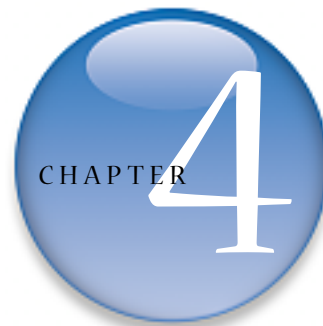


Describing the Relation between Two Variables



Outline

- 4.1 Scatter Diagrams and Correlation
- 4.2 Least-Squares Regression
- 4.3 The Coefficient of Determination
 - Chapter Review
 - Case Study: Thomas Malthus, Population, and Subsistence (On CD)

DECISIONS

You are still in the market to buy a car. Because cars lose value over time at different rates, you want to look at the depreciation rates of cars you are considering. After all, the higher the depreciation rate, the more value the car loses each year. See the Decision Project on page 204.



●●● Putting It All Together

In Chapters 2 and 3 we examined data in which a single variable was measured for each individual in the study (**univariate data**), such as the three-year rate of return (the variable) for various mutual funds (the individuals). We obtained descriptive measures for the variable that were both graphical and numerical.

However, much research is designed to describe the relation that may exist between two variables. For example, a researcher may be interested in the relationship between the club-head speed of a golf club and the distance

the golf ball travels. Here, each swing represents an individual, and the two variables are club-head speed and distance. This type of data is referred to as *bivariate data*. **Bivariate data** are data in which two variables are measured on an individual. To describe the relation between the two quantitative variables, we first graphically represent the data and then obtain some numerical descriptions of the data, just as we did when analyzing univariate data.

4.1 Scatter Diagrams and Correlation

Preparing for This Section Before getting started, review the following:

- Mean (Section 3.1, pp. 121–124)
- z-Scores (Section 3.4, pp. 165–166)
- Standard deviation (Section 3.2, pp. 143–144)

Objectives

- 1 Draw and interpret scatter diagrams
- 2 Understand the properties of the linear correlation coefficient
- 3 Compute and interpret the linear correlation coefficient
- 4 Determine whether there is a linear relation between two variables

Before we can graphically represent bivariate data, a fundamental question must be asked. Am I interested in using the value of one variable to predict the value of the other variable? For example, it seems reasonable to think that as the speed at which a golf club is swung increases, the distance the golf ball travels also increases. Therefore, we might use club-head speed to predict distance. We call distance the *response* (or *dependent*) *variable* and club-head speed the *explanatory* (or *predictor* or *independent*) *variable*.

Definition

The **response variable** is the variable whose value can be explained by the value of the **explanatory** or **predictor variable**.



In Other Words

We use the term *explanatory variable* because it helps to explain variability in the response variable.

It is important to recognize that, if the data used in the study are observational, we cannot conclude that there is a causal relationship between the two variables. We cannot say that changes in the level of the explanatory variable *cause* changes in the level of the response variable. In fact, it may be that the two are related through some *lurking variable*.

Recall that a **lurking variable** is a variable that may affect the response variable but is excluded from the analysis. For example, air-conditioning bills can be used to predict lemonade sales. As air-conditioning bills rise, the sales of lemonade rise. This relation does not mean that high air-conditioning bills cause high lemonade sales, because both high air-conditioning bills and high lemonade sales are associated with high summer temperatures. Therefore, air temperature is a lurking variable.



CAUTION

If bivariate data are observational, then we cannot conclude that any relation between the explanatory and response variables is due to cause and effect.

1 Draw and Interpret Scatter Diagrams

The first step in identifying the type of relation that might exist between two variables is to draw a picture. Bivariate data can be represented graphically through a *scatter diagram*.

Definition

A **scatter diagram** is a graph that shows the relationship between two quantitative variables measured on the same individual. Each individual in the data set is represented by a point in the scatter diagram. The explanatory variable is plotted on the horizontal axis and the response variable is plotted on the vertical axis. Do not connect the points when drawing a scatter diagram.

EXAMPLE 1

Drawing a Scatter Diagram

Problem: A golf pro wanted to learn the relation between the club-head speed of a golf club (measured in miles per hour) and the distance (in yards) that the ball will travel. He realized that there are other variables besides club-head speed

**Table 1**

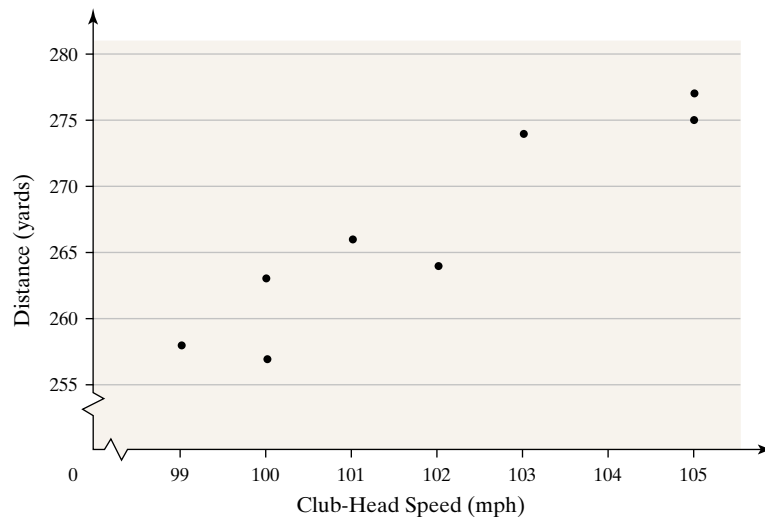
Club-Head Speed (mph)	Distance (yards)
100	257
102	264
103	274
101	266
105	277
100	263
99	258
105	275

Source: Paul Stephenson, student at Joliet Junior College

that determine the distance a ball will travel (such as club type, ball type, golfer, and weather conditions). To eliminate the variability due to these variables, the pro used a single model of club and ball. One golfer was chosen to swing the club on a clear, 70-degree day with no wind. The pro recorded the club-head speed and measured the distance that the ball traveled and collected the data in Table 1. Draw a scatter diagram of the data.

Approach: Because the pro wants to use club-head speed to predict the distance the ball travels, club-head speed is the explanatory variable (horizontal axis) and distance is the response variable (vertical axis). We plot the ordered pairs (100, 257), (102, 264), and so on, in a rectangular coordinate system.

Solution: The scatter diagram is shown in Figure 1.

Figure 1

It would appear from the graph that as club-head speed increases, the distance that the ball travels increases as well. _____

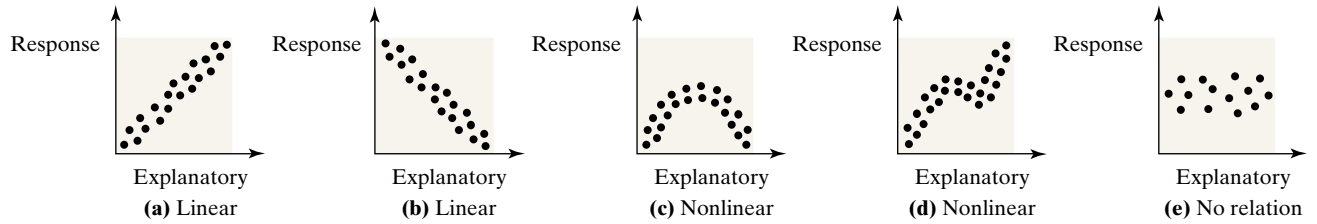
It is not always clear which variable should be considered the response variable and which should be considered the explanatory variable. For example, does high school GPA predict a student's SAT score or can the SAT score be used to predict GPA? The researcher must determine which variable plays the role of explanatory variable based on the questions he or she wants answered. For example, if the researcher is interested in predicting SAT scores on the basis of high school GPA, then high school GPA will play the role of explanatory variable.

Now Work Problems 23(a) and 23(b).

Scatter diagrams show the type of relation that exists between two variables. Our goal in interpreting scatter diagrams will be to distinguish scatter diagrams that imply a linear relation from those that imply a nonlinear relation or those that imply no relation. Figure 2 displays various scatter diagrams and the type of relation implied.

As we compare Figure 2(a) with Figure 2(b), we notice a distinct difference. In Figure 2(a), the data follow a linear pattern that slants upward to the right; the data in Figure 2(b) follow a linear pattern that slants downward to the right. Figures 2(c) and 2(d) show scatter diagrams of nonlinear relations. Figure 2(e) shows a scatter diagram in which there is no relation between the explanatory and response variables.

Figure 2



Definitions



In Other Words

If two variables that are linearly related are positively associated, then as one goes up the other also tends to go up. If two variables that are linearly related are negatively associated, then as one goes up the other tends to go down.

Two variables that are linearly related are said to be **positively associated** when above-average values of one variable are associated with above-average values of the other variable. That is, two variables are positively associated if, whenever the value of one variable increases, the value of the other variable also increases.

Two variables that are linearly related are said to be **negatively associated** when above-average values of one variable are associated with below-average values of the other variable. That is, two variables are negatively associated if, whenever the value of one variable increases, the value of the other variable decreases.

So the scatter diagram from Figure 1 implies that club-head speed is positively associated with the distance a golf ball travels.

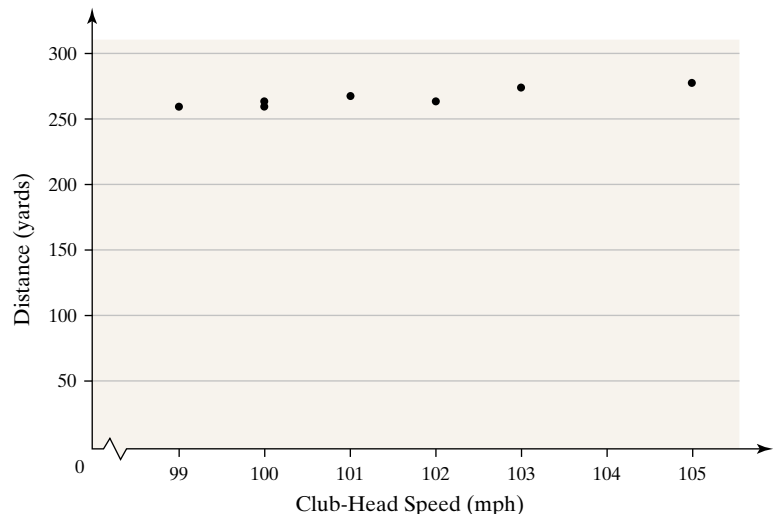
Now Work Problem 11.



Understand the Properties of the Linear Correlation Coefficient

It is dangerous to use only a scatter diagram to decide whether two variables follow a linear relation. Suppose we redraw the scatter diagram in Figure 1 using a different scale as shown in Figure 3.

Figure 3



CAUTION

The horizontal or vertical scale of a scatter diagram should be set so that the scatter diagram does not mislead a reader.

From Figure 3, we might conclude that club-head speed and distance are not related. The moral of the story is this: Just as we can manipulate the scale of graphs of univariate data, we can also manipulate the scale of the graphs of bivariate data, thereby encouraging incorrect conclusions. Therefore, numerical summaries of bivariate data should be used in conjunction with graphs to determine the type of relation, if any, that exists between two variables.

Definition

The **linear correlation coefficient** or **Pearson product moment correlation coefficient** is a measure of the strength of linear relation between two quantitative variables. We use the Greek letter ρ (rho) to represent the population correlation coefficient and r to represent the sample correlation coefficient. We present only the formula for the sample correlation coefficient.

**Historical Note**

Karl Pearson was born March 27, 1857. Pearson's proficiency as a statistician was recognized early in his life. It is said that his mother told him not to suck his thumb because otherwise his thumb would wither away. Pearson analyzed the size of each thumb and said to himself, "They look alike to me. I can't see that the thumb I suck is any smaller than the other. I wonder if she could be lying to me."

Karl Pearson graduated from Cambridge University in 1879. From 1893 to 1911, he wrote 18 papers on genetics and heredity. Through this work, he developed ideas regarding correlation and the chi-square test. (See Chapter 12.) In addition, Pearson came up with the term *standard deviation*.

Pearson and Ronald Fisher didn't get along. The dispute between the two was bad enough to have Fisher turn down the post of chief statistician at the Galton Laboratory in 1919 on the grounds that it would have meant working under Pearson. Pearson died on April 27, 1936.

Sample Correlation Coefficient*

$$r = \frac{\sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)}{n - 1} \quad (1)$$

where \bar{x} is the sample mean of the explanatory variable

s_x is the sample standard deviation of the explanatory variable

\bar{y} is the sample mean of the response variable

s_y is the sample standard deviation of the response variable

n is the number of individuals in the sample

The Pearson linear correlation coefficient is named in honor of Karl Pearson (1857–1936).

