value of  $Y$  and the prediction line. Equation (13.14) defines the residual.

### RESIDUAL

The residual is equal to the difference between the observed value of  $Y$  and the predicted value of  $Y$ .

$$e_i = Y_i - \hat{Y}_i \quad (13.14)$$

### Evaluating the Assumptions

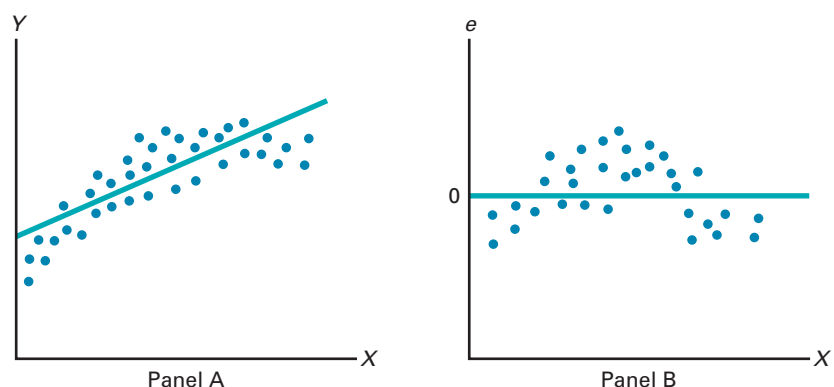
Recall from Section 13.4 that the four assumptions of regression (known by the acronym LINE) are linearity, independence, normality, and equal variance.

**Linearity** To evaluate linearity, you plot the residuals on the vertical axis against the corresponding  $X_i$  values of the independent variable on the horizontal axis. If the linear model is appropriate for the data, you will not see any apparent pattern in the plot. However, if the linear model is not appropriate, in the residual plot, there will be a relationship between the  $X_i$  values and the residuals,  $e_i$ .

You can see such a pattern in Figure 13.9. Panel A shows a situation in which, although there is an increasing trend in  $Y$  as  $X$  increases, the relationship seems curvilinear because the upward trend decreases for increasing values of  $X$ . This quadratic effect is highlighted in Panel B, where there is a clear relationship between  $X_i$  and  $e_i$ . By plotting the residuals, the linear trend of  $X$  with  $Y$  has been removed, thereby exposing the lack of fit in the simple linear model. Thus, a quadratic model is a better fit and should be used in place of the simple linear model. (See Section 15.1 for further discussion of fitting curvilinear models.)

**FIGURE 13.9**

Studying the appropriateness of the simple linear regression model



To determine whether the simple linear regression model is appropriate, return to the evaluation of the Sunflowers Apparel data. Figure 13.10 displays the predicted annual sales values and residuals.

**FIGURE 13.10**

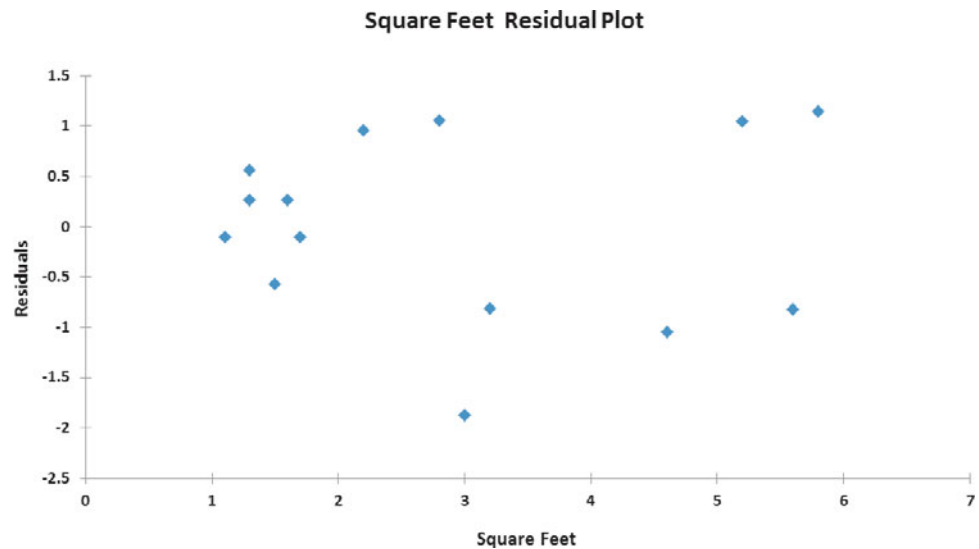
Table of residuals for the Sunflowers Apparel data

	A	B	C	D	E
1	Observation	Square Feet	Predicted Annual Sales	Annual Sales	Residuals
2	1	1.7	3.803239598	3.7	0.103239598
3	2	1.6	3.636253367	3.9	-0.263746633
4	3	2.8	5.640088147	6.7	-1.059911853
5	4	5.6	10.31570263	9.5	0.815702635
6	5	1.3	3.135294672	3.4	-0.264705328
7	6	2.2	4.638170757	5.6	-0.961829243
8	7	1.3	3.135294672	3.7	-0.564705328
9	8	1.1	2.801322208	2.7	0.101322208
10	9	3.2	6.308033074	5.5	0.808033074
11	10	1.5	3.469267135	2.9	0.569267135
12	11	5.2	9.647757708	10.7	-1.052242292
13	12	4.6	8.645840318	7.6	1.045840318
14	13	5.8	10.6496751	11.8	-1.150324902
15	14	3.0	5.974060611	4.1	1.874060611

To assess linearity, the residuals are plotted against the independent variable (store size, in thousands of square feet) in Figure 13.11. Although there is widespread scatter in the residual plot, there is no clear pattern or relationship between the residuals and  $X_i$ . The residuals appear to be evenly spread above and below 0 for different values of  $X$ . You can conclude that the linear model is appropriate for the Sunflowers Apparel data.

**FIGURE 13.11**

Plot of residuals against the square footage of a store for the Sunflowers Apparel data



**Independence** You can evaluate the assumption of independence of the errors by plotting the residuals in the order or sequence in which the data were collected. If the values of  $Y$  are part of a time series (see Section 2.6), one residual may sometimes be related to the previous residual. If this relationship exists between consecutive residuals (which violates the assumption of independence), the plot of the residuals versus the time in which the data were collected will often show a cyclical pattern. Because the Sunflowers Apparel data were collected during the same time period, you do not need to evaluate the independence assumption for these data.

**Normality** You can evaluate the assumption of normality in the errors by organizing the residuals into a frequency distribution as shown in Table 13.3. You cannot construct a meaningful histogram because the sample size is too small. And with such a small sample size ( $n = 14$ ), it can be difficult to evaluate the normality assumption by using a stem-and-leaf display (see Section 2.5), a boxplot (see Section 3.3), or a normal probability plot (see Section 6.3).

**TABLE 13.3**

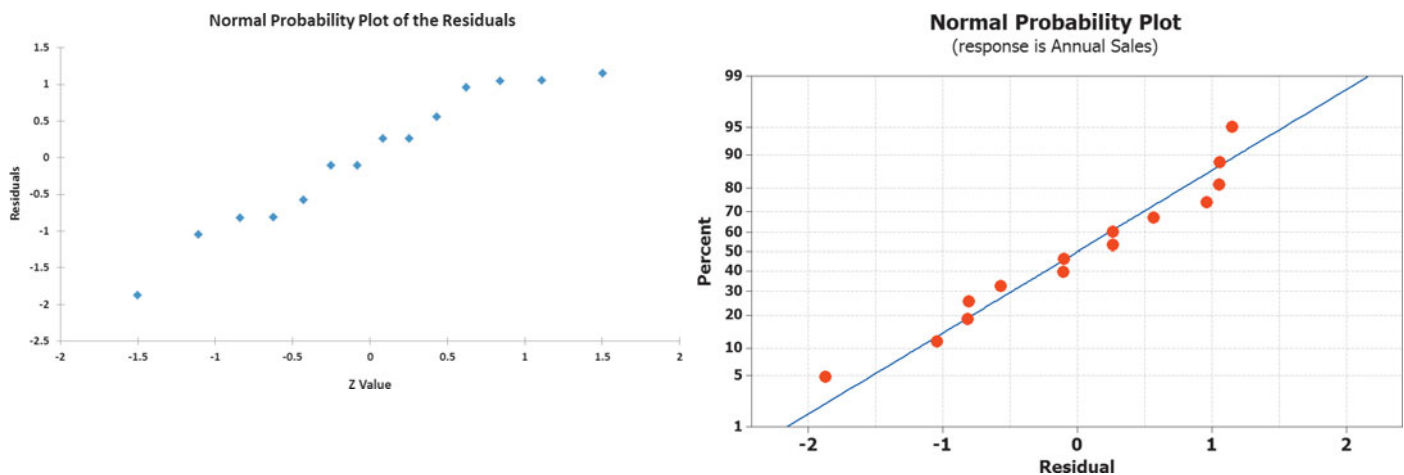
Frequency Distribution of 14 Residual Values for the Sunflowers Apparel Data

Residuals	Frequency
−2.25 but less than −1.75	1
−1.75 but less than −1.25	0
−1.25 but less than −0.75	3
−0.75 but less than −0.25	1
−0.25 but less than +0.25	2
+0.25 but less than +0.75	3
+0.75 but less than +1.25	4
	14

From the normal probability plot of the residuals in Figure 13.12, the data do not appear to depart substantially from a normal distribution. The robustness of regression analysis with modest departures from normality enables you to conclude that you should not be overly concerned about departures from this normality assumption in the Sunflowers Apparel data.

**FIGURE 13.12**

Excel and Minitab normal probability plots of the residuals for the Sunflowers Apparel data

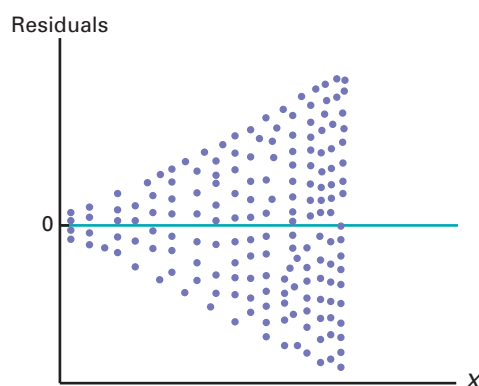


**Equal Variance** You can evaluate the assumption of equal variance from a plot of the residuals with  $X_i$ . For the Sunflowers Apparel data of Figure 13.11 on page 540, there do not appear to be major differences in the variability of the residuals for different  $X_i$  values. Thus, you can conclude that there is no apparent violation in the assumption of equal variance at each level of  $X$ .

To examine a case in which the equal-variance assumption is violated, observe Figure 13.13, which is a plot of the residuals with  $X_i$  for a hypothetical set of data. This plot is fan shaped because the variability of the residuals increases dramatically as  $X$  increases. Because this plot shows unequal variances of the residuals at different levels of  $X$ , the equal-variance assumption is invalid.

**FIGURE 13.13**

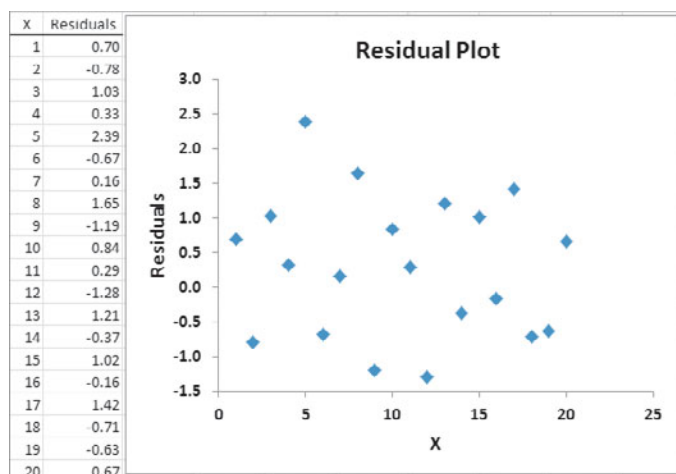
Violation of equal variance



## Problems for Section 13.5

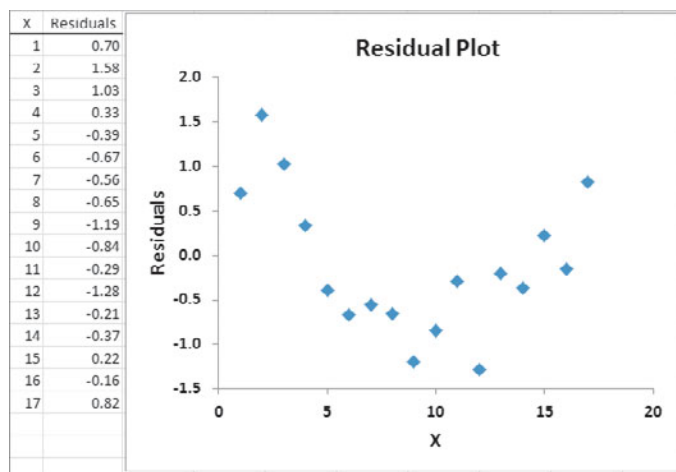
### LEARNING THE BASICS

**13.23** The following results provide the  $X$  values, residuals, and a residual plot from a regression analysis:



Is there any evidence of a pattern in the residuals? Explain.

**13.24** The following results show the  $X$  values, residuals, and a residual plot from a regression analysis:



Is there any evidence of a pattern in the residuals? Explain.

### APPLYING THE CONCEPTS

**13.25** In Problem 13.5 on page 531, you used the summated rating to predict the cost of a restaurant meal. Perform a residual analysis for these data (stored in **Restaurants**). Evaluate whether the assumptions of regression have been seriously violated.



**13.26** In Problem 13.4 on page 531, the marketing manager used shelf space for pet food to predict weekly sales. Perform a residual analysis for these data (stored in **Petfood**). Evaluate whether the assumptions of regression have been seriously violated.

**13.27** In Problem 13.7 on page 531, you used the number of customers to predict the waiting time at a supermarket checkout. Perform a residual analysis for these data (stored in **Supermarket**). Based on these results, evaluate whether the assumptions of regression have been seriously violated.

**13.28** In Problem 13.6 on page 531, the owner of a moving company wanted to predict labor hours based on the cubic feet moved. Perform a residual analysis for these data (stored in **Moving**). Based on these results, evaluate whether the assumptions of regression have been seriously violated.

**13.29** In Problem 13.9 on page 532, an agent for a real estate company wanted to predict the monthly rent for apartments, based on the size of the apartments. Perform a residual analysis for these data (stored in **Rent**). Based on these results, evaluate whether the assumptions of regression have been seriously violated.

**13.30** In Problem 13.8 on page 531, you used annual revenues to predict the value of a baseball franchise. Perform a residual analysis for these data (stored in **BBRevenue**). Based on these results, evaluate whether the assumptions of regression have been seriously violated.

**13.31** In Problem 13.10 on page 532, you used box office gross to predict DVD revenue. Perform a residual analysis for these data (stored in **Movie**). Based on these results, evaluate whether the assumptions of regression have been seriously violated.



## 13.6 Measuring Autocorrelation: The Durbin-Watson Statistic

One of the basic assumptions of the regression model is the independence of the errors. This assumption is sometimes violated when data are collected over sequential time periods because a residual at any one time period may tend to be similar to residuals at adjacent time periods. This pattern in the residuals is called **autocorrelation**. When a set of data has substantial autocorrelation, the validity of a regression model is in serious doubt.

### Residual Plots to Detect Autocorrelation

As mentioned in Section 13.5, one way to detect autocorrelation is to plot the residuals in time order. If a positive autocorrelation effect exists, there will be clusters of residuals with the same sign, and you will readily detect an apparent pattern. If negative autocorrelation exists, residuals will tend to jump back and forth from positive to negative to positive, and so on. This type of pattern is very rarely seen in regression analysis. Thus, the focus of this section is on positive autocorrelation. To illustrate positive autocorrelation, consider the following example.