Properties of the Linear Correlation Coefficient

1. The linear correlation coefficient is always between -1 and 1 , inclusive. That is, $-1 \leq r \leq 1$.
2. If $r = +1$, there is a perfect positive linear relation between the two variables. See Figure 4(a).
3. If $r = -1$, there is a perfect negative linear relation between the two variables. See Figure 4(d).
4. The closer r is to $+1$, the stronger is the evidence of positive association between the two variables. See Figures 4(b) and 4(c).
5. The closer r is to -1 , the stronger is the evidence of negative association between the two variables. See Figures 4(e) and 4(f).
6. If r is close to 0 , there is little or no evidence of a *linear* relation between the two variables. Because the linear correlation coefficient is a measure of the strength of the linear relation, **r close to 0 does not imply no relation, just no linear relation.** See Figures 4(g) and 4(h).
7. The linear correlation coefficient is a unitless measure of association. So the unit of measure for x and y plays no role in the interpretation of r .

**CAUTION**

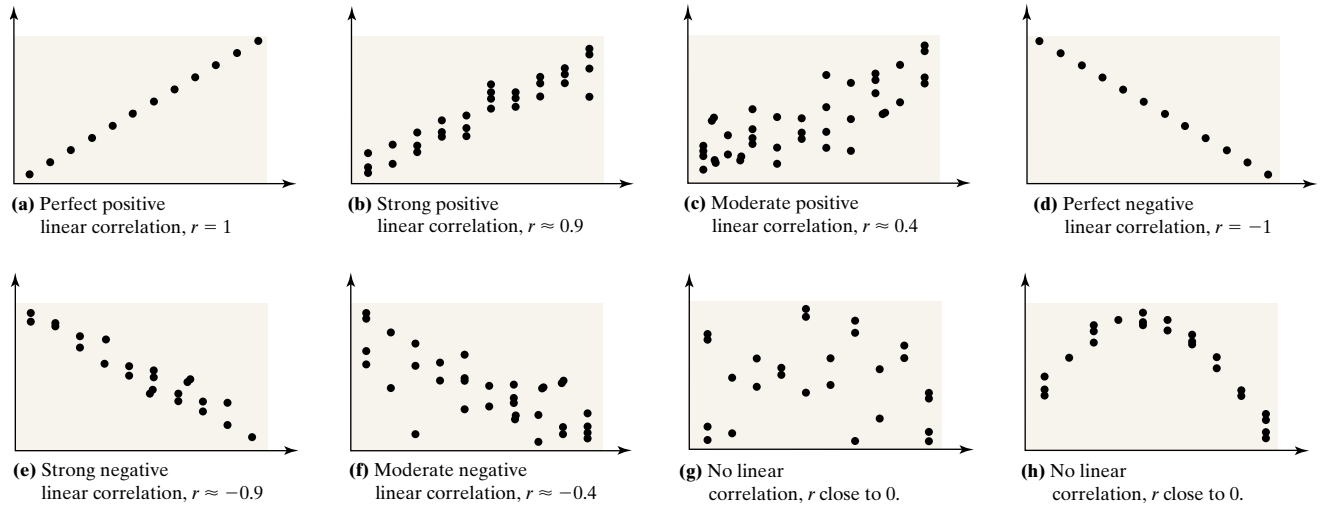
A linear correlation coefficient close to 0 does not imply that there is no relation, just no linear relation. For example, although the scatter diagram drawn in Figure 4(h) indicates that the two variables are related, the linear correlation coefficient of these data is close to 0 .

In looking carefully at Formula (1), we should notice that the numerator of the formula is the product of z -scores for the explanatory (x) and response (y) variables. A positive linear correlation coefficient means that the sum of the product of the z -scores for x and y must be positive. Under what circumstances

*An equivalent formula for the linear correlation coefficient is

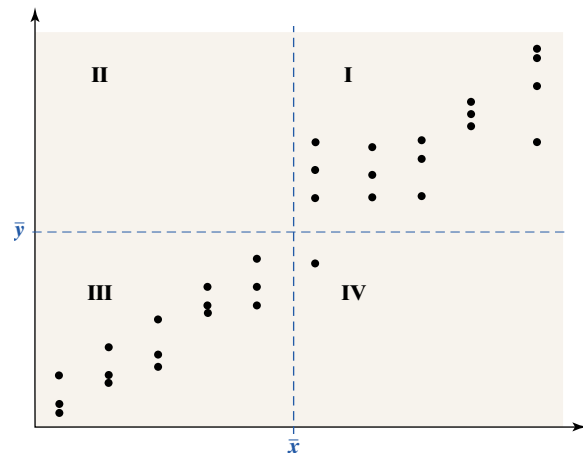
$$r = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sqrt{\left(\sum x_i^2 - \frac{(\sum x_i)^2}{n} \right) \left(\sum y_i^2 - \frac{(\sum y_i)^2}{n} \right)}} = \frac{S_{xy}}{\sqrt{S_{xx}} \sqrt{S_{yy}}}$$

Figure 4



does this occur? Figure 5 shows a scatter diagram that implies a positive association between x and y . The vertical dashed line represents the value of \bar{x} , and the horizontal dashed line represents the value of \bar{y} . These two dashed lines divide our scatter diagram into four quadrants, labeled I, II, III, and IV.

Figure 5



Consider the data in quadrants I and III. If a certain x -value is above its mean, \bar{x} , then the corresponding y -value will be above its mean, \bar{y} . If a certain x -value is below its mean, \bar{x} , then the corresponding y -value will be below its mean, \bar{y} . Therefore, for data in quadrant I, we have $\frac{x_i - \bar{x}}{s_x}$ positive and $\frac{y_i - \bar{y}}{s_y}$ positive, so their product is positive. For data in quadrant III, we have $\frac{x_i - \bar{x}}{s_x}$ negative and $\frac{y_i - \bar{y}}{s_y}$ negative, so their product is positive. The sum of these products is positive, and therefore we have a positive linear correlation coefficient. A similar argument can be made for negative correlation.

Now suppose the data are equally dispersed in the four quadrants. Then the negative products (resulting from data in quadrants II and IV) will offset the positive products (resulting from data in quadrants I and III). The result is a linear correlation coefficient close to 0.



In Other Words

The correlation coefficient describes the strength and the direction of the linear relationship between two variables.

Now Work Problem 15.

3 Compute and Interpret the Linear Correlation Coefficient

Now that we have an understanding of the properties of the linear correlation coefficient, we are ready to compute its value.

EXAMPLE 2

Computing and Interpreting the Correlation Coefficient

Problem: In Table 2, columns 1 and 2 represent the club-head speed (in miles per hour) and the distance the ball travels (in yards). Compute and interpret the linear correlation coefficient.

Approach: We treat club-head speed as the explanatory variable, x , and distance as the response variable, y .

Step 1: Compute \bar{x} , s_x , \bar{y} , and s_y .

Step 2: Determine $\frac{x_i - \bar{x}}{s_x}$ and $\frac{y_i - \bar{y}}{s_y}$ for each observation.

Step 3: Compute $\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$ for each observation.

Step 4: Determine $\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$ and substitute this value into Formula (1).

Solution

Step 1: We compute \bar{x} , s_x , \bar{y} , and s_y :

$$\bar{x} = 101.875, \quad s_x = 2.29518, \quad \bar{y} = 266.75, \quad s_y = 7.74135$$

To avoid round-off error when using Formula (1), do not round the statistics.

Step 2: We determine $\frac{x_i - \bar{x}}{s_x}$ and $\frac{y_i - \bar{y}}{s_y}$ in columns 3 and 4 in Table 2.

Step 3: We multiply the entries in columns 3 and 4 to obtain the entries in column 5.

Table 2

Club-Head Speed, x_i	Distance, y_i	$\frac{x_i - \bar{x}}{s_x}$	$\frac{y_i - \bar{y}}{s_y}$	$\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$
100	257	-0.816929	-1.259470	1.028898
102	264	0.054462	-0.355235	-0.0193347
103	274	0.490158	0.936529	0.459047
101	266	-0.381234	-0.096882	0.036935
105	277	1.361549	1.324058	1.802770
100	263	-0.816929	-0.484412	0.395726
99	258	-1.252625	-1.130294	1.415835
105	275	1.361549	1.065706	1.451011
				$\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$
				= 6.570887

Step 4: We add the entries in column 5 to obtain

$$\sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right) = 6.570887$$

Substitute this value into Formula (1) to obtain the correlation coefficient.

$$r = \frac{\sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)}{n - 1} = \frac{6.570887}{8 - 1} = 0.9387$$

The linear correlation between club-head speed and distance is 0.9387, indicating a strong positive association between the two variables. The higher the club-head speed, the farther the golf ball tends to travel.

Notice in Example 2 that we carry many decimal places in the computation of the correlation coefficient to avoid rounding error. Also, compare the signs of the entries in columns 3 and 4. Notice that negative values in column 3 correspond to negative values in column 4 and that positive values in column 3 correspond to positive values in column 4 (except for the second trial of the experiment). This means that above-average values of x are associated with above-average values of y , and below-average values of x are associated with below-average values of y . This is why the linear correlation coefficient is positive.

EXAMPLE 3

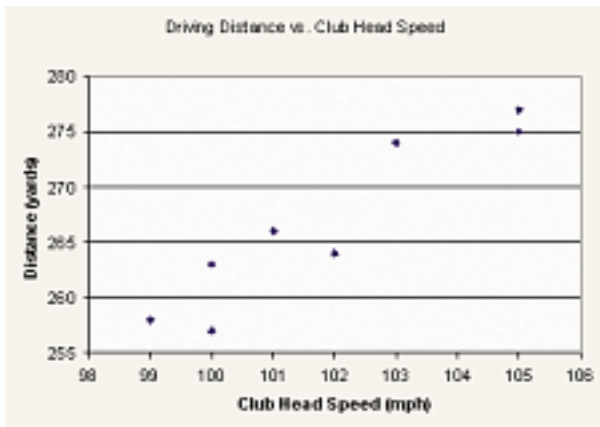
Drawing a Scatter Diagram and Determining the Linear Correlation Coefficient Using Technology

Problem: Use a statistical spreadsheet or a graphing calculator with advanced statistical features to draw a scatter diagram of the data in Table 1. Then determine the linear correlation between club-head speed and distance.

Approach: We will use Excel to draw the scatter diagram and obtain the linear correlation coefficient. The steps for drawing scatter diagrams and obtaining the linear correlation coefficient using MINITAB, Excel, or the TI-83 and TI-84 Plus graphing calculators are given in the Technology Step by Step on page 194.

Result: Figure 6(a) shows the scatter diagram and Figure 6(b) shows the linear correlation coefficient obtained from Excel. Notice that Excel provides a **correlation matrix**, which means that for every pair of columns in the spreadsheet it will compute and display the correlation in the bottom triangle of the matrix.

Figure 6



(a)

	Driving Distance	Club-Head Speed
Driving Distance	1	
Club-Head Speed	0.938695838	1

(b)

Now Work Problem 23(c).



CAUTION!

A linear correlation coefficient that implies a strong positive or negative association that is computed using observational data does not imply causation.

Correlation versus Causation

In Chapter 1 we stated that there are two types of studies: observational and experimental. The data given in Examples 1 through 3 are the result of an experiment. Therefore, we can claim that a higher club-head speed causes the golf ball to travel a longer distance. However, if data result from an observational study, we cannot claim causation. Consider the scatter diagram shown in Figure 7, which shows the relation between the birthrate (births per 1000 women) of teenagers and the homicide rate (homicides per 100,000 inhabitants) for the years 1993 to 2000.



The linear correlation coefficient between these two variables is 0.9987. Does this mean that higher birthrates among teenagers cause a higher homicide rate? Certainly not!

In Chapter 1, we introduced lurking variables. A lurking variable is one that has not been considered in your analysis but is related to both variables in the study. Can you think of any variables that might be related to teen birthrates and homicide rates? Perhaps there is an economic variable, such as poverty rate, proportion of homes with a single parent, or high school dropout rate, that is related to both teenage birthrate and homicide rate.

In-Class Activity: Correlation

Randomly select six students from the class and have them determine their at-rest pulse and then discuss the following:

1. When determining the at-rest pulse rate, would it be better to count beats for 30 seconds and multiply by 2 or count beats for 1 full minute? Explain. What are some other ways to find the at-rest pulse rate? Do any of these methods have an advantage?
2. What effect will physical activity have on pulse rate?
3. Do you think the at-rest pulse rate will have any effect on the pulse rate after physical activity? If so, how? If not, why not?

Have the same six students jog in place for 3 minutes and then immediately determine their pulse rate using the same technique as for the at-rest pulse rate.

4. Draw a scatter diagram for the pulse data using the at-rest data as the explanatory variable.
5. Comment on the relationship, if any, between the two variables. Is this consistent with your expectations?
6. Based on the graph, estimate the linear correlation coefficient for the data. Then compute the correlation coefficient using a graphing utility and compare to your estimate.

4 Determine Whether There Is a Linear Relation between Two Variables

A question you may be asking yourself is, “How do I know the correlation between two variables is strong enough for me to conclude that there is a linear relation between the variables?” While rigorous tests exist that can answer this question, for now we will be content with a simple comparison test that is based on the more rigorous approach.

To test whether the correlation between the explanatory and response variables is strong enough, determine the absolute value of the correlation coefficient. If the absolute value of the correlation coefficient is greater than the critical value in Table VIII in Appendix A for the given sample size, then we say there is a linear relation between the two variables. Otherwise, there is no linear relation.



Other Words

We use two vertical bars to denote absolute value, as in $|5|$ or $|-4|$. Recall, $|5| = 5$, $|-4| = 4$, and $|0| = 0$.

EXAMPLE 4 Is There a Linear Relation?

Problem: Using the data from Example 2, determine whether there is a linear relation between club-head speed and distance. Comment on the type of relation that appears to exist between club-head speed and distance.

Approach: We compare the absolute value of the linear correlation coefficient to the critical value in Table VIII with $n = 8$. If the absolute value of the linear correlation coefficient is greater than the critical value, we conclude that there is a linear relation between club-head speed and distance.

Solution: The linear correlation coefficient between club-head speed and distance was found to be 0.9387 in Example 2. The absolute value of 0.9387 is 0.9387. We find the critical value for correlation in Table VIII with $n = 8$ to be 0.707. Since 0.9387 is greater than 0.707, we conclude there is a positive linear relation between club-head speed and distance.

Now Work Problem 23(d).

4.1 ASSESS YOUR UNDERSTANDING

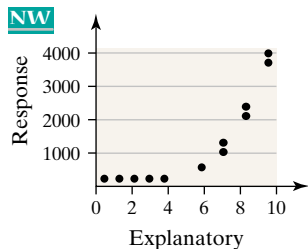
Concepts and Vocabulary

- Describe the difference between univariate and bivariate data.
- Explain what is meant by a lurking variable. Provide an example.
- What does it mean to say that two variables are positively associated?
- What does it mean to say that the linear correlation coefficient between two variables equals 1? What would the scatter diagram look like?
- What does it mean if $r = 0$?
- Is the linear correlation coefficient a resistant measure? Support your answer.
- Explain what is wrong with the following statement: “We have concluded that there is a high correlation between the gender of drivers and rates of automobile accidents.”
- Write a statement that explains the concept of correlation. Include a discussion of the role that $x_i - \bar{x}$ and $y_i - \bar{y}$ play in the computation.
- Explain what is wrong with the following statement: “A recent study showed that the correlation between the number of acres on a farm and the amount of corn produced was 0.93 bushel.”
- Explain the difference between correlation and causation. When does a linear correlation coefficient that implies a strong positive correlation also imply causation?

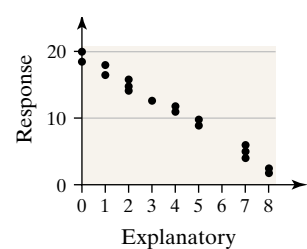
Skill Building

In Problems 11–14, determine whether the scatter diagram indicates that a linear relation may exist between the two variables. If the relation is linear, determine whether it indicates a positive or negative association between the variables.

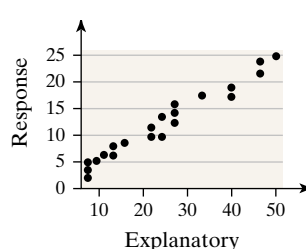
11.



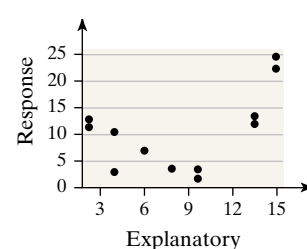
12.



13.



14.



15. Match the linear correlation coefficient to the scatter diagram. The scales on the x- and y-axes are the same for each scatter diagram.

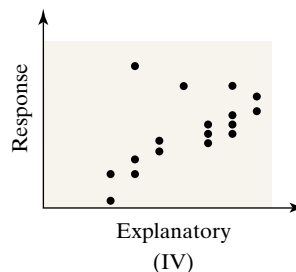
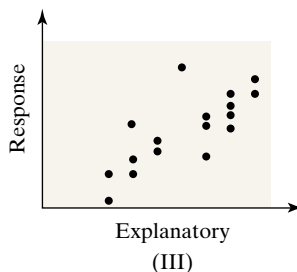
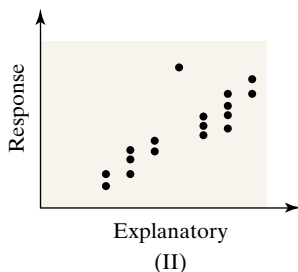
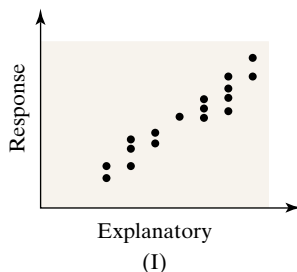
NW

(a) $r = 0.787$

(b) $r = 0.523$

(c) $r = 0.810$

(d) $r = 0.946$



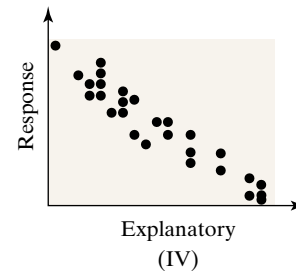
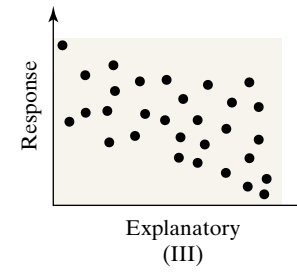
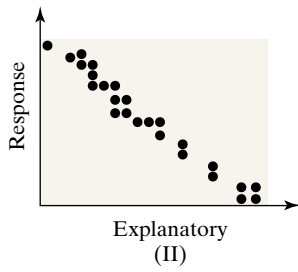
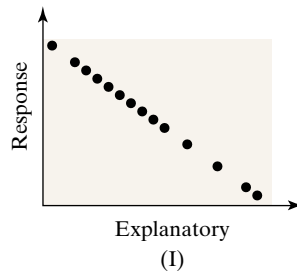
16. Match the linear correlation coefficient to the scatter diagram. The scales on the x- and y-axes are the same for each scatter diagram.

(a) $r = -0.969$

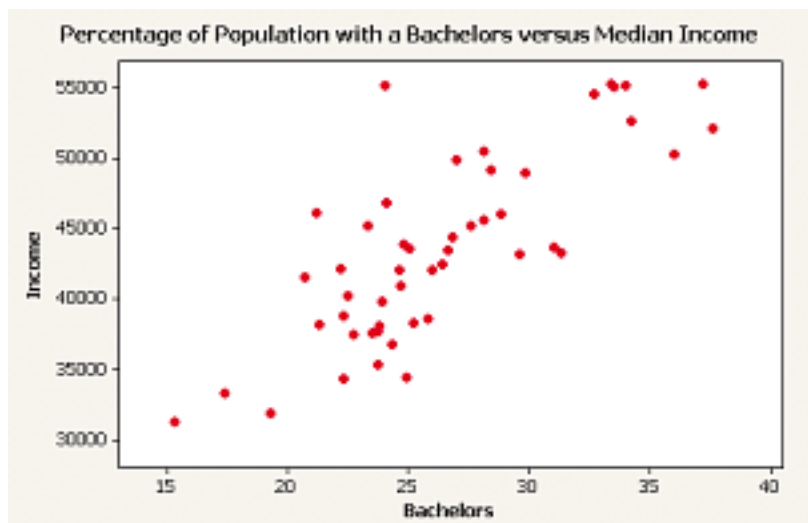
(b) $r = -0.049$

(c) $r = -1$

(d) $r = -0.992$

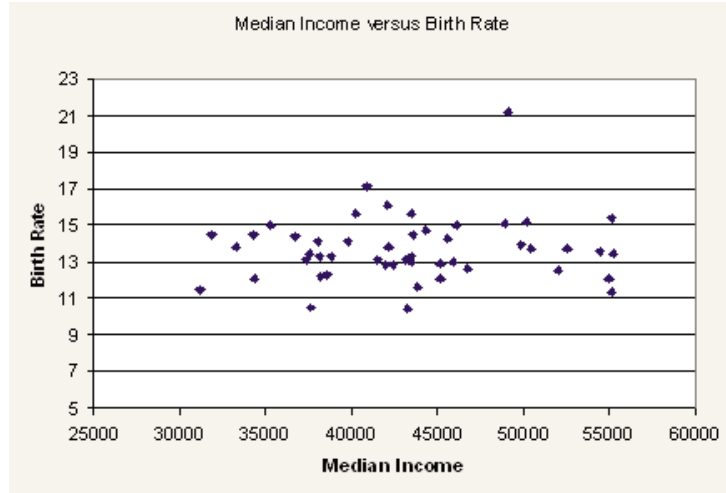


17. **Does Education Pay?** The following scatter diagram drawn in MINITAB shows the relation between the percentage of the population of a state that has at least a bachelor's degree and the median income (in dollars) of the state for 2003.



Source: U.S. Census Bureau

- (a) Describe the relation that appears to exist between level of education and median income.
 - (b) One observation appears to stick out from the rest. Which one? This particular observation is for the state of Alaska. Can you think of any reasons why the state of Alaska might have a high median income, given the proportion of the population that has at least a bachelor's degree?
- 18. Relation between Income and Birthrate?** The following scatter diagram drawn in Excel shows the relation between median income (in dollars) in a state and birthrate (births per 1000 women 15 to 44 years of age).



Source: U.S. Census Bureau

- (a) Does there appear to be any relation between median income and birthrate?
- (b) One observation sticks out from the rest. Which one? This particular observation is for the state of Utah. Are there any explanations for this result?

In Problems 19–22, (a) draw a scatter diagram of the data, (b) by hand, compute the correlation coefficient, and (c) comment on the type of relation that appears to exist between x and y .

19.

x	2	4	8	8	9
y	1	2	4	5	6

20.

x	2	3	5	6	6
y	10	9	8	3	1

21.

x	3	5	8	9	12	12
y	18	20	16	10	12	8

22.

x	0	1	1	2	4	7
y	3	5	4	6	8	9

Applying the Concepts

23. Height versus Head Circumference A pediatrician wants **NW** to determine the relation that may exist between a child's height and head circumference. She randomly selects 11 three-year-old children from her practice, measures their height and head circumference, and obtains the data shown in the table.



- (a) If the pediatrician wants to use height to predict head circumference, determine which variable is the explanatory variable and which is the response variable.
- (b) Draw a scatter diagram.
- (c) Compute the linear correlation coefficient between the height and head circumference of a child.
- (d) Comment on the type of relation that appears to exist between the height and head circumference of a child

on the basis of the scatter diagram and linear correlation coefficient.

Height (inches)	Head Circumference (inches)	Height (inches)	Head Circumference (inches)
27.75	17.5	26.5	17.3
24.5	17.1	27	17.5
25.5	17.1	26.75	17.3
26	17.3	26.75	17.5
25	16.9	27.5	17.5
27.75	17.6		

Source: Denise Slucki, student at Joliet Junior College

24. Gestation Period versus Life Expectancy A researcher wants to know if the gestation period of an animal can be used to predict life expectancy. She collects the following data:



Animal	Gestation (or Incubation) Period (days)	Life Expectancy (years)
Cat	63	11
Chicken	22	7.5
Dog	63	11
Duck	28	10
Goat	151	12
Lion	108	10
Parakeet	18	8
Pig	115	10
Rabbit	31	7
Squirrel	44	9

Source: Time Almanac 2000

- Suppose the researcher wants to use the gestation period of an animal to predict its life expectancy. Determine which variable is the explanatory variable and which is the response variable.
- Draw a scatter diagram.
- Compute the linear correlation coefficient between gestation period and life expectancy.
- Comment on the type of relation that appears to exist between gestation period and life expectancy based on the scatter diagram and linear correlation coefficient.
- Remove the goat from the data set, and recompute the linear correlation coefficient between the gestation period and life expectancy. What effect did the removal of the data value have on the linear correlation coefficient? Provide a justification for this result.

25. Weight of a Car versus Miles per Gallon An engineer wanted to determine how the weight of a car affects gas mileage. The following data represent the weight of various domestic cars and their gas mileage in the city for the 2005 model year.