The business problem faced by the manager of a package delivery store is to predict weekly sales. In approaching this problem, she has decided to develop a regression model to use the number of customers making purchases as an independent variable. Data are collected for a period of 15 weeks. Table 13.4 organizes the data (stored in **CustSale**).

**TABLE 13.4**

Customers and Sales  
for a Period of 15  
Consecutive Weeks

Week	Customers	Sales (\$Thousands)	Week	Customers	Sales (\$Thousands)
1	794	9.33	9	880	12.07
2	799	8.26	10	905	12.55
3	837	7.48	11	886	11.92
4	855	9.08	12	843	10.27
5	845	9.83	13	904	11.80
6	844	10.09	14	950	12.15
7	863	11.01	15	841	9.64
8	875	11.49			

Because the data are collected over a period of 15 consecutive weeks at the same store, you need to determine whether autocorrelation is present. Figure 13.14 presents results for these data.

**FIGURE 13.14**

Excel and Minitab regression results for the Table 13.4 package delivery store data

	A	B	C	D	E	F	G
1	Package Delivery Store Sales Analysis						
2							
3	Regression Statistics						
4	Multiple R	0.8108					
5	R Square	0.6574					
6	Adjusted R Square	0.6311					
7	Standard Error	0.9360					
8	Observations	15					
9							
10	ANOVA						
11		df	SS	MS	F	Significance F	
12	Regression	1	21.8604	21.8604	24.9501	0.0002	
13	Residual	13	11.3901	0.8762			
14	Total	14	33.2506				
15							
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
17	Intercept	-16.0322	5.3102	-3.0192	0.0099	-27.5041	-4.5603
18	Customers	0.0308	0.0062	4.9950	0.0002	0.0175	0.0441

### Regression Analysis: Sales versus Customers

The regression equation is

$$\text{Sales} = -16.0 + 0.0308 \text{ Customers}$$

Predictor	Coef	SE Coef	T	P
Constant	-16.032	5.310	-3.02	0.010
Customers	0.030760	0.006158	5.00	0.000

$$S = 0.936037 \quad R\text{-Sq} = 65.7\% \quad R\text{-Sq(adj)} = 63.1\%$$

### Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	21.860	21.860	24.95	0.000
Residual Error	13	11.390	0.876		
Total	14	33.251			

$$\text{Durbin-Watson statistic} = 0.883003$$

From Figure 13.14, observe that  $r^2$  is 0.6574, indicating that 65.74% of the variation in sales is explained by variation in the number of customers. In addition, the  $Y$  intercept,  $b_0$ , is  $-16.0322$ , and the slope,  $b_1$ , is  $0.0308$ . However, before using this model for prediction, you must perform a residual analysis. Because the data have been collected over a consecutive period of 15 weeks, in addition to checking the linearity, normality, and equal-variance assumptions, you must investigate the independence-of-errors assumption. To do this, you plot the residuals versus time in Figure 13.15 to help examine whether a pattern exists. In Figure 13.15, you can see that the residuals tend to fluctuate up and down in a cyclical pattern. This cyclical pattern provides strong cause for concern about the existence of autocorrelation in the residuals and, therefore, a violation of the independence-of-errors assumption.

**FIGURE 13.15**

Residual plot for the Table 13.4 package delivery store data



## The Durbin-Watson Statistic

The **Durbin-Watson statistic** is used to measure autocorrelation. This statistic measures the correlation between each residual and the residual for the previous time period. Equation (13.15) defines the Durbin-Watson statistic.

DURBIN-WATSON STATISTIC

$$D = \frac{\sum_{i=2}^n (e_i - e_{i-1})^2}{\sum_{i=1}^n e_i^2} \quad (13.15)$$

where

$e_i$  = residual at the time period  $i$

In Equation (13.15), the numerator,  $\sum_{i=2}^n (e_i - e_{i-1})^2$ , represents the squared difference between two successive residuals, summed from the second value to the  $n$ th value and the

denominator,  $\sum_{i=1}^n e_i^2$ , represents the sum of the squared residuals. This means that value of the Durbin-Watson statistic,  $D$ , will approach 0 if successive residuals are positively autocorrelated. If the residuals are not correlated, the value of  $D$  will be close to 2. (If the residuals are negatively autocorrelated,  $D$  will be greater than 2 and could even approach its maximum value of 4.) For the package delivery store data, the Durbin-Watson statistic,  $D$ , is 0.8830. (See the Figure 13.16 Excel results below or the Figure 13.14 Minitab results on page 543.)

**FIGURE 13.16**

Excel Durbin-Watson statistic worksheet for the package delivery store data

Minitab reports the Durbin-Watson statistic as part of its regression results. See Section MG13.6 for more information.

	A	B
1	Durbin-Watson Statistics	
2		
3	Sum of Squared Difference of Residuals	10.0575 =SUMXMY2(RESIDUALS!E3:E16, RESIDUALS!E2:E15)
4	Sum of Squared Residuals	11.3901 =SUMSQ(RESIDUALS!E2:E16)
5		
6	Durbin-Watson Statistic	0.8830 =B3/B4

You need to determine when the autocorrelation is large enough to conclude that there is significant positive autocorrelation. After computing  $D$ , you compare it to the critical values of the Durbin-Watson statistic found in Table E.8, a portion of which is presented in Table 13.5. The critical values depend on  $\alpha$ , the significance level chosen,  $n$ , the sample size, and  $k$ , the number of independent variables in the model (in simple linear regression,  $k = 1$ ).

**TABLE 13.5**

Finding Critical Values of the Durbin-Watson Statistic

$\alpha = .05$										
	$k = 1$		$k = 2$		$k = 3$		$k = 4$		$k = 5$	
$n$	$d_L$	$d_U$	$d_L$	$d_U$	$d_L$	$d_U$	$d_L$	$d_U$	$d_L$	$d_U$
15	1.08	1.36	.95	1.54	.82	1.75	.69	1.97	.56	2.21
16	1.10	1.37	.98	1.54	.86	1.73	.74	1.93	.62	2.15
17	1.13	1.38	1.02	1.54	.90	1.71	.78	1.90	.67	2.10
18	1.16	1.39	1.05	1.53	.93	1.69	.82	1.87	.71	2.06

In Table 13.5, two values are shown for each combination of  $\alpha$  (level of significance),  $n$  (sample size), and  $k$  (number of independent variables in the model). The first value,  $d_L$ , represents the lower critical value. If  $D$  is below  $d_L$ , you conclude that there is evidence of positive autocorrelation among the residuals. If this occurs, the least-squares method used in this chapter is inappropriate, and you should use alternative methods (see reference 4). The second value,  $d_U$ , represents the upper critical value of  $D$ , above which you would conclude that there is no evidence of positive autocorrelation among the residuals. If  $D$  is between  $d_L$  and  $d_U$ , you are unable to arrive at a definite conclusion.

For the package delivery store data, with one independent variable ( $k = 1$ ) and 15 values ( $n = 15$ ),  $d_L = 1.08$  and  $d_U = 1.36$ . Because  $D = 0.8830 < 1.08$ , you conclude that there is positive autocorrelation among the residuals. The least-squares regression analysis of the data is inappropriate because of the presence of significant positive autocorrelation among the residuals. In other words, the independence-of-errors assumption is invalid. You need to use alternative approaches, discussed in reference 4.

## Problems for Section 13.6

### LEARNING THE BASICS

**13.32** The residuals for 10 consecutive time periods are as follows:

Time Period	Residual	Time Period	Residual
1	-5	6	+1
2	-4	7	+2
3	-3	8	+3
4	-2	9	+4
5	-1	10	+5

- Plot the residuals over time. What conclusion can you reach about the pattern of the residuals over time?
- Based on (a), what conclusion can you reach about the autocorrelation of the residuals?

**13.33** The residuals for 15 consecutive time periods are as follows:

Time Period	Residual	Time Period	Residual
1	+4	9	+6
2	-6	10	-3
3	-1	11	+1
4	-5	12	+3
5	+2	13	0
6	+5	14	-4
7	-2	15	-7
8	+7		

- Plot the residuals over time. What conclusion can you reach about the pattern of the residuals over time?
- Compute the Durbin-Watson statistic. At the 0.05 level of significance, is there evidence of positive autocorrelation among the residuals?
- Based on (a) and (b), what conclusion can you reach about the autocorrelation of the residuals?

### APPLYING THE CONCEPTS

**13.34** In Problem 13.4 on page 531 concerning pet food sales, the marketing manager used shelf space for pet food to predict weekly sales.

- Is it necessary to compute the Durbin-Watson statistic in this case? Explain.
- Under what circumstances is it necessary to compute the Durbin-Watson statistic before proceeding with the least-squares method of regression analysis?

**13.35** What is the relationship between the price of crude oil and the price you pay at the pump for gasoline? The file **Oil & Gas** contains the price (\$) for a barrel of crude oil (Cushing, Oklahoma spot price) and a gallon of gasoline (New York Harbor Conventional spot price) for 104 weeks ending

June 25, 2010. (Data extracted from Energy Information Administration, U.S. Department of Energy, [www.eia.doe.gov](http://www.eia.doe.gov).)

- Construct a scatter plot with the price of oil on the horizontal axis and the price of gasoline on the vertical axis.
- Use the least-squares method to develop a simple linear regression equation to predict the price of a gallon of gasoline using the price of a barrel of crude oil as the independent variable.
- Interpret the meaning of the slope,  $b_1$ , in this problem.
- Plot the residuals versus the time period.
- Compute the Durbin-Watson statistic.
- At the 0.05 level of significance, is there evidence of positive autocorrelation among the residuals?
- Based on the results of (d) through (f), is there reason to question the validity of the model?

**SELF Test** **13.36** A mail-order catalog business that sells personal computer supplies, software, and hardware maintains a centralized warehouse for the distribution of products ordered. Management is currently examining the process of distribution from the warehouse and is interested in studying the factors that affect warehouse distribution costs. Currently, a small handling fee is added to the order, regardless of the amount of the order. Data that indicate the warehouse distribution costs and the number of orders received have been collected over the past 24 months and stored in **Warecost**. The results are

Months	Distribution Cost (\$Thousands)	Number of Orders
1	52.95	4,015
2	71.66	3,806
3	85.58	5,309
4	63.69	4,262
5	72.81	4,296
6	68.44	4,097
7	52.46	3,213
8	70.77	4,809
9	82.03	5,237
10	74.39	4,732
11	70.84	4,413
12	54.08	2,921
13	62.98	3,977
14	72.30	4,428
15	58.99	3,964
16	79.38	4,582
17	94.44	5,582
18	59.74	3,450
19	90.50	5,079
20	93.24	5,735
21	69.33	4,269
22	53.71	3,708
23	89.18	5,387
24	66.80	4,161

- Assuming a linear relationship, use the least-squares method to find the regression coefficients  $b_0$  and  $b_1$ .
- Predict the monthly warehouse distribution costs when the number of orders is 4,500.
- Plot the residuals versus the time period.
- Compute the Durbin-Watson statistic. At the 0.05 level of significance, is there evidence of positive autocorrelation among the residuals?
- Based on the results of (c) and (d), is there reason to question the validity of the model?

**13.37** A freshly brewed shot of espresso has three distinct components: the heart, body, and crema. The separation of these three components typically lasts only 10 to 20 seconds. To use the espresso shot in making a latte, a cappuccino, or another drink, the shot must be poured into the beverage during the separation of the heart, body, and crema. If the shot is used after the separation occurs, the drink becomes excessively bitter and acidic, ruining the final drink. Thus, a longer separation time allows the drink-maker more time to pour the shot and ensure that the beverage will meet expectations. An employee at a coffee shop hypothesized that the harder the espresso grounds were tamped down into the portafilter before brewing, the longer the separation time would be. An experiment using 24 observations was conducted to test this relationship. The independent variable **Tamp** measures the distance, in inches, between the espresso grounds and the top of the portafilter (i.e., the harder the tamp, the larger the distance). The dependent variable **Time** is the number of seconds the heart, body, and crema are separated (i.e., the amount of time after the shot is poured before it must be used for the customer's beverage). The data are stored in **Espresso** and are shown below:

- Use the least-squares method to develop a simple regression equation with **Time** as the dependent variable and **Tamp** as the independent variable.
- Predict the separation time for a tamp distance of 0.50 inch.
- Plot the residuals versus the time order of experimentation. Are there any noticeable patterns?
- Compute the Durbin-Watson statistic. At the 0.05 level of significance, is there evidence of positive autocorrelation among the residuals?

Shot	Tamp (Inches)	Time (Seconds)	Shot	Tamp (Inches)	Time (Seconds)
1	0.20	14	13	0.50	18
2	0.50	14	14	0.50	13
3	0.50	18	15	0.35	19
4	0.20	16	16	0.35	19
5	0.20	16	17	0.20	17
6	0.50	13	18	0.20	18
7	0.20	12	19	0.20	15
8	0.35	15	20	0.20	16
9	0.50	9	21	0.35	18
10	0.35	15	22	0.35	16
11	0.50	11	23	0.35	14
12	0.50	16	24	0.35	16

- Based on the results of (c) and (d), is there reason to question the validity of the model?

**13.38** The owner of a chain of ice cream stores has the business objective of improving the forecast of daily sales so that staffing shortages can be minimized during the summer season. The owner has decided to begin by developing a simple linear regression model to predict daily sales based on atmospheric temperature. A sample of 21 consecutive days is selected, and the results are stored in **IceCream**. (Hint: Determine which are the independent and dependent variables.)

- Assuming a linear relationship, use the least-squares method to compute the regression coefficients  $b_0$  and  $b_1$ .
- Predict the sales for a day in which the temperature is 83°F.
- Plot the residuals versus the time period.
- Compute the Durbin-Watson statistic. At the 0.05 level of significance, is there evidence of positive autocorrelation among the residuals?
- Based on the results of (c) and (d), is there reason to question the validity of the model?

## 13.7 Inferences About the Slope and Correlation Coefficient

In Sections 13.1 through 13.3, regression was used solely for descriptive purposes. You learned how the least-squares method determines the regression coefficients and how to predict  $Y$  for a given value of  $X$ . In addition, you learned how to compute and interpret the standard error of the estimate and the coefficient of determination.

When residual analysis, as discussed in Section 13.5, indicates that the assumptions of a least-squares regression model are not seriously violated and that the straight-line model is appropriate, you can make inferences about the linear relationship between the variables in the population.



## t Test for the Slope

To determine the existence of a significant linear relationship between the  $X$  and  $Y$  variables, you test whether  $\beta_1$  (the population slope) is equal to 0. The null and alternative hypotheses are as follows:

$$H_0: \beta_1 = 0 \text{ [There is no linear relationship (the slope is zero).]}$$

$$H_1: \beta_1 \neq 0 \text{ [There is a linear relationship (the slope is not zero).]}$$

If you reject the null hypothesis, you conclude that there is evidence of a linear relationship. Equation (13.16) defines the test statistic.

### TESTING A HYPOTHESIS FOR A POPULATION SLOPE, $\beta_1$ , USING THE $t$ TEST

The  $t_{STAT}$  test statistic equals the difference between the sample slope and hypothesized value of the population slope divided by the standard error of the slope.

$$t_{STAT} = \frac{b_1 - \beta_1}{S_{b_1}} \quad (13.16)$$

where

$$S_{b_1} = \frac{S_{YX}}{\sqrt{SSX}}$$

$$SSX = \sum_{i=1}^n (X_i - \bar{X})^2$$

The  $t_{STAT}$  test statistic follows a  $t$  distribution with  $n - 2$  degrees of freedom.