Car	Weight (pounds)	Miles per Gallon
Buick LeSabre	3565	20
Cadillac DeVille	3985	18
Chevrolet Corvette	3180	19
Chevrolet Monte Carlo	3340	21
Chrysler PT Cruiser	3100	21
Chrysler Sebring Sedan	3175	22
Dodge Neon	2580	27
Dodge Stratus Sedan	3175	22
Ford Focus	2655	26
Ford Mustang	3300	20
Lincoln LS	3680	20
Mercury Sable	3310	19
Pontiac Bonneville	3590	20
Pontiac Grand Am	3475	20
Pontiac Sunfire	2770	24
Saturn Ion	2690	26

Source: www.roadandtrack.com

- Determine which variable is the likely explanatory variable and which is the likely response variable.
- Draw a scatter diagram of the data.
- Compute the linear correlation coefficient between the weight of a car and its miles per gallon in the city.
- Comment on the type of relation that appears to exist between the weight of a car and its miles per gallon in the city based on the scatter diagram and the linear correlation coefficient.

26. Bone Length Research performed at NASA and led by Emily R. Morey-Holton measured the lengths of the right humerus and right tibia in 11 rats that were sent to space on Spacelab Life Sciences 2. The following data were collected.



Right Humerus (mm)	Right Tibia (mm)	Right Humerus (mm)	Right Tibia (mm)
24.8	36.05	25.9	37.38
24.59	35.57	26.11	37.96
24.59	35.57	26.63	37.46
24.29	34.58	26.31	37.75
23.81	34.2	26.84	38.5
24.87	34.73		

Source: NASA Life Sciences Data Archive

- Draw a scatter diagram, treating the length of the right humerus as the explanatory variable and the length of the right tibia as the response variable.
- Compute the linear correlation coefficient between the length of the right humerus and the length of the right tibia.
- Comment on the type of relation that appears to exist between the length of the right humerus and the length of the right tibia based on the scatter diagram and the linear correlation coefficient.
- Convert the data to inches (1 mm = 0.03937 inch), and recompute the linear correlation coefficient. What effect did the conversion from millimeters to inches have on the linear correlation coefficient?

27. **Attending Class** The following data represent the number of days absent and the final grade for a sample of college students in a general education course at a large midwestern state university.



Number of Absences	Final Grade
0	89.2
1	86.4
2	83.5
3	81.1
4	78.2
5	73.9
6	64.3
7	71.8
8	65.5
9	66.2

Source: *College Teaching*, Winter 2005, Vol. 53, Issue 1

- The researcher wants to use the number of days absent to predict the final grade. Determine which variable is the explanatory variable and which is the response variable.
- Draw a scatter diagram of the data.
- Compute the linear correlation coefficient between the number of days absent and the final grade.
- Comment on the type of relation that appears to exist between the number of days absent and the final grade.
- Will going to class every day guarantee a passing grade? What other factors might need to be taken into account?

28. **Antibiotics** A study on antibiotic use among children in Manitoba, Canada, gave the following data for the number of prescriptions per 1000 children x years after 1995.



Year, x	0	1	2	3	4	5	6
Prescriptions (per 1000 children)	1201	1070	944	964	909	949	864

Source: *Canadian Medical Association Journal*, Vol. 171, Issue 2

- Draw a scatter diagram of the data, treating year as the explanatory variable. What type of relation, if any, appears to exist between year and antibiotic prescriptions among children?
- Compute the linear correlation coefficient between year and antibiotic prescriptions among children.
- Comment on the type of relation that appears to exist between year and antibiotic prescriptions among children on the basis of the scatter diagram and the linear correlation coefficient.

29. **Age versus HDL Cholesterol** A doctor wanted to determine whether there was a relation between a male's age and his HDL (so-called good) cholesterol. He randomly selected 17 of his patients and determined their HDL cholesterol. He obtained the following data.



Age	HDL Cholesterol	Age	HDL Cholesterol
38	57	38	44
42	54	66	62
46	34	30	53
32	56	51	36
55	35	27	45
52	40	52	38
61	42	49	55
61	38	39	28
26	47		

Source: Data based on information obtained from the National Center for Health Statistics

- Draw a scatter diagram of the data, treating age as the explanatory variable. What type of relation, if any, appears to exist between age and HDL cholesterol?
- Compute the linear correlation coefficient between age and HDL cholesterol.
- Comment on the type of relation that appears to exist between age and HDL cholesterol on the basis of the scatter diagram and the linear correlation coefficient.

30. **Intensity of a Lightbulb** Cathy is conducting an experiment to measure the relation between a light bulb's intensity and the distance from the light source. She measures a 100-watt lightbulb's intensity 1 meter from the bulb and at 0.1-meter intervals up to 2 meters from the bulb and obtains the following data.



Distance (meters)	Intensity	Distance (meters)	Intensity
1.0	0.29645	1.6	0.11450
1.1	0.25215	1.7	0.10243
1.2	0.20547	1.8	0.09231
1.3	0.17462	1.9	0.08321
1.4	0.15342	2.0	0.07342
1.5	0.13521		

- Draw a scatter diagram of the data, treating distance as the explanatory variable.
- Do you think that it is appropriate to compute the linear correlation coefficient between distance and intensity? Why?

31. Does Size Matter? Researchers wondered whether the size of a person’s brain was related to the individual’s mental capacity. They selected a sample of right-handed introductory psychology students who had SAT scores higher than 1350. The subjects took the Wechsler (1981)



Gender	MRI Count	IQ	Gender	MRI Count	IQ
Female	816,932	133	Male	949,395	140
Female	951,545	137	Male	1,001,121	140
Female	991,305	138	Male	1,038,437	139
Female	833,868	132	Male	965,353	133
Female	856,472	140	Male	955,466	133
Female	852,244	132	Male	1,079,549	141
Female	790,619	135	Male	924,059	135
Female	866,662	130	Male	955,003	139
Female	857,782	133	Male	935,494	141
Female	948,066	133	Male	949,589	144

Source: Willerman, L., Schultz, R., Rutledge, J. N., and Bigler, E. (1991). “In Vivo Brain Size and Intelligence,” *Intelligence*, 15, 223–228

Adult Intelligence Scale-Revised exam to obtain their IQ scores. Magnetic resonance imaging (MRI) scans were performed at the same facility for the subjects. The scans consisted of 18 horizontal magnetic resonance images. The computer counted all pixels with nonzero gray scale in each of the 18 images, and the total count served as an index for brain size.

- Draw a scatter diagram, treating MRI count as the explanatory variable and IQ as the response variable. Comment on what you see.
- Compute the linear correlation coefficient between MRI count and IQ. Do you think that MRI count and IQ are linearly related?
- A lurking variable in the analysis is gender. Draw a scatter diagram, treating MRI count as the explanatory variable and IQ as the response variable, but use a different plotting symbol for each gender. For example, use a circle for males and a triangle for females. What do you notice?
- Compute the linear correlation coefficient between MRI count and IQ for females. Compute the linear correlation coefficient between MRI count and IQ for males. Do you believe that MRI count and IQ are linearly related? What is the moral?

32. Male versus Female Drivers The following data represent the number of licensed drivers in various age groups and the number of accidents within the age group by gender.



Age Group	Number of Male Licensed Drivers (000s)	Number of Crashes Involving a Male (000s)	Number of Female Licensed Drivers (000s)	Number of Crashes Involving a Female (000s)
16	816	244	764	178
17	1,198	233	1,115	175
18	1,342	243	1,212	164
19	1,454	229	1,333	145
20–24	7,866	951	7,394	618
25–29	9,356	899	8,946	595
30–34	10,121	875	9,871	571
35–39	10,521	901	10,439	566
40–44	9,776	692	9,752	455
45–49	8,754	667	8,710	390
50–54	6,840	390	6,763	247
55–59	5,341	290	5,258	165
60–64	4,565	218	4,486	133
65–69	4,234	191	4,231	121
70–74	3,604	167	3,749	104
75–79	2,563	118	2,716	77
80–84	1,400	61	1,516	45
≥85	767	34	767	20

Source: National Highway and Traffic Safety Institute

- (a) On the same graph, draw a scatter diagram for both males and females. Be sure to use a different plotting symbol for each group. For example, use a square (\square) or an M for males and a plus sign (+) or an F for females. Treat number of licensed drivers as the explanatory variable.
- (b) Based on the scatter diagrams, do you think that insurance companies are justified in charging different insurance rates for males and females? Why?
- (c) Compute the linear correlation coefficient between number of licensed drivers and number of crashes for males.
- (d) Compute the linear correlation coefficient between number of licensed drivers and number of crashes for females.
- (e) Which gender has the stronger linear relation between number of licensed drivers and number of crashes. Why?
- 33. Weight of a Car versus Miles per Gallon** Suppose we add the Ford Taurus to the data in Problem 25. A Ford Taurus weighs 3305 pounds and gets 19 miles per gallon.
- (a) Redraw the scatter diagram with the Taurus included.
- (b) Recompute the linear correlation coefficient with the Taurus included.
- (c) Compare the results of parts (a) and (b) with the results of Problem 25. Why are the results here reasonable?
- (d) Now suppose we add the Toyota Prius to the data in Problem 25 (remove the Taurus). A Toyota Prius weighs 2890 pounds and gets 60 miles per gallon. Redraw the scatter diagram with the Prius included. What do you notice?
- (e) Recompute the linear correlation coefficient with the Prius included. How did this new value affect your result?
- (f) Why does this observation not follow the pattern of the data?
- 34. Gestation Period versus Life Expectancy** Suppose we add humans to the data in Problem 24. Humans have a gestation period of 268 days and a life expectancy of 76.5 years.
- (a) Redraw the scatter diagram with humans included.
- (b) Recompute the linear correlation coefficient with humans included.
- (c) Compare the results of (a) and (b) with the results of Problem 24. Provide a statement that explains the results.



- 35.** Consider the following four data sets:

Data Set 1	
x	y
10	8.04
8	6.95
13	7.58
9	8.81
11	8.33
14	9.96
6	7.24
4	4.26
12	10.84
7	4.82
5	5.68

Data Set 2	
x	y
10	9.14
8	8.14
13	8.74
9	8.77
11	9.26
14	8.10
6	6.13
4	3.10
12	9.13
7	7.26
5	4.47

Data Set 3	
x	y
10	7.46
8	6.77
13	12.74
9	7.11
11	7.81
14	8.84
6	6.08
4	5.39
12	8.15
7	6.42
5	5.73

Data Set 4	
x	y
8	6.58
8	5.76
8	7.71
8	8.84
8	8.47
8	7.04
8	5.25
8	5.56
8	7.91
8	6.89
19	12.50

Source: Anscombe, Frank, Graphs in statistical analysis, *American Statistician*, 27 (1973):17–21

- (a) Compute the linear correlation coefficient for each data set.
- (b) Draw a scatter diagram for each data set. Conclude that linear correlation coefficients and scatter diagrams must be used together in any statistical analysis of bivariate data.

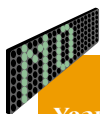
- 36. The Best Predictor of the Winning Percentage** The ultimate goal in any sport (besides having fun) is to win. One measure of how well a team does is the winning percentage. In baseball, a lot of effort goes into figuring out the variable that best predicts a team's winning percentage. The following data represent the winning percentages of teams in the National League along with potential explanatory variables. Which variable do you think is the best predictor of winning percentage? Why?



Team	Winning Percentage	Runs Scored	Home Runs	Team Batting Average	On-Base Percentage	Batting Average against Team	Team Earned-Run Average
Arizona	0.315	615	135	0.253	0.310	0.266	4.98
Atlanta	0.593	803	178	0.270	0.343	0.265	3.74
Chicago Cubs	0.549	789	235	0.268	0.328	0.247	3.81
Cincinnati	0.469	750	194	0.250	0.331	0.280	5.19
Colorado	0.420	833	202	0.275	0.345	0.290	5.54
Florida	0.512	718	148	0.264	0.329	0.256	4.10
Houston	0.568	803	187	0.267	0.342	0.258	4.05
Los Angeles	0.574	761	203	0.262	0.332	0.254	4.01
Milwaukee	0.416	634	135	0.248	0.321	0.259	4.24
Montreal	0.414	635	151	0.249	0.313	0.266	4.33
New York Mets	0.438	684	185	0.249	0.317	0.261	4.09
Philadelphia	0.531	840	215	0.267	0.345	0.264	4.45
Pittsburgh	0.447	680	142	0.260	0.321	0.267	4.29
San Diego	0.537	768	139	0.273	0.342	0.263	4.03
San Francisco	0.562	850	183	0.270	0.357	0.265	4.29
St. Louis	0.648	855	214	0.278	0.344	0.251	3.75

Source: espn.com

- 37. Diversification** One basic theory of investing is diversification. The idea is that you want to have a basket of stocks that do not all “move in the same direction.” In other words, if one investment goes down, you don't want a second investment in your portfolio that is also likely to go down. One hallmark of a good portfolio is a low correlation between investments. The following data represent the annual rates of return for various stocks. If you only wish to invest in two of the stocks, which two would you select if your goal is to have low correlation between the two investments? Which two would you select if your goal is to have one stock go up when the other goes down?



Year	Rate of Return				
	Cisco Systems	Walt Disney	General Electric	Exxon Mobil	TECO Energy
1996	0.704	0.204	0.565	0.405	-0.012
1997	0.314	0.448	0.587	0.342	0.223
1998	1.50	-0.080	0.451	0.254	0.050
1999	1.31	-0.015	0.574	0.151	-0.303
2000	-0.286	-0.004	-0.055	0.127	0.849
2001	-0.527	-0.277	-0.151	-0.066	-0.150
2002	-0.277	-0.203	-0.377	-0.089	-0.369
2003	0.850	0.444	0.308	0.206	0.004
2004	-0.203	0.202	0.207	0.281	0.128

Source: Yahoo!Finance

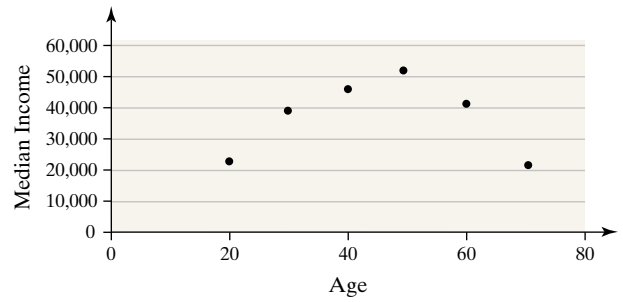
- 38. Lyme Disease versus Drownings** Lyme disease is an inflammatory disease that results in skin rash and flulike symptoms. It is transmitted through the bite of an infected deer tick. The following data represent the number of reported cases of Lyme disease and the number of drowning deaths for a rural county in the United States.

Month	J	F	M	A	M	J	J	A	S	O	N	D
Cases of Lyme Disease	3	2	2	4	5	15	22	13	6	5	4	1
Drowning Deaths	0	1	2	1	2	9	16	5	3	3	1	0

- (a) Draw a scatter diagram of the data using cases of Lyme disease as the explanatory variable.
- (b) Compute the correlation coefficient for the data.
- (c) Based on your results from parts (a) and (b), what type of relation appears to exist between the number of reported cases of Lyme disease and drowning deaths? Do you believe that an increase in cases of Lyme disease causes an increase in drowning deaths?
- 39. Television Stations and Life Expectancy** Based on data obtained from the *CIA World Factbook*, the linear correlation coefficient between number of television stations in a country and life expectancy of residents of the country is 0.599. What does this correlation imply? Do you believe that the more television stations a country has, the longer its population can expect to live? Why or why not?
- 40. Caffeine and SIDS** A study on the relationship between caffeine consumption during pregnancy and sudden infant death syndrome (SIDS) showed that heavy caffeine consumption during pregnancy was associated with a significant risk of SIDS. The study was later criticized on the claim that parental smoking was not properly assessed. Explain why this might be a concern.
- 41. Influential** Consider the following set of data:

x	2.2	3.7	3.9	4.1	2.6	4.1	2.9	4.7
y	3.9	4.0	1.4	2.8	1.5	3.3	3.6	4.9

- (a) Draw a scatter diagram of the data and compute the linear correlation coefficient.
- (b) Draw a scatter diagram of the data and compute the linear correlation coefficient with the additional data point (10.4, 9.3). Comment on the effect the additional data point has on the linear correlation coefficient. Explain why correlations should always be reported with scatter diagrams.
- 42. Faulty Use of Correlation** On the basis of the accompanying scatter diagram, explain what is wrong with the following statement: "Because the linear correlation coefficient between age and median income is 0.012, there is no relation between age and median income."



- 43. Name the Relation, Part I** For each of the following statements, explain whether you think the variables will have positive correlation, negative correlation, or no correlation. Support your opinion.
- Number of children in the household under the age of 3 and expenditures on diapers
 - Interest rates on car loans and number of cars sold
 - Number of hours per week on the treadmill and cholesterol level
 - Price of a Big Mac and number of McDonald's french fries sold in a week
 - Shoe size and IQ
- 44. Name the Relation, Part II** For each of the following statements, explain whether you think the variables will have positive correlation, negative correlation, or no correlation. Support your opinion.
- Number of cigarettes smoked by a pregnant woman each week and birth weight of her baby
 - Annual salary and years of education
 - Number of doctors on staff at a hospital and number of administrators on staff.
 - Head circumference and IQ
 - Number of moviegoers and movie ticket price
- 45. Transformations** Consider the following data set:
- | | | | | | | | | |
|-----|-----|---|-----|-----|-----|-----|-----|-----|
| x | 5 | 6 | 7 | 7 | 8 | 8 | 8 | 8 |
| y | 4.2 | 5 | 5.2 | 5.9 | 6 | 6.2 | 6.1 | 6.9 |
| x | 9 | 9 | 10 | 10 | 11 | 11 | 12 | 12 |
| y | 7.2 | 8 | 8.3 | 7.4 | 8.4 | 7.8 | 8.5 | 9.5 |
- Draw a scatter diagram with the x -axis starting at 0 and ending at 30 and with the y -axis starting at 0 and ending at 20.
 - Compute the linear correlation coefficient.
 - Now multiply both x and y by 2.
 - Draw a scatter diagram of the new data with the x -axis starting at 0 and ending at 30 and with the y -axis starting at 0 and ending at 20. Compare the scatter diagrams.
 - Compute the linear correlation coefficient.
 - Conclude that multiplying each value in the data set does not affect the correlation between the variables. Explain why this is the case.

46. Obesity In a study published in the *Journal of the American Medical Association* (May 16, 2001), researchers found that breast-feeding may help to prevent obesity in kids. In an interview, the head investigator stated, “It’s not clear whether breast milk has obesity-preventing properties or the women who are breast-feeding are less likely to have fat kids because they are less likely to be fat themselves and may be more health conscious.” Using this researcher’s statement, explain what might be wrong with the conclusion that breast-feeding prevents obesity. Identify some lurking variables in the study.

47. How Well Will You Do in College? The College Board is a membership association composed of schools, colleges, universities, and other educational organizations. One of its better-known programs is the administration of the SAT college entrance exam. In a recent study, the College Board wanted to learn what the best predictor of college grade-point average (GPA) was. The following correlations were obtained based on 48,039 students.

Correlation between College GPA and:	Correlation
--------------------------------------	-------------

SAT score combined with high school GPA	0.61
SAT verbal score	0.47
SAT math score	0.48
SAT combined verbal and math score	0.52
High school GPA	0.54

Source: The College Board

- Which variable is the best predictor of college GPA?
- Which variable is the worst predictor of college GPA?

48. Correlation Applet Load the correlation by eye applet.

- In the lower-left corner of the applet, add 10 points that line up with a positive slope so that the linear correlation between the points is about 0.8. Click “show r” to show the correlation.
- Add another point in the upper-right corner of the applet that roughly lines up with the 10 points you have in the lower-left corner. Comment on how the linear correlation coefficient changes.
- Drag the point in the upper-right corner straight down. Take note of the change in the linear correlation coefficient. Notice how a single point can have a substantial impact on the linear correlation coefficient.

49. Correlation Applet Load the correlation by eye applet.

- Add about 10 points that form an upside-down U. Certainly, there is a relation between x and y , but what is the value of the linear correlation coefficient? Conclude that a low linear correlation coefficient does not imply there is no relation between two variables; it means there may be no linear relation between two variables.

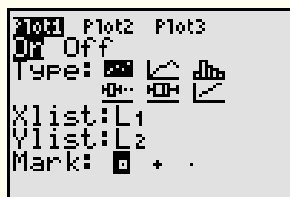
50. Correlation Applet Load the correlation by eye applet.

- Plot about 10 points that follow a linear trend and have a linear correlation coefficient that is close to 0.8.
- Clear the applet. Plot about 6 points vertically on top of each other on the left side of the applet. Add a seventh point to the right of the applet. Move the point until the linear correlation coefficient is close to 0.8.
- Clear the applet. Plot about 7 points in a U-shaped curve. Add an eighth point and move it around the applet until the linear correlation coefficient is close to 0.8.
- Conclude that a linear correlation coefficient can result from data that have many patterns and so you should always plot your data.

Technology Step by Step

Drawing Scatter Diagrams and Determining the Correlation Coefficient

TI-83/84 Plus Scatter Diagrams



Step 1: Enter the explanatory variable in L1 and the response variable in L2.

Step 2: Press 2^{nd} $Y =$ to bring up the StatPlot menu. Select 1: Plot1.

Step 3: Turn Plot 1 ON by putting the cursor on the ON button and pressing ENTER.

Step 4: Highlight the scatter diagram icon (see the figure) and press ENTER. Be sure that Xlist is L1 and Ylist is L2.

Step 5: Press ZOOM and select 9: ZoomStat.

Correlation Coefficient

Step 1: Turn the diagnostics on by selecting the catalog (2^{nd} θ). Scroll down and select **DiagnosticOn**. Hit ENTER to activate diagnostics.

Step 2: With the explanatory variable in L1 and the response variable in L2, press STAT, highlight CALC, and select 4: **LinReg (ax + b)**. With **LinReg** on the HOME screen, press ENTER.

MINITAB Scatter Diagrams

Step 1: Enter the explanatory variable in C1 and the response variable in C2. You may want to name the variables.

Step 2: Select the **Graph** menu and highlight **Plot . . .**

Step 3: With the cursor in the Y column, select the response variable. With the cursor in the X column, select the explanatory variable. Click OK.

Correlation Coefficient

Step 1: With the explanatory variable in C1 and the response variable in C2, select the **Stat** menu and highlight **Basic Statistics**. Highlight **Correlation**.

Step 2: Select the variables whose correlation you wish to determine and click OK.

Excel Scatter Diagrams

Step 1: Enter the explanatory variable in column A and the response variable in column B. Select the Chart Wizard icon.

Step 2: Follow the instructions in the Chart Wizard.

Correlation Coefficient

Step 1: Be sure the Data Analysis Tool Pak is activated by selecting the **Tools** menu and highlighting **Add-Ins . . .** Check the box for the Analysis ToolPak and select OK.

Step 2: Select **Tools** and highlight **Data Analysis . . .** Highlight **Correlation** and select OK.

Step 3: With the cursor in the Input Range, highlight the data. Select OK.

4.2 Least-Squares Regression

Preparing for This Section Before getting started, review the following:

- Lines (Section C.1 on CD, pp. C1–C6)

Objectives

- 1 Find the least-squares regression line and use the line to make predictions**
- 2 Interpret the slope and the y -intercept of the least-squares regression line**
- 3 Compute the sum of squared residuals**

Once the scatter diagram and linear correlation coefficient indicate that a linear relation exists between two variables, we proceed to find a linear equation that describes the relation between the two variables. One way to obtain a line that describes the relation is to select two points from the data that appear to provide a good fit and find the equation of the line through these points.

EXAMPLE 1**Finding an Equation That Describes Linearly Related Data****CAUTION**

The method for obtaining an equation that describes the relation between two variables discussed in Example 1 is *not* the least-squares method. It is used to illustrate a point.

Problem: The data in Table 3 represent the club-head speed and the distance a golf ball travels for eight swings of the club. We determined that these data are linearly related in the last section.

- Find a linear equation that relates club-head speed, x (the explanatory variable), and distance, y (the response variable), by selecting two points and finding the equation of the line containing the points.
- Graph the line on the scatter diagram.
- Use the equation to predict the distance a golf ball will travel if the club-head speed is 104 miles per hour.

**Table 3**

Club-Head Speed (mph) x	Distance (yards) y	(x, y)
100	257	(100, 257)
102	264	(102, 264)
103	274	(103, 274)
101	266	(101, 266)
105	277	(105, 277)
100	263	(100, 263)
99	258	(99, 258)
105	275	(105, 275)

Source: Paul Stephenson, student at Joliet Junior College

Approach

- To answer part (a), we perform the following steps:

Step 1: Select two points from Table 3 so that a line drawn through the points appears to give a good fit. Call the points (x_1, y_1) and (x_2, y_2) . Refer to Figure 1 for the scatter diagram.

Step 2: Find the slope of the line containing these two points using

$$m = \frac{y_2 - y_1}{x_2 - x_1}$$

Step 3: Use the point-slope formula, $y - y_1 = m(x - x_1)$, to find the line through the points selected in Step 1. Express the line in the form $y = mx + b$, where m is the slope and b is the y -intercept.

- For part (b), draw a line through the points selected in Step 1 of part (a).
- Finally, for part (c), we let $x = 104$ in the equation found in part (a).