Return to the Sunflowers Apparel scenario on page 521. To test whether there is a significant linear relationship between the size of the store and the annual sales at the 0.05 level of significance, refer to the  $t$  test results shown in Figure 13.17.

**FIGURE 13.17**

Excel and Minitab  $t$  test results for the slope for the Sunflowers Apparel data

	A	B	C	D	E	F	G	H	I
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
17	Intercept	0.9645	0.5262	1.8329	0.0917	-0.1820	2.1110	-0.1820	2.11095
18	Square Feet	1.6699	0.1569	10.6411	0.0000	1.3280	2.0118	1.3280	2.01177

Predictor	Coef	SE Coef	T	P
Constant	0.9645	0.5262	1.83	0.092
Square Feet	1.6699	0.1569	10.64	0.000

From Figure 13.17,

$$b_1 = +1.6699 \quad n = 14 \quad S_{b_1} = 0.1569$$

and

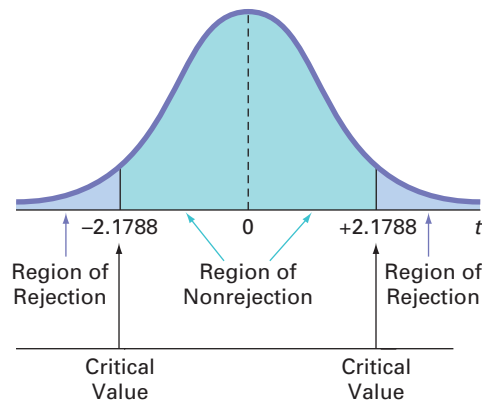
$$t_{STAT} = \frac{b_1 - \beta_1}{S_{b_1}}$$

$$= \frac{1.6699 - 0}{0.1569} = 10.6411$$

Using the 0.05 level of significance, the critical value of  $t$  with  $n - 2 = 12$  degrees of freedom is 2.1788. Because  $t_{STAT} = 10.6411 > 2.1788$  or because the  $p$ -value is approximately 0, which is less than  $\alpha = 0.05$ , you reject  $H_0$  (see Figure 13.18). Hence, you can conclude that there is a significant linear relationship between mean annual sales and the size of the store.

**FIGURE 13.18**

Testing a hypothesis about the population slope at the 0.05 level of significance, with 12 degrees of freedom



## F Test for the Slope

As an alternative to the  $t$  test, in simple linear regression, you can use an  $F$  test to determine whether the slope is statistically significant. In Section 10.4, you used the  $F$  distribution to test the ratio of two variances. Equation (13.17) defines the  $F$  test for the slope as the ratio of the variance that is due to the regression ( $MSR$ ) divided by the error variance ( $MSE = S^2_{YX}$ ).

### TESTING A HYPOTHESIS FOR A POPULATION SLOPE, $\beta_1$ , USING THE $F$ TEST

The  $F_{STAT}$  test statistic is equal to the regression mean square ( $MSR$ ) divided by the mean square error ( $MSE$ ).

$$F_{STAT} = \frac{MSR}{MSE} \quad (13.17)$$

where

$$MSR = \frac{SSR}{1} = SSR$$

$$MSE = \frac{SSE}{n - 2}$$

The  $F_{STAT}$  test statistic follows an  $F$  distribution with 1 and  $n - 2$  degrees of freedom.

Using a level of significance  $\alpha$ , the decision rule is

Reject  $H_0$  if  $F_{STAT} > F_\alpha$ ;

otherwise, do not reject  $H_0$ .

Table 13.6 organizes the complete set of results into an analysis of variance (ANOVA) table.

**TABLE 13.6**

ANOVA Table for Testing the Significance of a Regression Coefficient

Source	df	Sum of Squares	Mean Square (Variance)	F
Regression	1	SSR	$MSR = \frac{SSR}{1} = SSR$	$F_{STAT} = \frac{MSR}{MSE}$
Error	$n - 2$	SSE	$MSE = \frac{SSE}{n - 2}$	
Total	$n - 1$	SST		

Figure 13.19, a completed ANOVA table for the Sunflowers sales data, shows that the computed  $F_{STAT}$  test statistic is 113.2335 and the  $p$ -value is approximately 0.

**FIGURE 13.19**Excel and Minitab  $F$  test results for the Sunflowers Apparel data

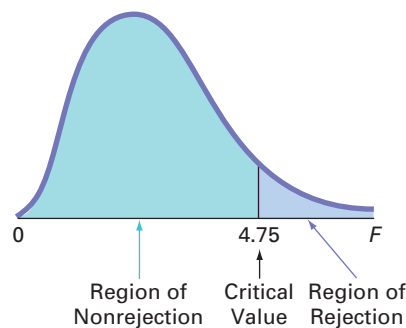
	A	B	C	D	E	F
10	ANOVA					
11		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
12	Regression	1	105.7476	105.7476	113.2335	0.0000
13	Residual	12	11.2067	0.9339		
14	Total	13	116.9543			

Analysis of Variance					
Source	DF	SS	MS	F	P
Regression	1	105.75	105.75	113.23	0.000
Residual Error	12	11.21	0.93		
Total	13	116.95			

Using a level of significance of 0.05, from Table E.5, the critical value of the  $F$  distribution, with 1 and 12 degrees of freedom, is 4.75 (see Figure 13.20). Because  $F_{STAT} = 113.2335 > 4.75$  or because the  $p$ -value = 0.0000 < 0.05, you reject  $H_0$  and conclude that there is a significant linear relationship between the size of the store and annual sales. Because the  $F$  test in Equation (13.17) on page 549 is equivalent to the  $t$  test in Equation (13.16) on page 548, you reach the same conclusion.

**FIGURE 13.20**

Regions of rejection and nonrejection when testing for the significance of the slope at the 0.05 level of significance, with 1 and 12 degrees of freedom



## Confidence Interval Estimate for the Slope

As an alternative to testing for the existence of a linear relationship between the variables, you can construct a confidence interval estimate of  $\beta_1$  using Equation (13.18).

### CONFIDENCE INTERVAL ESTIMATE OF THE SLOPE, $\beta_1$

The confidence interval estimate for the population slope can be constructed by taking the sample slope,  $b_1$ , and adding and subtracting the critical  $t$  value multiplied by the standard error of the slope.

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

(or)

$$b_1 - t_{\alpha/2} S_{b_1} \leq \beta_1 \leq b_1 + t_{\alpha/2} S_{b_1} \quad (13.18)$$

where

$t_{\alpha/2}$  = critical value corresponding to an upper-tail probability of  $\alpha/2$  from the  $t$  distribution with  $n - 2$  degrees of freedom (i.e., a cumulative area of  $1 - \alpha/2$ ).

From the Figure 13.17 results on page 548,

$$b_1 = 1.6699 \quad n = 14 \quad S_{b_1} = 0.1569$$

To construct a 95% confidence interval estimate,  $\alpha/2 = 0.025$ , and from Table E.3,  $t_{\alpha/2} = 2.1788$ . Thus,

$$\begin{aligned} b_1 \pm t_{\alpha/2} S_{b_1} &= 1.6699 \pm (2.1788)(0.1569) \\ &= 1.6699 \pm 0.3419 \\ 1.3280 &\leq \beta_1 \leq 2.0118 \end{aligned}$$

Therefore, you estimate with 95% confidence that the population slope is between 1.3280 and 2.0118. Because these values are both above 0, you conclude that there is a significant linear relationship between annual sales and the size of the store. Had the interval included 0, you would have concluded that no significant relationship exists between the variables. The confidence interval indicates that for each increase of 1,000 square feet, predicted annual sales are estimated to increase by at least \$1,328,000 but no more than \$2,011,800.

## t Test for the Correlation Coefficient

In Section 3.5 on page 127, the strength of the relationship between two numerical variables was measured using the **correlation coefficient**,  $r$ . The values of the coefficient of correlation range from  $-1$  for a perfect negative correlation to  $+1$  for a perfect positive correlation. You can use the correlation coefficient to determine whether there is a statistically significant linear relationship between  $X$  and  $Y$ . To do so, you hypothesize that the population correlation coefficient,  $\rho$ , is 0. Thus, the null and alternative hypotheses are

$$H_0: \rho = 0 \text{ (no correlation)}$$

$$H_1: \rho \neq 0 \text{ (correlation)}$$

Equation (13.19) defines the test statistic for determining the existence of a significant correlation.

### TESTING FOR THE EXISTENCE OF CORRELATION

$$t_{STAT} = \frac{r - \rho}{\sqrt{\frac{1 - r^2}{n - 2}}} \quad (13.19a)$$

where

$$r = +\sqrt{r^2} \text{ if } b_1 > 0$$

$$r = -\sqrt{r^2} \text{ if } b_1 < 0$$

The  $t_{STAT}$  test statistic follows a  $t$  distribution with  $n - 2$  degrees of freedom.  $r$  is calculated as follows:

$$r = \frac{\text{cov}(X, Y)}{S_X S_Y} \quad (13.19b)$$

where

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$

$$S_X = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}}$$

$$S_Y = \sqrt{\frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n - 1}}$$

In the Sunflowers Apparel problem,  $r^2 = 0.9042$  and  $b_1 = +1.6699$  (see Figure 13.4 on page 526). Because  $b_1 > 0$ , the correlation coefficient for annual sales and store size is the

positive square root of  $r^2$ , that is,  $r = +\sqrt{0.9042} = +0.9509$ . Using Equation (13.19a) to test the null hypothesis that there is no correlation between these two variables results in the following observed  $t$  statistic:

$$\begin{aligned} t_{STAT} &= \frac{r - 0}{\sqrt{\frac{1 - r^2}{n - 2}}} \\ &= \frac{0.9509 - 0}{\sqrt{\frac{1 - (0.9509)^2}{14 - 2}}} = 10.6411 \end{aligned}$$

Using the 0.05 level of significance, because  $t_{STAT} = 10.6411 > 2.1788$ , you reject the null hypothesis. You conclude that there is a significant association between annual sales and store size. This  $t_{STAT}$  test statistic is equivalent to the  $t_{STAT}$  test statistic found when testing whether the population slope,  $\beta_1$ , is equal to zero.

## Problems for Section 13.7

### LEARNING THE BASICS

**13.39** You are testing the null hypothesis that there is no linear relationship between two variables,  $X$  and  $Y$ . From your sample of  $n = 10$ , you determine that  $r = 0.80$ .

- What is the value of the  $t$  test statistic  $t_{STAT}$ ?
- At the  $\alpha = 0.05$  level of significance, what are the critical values?
- Based on your answers to (a) and (b), what statistical decision should you make?


**13.40** You are testing the null hypothesis that there is no linear relationship between two variables,  $X$  and  $Y$ . From your sample of  $n = 18$ , you determine that  $b_1 = +4.5$  and  $S_{b_1} = 1.5$ .

- What is the value of  $t_{STAT}$ ?
- At the  $\alpha = 0.05$  level of significance, what are the critical values?
- Based on your answers to (a) and (b), what statistical decision should you make?
- Construct a 95% confidence interval estimate of the population slope,  $\beta_1$ .

**13.41** You are testing the null hypothesis that there is no linear relationship between two variables,  $X$  and  $Y$ . From your sample of  $n = 20$ , you determine that  $SSR = 60$  and  $SSE = 40$ .

- What is the value of  $F_{STAT}$ ?
- At the  $\alpha = 0.05$  level of significance, what is the critical value?
- Based on your answers to (a) and (b), what statistical decision should you make?
- Compute the correlation coefficient by first computing  $r^2$  and assuming that  $b_1$  is negative.
- At the 0.05 level of significance, is there a significant correlation between  $X$  and  $Y$ ?

### APPLYING THE CONCEPTS

 **13.42** In Problem 13.4 on page 531, the marketing manager used shelf space for pet food to predict weekly sales. The data are stored in **Petfood**. From the results of that problem,  $b_1 = 7.4$  and  $S_{b_1} = 1.59$ .

- At the 0.05 level of significance, is there evidence of a linear relationship between shelf space and sales?
- Construct a 95% confidence interval estimate of the population slope,  $\beta_1$ .

**13.43** In Problem 13.5 on page 531, you used the summated rating of a restaurant to predict the cost of a meal. The data are stored in **Restaurants**. Using the results of that problem,  $b_1 = 1.2409$  and  $S_{b_1} = 0.1421$ .

- At the 0.05 level of significance, is there evidence of a linear relationship between the summated rating of a restaurant and the cost of a meal?
- Construct a 95% confidence interval estimate of the population slope,  $\beta_1$ .

**13.44** In Problem 13.6 on page 531, the owner of a moving company wanted to predict labor hours, based on the number of cubic feet moved. The data are stored in **Moving**. Use the results of that problem.

- At the 0.05 level of significance, is there evidence of a linear relationship between the number of cubic feet moved and labor hours?
- Construct a 95% confidence interval estimate of the population slope,  $\beta_1$ .

**13.45** In Problem 13.7 on page 531, you used the number of customers to predict the waiting time on the checkout line. The data are stored in **Supermarket**. Use the results of that problem.

- At the 0.05 level of significance, is there evidence of a linear relationship between the number of customers and the waiting time on the checkout line?
- Construct a 95% confidence interval estimate of the population slope,  $\beta_1$ .

**13.46** In Problem 13.8 on page 531, you used annual revenues to predict the value of a baseball franchise. The data are stored in **BBRevenue**. Use the results of that problem.

- At the 0.05 level of significance, is there evidence of a linear relationship between annual revenue and franchise value?
- Construct a 95% confidence interval estimate of the population slope,  $\beta_1$ .

**13.47** In Problem 13.9 on page 532, an agent for a real estate company wanted to predict the monthly rent for apartments, based on the size of the apartment. The data are stored in **Rent**. Use the results of that problem.

- At the 0.05 level of significance, is there evidence of a linear relationship between the size of the apartment and the monthly rent?
- Construct a 95% confidence interval estimate of the population slope,  $\beta_1$ .

**13.48** In Problem 13.10 on page 532, you used box office gross to predict DVD revenue. The data are stored in **Movie**. Use the results of that problem.

- At the 0.05 level of significance, is there evidence of a linear relationship between box office gross and DVD revenue?
- Construct a 95% confidence interval estimate of the population slope,  $\beta_1$ .

**13.49** The volatility of a stock is often measured by its beta value. You can estimate the beta value of a stock by developing a simple linear regression model, using the percentage weekly change in the stock as the dependent variable and the percentage weekly change in a market index as the independent variable. The S&P 500 Index is a common index to use. For example, if you wanted to estimate the beta for Disney, you could use the following model, which is sometimes referred to as a *market model*:

$$(\% \text{ weekly change in Disney}) = \beta_0 + \beta_1(\% \text{ weekly change in S \& P 500 index}) + \varepsilon$$

The least-squares regression estimate of the slope  $b_1$  is the estimate of the beta value for Disney. A stock with a beta value of 1.0 tends to move the same as the overall market. A stock with a beta value of 1.5 tends to move 50% more than the overall market, and a stock with a beta value of 0.6 tends to move only 60% as much as the overall market. Stocks with negative beta values tend to move in the opposite direction of the overall market. The following table gives some beta values for some widely held stocks as of July 7, 2010:

- For each of the six companies, interpret the beta value.
- How can investors use the beta value as a guide for investing?

Company	Ticker Symbol	Beta
Procter & Gamble	PG	0.53
AT&T	T	0.65
Disney	DIS	1.25
Apple	AAPL	1.43
eBay	EBAY	1.75
Ford	F	2.75

Source: Data extracted from **finance.yahoo.com**, July 7, 2010.

**13.50** Index funds are mutual funds that try to mimic the movement of leading indexes, such as the S&P 500 or the Russell 2000. The beta values (as described in Problem 13.49) for these funds are therefore approximately 1.0, and the estimated market models for these funds are approximately

$$(\% \text{ weekly change in index fund}) = 0.0 + 1.0(\% \text{ weekly change in the index})$$

Leveraged index funds are designed to magnify the movement of major indexes. Direxion Funds is a leading provider of leveraged index and other alternative-class mutual fund products for investment advisors and sophisticated investors. Two of the company's funds are shown in the following table. (Data extracted from **www.direxionfunds.com**, July 7, 2010.)

Name	Ticker Symbol	Description
Daily Small Cap 3x Fund	TNA	300% of the Russell 2000 Index
Daily India Bull 2x Fund	INDL	200% of the Indus India Index

The estimated market models for these funds are approximately

$$(\% \text{ weekly change in TNA}) = 0.0 + 3.0(\% \text{ weekly change in the Russell 2000})$$

$$(\% \text{ weekly change in INDL}) = 0.0 + 2.0(\% \text{ weekly change in the Indus India Index})$$

Thus, if the Russell 2000 Index gains 10% over a period of time, the leveraged mutual fund TNA gains approximately 30%. On the downside, if the same index loses 20%, TNA loses approximately 60%.

- The objective of the Direxion Funds Large Cap Bull 3x fund, BGU, is 300% of the performance of the Russell 1000 Index. What is its approximate market model?
- If the Russell 1000 Index gains 10% in a year, what return do you expect BGU to have?
- If the Russell 1000 Index loses 20% in a year, what return do you expect BGU to have?
- What type of investors should be attracted to leveraged index funds? What type of investors should stay away from these funds?



**13.51** The file **Cereals** contains the calories and sugar, in grams, in one serving of seven breakfast cereals:

Cereal	Calories	Sugar
Kellogg's All Bran	80	6
Kellogg's Corn Flakes	100	2
Wheaties	100	4
Nature's Path Organic Multigrain Flakes	110	4
Kellogg's Rice Krispies	130	4
Post Shredded Wheat	190	11
Vanilla Almond Kellogg Mini Wheats	200	10

- Compute and interpret the coefficient of correlation,  $r$ .
- At the 0.05 level of significance, is there a significant linear relationship between calories and sugar?

**13.52** Movie companies need to predict the gross receipts of an individual movie once the movie has debuted. The following results (stored in **PotterMovies**) are the first weekend gross, the U.S. gross, and the worldwide gross (in \$millions) of the six Harry Potter movies that debuted from 2001 to 2009:

Title	First Weekend	U.S. Gross	Worldwide Gross
<i>Sorcerer's Stone</i>	90.295	317.558	976.458
<i>Chamber of Secrets</i>	88.357	261.988	878.988
<i>Prisoner of Azkaban</i>	93.687	249.539	795.539
<i>Goblet of Fire</i>	102.335	290.013	896.013
<i>Order of the Phoenix</i>	77.108	292.005	938.469
<i>Half-Blood Prince</i>	77.836	301.460	934.601

Source: Data extracted from [www.the-numbers.com/interactive/comp-Harry-Potter.php](http://www.the-numbers.com/interactive/comp-Harry-Potter.php).

- Compute the coefficient of correlation between first weekend gross and the U.S. gross, first weekend gross and the worldwide gross, and the U.S. gross and worldwide gross.
- At the 0.05 level of significance, is there a significant linear relationship between first weekend gross and the U.S. gross, first weekend gross and the worldwide gross, and the U.S. gross and worldwide gross?

**13.53** College basketball is big business, with coaches' salaries, revenues, and expenses in millions of dollars. The file **College Basketball** contains the coaches' salary and revenue for college basketball at 60 of the 65 schools that played in the 2009 NCAA men's basketball tournament. (Data extracted from "Compensation for Division I Men's Basketball Coaches," *USA Today*, April 2, 2010, p. 8C; and C. Isadore, "Nothing but Net: Basketball Dollars by School," [money.cnn.com/2010/03/18/news/companies/basketball\\_profits/](http://money.cnn.com/2010/03/18/news/companies/basketball_profits/).)

- Compute and interpret the coefficient of correlation,  $r$ .
- At the 0.05 level of significance, is there a significant linear relationship between a coach's salary and revenue?

**13.54** College football players trying out for the NFL are given the Wonderlic standardized intelligence test. The file **Wonderlic** lists the average Wonderlic scores of football players trying out for the NFL and the graduation rates for football players at the schools they attended. (Data extracted from S. Walker, "The NFL's Smartest Team," *The Wall Street Journal*, September 30, 2005, pp. W1, W10.)

- Compute and interpret the coefficient of correlation,  $r$ .
- At the 0.05 level of significance, is there a significant linear relationship between the average Wonderlic score of football players trying out for the NFL and the graduation rates for football players at selected schools?
- What conclusions can you reach about the relationship between the average Wonderlic score of football players trying out for the NFL and the graduation rates for football players at selected schools?

## 13.8 Estimation of Mean Values and Prediction of Individual Values

In Chapter 8, you studied the concept of the confidence interval estimate of the population mean. In Example 13.2 on page 527, you used the prediction line to predict the mean value of  $Y$  for a given  $X$ . The annual sales for stores with 4,000 square feet was predicted to be 7.644 millions of dollars (\$7,644,000). This estimate, however, is a *point estimate* of the population mean. This section presents methods to develop a confidence interval estimate for the mean response for a given  $X$  and for developing a prediction interval for an individual response,  $Y$ , for a given value of  $X$ .

### The Confidence Interval Estimate

Equation (13.20) defines the **confidence interval estimate for the mean response** for a given  $X$ .

## CONFIDENCE INTERVAL ESTIMATE FOR THE MEAN OF Y

$$\hat{Y}_i \pm t_{\alpha/2} S_{YX} \sqrt{h_i}$$

$$\hat{Y}_i - t_{\alpha/2} S_{YX} \sqrt{h_i} \leq \mu_{Y|X=X_i} \leq \hat{Y}_i + t_{\alpha/2} S_{YX} \sqrt{h_i} \quad (13.20)$$

where

$$h_i = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{SSX}$$

$$\hat{Y}_i = \text{predicted value of } Y; \hat{Y}_i = b_0 + b_1 X_i$$

$S_{YX}$  = standard error of the estimate

$n$  = sample size

$X_i$  = given value of  $X$

$\mu_{Y|X=X_i}$  = mean value of  $Y$  when  $X = X_i$

$$SSX = \sum_{i=1}^n (X_i - \bar{X})^2$$

$t_{\alpha/2}$  = critical value corresponding to an upper-tail probability of  $\alpha/2$  from the  $t$  distribution with  $n - 2$  degrees of freedom (i.e., a cumulative area of  $1 - \alpha/2$ ).

The width of the confidence interval in Equation (13.20) depends on several factors. Increased variation around the prediction line, as measured by the standard error of the estimate, results in a wider interval. As you would expect, increased sample size reduces the width of the interval. In addition, the width of the interval varies at different values of  $X$ . When you predict  $Y$  for values of  $X$  close to  $\bar{X}$ , the interval is narrower than for predictions for  $X$  values further away from  $\bar{X}$ .

In the Sunflowers Apparel example, suppose you want to construct a 95% confidence interval estimate of the mean annual sales for the entire population of stores that contain 4,000 square feet ( $X = 4$ ). Using the simple linear regression equation,

$$\begin{aligned} \hat{Y}_i &= 0.9645 + 1.6699X_i \\ &= 0.9645 + 1.6699(4) = 7.6439 \text{ (millions of dollars)} \end{aligned}$$

Also, given the following:

$$\bar{X} = 2.9214 \quad S_{YX} = 0.9664$$

$$SSX = \sum_{i=1}^n (X_i - \bar{X})^2 = 37.9236$$

From Table E.3,  $t_{\alpha/2} = 2.1788$ . Thus,

$$\hat{Y}_i \pm t_{\alpha/2} S_{YX} \sqrt{h_i}$$

where

$$h_i = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{SSX}$$

so that

$$\begin{aligned} \hat{Y}_i \pm t_{\alpha/2} S_{YX} \sqrt{\frac{1}{n} + \frac{(X_i - \bar{X})^2}{SSX}} \\ &= 7.6439 \pm (2.1788)(0.9664) \sqrt{\frac{1}{14} + \frac{(4 - 2.9214)^2}{37.9236}} \\ &= 7.6439 \pm 0.6728 \end{aligned}$$

so

$$6.9711 \leq \mu_{Y|X=4} \leq 8.3167$$

Therefore, the 95% confidence interval estimate is that the mean annual sales are between \$6,971,100 and \$8,316,700 for the population of stores with 4,000 square feet.

## The Prediction Interval

In addition to constructing a confidence interval for the mean value of  $Y$ , you can also construct a prediction interval for an individual value of  $Y$ . Although the form of this interval is similar to that of the confidence interval estimate of Equation (13.20), the prediction interval is predicting an individual value, not estimating a mean. Equation (13.21) defines the **prediction interval for an individual response,  $Y$** , at a given value,  $X_i$ , denoted by  $Y_{X=X_i}$ .

PREDICTION INTERVAL FOR AN INDIVIDUAL RESPONSE,  $Y$

$$\hat{Y}_i \pm t_{\alpha/2} S_{YX} \sqrt{1 + h_i} \quad (13.21)$$

$$\hat{Y}_i - t_{\alpha/2} S_{YX} \sqrt{1 + h_i} \leq Y_{X=X_i} \leq \hat{Y}_i + t_{\alpha/2} S_{YX} \sqrt{1 + h_i}$$

where

$Y_{X=X_i}$  = future value of  $Y$  when  $X = X_i$

$t_{\alpha/2}$  = critical value corresponding to an upper-tail probability of  $\alpha/2$  from the  $t$  distribution with  $n - 2$  degrees of freedom (i.e., a cumulative area of  $1 - \alpha/2$ )

In addition,  $h_i$ ,  $\hat{Y}_i$ ,  $S_{YX}$ ,  $n$ , and  $X_i$  are defined as in Equation (13.20) on page 555.

To construct a 95% prediction interval of the annual sales for an individual store that contains 4,000 square feet ( $X = 4$ ), you first compute  $\hat{Y}_i$ . Using the prediction line:

$$\begin{aligned} \hat{Y}_i &= 0.9645 + 1.6699X_i \\ &= 0.9645 + 1.6699(4) \\ &= 7.6439 \text{ (millions of dollars)} \end{aligned}$$

Also, given the following:

$$\begin{aligned} \bar{X} &= 2.9214 \quad S_{YX} = 0.9664 \\ SSX &= \sum_{i=1}^n (X_i - \bar{X})^2 = 37.9236 \end{aligned}$$

From Table E.3,  $t_{\alpha/2} = 2.1788$ . Thus,

$$\hat{Y}_i \pm t_{\alpha/2} S_{YX} \sqrt{1 + h_i}$$

where

$$h_i = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

so that

$$\begin{aligned}\hat{Y}_i \pm t_{\alpha/2} S_{YX} \sqrt{1 + \frac{1}{n} + \frac{(X_i - \bar{X})^2}{SSX}} \\ = 7.6439 \pm (2.1788)(0.9664) \sqrt{1 + \frac{1}{14} + \frac{(4 - 2.9214)^2}{37.9236}} \\ = 7.6439 \pm 2.2104\end{aligned}$$

so

$$5.4335 \leq Y_{X=4} \leq 9.8543$$

Therefore, with 95% confidence, you predict that the annual sales for an individual store with 4,000 square feet is between \$5,433,500 and \$9,854,300.

Figure 13.21 presents results for the confidence interval estimate and the prediction interval for the Sunflowers Apparel data. If you compare the results of the confidence interval estimate and the prediction interval, you see that the width of the prediction interval for an individual store is much wider than the confidence interval estimate for the mean. Remember that there is much more variation in predicting an individual value than in estimating a mean value.

**FIGURE 13.21**

Excel and Minitab confidence interval estimate and prediction interval results for the Sunflowers Apparel data

	A	B
1	Confidence Interval Estimate and Prediction Interval	
2		
3	Data	
4	X Value	4
5	Confidence Level	95%
6		
7	Intermediate Calculations	
8	Sample Size	14 =COUNT(SLRData!A:A)
9	Degrees of Freedom	12 =B8 - 2
10	t Value	2.1788 =TINV(1 - B5, B9)
11	Sample Mean	2.9214 =AVERAGE(SLRData!A:A)
12	Sum of Squared Difference	37.9236 =DEVSQ(SLRData!A:A)
13	Standard Error of the Estimate	0.9664 =COMPUTE!B7
14	h Statistic	0.1021 =1/B8 + (B4 - B11)^2/B12
15	Predicted Y (YHat)	7.6439 =TREND(SLRData!B2:B15, SLRData!A2:A15, B4)
16		
17	For Average Y	
18	Interval Half Width	0.6728 =B10 * B13 * SQRT(B14)
19	Confidence Interval Lower Limit	6.9711 =B15 - B18
20	Confidence Interval Upper Limit	8.3167 =B15 + B18
21		
22	For Individual Response Y	
23	Interval Half Width	2.2104 =B10 * B13 * SQRT(1 + B14)
24	Prediction Interval Lower Limit	5.4335 =B15 - B23
25	Prediction Interval Upper Limit	9.8544 =B15 + B23

#### Predicted Values for New Observations

New Obs	Fit	SE Fit	95% CI	95% PI
1	7.644	0.309	(6.971, 8.317)	(5.433, 9.854)

#### Values of Predictors for New Observations

New Obs	Square Feet
1	4.00

## Problems for Section 13.8

### LEARNING THE BASICS

**13.55** Based on a sample of  $n = 20$ , the least-squares method was used to develop the following prediction line:

$$\hat{Y}_i = 5 + 3X_i.$$

In addition,

$$S_{YX} = 1.0 \quad \bar{X} = 2 \quad \sum_{i=1}^n (X_i - \bar{X})^2 = 20$$

- Construct a 95% confidence interval estimate of the population mean response for  $X = 2$ .
- Construct a 95% prediction interval of an individual response for  $X = 2$ .

**13.56** Based on a sample of  $n = 20$ , the least-squares method was used to develop the following prediction line:

$$\hat{Y}_i = 5 + 3X_i.$$

In addition,

$$S_{YX} = 1.0 \quad \bar{X} = 2 \quad \sum_{i=1}^n (X_i - \bar{X})^2 = 20$$


- Construct a 95% confidence interval estimate of the population mean response for  $X = 4$ .
- Construct a 95% prediction interval of an individual response for  $X = 4$ .

- c. Compare the results of (a) and (b) with those of Problem 13.55 (a) and (b). Which intervals are wider? Why?

### APPLYING THE CONCEPTS

**13.57** In Problem 13.5 on page 531, you used the summated rating of a restaurant to predict the cost of a meal. The data are stored in **Restaurants**. For these data,  $S_{YX} = 9.5505$  and  $h_i = 0.026844$  when  $X = 50$ .

- Construct a 95% confidence interval estimate of the mean cost of a meal for restaurants that have a summated rating of 50.
- Construct a 95% prediction interval of the cost of a meal for an individual restaurant that has a summated rating of 50.
- Explain the difference in the results in (a) and (b).

 **13.58** In Problem 13.4 on page 531, the marketing manager used shelf space for pet food to predict weekly sales. The data are stored in **Petfood**. For these data,  $S_{YX} = 30.81$  and  $h_i = 0.1373$  when  $X = 8$ .

- Construct a 95% confidence interval estimate of the mean weekly sales for all stores that have 8 feet of shelf space for pet food.
- Construct a 95% prediction interval of the weekly sales of an individual store that has 8 feet of shelf space for pet food.
- Explain the difference in the results in (a) and (b).