Solution

- Step 1:** We will select $(x_1, y_1) = (99, 258)$ and $(x_2, y_2) = (105, 275)$, because a line drawn through these two points seems to give a good fit.

$$\text{Step 2: } m = \frac{y_2 - y_1}{x_2 - x_1} = \frac{275 - 258}{105 - 99} = \frac{17}{6} = 2.8333$$

Step 3: We use the point-slope form of a line to find the equation of the line.

**In Other Words**

A *good fit* means that the line drawn appears to describe the relation between the two variables well.

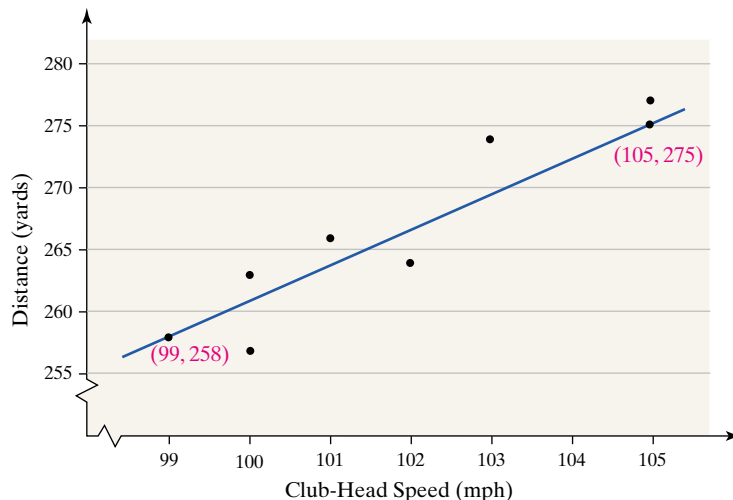
**CAUTION**

The line found in Step 3 of Example 1 is not the least-squares regression line.

$$\begin{aligned} y - y_1 &= m(x - x_1) \\ y - 258 &= 2.8333(x - 99) & m = 2.8333, x_1 = 99, y_1 = 258 \\ y - 258 &= 2.8333x - 280.4967 \\ y &= 2.8333x - 22.4967 \end{aligned} \quad (1)$$

The slope of the line is 2.8333 and the y -intercept is -22.4967 .

(b) Figure 8 shows the scatter diagram along with the line drawn through the points $(99, 258)$ and $(105, 275)$.

Figure 8**CAUTION**

Unless otherwise noted, we will round to four decimal places. As always, do not round until the last computation.

(c) We let $x = 104$ in equation (1) to predict the distance a golf ball travels when hit with a club-head speed of 104 miles per hour.

$$\begin{aligned} y &= 2.8333(104) - 22.4967 \\ &= 272.2 \text{ yards} \end{aligned}$$

We predict that a golf ball will travel 272.2 yards when it is hit with a club-head speed of 104 miles per hour.

Now Work Problems 11(a), 11(b), and 11(c).

1**Find the Least-Squares Regression Line and Use the Line to Make Predictions**

Although the line that we found in Example 1 appears to describe the relation between club-head speed and distance well, is there a line that fits the data better? Is there a line that fits the data *best*?

Consider Figure 9. Each y -coordinate on the line corresponds to a predicted distance for a given club-head speed. For example, if club-head speed is 103 miles per hour, the predicted distance is $2.8333(103) - 22.4967 = 269.3$ yards. The observed distance for this club-head speed is 274 yards. The difference between the observed value of y and the predicted value of y is the error or **residual**. For a swing speed of 103 mph the residual is

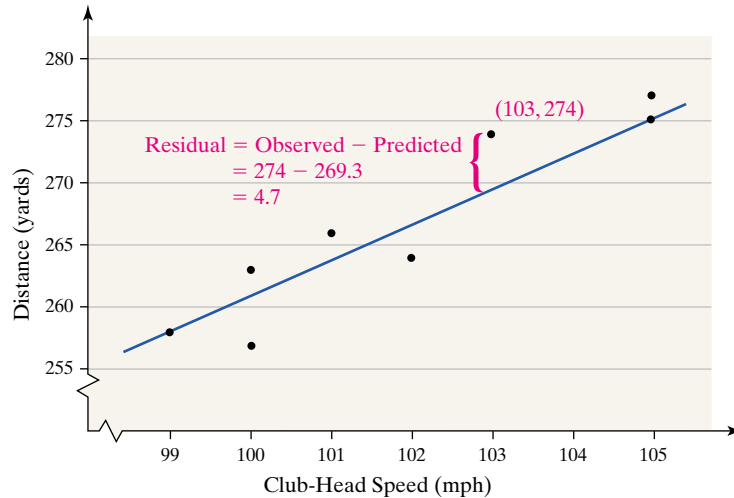
$$\begin{aligned} \text{Residual} &= \text{observed } y - \text{predicted } y \\ &= 274 - 269.3 \\ &= 4.7 \text{ yards} \end{aligned}$$

**In Other Words**

The residual represents how close our prediction comes to actual observation. The smaller the residual, the better the prediction.

The residual for a club-head speed of 103 miles per hour is labeled in Figure 9.

Figure 9



The line that *best* describes the relation between two variables is the one that minimizes the distance between the points and the line. The most popular technique for making the residuals as small as possible is the *method of least squares*, discovered by Adrien Marie Legendre.



Definition

Historical Note

Adrien Marie Legendre was born on September 18, 1752, into a wealthy family and was educated in mathematics and physics at the College Mazarin in Paris. From 1775 to 1780, he taught at Ecole Militaire. On March 30, 1783, Legendre was appointed an adjoint in the Académie des Sciences. On May 13, 1791, he became a member of the committee of the Académie des Sciences and was charged with the task of standardizing weights and measures. The committee worked to compute the length of the meter. During the revolution, Legendre lost his small fortune. In 1794, Legendre published *Éléments de géométrie*, which was the leading elementary text in geometry for around 100 years. In 1806, Legendre published a book on orbits, in which he developed the theory of least squares. He died on January 10, 1833.

Least-Squares Regression Criterion

The **least-squares regression line** is the one that minimizes the sum of the squared errors (or residuals). It is the line that minimizes the square of the vertical distance between the observed values of y and those predicted by the line, \hat{y} (read “y-hat”). We represent this as

$$\text{Minimize } \sum \text{residuals}^2$$

The advantage of the least-squares criterion is that it allows for statistical inference on the predicted value and slope (Chapter 12). Another advantage of the least-squares criterion is explained by Legendre in his text *Nouvelles méthodes pour la détermination des orbites des comètes*, published in 1806.

Of all the principles that can be proposed for this purpose, I think there is none more general, more exact, or easier to apply, than that which we have used in this work; it consists of making the sum of squares of the errors a *minimum*. By this method, a kind of equilibrium is established among the errors which, since it prevents the extremes from dominating, is appropriate for revealing the state of the system which most nearly approaches the truth.

Equation of the Least-Squares Regression Line

The equation of the least-squares regression line is given by

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

where

$$b_1 = r \cdot \frac{s_y}{s_x} \text{ is the } \mathbf{slope} \text{ of the least-squares regression line}^* \quad (2)$$

and

$$b_0 = \bar{y} - b_1\bar{x} \text{ is the } \mathbf{y-intercept} \text{ of the least-squares regression line} \quad (3)$$

Note: \bar{x} is the sample mean and s_x is the sample standard deviation of the explanatory variable x ; \bar{y} is the sample mean and s_y is the sample standard deviation of the response variable y .

*An equivalent formula is

$$b_1 = \frac{s_{xy}}{s_{xx}} = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}$$

The notation \hat{y} is used in the least-squares regression line to serve as a reminder that it is a predicted value of y for a given value of x . An interesting property of the least-squares regression line, $\hat{y} = b_1x + b_0$, is that the line always contains the point (\bar{x}, \bar{y}) . This property can be useful when drawing the least-squares regression line by hand.

Since s_y and s_x must both be positive, the sign of the linear correlation coefficient and the sign of the slope of the least-squares regression line are the same. For example, if r is positive, then the slope of the least-squares regression line will also be positive.

EXAMPLE 2

Finding the Least-Squares Regression Line



Historical Note

Sir Francis Galton was born on February 16, 1822. Galton came from a wealthy and well-known family. Charles Darwin was his first cousin. Galton studied medicine at Cambridge. After receiving a large inheritance, he left the medical field and traveled the world. He explored Africa from 1850 to 1852. In the 1860s, his study of meteorology led him to discover anticyclones. Influenced by Darwin, Galton always had an interest in genetics and heredity. He studied heredity through experiments with sweet peas. He noticed that the weight of the “children” of the “parent” peas reverted or *regressed* to the mean weight of all peas. Hence, the term *regression analysis*. Galton died on January 17, 1911.

Problem: For the data in Table 3 on page 196,

- Find the least-squares regression line.
- Predict the distance a golf ball will travel when hit with a club-head speed of 103 miles per hour.
- Compute the residual for the prediction made in part (b).
- Draw the least-squares regression line on the scatter diagram of the data.

Approach

- From Example 2 in Section 4.1, we have the following:

$$r = 0.9387, \bar{x} = 101.875, s_x = 2.2952, \bar{y} = 266.75, \text{ and } s_y = 7.74135$$

We substitute these values into Formula (2) to find the slope of the least-squares regression line. We use Formula (3) to find the intercept of the least-squares regression line.

- Substitute $x = 103$ into the least-squares regression line found in part (a) to find \hat{y} .
- The residual is the difference between the observed y and the predicted y . That is, $\text{residual} = y - \hat{y}$.
- To draw the least-squares regression line, select two values of x and use the equation to find the predicted values of y . Plot these points on the scatter diagram and draw a line through the points.

Solution

- Substituting $r = 0.9387$, $s_x = 2.2952$, and $s_y = 7.74135$ into Formula (2), we obtain

$$b_1 = r \cdot \frac{s_y}{s_x} = 0.9387 \cdot \frac{7.74135}{2.2952} = 3.1661$$

We have that $\bar{x} = 101.875$ and $\bar{y} = 266.75$. Substituting these values into Formula (3), we obtain

$$b_0 = \bar{y} - b_1\bar{x} = 266.75 - 3.1661(101.875) = -55.7964$$

The least-squares regression line is

$$\hat{y} = 3.1661x - 55.7964$$

- We let $x = 103$ in the equation $y = 3.1661x - 55.7964$ to predict the distance a golf ball hit with a club-head speed of 103 miles per hour will travel.

$$\hat{y} = 3.1661(103) - 55.7964$$

$$= 270.3 \text{ yards}$$

We predict that the distance the ball will travel is 270.3 yards.



CAUTION

Throughout the text, we will round the slope and y -intercept values to four decimal places. The predictions should be made to one more decimal place than the response variable.

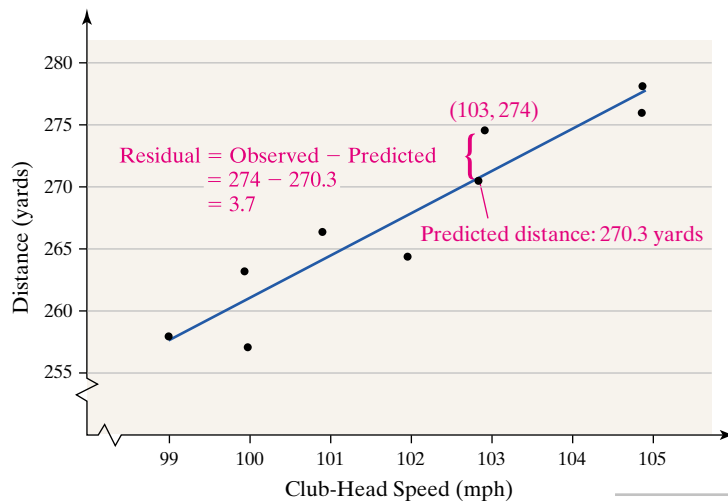
- (c) The actual distance the ball traveled is 274 yards. The residual is

$$\begin{aligned}\text{Residual} &= \text{observed } y - \text{predicted } y \\ &= y - \hat{y} \\ &= 274 - 270.3 \\ &= 3.7 \text{ yards}\end{aligned}$$

We underestimated the distance by 3.7 yards.

- (d) Figure 10 shows the graph of the least-squares regression line drawn on the scatter diagram with the residual labeled.

Figure 10



In Other Words

An underestimate means the residual is positive; an overestimate means the residual is negative. A residual of zero means the prediction is right on!

Notice that an underestimate results in a positive residual, while an overestimate results in a negative residual.

In-Class Activity: Paper Thin (Regression)

Each student, or small group of students, should receive a ruler with both inches and centimeters.