**13.59** In Problem 13.7 on page 531, you used the total number of customers in the store to predict the waiting time at the checkout counter. The data are stored in **Supermarket**.

- Construct a 95% confidence interval estimate of the mean waiting time for all customers when there are 20 customers in the store.
- Construct a 95% prediction interval of the waiting time for an individual customer when there are 20 customers in the store.
- Why is the interval in (a) narrower than the interval in (b)?

**13.60** In Problem 13.6 on page 531, the owner of a moving company wanted to predict labor hours based on the number of cubic feet moved. The data are stored in **Moving**.

- Construct a 95% confidence interval estimate of the mean labor hours for all moves of 500 cubic feet.
- Construct a 95% prediction interval of the labor hours of an individual move that has 500 cubic feet.
- Why is the interval in (a) narrower than the interval in (b)?

**13.61** In Problem 13.9 on page 532, an agent for a real estate company wanted to predict the monthly rent for apartments, based on the size of an apartment. The data are stored in **Rent**.

- Construct a 95% confidence interval estimate of the mean monthly rental for all apartments that are 1,000 square feet in size.
- Construct a 95% prediction interval of the monthly rental for an individual apartment that is 1,000 square feet in size.
- Explain the difference in the results in (a) and (b).

**13.62** In Problem 13.8 on page 531, you predicted the value of a baseball franchise, based on current revenue. The data are stored in **BBRevenue**.

- Construct a 95% confidence interval estimate of the mean value of all baseball franchises that generate \$200 million of annual revenue.
- Construct a 95% prediction interval of the value of an individual baseball franchise that generates \$200 million of annual revenue.
- Explain the difference in the results in (a) and (b).

**13.63** In Problem 13.10 on page 532, you used box office gross to predict DVD revenue. The data are stored in **Movie**. The company is about to release a movie on DVD that had a box office gross of \$75 million.

- What is the predicted DVD revenue?
- Which interval is more useful here, the confidence interval estimate of the mean or the prediction interval for an individual response? Explain.
- Construct and interpret the interval you selected in (b).

## 13.9 Pitfalls in Regression

Some of the pitfalls involved in using regression analysis are as follows:

- Lacking awareness of the assumptions of least-squares regression
- Not knowing how to evaluate the assumptions of least-squares regression
- Not knowing what the alternatives are to least-squares regression if a particular assumption is violated
- Using a regression model without knowledge of the subject matter
- Extrapolating outside the relevant range
- Concluding that a significant relationship identified in an observational study is due to a cause-and-effect relationship

The widespread availability of spreadsheet and statistical applications has made regression analysis much more feasible today than it once was. However, many users with access to such applications do not understand how to use regression analysis properly. Someone who is

not familiar with either the assumptions of regression or how to evaluate the assumptions cannot be expected to know what the alternatives to least-squares regression are if a particular assumption is violated.

The data in Table 13.7 (stored in **Anscombe**) illustrate the importance of using scatter plots and residual analysis to go beyond the basic number crunching of computing the  $Y$  intercept, the slope, and  $r^2$ .

**TABLE 13.7**

Four Sets of Artificial Data

Data Set A		Data Set B		Data Set C		Data Set D	
$X_i$	$Y_i$	$X_i$	$Y_i$	$X_i$	$Y_i$	$X_i$	$Y_i$
10	8.04	10	9.14	10	7.46	8	6.58
14	9.96	14	8.10	14	8.84	8	5.76
5	5.68	5	4.74	5	5.73	8	7.71
8	6.95	8	8.14	8	6.77	8	8.84
9	8.81	9	8.77	9	7.11	8	8.47
12	10.84	12	9.13	12	8.15	8	7.04
4	4.26	4	3.10	4	5.39	8	5.25
7	4.82	7	7.26	7	6.42	19	12.50
11	8.33	11	9.26	11	7.81	8	5.56
13	7.58	13	8.74	13	12.74	8	7.91
6	7.24	6	6.13	6	6.08	8	6.89

Source: Data extracted from F. J. Anscombe, "Graphs in Statistical Analysis," *The American Statistician*, 27 (1973), 17–21.

Anscombe (reference 1) showed that all four data sets given in Table 13.7 have the following identical results:

$$\hat{Y}_i = 3.0 + 0.5X_i$$

$$S_{YX} = 1.237$$

$$S_{b_1} = 0.118$$

$$r^2 = 0.667$$

$$SSR = \text{Explained variation} = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 = 27.51$$

$$SSE = \text{Unexplained variation} = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = 13.76$$

$$SST = \text{Total variation} = \sum_{i=1}^n (Y_i - \bar{Y})^2 = 41.27$$

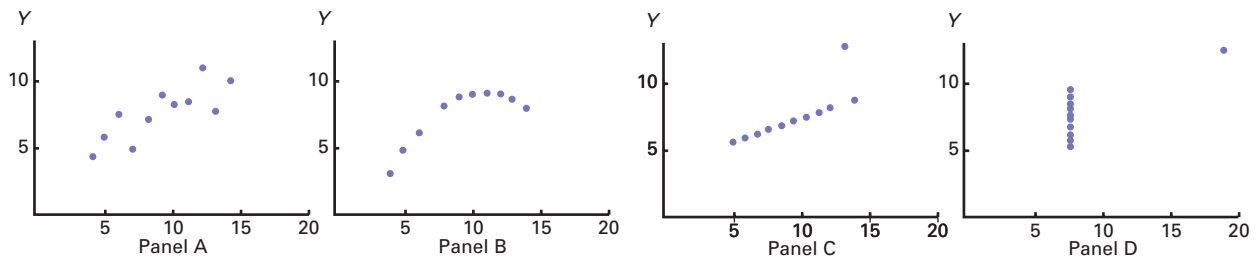
If you stopped the analysis at this point, you would fail to observe the important differences among the four data sets.

From the scatter plots of Figure 13.22 and the residual plots of Figure 13.23 on page 560, you see how different the data sets are. Each has a different relationship between  $X$  and  $Y$ . The only data set that seems to approximately follow a straight line is data set A. The residual plot for data set A does not show any obvious patterns or outlying residuals. This is certainly not true for data sets B, C, and D. The scatter plot for data set B shows that a curvilinear regression model is more appropriate. This conclusion is reinforced by the residual plot for data set B. The scatter plot and the residual plot for data set C clearly show an outlying observation. In this case, one approach used is to remove the outlier and reestimate the regression model (see reference 4). The scatter plot for data set D represents a situation in which the model is heavily dependent on the outcome of a single data point ( $X_8 = 19$  and  $Y_8 = 12.50$ ). Any regression model with this characteristic should be used with caution.

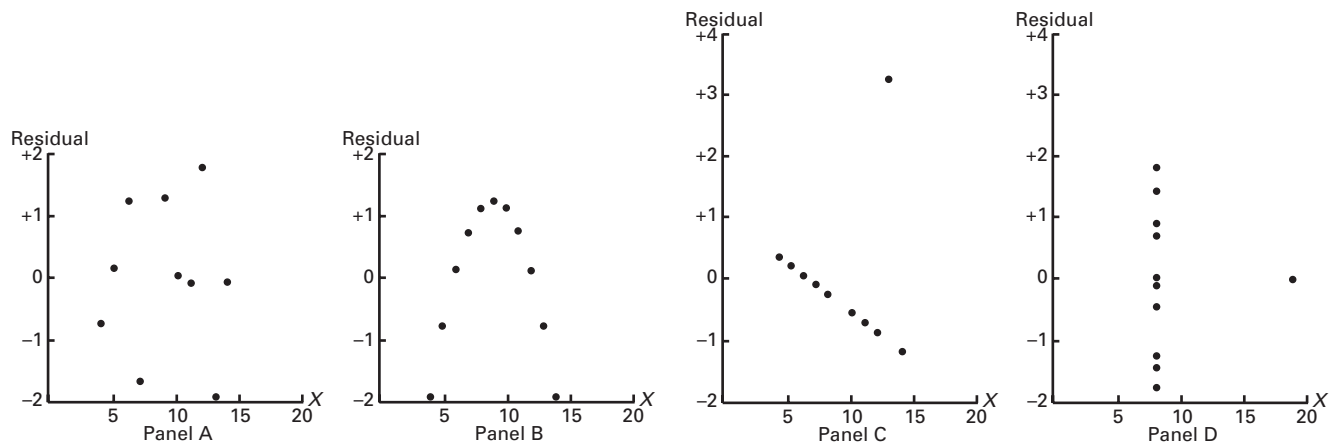


**FIGURE 13.22**

Scatter plots for four data sets

**FIGURE 13.23**

Residual plots for four data sets



In summary, scatter plots and residual plots are of vital importance to a complete regression analysis. The information they provide is so basic to a credible analysis that you should always include these graphical methods as part of a regression analysis. Thus, a strategy you can use to help avoid the pitfalls of regression is as follows:

1. Start with a scatter plot to observe the possible relationship between  $X$  and  $Y$ .
2. Check the assumptions of regression (**l**inearity, **i**ndependence, **n**ormality, **e**qual variance) by performing a residual analysis that includes the following:
  - a. Plotting the residuals versus the independent variable to determine whether the linear model is appropriate and to check for equal variance
  - b. Constructing a histogram, stem-and-leaf display, boxplot, or normal probability plot of the residuals to check for normality
  - c. Plotting the residuals versus time to check for independence (this step is necessary only if the data are collected over time)
3. If there are violations of the assumptions, use alternative methods to least-squares regression or alternative least-squares models (see reference 4).
4. If there are no violations of the assumptions, carry out tests for the significance of the regression coefficients and develop confidence and prediction intervals.
5. Avoid making predictions and forecasts outside the relevant range of the independent variable.

6. Keep in mind that the relationships identified in observational studies may or may not be due to cause-and-effect relationships. Remember that, although causation implies correlation, correlation does not imply causation.

## THINK ABOUT THIS By Any Other Name

You may not have frequently heard the phrase “regression model” outside a classroom, but the basic concepts of regression can be found under a variety of names in many sectors of the economy:

- **Advertising and marketing** Managers use econometric models (in other words, regression models) to determine the effect of an advertisement on sales, based on a set of factors. In one recent example, the number of tweets that mention specific products was used to make accurate prediction of sales trends. (See H. Rui, A. Whinston, and E. Winkler, “Follow the Tweets,” *The Wall Street Journal*, November 30, 2009, p. R4.) Also, managers use data mining to predict patterns of behavior of what customers will buy in the future, based on historic information about the consumer.
- **Finance** Any time you read about a financial “model,” you should assume that some type of regression model is being used. For example, a *New York Times* article on June 18, 2006, titled “An Old Formula That Points to New Worry” by Mark Hulbert (p. BU8), discusses a market timing model that predicts the returns of stocks in the next three to five years, based on the dividend yield of the stock market and the interest rate of 90-day Treasury bills.

- **Food and beverage** Enologix, a California consulting company, has developed a “formula” (a regression model) that predicts a wine’s quality index, based on a set of chemical compounds found in the wine. (See D. Darlington, “The Chemistry of a 90+ Wine,” *The New York Times Magazine*, August 7, 2005, pp. 36–39.)
- **Government** The Bureau of Labor Statistics uses hedonic models, a type of regression model, to adjust and manage its consumer price index (“Hedonic Quality Adjustment in the CPI,” *Consumer Price Index*, [stat.bls.gov/cpi/cpihqaitem.htm](http://stat.bls.gov/cpi/cpihqaitem.htm)).
- **Transportation** Bing Travel uses data mining and predictive technologies to objectively predict airfare pricing. (See C. Elliott, “Bing Travel’s Crean: We Save the Average Couple \$50 per Trip,” *Elliott Blog*, [www.elliott.org/first-person/bing-travel-we-save-the-average-couple-50-per-trip/](http://www.elliott.org/first-person/bing-travel-we-save-the-average-couple-50-per-trip/).)
- **Real estate** Zillow.com uses information about the features contained in a home and its location to develop estimates about the market value of the home, using a “formula” built with a proprietary model.

In a more general way, regression models are part of the “quants” movement that revolutionized Wall Street investing before moving on to

other fields (see S. Baker, “Why Math Will Rock Your World: More Math Geeks Are Calling the Shots in Business. Is Your Industry Next?” *BusinessWeek*, January 23, 2006, pp. 54–62). While the methods, including advanced regression models, that the quants used in Wall Street operations have been seen by some as the cause of the 2007 economic meltdown (see S. Patterson, *The Quants: How a New Breed of Math Whizzes Conquered Wall Street and Nearly Destroyed It*, New York: Crown Business, 2010), the rise of quants reflects a growing use of regression and other statistical techniques in business.

In his landmark 2006 *BusinessWeek* article, Baker predicted that statistics and probability will become core skills for businesspeople and consumers. He claimed that those who would become successful would know how to use statistics, whether they are building financial models or making marketing plans. More recent articles, including S. Lohr’s “For Today’s Graduate, Just One Word: Statistics” (*The New York Times*, August 6, 2009, pp. A1, A3) confirm Baker’s prediction and discussed how statistics is being used to “mine” large data sets to discover patterns, often using regression models. Hal Varian, the chief economist at Google, is quoted in that article as saying, “I keep saying that the sexy job in the next ten years will be statisticians.”



## USING STATISTICS

## @Sunflowers Apparel Revisited

In the Sunflowers Apparel scenario, you were the director of planning for a chain of upscale clothing stores for women. Until now, Sunflowers managers selected sites based on factors such as the availability of a good lease or a subjective opinion that a location seemed like a good place for a store. To make more objective decisions, you developed a regression model to analyze the relationship between the size of a store and its annual sales. The model indicated that about 90.4% of the variation in

sales was explained by the size of the store. Furthermore, for each increase of 1,000 square feet, mean annual sales were estimated to increase by \$1.67 million. You can now use your model to help make better decisions when selecting new sites for stores as well as to forecast sales for existing stores.

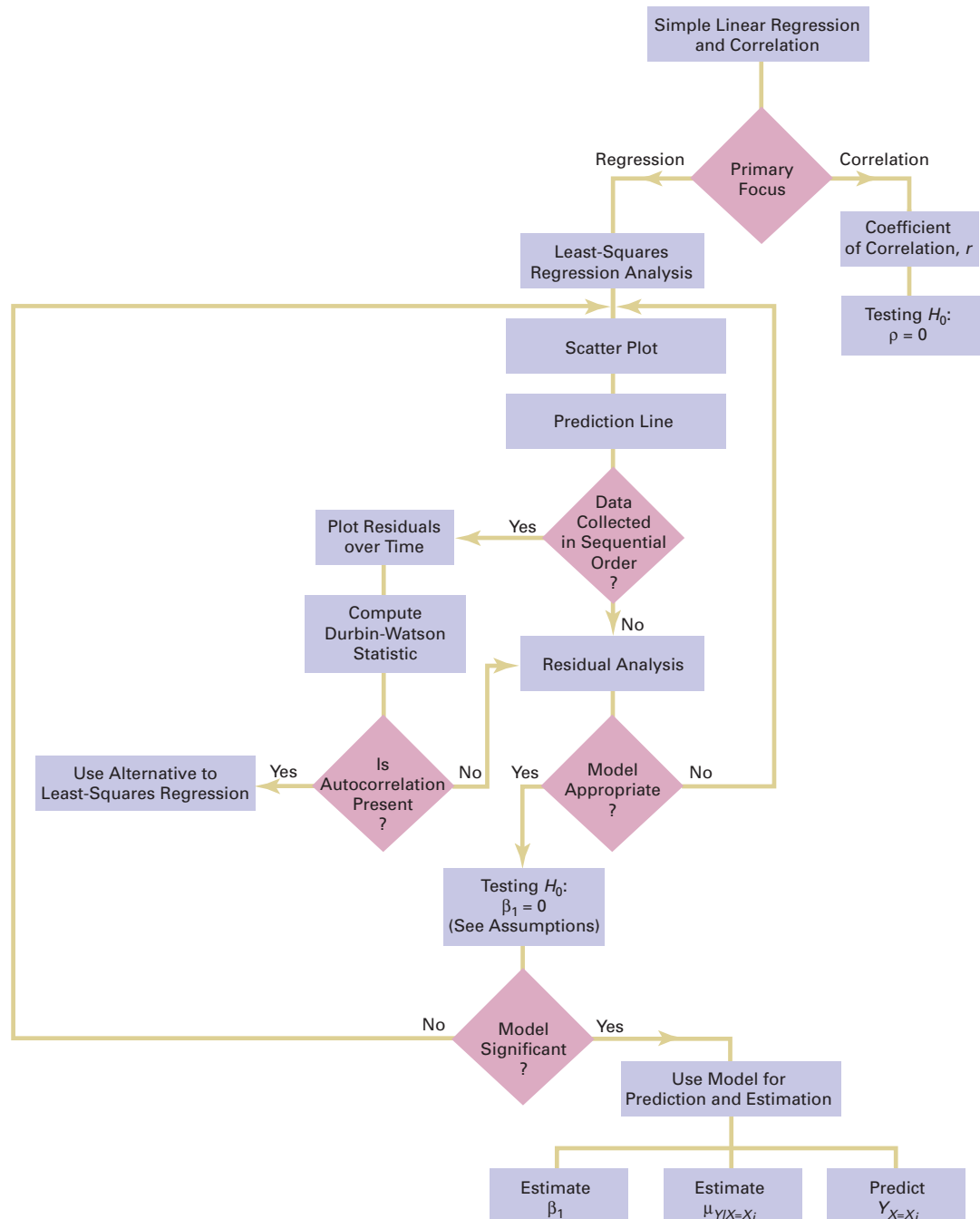
## SUMMARY

As you can see from the chapter roadmap in Figure 13.24, this chapter develops the simple linear regression model and discusses the assumptions and how to evaluate them. Once you are assured that the model is appropriate, you can predict

values by using the prediction line and test for the significance of the slope. In Chapter 14, regression analysis is extended to situations in which more than one independent variable is used to predict the value of a dependent variable.

**FIGURE 13.24**

Roadmap for simple linear regression



## KEY EQUATIONS

### Simple Linear Regression Model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad (13.1)$$

### Simple Linear Regression Equation: The Prediction Line

$$\hat{Y}_i = b_0 + b_1 X_i \quad (13.2)$$

### Computational Formula for the Slope, $b_1$

$$b_1 = \frac{SSXY}{SSX} \quad (13.3)$$

### Computational Formula for the $Y$ Intercept, $b_0$

$$b_0 = \bar{Y} - b_1 \bar{X} \quad (13.4)$$

### Measures of Variation in Regression

$$SST = SSR + SSE \quad (13.5)$$

### Total Sum of Squares ( $SST$ )

$$SST = \text{Total sum of squares} = \sum_{i=1}^n (Y_i - \bar{Y})^2 \quad (13.6)$$

### Regression Sum of Squares ( $SSR$ )

$$\begin{aligned} SSR &= \text{Explained variation or regression sum of squares} \\ &= \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 \end{aligned} \quad (13.7)$$

### Error Sum of Squares ( $SSE$ )

$$\begin{aligned} SSE &= \text{Unexplained variation or error sum of squares} \\ &= \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \end{aligned} \quad (13.8)$$

### Coefficient of Determination

$$r^2 = \frac{\text{Regression sum of squares}}{\text{Total sum of squares}} = \frac{SSR}{SST} \quad (13.9)$$

### Computational Formula for $SST$

$$SST = \sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n Y_i^2 - \frac{\left(\sum_{i=1}^n Y_i\right)^2}{n} \quad (13.10)$$

### Computational Formula for $SSR$

$$\begin{aligned} SSR &= \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 \\ &= b_0 \sum_{i=1}^n Y_i + b_1 \sum_{i=1}^n X_i Y_i - \frac{\left(\sum_{i=1}^n Y_i\right)^2}{n} \end{aligned} \quad (13.11)$$

### Computational Formula for $SSE$

$$SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n Y_i^2 - b_0 \sum_{i=1}^n Y_i - b_1 \sum_{i=1}^n X_i Y_i \quad (13.12)$$

### Standard Error of the Estimate

$$S_{YX} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2}} \quad (13.13)$$

### Residual

$$e_i = Y_i - \hat{Y}_i \quad (13.14)$$

### Durbin-Watson Statistic

$$D = \frac{\sum_{i=2}^n (e_i - e_{i-1})^2}{\sum_{i=1}^n e_i^2} \quad (13.15)$$

### Testing a Hypothesis for a Population Slope, $\beta_1$ , Using the $t$ Test

$$t_{STAT} = \frac{b_1 - \beta_1}{S_{b_1}} \quad (13.16)$$

### Testing a Hypothesis for a Population Slope, $\beta_1$ , Using the $F$ Test

$$F_{STAT} = \frac{MSR}{MSE} \quad (13.17)$$

### Confidence Interval Estimate of the Slope, $\beta_1$

$$\begin{aligned} &b_1 \pm t_{\alpha/2} S_{b_1} \\ &b_1 - t_{\alpha/2} S_{b_1} \leq \beta_1 \leq b_1 + t_{\alpha/2} S_{b_1} \end{aligned} \quad (13.18)$$

**Testing for the Existence of Correlation**

$$t_{STAT} = \frac{r - \rho}{\sqrt{\frac{1 - r^2}{n - 2}}} \quad (13.19a)$$

$$r = \frac{cov(X, Y)}{S_X S_Y} \quad (13.19b)$$

**Confidence Interval Estimate for the Mean of  $Y$** 

$$\hat{Y}_i \pm t_{\alpha/2} S_{YX} \sqrt{h_i}$$

$$\hat{Y}_i - t_{\alpha/2} S_{YX} \sqrt{h_i} \leq \mu_{Y|X=X_i} \leq \hat{Y}_i + t_{\alpha/2} S_{YX} \sqrt{h_i} \quad (13.20)$$

**Prediction Interval for an Individual Response,  $Y$** 

$$\hat{Y}_i \pm t_{\alpha/2} S_{YX} \sqrt{1 + h_i}$$

$$\hat{Y}_i - t_{\alpha/2} S_{YX} \sqrt{1 + h_i} \leq Y_{X=X_i} \leq \hat{Y}_i + t_{\alpha/2} S_{YX} \sqrt{1 + h_i} \quad (13.21)$$

**KEY TERMS**

assumptions of regression 538	independent variable 522	residual analysis 539
autocorrelation 543	least-squares method 525	response variable 522
coefficient of determination 534	linearity 538	scatter diagram 522
confidence interval estimate for the mean response 554	linear relationship 522	scatter plot 522
correlation coefficient 551	normality 538	simple linear regression 522
dependent variable 522	prediction interval for an individual response, $Y$ 556	simple linear regression equation 525
Durbin-Watson statistic 544	prediction line 525	slope 523
equal variance 538	regression analysis 522	standard error of the estimate 536
error sum of squares ( $SSE$ ) 533	regression coefficient 525	total sum of squares ( $SST$ ) 533
explained variation 533	regression sum of squares ( $SSR$ ) 533	total variation 533
explanatory variable 522	relevant range 527	unexplained variation 533
homoscedasticity 538	residual 539	$Y$ intercept 523
independence of errors 538		

**CHAPTER REVIEW PROBLEMS****CHECKING YOUR UNDERSTANDING**

**13.64** What is the interpretation of the  $Y$  intercept and the slope in the simple linear regression equation?

**13.65** What is the interpretation of the coefficient of determination?

**13.66** When is the unexplained variation (i.e., error sum of squares) equal to 0?

**13.67** When is the explained variation (i.e., regression sum of squares) equal to 0?

**13.68** Why should you always carry out a residual analysis as part of a regression model?

**13.69** What are the assumptions of regression analysis?

**13.70** How do you evaluate the assumptions of regression analysis?

**13.71** When and how do you use the Durbin-Watson statistic?

**13.72** What is the difference between a confidence interval estimate of the mean response,  $\mu_{Y|X=X_i}$ , and a prediction interval of  $Y_{X=X_i}$ ?

**APPLYING THE CONCEPTS**

**13.73** Researchers from the Pace University Lubin School of Business conducted a study on Internet-supported courses. In one part of the study, four numerical variables were collected on 108 students in an introductory management course that met once a week for an entire semester. One variable collected was *hit consistency*. To measure hit consistency,

tency, the researchers did the following: If a student did not visit the Internet site between classes, the student was given a 0 for that time period. If a student visited the Internet site one or more times between classes, the student was given a 1 for that time period. Because there were 13 time periods, a student's score on hit consistency could range from 0 to 13.

The other three variables included the student's course average, the student's cumulative grade point average (GPA), and the total number of hits the student had on the Internet site supporting the course. The following table gives the correlation coefficient for all pairs of variables. Note that correlations marked with an \* are statistically significant, using  $\alpha = 0.001$ :

Variable	Correlation
Course Average, Cumulative GPA	0.72*
Course Average, Total Hits	0.08
Course Average, Hit Consistency	0.37*
Cumulative GPA, Total Hits	0.12
Cumulative GPA, Hit Consistency	0.32*
Total Hits & Hit Consistency	0.64*

Source: Data extracted from D. Baugher, A. Varanelli, and E. Weisbord, "Student Hits in an Internet-Supported Course: How Can Instructors Use Them and What Do They Mean?" *Decision Sciences Journal of Innovative Education*, 1 (Fall 2003), 159–179.

- What conclusions can you reach from this correlation analysis?
- Are you surprised by the results, or are they consistent with your own observations and experiences?

**13.74** Management of a soft-drink bottling company has the business objective of developing a method for allocating delivery costs to customers. Although one cost clearly relates to travel time within a particular route, another variable cost reflects the time required to unload the cases of soft drink at the delivery point. To begin, management decided to develop a regression model to predict delivery time based on the number of cases delivered. A sample of 20 deliveries within a territory was selected. The delivery times and the number of cases delivered were organized in the following table (and stored in **Delivery**):

Customer	Number of Cases	Delivery Time (Minutes)	Customer	Number of Cases	Delivery Time (Minutes)
1	52	32.1	11	161	43.0
2	64	34.8	12	184	49.4
3	73	36.2	13	202	57.2
4	85	37.8	14	218	56.8
5	95	37.8	15	243	60.6
6	103	39.7	16	254	61.2
7	116	38.5	17	267	58.2
8	121	41.9	18	275	63.1
9	143	44.2	19	287	65.6
10	157	47.1	20	298	67.3

- Use the least-squares method to compute the regression coefficients  $b_0$  and  $b_1$ .
- Interpret the meaning of  $b_0$  and  $b_1$  in this problem.
- Predict the delivery time for 150 cases of soft drink.
- Should you use the model to predict the delivery time for a customer who is receiving 500 cases of soft drink? Why or why not?
- Determine the coefficient of determination,  $r^2$ , and explain its meaning in this problem.
- Perform a residual analysis. Is there any evidence of a pattern in the residuals? Explain.
- At the 0.05 level of significance, is there evidence of a linear relationship between delivery time and the number of cases delivered?
- Construct a 95% confidence interval estimate of the mean delivery time for 150 cases of soft drink and a 95% prediction interval of the delivery time for a single delivery of 150 cases of soft drink.

**13.75** Measuring the height of a California redwood tree is a very difficult undertaking because these trees grow to heights of over 300 feet. People familiar with these trees understand that the height of a California redwood tree is related to other characteristics of the tree, including the diameter of the tree at the breast height of a person. The data in **Redwood** represent the height (in feet) and diameter (in inches) at the breast height of a person for a sample of 21 California redwood trees.

- Assuming a linear relationship, use the least-squares method to compute the regression coefficients  $b_0$  and  $b_1$ . State the regression equation that predicts the height of a tree based on the tree's diameter at breast height of a person.
- Interpret the meaning of the slope in this equation.
- Predict the height for a tree that has a breast diameter of 25 inches.
- Interpret the meaning of the coefficient of determination in this problem.
- Perform a residual analysis on the results and determine the adequacy of the model.
- Determine whether there is a significant relationship between the height of redwood trees and the breast height diameter at the 0.05 level of significance.
- Construct a 95% confidence interval estimate of the population slope between the height of the redwood trees and breast diameter.

**13.76** You want to develop a model to predict the selling price of homes based on assessed value. A sample of 30 recently sold single-family houses in a small city is selected to study the relationship between selling price (in thousands of dollars) and assessed value (in thousands of dollars). The houses in the city were reassessed at full value one year prior to the study. The results are in **House1**. (Hint: First, determine which are the independent and dependent variables.)



- Construct a scatter plot and, assuming a linear relationship, use the least-squares method to compute the regression coefficients  $b_0$  and  $b_1$ .
- Interpret the meaning of the  $Y$  intercept,  $b_0$ , and the slope,  $b_1$ , in this problem.
- Use the prediction line developed in (a) to predict the selling price for a house whose assessed value is \$170,000.
- Determine the coefficient of determination,  $r^2$ , and interpret its meaning in this problem.
- Perform a residual analysis on your results and evaluate the regression assumptions.
- At the 0.05 level of significance, is there evidence of a linear relationship between selling price and assessed value?
- Construct a 95% confidence interval estimate of the population slope.

**13.77** You want to develop a model to predict the assessed value of houses, based on heating area. A sample of 15 single-family houses in a city is selected. The assessed value (in thousands of dollars) and the heating area of the houses (in thousands of square feet) are recorded; the results are stored in **House2**. (Hint: First, determine which are the independent and dependent variables.)

- Construct a scatter plot and, assuming a linear relationship, use the least-squares method to compute the regression coefficients  $b_0$  and  $b_1$ .
- Interpret the meaning of the  $Y$  intercept,  $b_0$ , and the slope,  $b_1$ , in this problem.
- Use the prediction line developed in (a) to predict the assessed value for a house whose heating area is 1,750 square feet.
- Determine the coefficient of determination,  $r^2$ , and interpret its meaning in this problem.
- Perform a residual analysis on your results and evaluate the regression assumptions.
- At the 0.05 level of significance, is there evidence of a linear relationship between assessed value and heating area?

**13.78** The director of graduate studies at a large college of business would like to predict the grade point average (GPA) of students in an MBA program based on Graduate Management Admission Test (GMAT) score. A sample of 20 students who have completed two years in the program is selected. The results are stored in **GPIGMAT**. (Hint: First, determine which are the independent and dependent variables.)

- Construct a scatter plot and, assuming a linear relationship, use the least-squares method to compute the regression coefficients  $b_0$  and  $b_1$ .
- Interpret the meaning of the  $Y$  intercept,  $b_0$ , and the slope,  $b_1$ , in this problem.
- Use the prediction line developed in (a) to predict the GPA for a student with a GMAT score of 600.
- Determine the coefficient of determination,  $r^2$ , and interpret its meaning in this problem.

- Perform a residual analysis on your results and evaluate the regression assumptions.
- At the 0.05 level of significance, is there evidence of a linear relationship between GMAT score and GPA?
- Construct a 95% confidence interval estimate of the mean GPA of students with a GMAT score of 600 and a 95% prediction interval of the GPA for a particular student with a GMAT score of 600.
- Construct a 95% confidence interval estimate of the population slope.