1. Use the ruler provided to measure the thickness of only one page of the text. Which unit of measurement did you use and why?
2. Grouping pages (one page is one sheet), complete the following table:

No. of pages	25	50	75	150	200	225
Thickness						

3. Compute the least-squares regression line for your data.
4. Compare your data and your regression line to those around you. Did everyone get the same measurements and model? Explain why or why not.
5. Use your model to estimate the thickness of one page of the text. Is the value you obtained reasonable?
6. Use your model to estimate the thickness of 0 pages of the text. Is the value you obtained reasonable?
7. What, if anything, could you do to improve the model?

We can think of any point on the least-squares regression line as an estimate of the mean value of the response variable for a given value of the explanatory variable. For example, the mean distance a golf ball will travel when hit with a club-head speed of 103 miles per hour is 270.3 yards. In our experiment, when we

hit the ball with a club-head speed of 103 miles per hour, the ball traveled 274 yards. So the distance the ball traveled was above the mean. Perhaps we hit the ball in the “sweet spot” of the club face or a breeze kicked up at our back.

2 Interpret the Slope and y -Intercept of the Least-Squares Regression Line

The definition of the slope of a line is $\frac{\text{Rise}}{\text{Run}}$ or $\frac{\text{Change in } y}{\text{Change in } x}$. For a line whose slope is $\frac{2}{3}$, if x increases by 3, y will *increase* by 2. Or, if the slope of a line is $-4 = \frac{-4}{1}$, if x increases by 1, y will *decrease* by 4.

The y -intercept of any line is the point where the graph intersects the vertical axis. It is found by letting $x = 0$ in an equation and solving for y .

We found the regression equation in Example 2 to be $\hat{y} = 3.1661x - 55.7964$. So the slope of the line is 3.1661. We interpret this slope as follows: If the club-head speed increases by 1 mile per hour, the distance the ball travels increases by 3.1661 yards, on average. To interpret the y -intercept, we must first ask two questions:

1. Is 0 a reasonable value for the explanatory variable?
2. Do any observations near $x = 0$ exist in the data set?

If the answer to either of these questions is no, we do not give an interpretation to the y -intercept. In the regression equation of Example 2, a swing speed of 0 miles per hour does not make sense, so an interpretation of the y -intercept is unreasonable. To interpret a y -intercept, we would say that it is the value of the response variable when the value of the explanatory variable is 0.

The second condition for interpreting the y -intercept is especially important because we should not use the regression model to make predictions **outside the scope of the model**. If this cannot be avoided, be cautious when using the regression model to make predictions for values of the explanatory variable that are much larger or much smaller than those observed, because we cannot be certain of the behavior of data for which we have no observations.

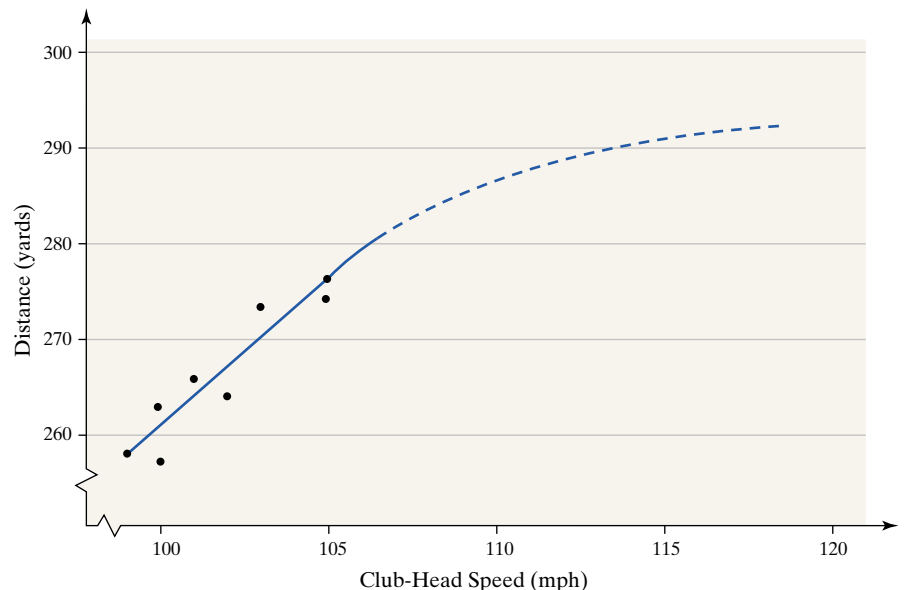
For example, it is inappropriate to use the line we determined in Example 2 to predict distance when club-head speed is 140 miles per hour. The highest observed club-head speed in our data set is 105 miles per hour. We cannot be certain that the linear relation between distance and club-head speed will continue. See Figure 11.



CAUTION

Be careful when using the least-squares regression line to make predictions for values of the explanatory variable that are much larger or much smaller than those observed.

Figure 11



We have presented the procedure for determining the least-squares regression line by hand. In practice, however, a statistical spreadsheet or calculator with advanced statistical features is used to determine the least-squares regression line.

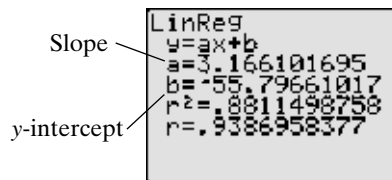
EXAMPLE 3 Finding the Least-Squares Regression Line Using Technology

Problem: Use a statistical spreadsheet or a graphing calculator with advanced statistical features to find the least-squares regression line of the data in Table 3.

Approach: Because technology plays a major role in obtaining the least-squares regression line, we will use a TI-84 Plus graphing calculator, MINITAB, and Excel to obtain the least-squares regression line. The steps for obtaining these lines are given in the Technology Step by Step on page 208.

Result: Figure 12(a) shows the output obtained from a TI-84 Plus graphing calculator, Figure 12(b) shows the output obtained from MINITAB with the slope and y-intercept highlighted, and Figure 12(c) shows partial output from Excel with the slope and y-intercept highlighted.

Figure 12



(a) TI-84 Plus output

The regression equation is

$$\text{Distance (yards)} = -55.8 + 3.17 \text{ Club Head Speed (mph)}$$

Predictor	Coef	SE Coef	T	P
Constant	-55.80	48.37	-1.15	0.293
Club Head Speed (mph)	3.1661	0.4747	6.67	0.001

$$S = 2.88264 \quad R - Sq = 88.1\% \quad R - Sq(\text{adj}) = 86.1\%$$

(b) MINITAB output

	Coefficients	Standard Error	t Stat	P-value
Intercept	-55.79661017	48.37134953	-1.153505344	0.29257431
Club Head Speed (mph)	3.166101695	0.47470539	6.669613957	0.00054983

(c) Excel output

Note: To get the linear correlation coefficient from MINITAB, use

$$r = \sqrt{R - Sq}. \text{ So } r = \sqrt{0.881} = 0.9386.$$

Now Work Problems 11(d) and 11(e).

3

Compute the Sum of Squared Residuals

Recall that the least-squares regression line is the line that minimizes the sum of the squared residuals. This means that the sum of the squared residuals, $\sum \text{residuals}^2$, for the least-squares line will be smaller than for any other line that may describe the relation between the two variables. In particular, the sum of the squared residuals for the line obtained in Example 2 using the method of least squares will be smaller than the sum of the squared residuals for the line obtained in Example 1. It is worthwhile to verify this result.

EXAMPLE 4 Comparing the Sum of Squared Residuals

Problem: Compare the sum of squared residuals for the lines obtained in Examples 1 and 2.

Approach: We compute Σ residuals² using the predicted values of y , \hat{y} , for the lines obtained in Examples 1 and 2. This is best done by creating a table of values.

Solution: We create Table 4, which contains the value of the explanatory variable in column 1. Column 2 contains the corresponding response variable. Column 3 contains the predicted value using the equation obtained in Example 1, $\hat{y} = 2.8333x - 22.4967$. In column 4, we compute the residuals for each observation: residual = observed y - predicted y . For example, the first residual using the equation found in Example 1 is observed y - predicted $y = 257 - 260.8 = -3.8$. Column 5 contains the squares of the residuals obtained in column 4. Column 6 contains the predicted value using the least-squares regression equation obtained in Example 2: $\hat{y} = 3.1661x - 55.7964$. Column 7 represents the residuals for each observation and column 8 represents the squared residuals.

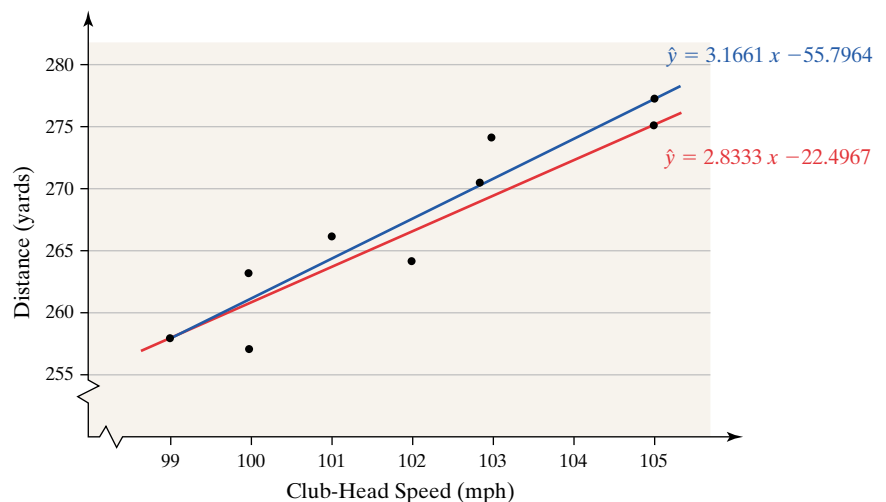
Table 4

Club-Head Speed (mph)	Distance (yards)	Example 1		Example 2			
		($\hat{y} = 2.8333x - 22.4967$)	Residual	Residual ²	($\hat{y} = 3.1661x - 55.7964$)	Residual	Residual ²
100	257	260.8	-3.8	14.44	260.8	-3.8	14.44
102	264	266.5	-2.5	6.25	267.2	-3.2	10.24
103	274	269.3	4.7	22.09	270.3	3.7	13.69
101	266	263.7	2.3	5.29	264.0	2.0	4.00
105	277	275.0	2.0	4.00	276.6	0.4	0.16
100	263	260.8	2.2	4.84	260.8	2.2	4.84
99	258	258.0	0.0	0.00	257.6	0.4	0.16
105	275	275.0	0.0	0.00	276.6	-1.6	2.56
Σ residual ²						Σ residual ²	
= 56.91						= 50.09	

The sum of the squared residuals for the line found in Example 1 is 56.91; the sum of the squared residuals for the least-squares regression line is 50.09. Again, any line that describes the relation between distance and club-head speed will have a sum of squared residuals that is greater than 50.09.

Now Work Problems 11(f), (g) and (h)

We draw the graphs of the two lines obtained in Examples 1 and 2 on the same scatter diagram in Figure 13 to help the reader visualize the difference.

Figure 13

MAKING AN INFORMED DECISION

**What Car Should I Buy?**

You are still in the market to buy a car. As we all know, cars lose value over time. Therefore, another item to consider when purchasing a car is its depreciation rate. The higher the depreciation rate, the more value the car loses each year. Using the same three cars that you used in the Chapter 3 Decision, answer the following questions to determine depreciation rate.

1. Collect information regarding the three cars in your list by finding at least 12 cars of each car model that are for sale. Obtain the asking price and age of the car. Sources of data include your local newspaper classified ads or car Web sites, such as www.cars.com and www.vehix.com.
2. For each car type, draw a scatter diagram, treating age of the car as the explanatory variable and asking price as the response variable. Does the relation between the two variables appear linear?
3. The asking price of a car and the age of a car are related through the exponential equation $y = ab^x$, where y is the price of the car and x is the age of the car. The depreciation rate is $1 - b$. To estimate the values of a and b using least-squares regres-

sion, we need to transform the equation to a linear equation. This is accomplished by computing the logarithm of the asking price. For example, if the asking price of a car is \$8000, compute $\log(8000) = 3.9031$. Compute the logarithm of each asking price for each car.

4. For each car model, draw a scatter diagram, treating age of the car as the explanatory variable and the logarithm of asking price as the response variable. Does the relation between age and the logarithm of asking price appear to be linear?
5. For each car model, find the least-squares regression line, treating age of the car as the explanatory variable and the logarithm of asking price as the response variable. The line will be $Y = \log a + (\log b)x = A + Bx$.
6. To determine a and b for the exponential equation $y = ab^x$, let $a = 10^A$ and $b = 10^B$. Graph the exponential equation on each scatter diagram from question 2.
7. For each exponential equation determined in question 6, the depreciation rate is $1 - b$. What are the depreciation rates for the three cars considered? Will this result affect your decision about which car to buy?