**13.79** An accountant for a large department store would like to develop a model to predict the amount of time it takes to process invoices. Data are collected from the past 32 working days, and the number of invoices processed and completion time (in hours) are stored in **Invoice**. (Hint: First, determine which are the independent and dependent variables.)

- Assuming a linear relationship, use the least-squares method to compute the regression coefficients  $b_0$  and  $b_1$ .
- Interpret the meaning of the  $Y$  intercept,  $b_0$ , and the slope,  $b_1$ , in this problem.
- Use the prediction line developed in (a) to predict the amount of time it would take to process 150 invoices.
- Determine the coefficient of determination,  $r^2$ , and interpret its meaning.
- Plot the residuals against the number of invoices processed and also against time.
- Based on the plots in (e), does the model seem appropriate?
- Based on the results in (e) and (f), what conclusions can you make about the validity of the prediction made in (c)?

**13.80** On January 28, 1986, the space shuttle *Challenger* exploded, and seven astronauts were killed. Prior to the launch, the predicted atmospheric temperature was for freezing weather at the launch site. Engineers for Morton Thiokol (the manufacturer of the rocket motor) prepared charts to make the case that the launch should not take place due to the cold weather. These arguments were rejected, and the launch tragically took place. Upon investigation after the tragedy, experts agreed that the disaster occurred because of leaky rubber O-rings that did not seal properly due to the cold temperature. Data indicating the atmospheric temperature at the time of 23 previous launches and the O-ring damage index are stored in **O-Ring**.

Note: Data from flight 4 is omitted due to unknown O-ring condition.

Sources: Data extracted from *Report of the Presidential Commission on the Space Shuttle Challenger Accident*, Washington, DC, 1986, Vol. II (H1–H3); and Vol. IV (664), and *Post Challenger Evaluation of Space Shuttle Risk Assessment and Management*, Washington, DC, 1988, pp. 135–136.

- Construct a scatter plot for the seven flights in which there was O-ring damage (O-ring damage index  $\neq 0$ ). What conclusions, if any, can you reach about the relationship between atmospheric temperature and O-ring damage?

- b. Construct a scatter plot for all 23 flights.
- c. Explain any differences in the interpretation of the relationship between atmospheric temperature and O-ring damage in (a) and (b).
- d. Based on the scatter plot in (b), provide reasons why a prediction should not be made for an atmospheric temperature of 31°F, the temperature on the morning of the launch of the *Challenger*.
- e. Although the assumption of a linear relationship may not be valid for the set of 23 flights, fit a simple linear regression model to predict O-ring damage, based on atmospheric temperature.
- f. Include the prediction line found in (e) on the scatter plot developed in (b).
- g. Based on the results in (f), do you think a linear model is appropriate for these data? Explain.
- h. Perform a residual analysis. What conclusions do you reach?

**13.81** Crazy Dave, a well-known baseball analyst, would like to study various team statistics for the 2009 baseball season to determine which variables might be useful in predicting the number of wins achieved by teams during the season. He has decided to begin by using a team's earned run average (ERA), a measure of pitching performance, to predict the number of wins. The data for the 30 Major League Baseball teams are stored in [BB2009](#). (Hint: First, determine which are the independent and dependent variables.)

- a. Assuming a linear relationship, use the least-squares method to compute the regression coefficients  $b_0$  and  $b_1$ .
- b. Interpret the meaning of the  $Y$  intercept,  $b_0$ , and the slope,  $b_1$ , in this problem.
- c. Use the prediction line developed in (a) to predict the number of wins for a team with an ERA of 4.50.
- d. Compute the coefficient of determination,  $r^2$ , and interpret its meaning.
- e. Perform a residual analysis on your results and determine the adequacy of the fit of the model.
- f. At the 0.05 level of significance, is there evidence of a linear relationship between the number of wins and the ERA?
- g. Construct a 95% confidence interval estimate of the mean number of wins expected for teams with an ERA of 4.50.
- h. Construct a 95% prediction interval of the number of wins for an individual team that has an ERA of 4.50.
- i. Construct a 95% confidence interval estimate of the population slope.
- j. The 30 teams constitute a population. In order to use statistical inference, as in (f) through (i), the data must be assumed to represent a random sample. What "population" would this sample be drawing conclusions about?
- k. What other independent variables might you consider for inclusion in the model?

**13.82** Can you use the annual revenues generated by National Basketball Association (NBA) franchises to predict franchise values? Figure 2.15 on page 57 shows a scatter plot of revenue with franchise value, and Figure 3.10 on page 129, shows the correlation coefficient. Now, you want to develop a simple linear regression model to predict franchise values based on revenues. (Franchise values and revenues are stored in [NBAValues](#).)

- a. Assuming a linear relationship, use the least-squares method to compute the regression coefficients  $b_0$  and  $b_1$ .
- b. Interpret the meaning of the  $Y$  intercept,  $b_0$ , and the slope,  $b_1$ , in this problem.
- c. Predict the value of an NBA franchise that generates \$150 million of annual revenue.
- d. Compute the coefficient of determination,  $r^2$ , and interpret its meaning.
- e. Perform a residual analysis on your results and evaluate the regression assumptions.
- f. At the 0.05 level of significance, is there evidence of a linear relationship between the annual revenues generated and the value of an NBA franchise?
- g. Construct a 95% confidence interval estimate of the mean value of all NBA franchises that generate \$150 million of annual revenue.
- h. Construct a 95% prediction interval of the value of an individual NBA franchise that generates \$150 million of annual revenue.
- i. Compare the results of (a) through (h) to those of baseball franchises in Problems 13.8, 13.20, 13.30, 13.46, and 13.62 and European soccer teams in Problem 13.83.

**13.83** In Problem 13.82 you used annual revenue to develop a model to predict the franchise value of National Basketball Association (NBA) teams. Can you also use the annual revenues generated by European soccer teams to predict franchise values? (European soccer team values and revenues are stored in [SoccerValues](#).)

- a. Repeat Problem 13.82 (a) through (h) for the European soccer teams.
- b. Compare the results of (a) to those of baseball franchises in Problems 13.8, 13.20, 13.30, 13.46, and 13.62 and NBA franchises in Problem 13.82.

**13.84** During the fall harvest season in the United States, pumpkins are sold in large quantities at farm stands. Often, instead of weighing the pumpkins prior to sale, the farm stand operator will just place the pumpkin in the appropriate circular cutout on the counter. When asked why this was done, one farmer replied, "I can tell the weight of the pumpkin from its circumference." To determine whether this was really true, a sample of 23 pumpkins were measured for circumference and weighed; the results are stored in [Pumpkin](#).

- a. Assuming a linear relationship, use the least-squares method to compute the regression coefficients  $b_0$  and  $b_1$ .
- b. Interpret the meaning of the slope,  $b_1$ , in this problem.

- c. Predict the weight for a pumpkin that is 60 centimeters in circumference.
- d. Do you think it is a good idea for the farmer to sell pumpkins by circumference instead of weight? Explain.
- e. Determine the coefficient of determination,  $r^2$ , and interpret its meaning.
- f. Perform a residual analysis for these data and evaluate the regression assumptions.
- g. At the 0.05 level of significance, is there evidence of a linear relationship between the circumference and weight of a pumpkin?
- h. Construct a 95% confidence interval estimate of the population slope,  $\beta_1$ .

**13.85** Can demographic information be helpful in predicting sales at sporting goods stores? The file **Sporting** contains the monthly sales totals from a random sample of 38 stores in a large chain of nationwide sporting goods stores. All stores in the franchise, and thus within the sample, are approximately the same size and carry the same merchandise. The county or, in some cases, counties in which the store draws the majority of its customers is referred to here as the customer base. For each of the 38 stores, demographic information about the customer base is provided. The data are real, but the name of the franchise is not used, at the request of the company. The data set contains the following variables:

Sales—Latest one-month sales total (dollars)  
 Age—Median age of customer base (years)  
 HS—Percentage of customer base with a high school diploma  
 College—Percentage of customer base with a college diploma  
 Growth—Annual population growth rate of customer base over the past 10 years  
 Income—Median family income of customer base (dollars)

- a. Construct a scatter plot, using sales as the dependent variable and median family income as the independent variable. Discuss the scatter plot.
- b. Assuming a linear relationship, use the least-squares method to compute the regression coefficients  $b_0$  and  $b_1$ .
- c. Interpret the meaning of the  $Y$  intercept,  $b_0$ , and the slope,  $b_1$ , in this problem.
- d. Compute the coefficient of determination,  $r^2$ , and interpret its meaning.
- e. Perform a residual analysis on your results and determine the adequacy of the fit of the model.
- f. At the 0.05 level of significance, is there evidence of a linear relationship between the independent variable and the dependent variable?
- g. Construct a 95% confidence interval estimate of the population slope and interpret its meaning.

**13.86** For the data of Problem 13.85, repeat (a) through (g), using Age as the independent variable.

**13.87** For the data of Problem 13.85, repeat (a) through (g), using HS as the independent variable.

**13.88** For the data of Problem 13.85, repeat (a) through (g), using College as the independent variable.

**13.89** For the data of Problem 13.85, repeat (a) through (g), using Growth as the independent variable.

**13.90** The file **CEO-Compensation** includes the total compensation (in \$) of CEOs of 197 large public companies and their investment return in 2009.

Source: Data extracted from D. Leonard, "Bargains in the Boardroom," *The New York Times*, April 4, 2010, pp. BU1, BU7, BU10, BU11.

- a. Compute the correlation coefficient between compensation and the investment return in 2009.
- b. At the 0.05 level of significance, is the correlation between compensation and the investment return in 2009 statistically significant?
- c. Write a short summary of your findings in (a) and (b). Do the results surprise you?

**13.91** Refer to the discussion of beta values and market models in Problem 13.49 on page 553. The S&P 500 Index tracks the overall movement of the stock market by considering the stock prices of 500 large corporations. The file **StockPrices** contains 2009 weekly data for the S&P 500 and three companies. The following variables are included:

WEEK—Week ending on date given  
 S&P—Weekly closing value for the S&P 500 Index  
 GE—Weekly closing stock price for General Electric  
 DISC—Weekly closing stock price for Discovery Communications  
 AAPL—Weekly closing stock price for Apple

Source: Data extracted from [finance.yahoo.com](http://finance.yahoo.com), June 3, 2010.

- a. Estimate the market model for GE. (Hint: Use the percentage change in the S&P 500 Index as the independent variable and the percentage change in GE's stock price as the dependent variable.)
- b. Interpret the beta value for GE.
- c. Repeat (a) and (b) for Discovery.
- d. Repeat (a) and (b) for Apple.
- e. Write a brief summary of your findings.

## REPORT WRITING EXERCISE

**13.92** In Problems 13.85 through 13.89, you developed regression models to predict monthly sales at a sporting goods store. Now, write a report based on the models you developed. Append to your report all appropriate charts and statistical information.

## MANAGING ASHLAND MULTICOMM SERVICES

To ensure that as many trial subscriptions to the *3-For-All* service as possible are converted to regular subscriptions, the marketing department works closely with the customer support department to accomplish a smooth initial process for the trial subscription customers. To assist in this effort, the marketing department needs to accurately forecast the monthly total of new regular subscriptions.

A team consisting of managers from the marketing and customer support departments was convened to develop a better method of forecasting new subscriptions. Previously, after examining new subscription data for the prior three months, a group of three managers would develop a subjective forecast of the number of new subscriptions. Livia Salvador, who was recently hired by the company to provide expertise in quantitative forecasting methods, suggested that the department look for factors that might help in predicting new subscriptions.

Members of the team found that the forecasts in the past year had been particularly inaccurate because in some months, much more time was spent on telemarketing than in other months. Livia collected data (stored in **AMS13**) for the number of new subscriptions and hours spent on telemarketing for each month for the past two years.

### EXERCISES

1. What criticism can you make concerning the method of forecasting that involved taking the new subscriptions data for the prior three months as the basis for future projections?
2. What factors other than number of telemarketing hours spent might be useful in predicting the number of new subscriptions? Explain.
3.
  - a. Analyze the data and develop a regression model to predict the number of new subscriptions for a month, based on the number of hours spent on telemarketing for new subscriptions.
  - b. If you expect to spend 1,200 hours on telemarketing per month, estimate the number of new subscriptions for the month. Indicate the assumptions on which this prediction is based. Do you think these assumptions are valid? Explain.
  - c. What would be the danger of predicting the number of new subscriptions for a month in which 2,000 hours were spent on telemarketing?

## DIGITAL CASE

*Apply your knowledge of simple linear regression in this Digital Case, which extends the Sunflowers Apparel Using Statistics scenario from this chapter.*

Leasing agents from the Triangle Mall Management Corporation have suggested that Sunflowers consider several locations in some of Triangle's newly renovated lifestyle malls that cater to shoppers with higher-than-mean disposable income. Although the locations are smaller than the typical Sunflowers location, the leasing agents argue that higher-than-mean disposable income in the surrounding community is a better predictor than store size of higher sales. The leasing agents maintain that sample data from 14 Sunflowers stores prove that this is true.

Open **Triangle\_Sunflower.pdf** and review the leasing agents' proposal and supporting documents. Then answer the following questions:

1. Should mean disposable income be used to predict sales based on the sample of 14 Sunflowers stores?
2. Should the management of Sunflowers accept the claims of Triangle's leasing agents? Why or why not?
3. Is it possible that the mean disposable income of the surrounding area is not an important factor in leasing new locations? Explain.
4. Are there any other factors not mentioned by the leasing agents that might be relevant to the store leasing decision?

## REFERENCES

1. Anscombe, F. J., “Graphs in Statistical Analysis,” *The American Statistician*, 27 (1973), 17–21.
2. Hoaglin, D. C., and R. Welsch, “The Hat Matrix in Regression and ANOVA,” *The American Statistician*, 32 (1978), 17–22.
3. Hocking, R. R., “Developments in Linear Regression Methodology: 1959–1982,” *Technometrics*, 25 (1983), 219–250.
4. Kutner, M. H., C. J. Nachtsheim, J. Neter, and W. Li, *Applied Linear Statistical Models*, 5th ed. (New York: McGraw-Hill/Irwin, 2005).
5. *Microsoft Excel 2010* (Redmond, WA: Microsoft Corp., 2010).
6. *Minitab Release 16* (State College, PA: Minitab, Inc., 2010).



# CHAPTER 13 EXCEL GUIDE

## EG13.1 TYPES of REGRESSION MODELS

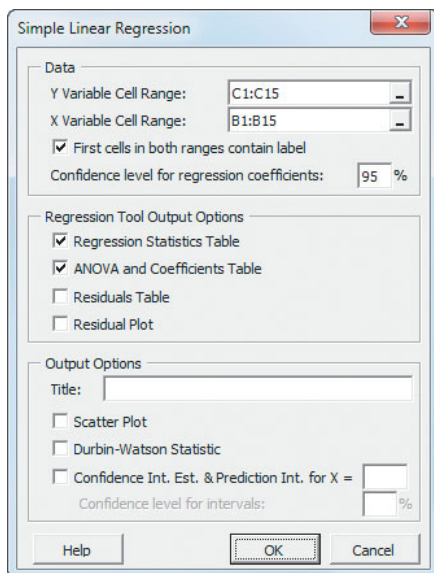
There are no Excel Guide instructions for this section.

## EG13.2 DETERMINING the SIMPLE LINEAR REGRESSION EQUATION

**PHStat2** Use **Simple Linear Regression** to perform a simple linear regression analysis. For example, to perform the Figure 13.4 analysis of the Sunflowers Apparel data on page 526, open to the **DATA** worksheet of the **Site workbook**. Select **PHStat** → **Regression** → **Simple Linear Regression**. In the procedure's dialog box (shown below):

1. Enter **C1:C15** as the **Y Variable Cell Range**.
2. Enter **B1:B15** as the **X Variable Cell Range**.
3. Check **First cells in both ranges contain label**.
4. Enter **95** as the **Confidence level for regression coefficients**.
5. Check **Regression Statistics Table** and **ANOVA and Coefficients Table**.
6. Enter a **Title** and click **OK**.

The procedure creates a worksheet that contains a copy of your data as well as the worksheet shown in Figure 13.4.



For more information about these worksheets, read the following *In-Depth Excel* section.

To create a scatter plot that contains a prediction line and regression equation similar to Figure 13.5 on page 526,

modify step 6 by checking the **Scatter Plot** output option before clicking **OK**.

**In-Depth Excel** Use the **COMPUTE** worksheet of the **Simple Linear Regression** workbook, shown in Figure 13.4 on page 526, as a template for performing simple linear regression. Columns A through I of this worksheet duplicate the visual design of the Analysis ToolPak regression worksheet. The worksheet uses the regression data in the **SLRDATA** worksheet to perform the regression analysis for the Table 13.1 Sunflowers Apparel data.

Not shown in Figure 13.4 is the Calculations area in columns K through M. This area contains an array formula in the cell range L2:M6 that contains the expression **LINEST(cell range of Y variable, cell range of X variable, True, True)** to compute the  $b_1$  and  $b_0$  coefficients in cells L2 and M2, the  $b_1$  and  $b_0$  standard errors in cells L3 and M3,  $r^2$  and the standard error of the estimate in cells L4 and M4, the  $F$  test statistic and error  $df$  in cells L5 and M5, and  $SSR$  and  $SSE$  in cells L6 and M6. In cell L9, the expression **TINV(1 – confidence level, Error degrees of freedom)** computes the critical value for the  $t$  test.

To perform simple linear regression for other data, paste the regression data into the SLRDATA worksheet. Paste the values for the  $X$  variable into column A and the values for the  $Y$  variable into column B. Open to the **COMPUTE** worksheet. First, enter the confidence level in cell L8. Then edit the array formula: Select the cell range L2:M6, edit the cell ranges in the formulas, and then, while holding down the **Control** and **Shift** keys (or the **Apple** key on a Mac), press the **Enter** key. (Open the **COMPUTE\_FORMULAS** worksheet to examine all the formulas in the worksheet, some of which are discussed in later sections of this Excel Guide.)

To create a scatter plot that contains a prediction line and regression equation similar to Figure 13.5 on page 526, first use the Section EG2.6 *In-Depth Excel* scatter plot instructions with the Table 13.1 Sunflowers Apparel data to create a basic plot. Then select the plot and:

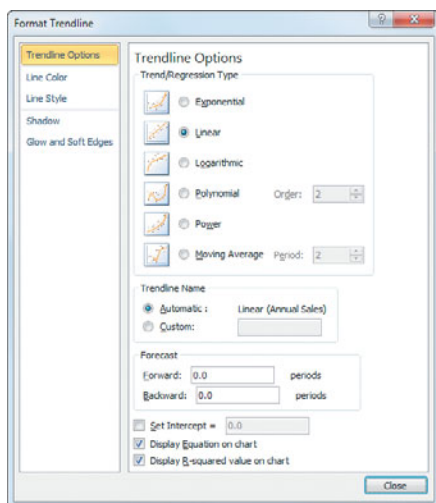
1. Select **Layout** → **Trendline** and select **More Trendline Options** from the Trendline gallery.

In the Format Trendline dialog box (shown on page 572):

2. Click **Trendline Options** in the left pane. In the Trendline Options pane on the right, click **Linear**, check **Display Equation on chart**, check **Display R-squared value on chart**, and then click **Close**.

For scatter plots of other data, if the  $X$  axis does not appear at the bottom of the plot, right-click the **Y** axis and





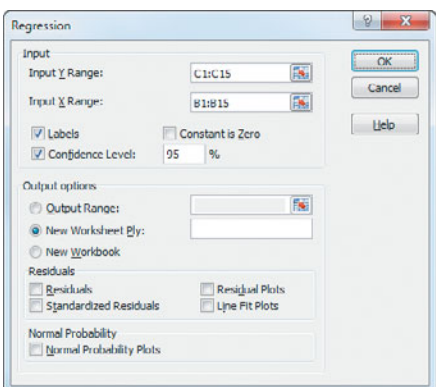
click **Format Axis** from the shortcut menu. In the Format Axis dialog box, click **Axis Options** in the left pane. In the Axis Options pane on the right, click **Axis value** and in its box enter the value shown in the dimmed **Minimum** box at the top of the pane. Then click **Close**.

**Analysis ToolPak** Use **Regression** to perform simple linear regression. For example, to perform the Figure 13.4 analysis of the Sunflowers Apparel data (see page 526), open to the **DATA worksheet** of the **Site workbook** and:

1. Select **Data → Data Analysis**.
2. In the Data Analysis dialog box, select **Regression** from the **Analysis Tools** list and then click **OK**.

In the Regression dialog box (see below):

3. Enter **C1:C15** as the **Input Y Range** and enter **B1:B15** as the **Input X Range**.
4. Check **Labels** and check **Confidence Level** and enter **95** in its box.
5. Click **New Worksheet Ply** and then click **OK**.



## EG13.3 MEASURES of VARIATION

The measures of variation are computed as part of creating the simple linear regression worksheet using the Section EG13.2 instructions.

If you use either Section EG13.2 *PHStat2* or *In-Depth Excel* instructions, formulas used to compute these measures are in the **COMPUTE worksheet** that is created. Formulas in cells B5, B7, B13, C12, C13, D12, and E12 copy values computed by the array formula in cell range L2:M6. The cell F12 formula, in the form **=FDIST(*F test statistic*, 1, *error degrees of freedom*)**, computes the *p*-value for the *F* test for the slope, discussed in Section 13.7.

## EG13.4 ASSUMPTIONS

There are no Excel Guide instructions for this section.

## EG13.5 RESIDUAL ANALYSIS

**PHStat2** Use the Section EG13.2 *PHStat2* instructions. Modify step 5 by checking **Residuals Table** and **Residual Plot** in addition to checking **Regression Statistics Table** and **ANOVA and Coefficients Table**.

**In-Depth Excel** Use the **RESIDUALS worksheet** of the **Simple Linear Regression workbook**, shown in Figure 13.10 on page 540, as a template for creating a residuals worksheet. This worksheet computes the residuals for the regression analysis for the Table 13.1 Sunflowers Apparel data. In column C, the worksheet computes the predicted *Y* values (labeled Predicted Annual Sales in Figure 13.10) by first multiplying the *X* values by the  $b_1$  coefficient in cell B18 of the **COMPUTE worksheet** and then adding the  $b_0$  coefficient (in cell B17 of **COMPUTE**). In column E, the worksheet computes residuals by subtracting the predicted *Y* values from the *Y* values.

For other problems, modify this worksheet by pasting the *X* values into column B and the *Y* values into column D. Then, for sample sizes smaller than 14, delete the extra rows. For sample sizes greater than 14, copy the column C and E formulas down through the row containing the last pair and *X* and *Y* values and add the new observation numbers in column A.

**Analysis ToolPak** Use the Section EG13.2 *Analysis ToolPak* instructions. Modify step 5 by checking **Residuals** and **Residual Plots** before clicking **New Worksheet Ply** and then **OK**.

To create a scatter plot similar to Figure 13.11, use the original *X* variable and the residuals (plotted as the *Y* variable) as the chart data.

## EG13.6 MEASURING AUTOCORRELATION: the DURBIN-WATSON STATISTIC

**PHStat2** Use the *PHStat2* instructions at the beginning of Section EG13.2. Modify step 6 by checking the **Durbin-Watson Statistic** output option before clicking **OK**.

**In-Depth Excel** Use the **DURBIN\_WATSON** worksheet of the **Simple Linear Regression workbook**, similar to the worksheet shown in Figure 13.16 on page 545, as a template for computing the Durbin-Watson statistic. The worksheet computes the statistic for the package delivery simple linear regression model. In cell B3, the worksheet uses the expression **SUMXMY2(cell range of the second through last residual, cell range of the first through the second-to-last residual)** to compute the sum of squared difference of the residuals, the numerator in Equation (13.15) on page 544, and in cell B4 uses **SUMSQ(cell range of the residuals)** to compute the sum of squared residuals, the denominator in Equation (13.15).

To compute the Durbin-Watson statistic for other problems, first create the simple linear regression model and the **RESIDUALS** worksheet for the problem, using the instructions in Sections EG13.2 and EG13.5. Then open the **DURBIN\_WATSON** worksheet and edit the formulas in cell B3 and B4 to point to the proper cell ranges of the new residuals.

### EG13.7 INFERENCES ABOUT the SLOPE and CORRELATION COEFFICIENT

The  $t$  test for the slope and  $F$  test for the slope are included in the worksheet created by using the Section EG13.2 instructions. The  $t$  test computations in the worksheets created by using the **PHStat2** and *In-Depth Excel* instructions are discussed in Section EG13.2. The  $F$  test computations are discussed in Section EG13.3.

### EG13.8 ESTIMATION of MEAN VALUES and PREDICTION of INDIVIDUAL VALUES

**PHStat2** Use the Section EG13.2 **PHStat2** instructions but replace step 6 with these steps 6 and 7:

6. Check **Confidence Int. Est. & Prediction Int. for X =** and enter **4** in its box. Enter **95** as the percentage for **Confidence level for intervals**.

7. Enter a **Title** and click **OK**.

The additional worksheet created is discussed in the following *In-Depth Excel* instructions.

**In-Depth Excel** Use the **CIEandPI** worksheet of the **Simple Linear Regression workbook**, shown in Figure 13.21 on page 557, as a template for computing confidence interval estimates and prediction intervals. The worksheet contains the data and formulas for the Section 13.8 examples that use the Table 13.1 Sunflowers Apparel data. The worksheet uses the expression **TINV(1 – confidence level, degrees of freedom)** to compute the  $t$  critical value in cell B10 and the expression **TREND(Y variable cell range, X variable cell range, X value)** to compute the predicted  $Y$  value for the  $X$  value in cell B15. In cell B12, the expression **DEVSQ(X variable cell range)** computes the  $SSX$  value that is used, in turn, to help compute the  $h$  statistic.

To compute a confidence interval estimate and prediction interval for other problems:

1. Paste the regression data into the **SLRData** worksheet. Use column A for the  $X$  variable data and column B for the  $Y$  variable data.
2. Open to the **CIEandPI** worksheet.

In the **CIEandPI** worksheet:

3. Change values for the **X Value** and **Confidence Level**, as is necessary.
4. Edit the cell ranges used in the cell B15 formula that uses the **TREND** function to refer to the new cell ranges for the  $Y$  and  $X$  variables.

To create a scatter plot similar to Figure 13.11 on page 540, use the original  $X$  variable and the residuals (plotted as the  $Y$  variable) as the chart data.

# CHAPTER 13 MINITAB GUIDE

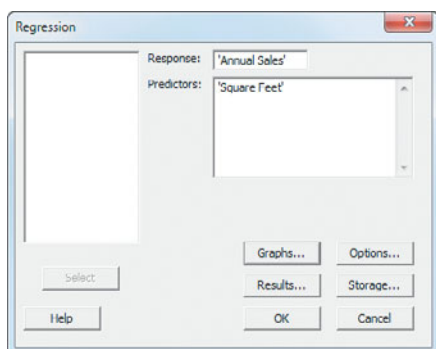
## MG13.1 TYPES of REGRESSION MODELS

There are no Minitab Guide instructions for this section.

## MG13.2 DETERMINING the SIMPLE LINEAR REGRESSION EQUATION

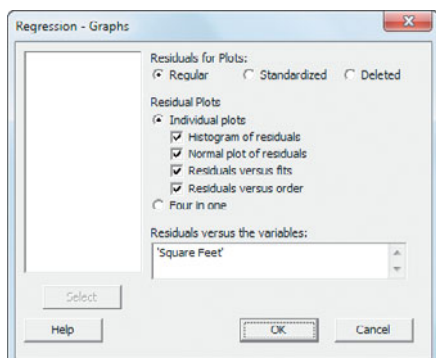
Use **Regression** to perform a simple linear regression analysis. For example, to perform the Figure 13.4 analysis of the Sunflowers Apparel data on page 526, open to the **Site worksheet**. Select **Stat** → **Regression** → **Regression**. In the Regression dialog box (shown below):

1. Double-click **C3 Annual Sales** in the variables list to add 'Annual Sales' to the **Response** box.
2. Double-click **C2 Square Feet** in the variables list to add 'Square Feet' to the **Predictors** box.
3. Click **Graphs**.



In the Regression - Graphs dialog box (shown below):

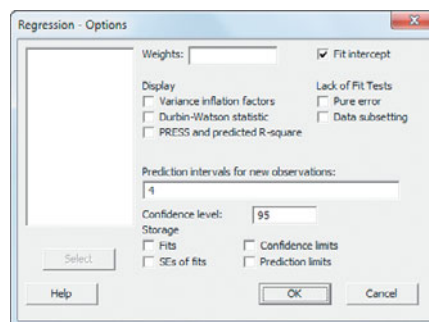
4. Click **Regular** (in Residuals for Plots) and **Individual Plots** (in Residual Plots).
5. Check **Histogram of residuals**, **Normal plot of residuals**, **Residuals versus fits**, and **Residuals versus order** and then press **Tab**.
6. Double-click **C2 Square Feet** in the variables list to add 'Square Feet' in the **Residuals versus the variables** box.
7. Click **OK**.



8. Back in the Regression dialog box, click **Results**.

In the Regression - Results dialog box (not shown):

9. Click **Regression equation**, **table of coefficients**, **s**, **R-squared**, and **basic analysis of variance** and then click **OK**.
10. Back in the Regression dialog box, click **Options**. In the Regression - Options dialog box (shown below):
11. Check **Fit Intercept**.
12. Clear all the **Display** and **Lack of Fit Test** check boxes.
13. Enter **4** in the **Prediction intervals for new observations** box.
14. Enter **95** in the **Confidence level** box.
15. Click **OK**.
16. Back in the Regression dialog box, click **OK**.



To create a scatter plot that contains a prediction line and regression equation similar to Figure 13.5 on page 526, use the Section MG2.6 scatter plot instructions with the Table 13.1 Sunflowers Apparel data.

## MG13.3 MEASURES of VARIATION

The measures of variation are computed in the Analysis of Variance table that is part of the simple linear regression results created using the Section MG13.2 instructions.

## MG13.4 ASSUMPTIONS

There are no Minitab Guide instructions for this section.

## MG13.5 RESIDUAL ANALYSIS

Selections in step 5 of the Section MG13.2 instructions create the residual plots and normal probability plots necessary for residual analysis. To create the list of residual values similar to column E in Figure 13.10 on page 540, replace step

15 of the Section MG13.2 instructions with these steps 15 through 17:

15. Click **Storage**.
16. In the Regression - Storage dialog box, check **Residuals** and then click **OK**.
17. Back in the Regression dialog box, click **OK**.

### MG13.6 MEASURING AUTOCORRELATION: the DURBIN-WATSON STATISTIC

To compute the Durbin-Watson statistic, use the Section MG13.2 instructions but check **Durbin-Watson statistic** (in the Regression - Options dialog box) as part of step 12.

### MG13.7 INFERENCES ABOUT the SLOPE and CORRELATION COEFFICIENT

The  $t$  test for the slope and  $F$  test for the slope are included in the results created by using the Section MG13.2 instructions.

### MG13.8 ESTIMATION of MEAN VALUES and PREDICTION of INDIVIDUAL VALUES

The confidence interval estimate and prediction interval are included in the results created by using the Section MG13.2 instructions.