4.2 ASSESS YOUR UNDERSTANDING

Concepts and Vocabulary

1. Explain the least-squares regression criterion.
2. What is a residual? What does it mean when a residual is positive?
3. Explain the phrase *outside the scope of the model*. Why is it dangerous to make predictions outside the scope of the model?
4. If the linear correlation between two variables is negative, what can be said about the slope of the regression line?
5. In your own words, explain the meaning of Legendre's quote given on page 198.
6. *True or False:* The least-squares regression line always travels through the point (\bar{x}, \bar{y}) .
7. In your own words, explain what each point on the least-squares regression line represents.
8. If the linear correlation coefficient is 0, what is the equation of the least-squares regression line?

Skill Building

9. For the data set

x	0	2	3	5	6	6
y	5.8	5.7	5.2	2.8	1.9	2.2

- (a) Draw a scatter diagram. Comment on the type of relation that appears to exist between x and y .
- (b) Given that $\bar{x} = 3.667$, $s_x = 2.42212$, $\bar{y} = 3.933$, $s_y = 1.8239152$, and $r = -0.9476938$, determine the least-squares regression line.
- (c) Graph the least-squares regression line on the scatter diagram drawn in part (a).

10. For the data set

x	2	4	8	8	9
y	1.4	1.8	2.1	2.3	2.6

- (a) Draw a scatter diagram. Comment on the type of relation that appears to exist between x and y .
- (b) Given $\bar{x} = 6.2$, $s_x = 3.03315$, $\bar{y} = 2.04$, that $s_y = 0.461519$, and $r = 0.957241$, determine the least-squares regression line.
- (c) Graph the least-squares regression line on the scatter diagram drawn in part (a).

In Problems 11–16,

- (a) Draw a scatter diagram treating x as the explanatory variable and y as the response variable.
- (b) Select two points from the scatter diagram and find the equation of the line containing the points selected.
- (c) Graph the line found in part (b) on the scatter diagram.
- (d) Determine the least-squares regression line.
- (e) Graph the least-squares regression line on the scatter diagram.
- (f) Compute the sum of the squared residuals for the line found in part (b).
- (g) Compute the sum of the squared residuals for the least-squares regression line found in part (d).
- (h) Comment on the fit of the line found in part (b) versus the least-squares regression line found in part (d).

11. NW

x	3	4	5	7	8
y	4	6	7	12	14

12.

x	3	5	7	9	11
y	0	2	3	6	9

13.

x	-2	-1	0	1	2
y	-4	0	1	4	5

14.

x	-2	-1	0	1	2
y	7	6	3	2	0

15.

x	20	30	40	50	60
y	100	95	91	83	70

16.

x	5	10	15	20	25
y	2	4	7	11	18

Applying the Concepts

Problems 17–22 use the results from Problems 23–28 in Section 4.1.

- 17. Height versus Head Circumference** (Refer to Problem 23, Section 4.1) A pediatrician wants to determine the relation that exists between a child's height, x , and head circumference, y . She randomly selects 11 children from her practice, measures their height and head circumference, and obtains the following data.

Height, x (inches)	Head Circumference, y (inches)	Height, x (inches)	Head Circumference, y (inches)
27.75	17.5	26.5	17.3
24.5	17.1	27	17.5
25.5	17.1	26.75	17.3
26	17.3	26.75	17.5
25	16.9	27.5	17.5
27.75	17.6		

Source: Denise Slucki, student at Joliet Junior College

- (a) Find the least-squares regression line, treating height as the explanatory variable and head circumference as the response variable.
- (b) Interpret the slope and intercept, if appropriate.
- (c) Use the regression equation to predict the head circumference of a child who is 25 inches tall.
- (d) Compute the residual based on the observed head circumference of the 25-inch-tall child in the table. Is the head circumference of this child above average or below average?
- (e) Draw the least-squares regression line on the scatter diagram of the data and label the residual from part (d).
- (f) Notice that two children are 26.75 inches tall. One has a head circumference of 17.3 inches; the other has a head circumference of 17.5 inches. How can this be?
- (g) Would it be reasonable to use the least-squares regression line to predict the head circumference of a child who was 32 inches tall? Why?

- 18. Gestation Period versus Life Expectancy** (Refer to Problem 24, Section 4.1) The following data represent the gestation period, x , of various animals along with their life expectancy, y .



Animal	Gestation (or Incubation) Period (days), x	Life Expectancy (years), y
Cat	63	11
Chicken	22	7.5
Dog	63	11
Duck	28	10
Goat	151	12
Lion	108	10
Parakeet	18	8
Pig	115	10
Rabbit	31	7
Squirrel	44	9

Source: Time Almanac 2000

- (a) Find the least-squares regression line, treating gestation period as the explanatory variable and life expectancy as the response variable.
- (b) Interpret the slope and intercept, if appropriate.
- (c) Suppose a new animal species has been discovered. After breeding the species in captivity, it is determined that the gestation period is 95 days. Use the least-squares regression line to predict the life expectancy of the animal.
- (d) Use the regression equation to predict the life expectancy of a parakeet.
- (e) Use the regression equation to predict the life expectancy of a rabbit.
- (f) Compute the residual of the prediction made in part (e). Conclude that the least-squares regression

line sometimes provides accurate predictions (as in the case of the parakeet) and sometimes provides inaccurate predictions (as in the case of the rabbit). Unfortunately, when the value of the response variable is unknown, as in the case of the new animal species from part (c), we don't know the accuracy of the prediction.

19. Weight of a Car versus Miles per Gallon (Refer to Problem 25, Section 4.1) An engineer wants to determine how the weight of a car, x , affects gas mileage, y . The following data represent the weight of various domestic cars and their miles per gallon in the city for the 2005 model year.



Car	Weight (pounds), x	Miles per Gallon, y
Buick LeSabre	3565	20
Cadillac DeVille	3985	18
Chevrolet Corvette	3180	19
Chevrolet Monte Carlo	3340	21
Chrysler PT Cruiser	3100	21
Chrysler Sebring Sedan	3175	22
Dodge Neon	2580	27
Dodge Stratus Sedan	3175	22
Ford Focus	2655	26
Ford Mustang	3300	20
Lincoln LS	3680	20
Mercury Sable	3310	19
Pontiac Bonneville	3590	20
Pontiac Grand Am	3475	20
Pontiac Sunfire	2770	24
Saturn Ion	2690	26

Source: www.roadandtrack.com

- Find the least-squares regression line treating weight as the explanatory variable and miles per gallon as the response variable.
- Interpret the slope and intercept, if appropriate.
- Predict the miles per gallon of a Ford Mustang and compute the residual. Is the miles per gallon of a Mustang above average or below average for cars of this weight?
- Draw the least-squares regression line on the scatter diagram of the data and label the residual.
- Would it be reasonable to use the least-squares regression line to predict the miles per gallon of a Toyota Prius, a hybrid gas and electric car? Why or why not?

20. Bone Length (Refer to Problem 26, Section 4.1) Research performed at NASA and led by Emily R. Morey-Holton measured the lengths of the right humerus and right tibia in 11 rats that were sent to space on SpaceLab Life Sciences 2. The following data were collected.



Right Humerus (mm)	Right Tibia (mm)	Right Humerus (mm)	Right Tibia (mm)
24.80	36.05	25.90	37.38
24.59	35.57	26.11	37.96
24.59	35.57	26.63	37.46
24.29	34.58	26.31	37.75
23.81	34.20	26.84	38.50
24.87	34.73		

Source: NASA Life Sciences Data Archive

- Find the least-squares regression line, treating the length of the right humerus, x , as the explanatory variable and the length of the right tibia, y , as the response variable.
- Interpret the slope and intercept, if appropriate.
- Determine the residual if the length of the right humerus is 26.11 mm and the actual length of the right tibia is 37.96 mm. Is the length of this tibia above or below average?
- Draw the least-squares regression line on the scatter diagram and label the residual from part (c).
- Suppose one of the rats sent to space experienced a broken right tibia due to a severe landing. The length of the right humerus is determined to be 25.31 mm. Use the least-squares regression line to estimate the length of the right tibia.

21. Attending Class (Refer to Problem 27, Section 4.1) The following data represent the number of days absent, x , and the final grade, y , for a sample of college students in a general education course at a large state university.

No. of absences, x	0	1	2	3	4	5	6	7	8	9
Final grade, y	89.2	86.4	83.5	81.1	78.2	73.9	64.3	71.8	65.5	66.2

Source: College Teaching, Winter 2005, Vol. 53, Issue 1

- Find the least-squares regression line, treating number of absences as the explanatory variable and final grade as the response variable.
- Interpret the slope and intercept, if appropriate.
- Predict the final grade for a student who misses five class periods and compute the residual. Is the final grade above or below average for this number of absences?
- Draw the least-squares regression line on the scatter diagram of the data and label the residual.
- Would it be reasonable to use the least-squares regression line to predict the final grade for a student who has missed 15 class periods? Why or why not?

22. Antibiotics (Refer to Problem 28, Section 4.1) A study on antibiotic use among children in Manitoba, Canada, gave the following data for the number of prescriptions per 1000 children x years after 1995.

Year, x	0	1	2	3	4	5	6
Prescriptions (per 1000 children), y	1201	1070	944	964	909	949	864

Source: Canadian Medical Association Journal, Vol. 171, Issue 2

- (a) Find the least-squares regression line, treating year as the explanatory variable and prescriptions as the response variable.
- (b) Interpret the slope and intercept, if appropriate.
- (c) Predict the number of prescriptions per 1000 children in Manitoba, Canada, in 2002 ($x = 7$).
- (d) Draw the least-squares regression line on the scatter diagram of the data.
- (e) Would it be reasonable to use the least-squares regression line to predict the number of prescriptions in Manitoba, Canada, in 2010? Why or why not?

23. Does Size Matter? Researchers wondered whether the size of a person's brain was related to the individual's mental capacity. They selected a sample of right-handed introductory psychology students who had SAT scores higher than 1350. The subjects were administered the Wechsler (1981) Adult Intelligence Scale-Revised exam to obtain their IQ scores. MRI scans, performed at the same facility, consisted of 18 horizontal magnetic resonance images. The computer counted all pixels with nonzero gray scale in each of the 18 images, and the total count served as an index for brain size. The resulting data are presented in the table.



MRI			MRI		
Gender	Count, x	IQ, y	Gender	Count, x	IQ, y
Female	816,932	133	Male	949,395	140
Female	951,545	137	Male	1,001,121	140
Female	991,305	138	Male	1,038,437	139
Female	833,868	132	Male	965,353	133
Female	856,472	140	Male	955,466	133
Female	852,244	132	Male	1,079,549	141
Female	790,619	135	Male	924,059	135
Female	866,662	130	Male	955,003	139
Female	857,782	133	Male	935,494	141
Female	948,066	133	Male	949,589	144

Source: Willerman, L., Schultz, R., Rutledge, J. N., and Bigler, E. (1991). "In Vivo Brain Size and Intelligence." Intelligence, 15, 223–228

- (a) Find the least-squares regression line treating MRI count as the explanatory variable and IQ as the response variable.
- (b) What do you notice about the value of the slope? Why does this result seem reasonable based on the scatter diagram and linear correlation coefficient obtained in Problem 31 of Section 4.1?
- (c) When there is no relation between the explanatory and response variable, we use the mean value of the response variable, \bar{y} , to predict. Predict the IQ of an individual whose MRI count is 1,000,000. Predict the IQ of an individual whose MRI count is 830,000.

24. Male versus Female Drivers The following data represent the number of licensed drivers in various age groups and the number of accidents within the age group by gender.



Age Group	Number of Male Licensed Drivers (000s)	Number of Crashes Involving a Male (000s)	Number of Female Licensed Drivers (000s)	Number of Crashes Involving a Female (000s)
16	816	244	764	178
17	1,198	233	1,115	175
18	1,342	243	1,212	164
19	1,454	229	1,333	145
20–24	7,866	951	7,394	618
25–29	9,356	899	8,946	595
30–34	10,121	875	9,871	571
35–39	10,521	901	10,439	566
40–44	9,776	692	9,752	455
45–49	8,754	667	8,710	390
50–54	6,840	390	6,763	247
55–59	5,341	290	5,258	165
60–64	4,565	218	4,486	133
65–69	4,234	191	4,231	121
70–74	3,604	167	3,749	104
75–79	2,563	118	2,716	77
80–84	1,400	61	1,516	45
≥85	767	34	767	20

Source: National Highway and Traffic Safety Institute

- (a) Find the least-squares regression line for males, treating number of licensed drivers as the explanatory variable, x , and number of crashes, y , as the response variable. Repeat this procedure for females.
- (b) Interpret the slope of the least-squares regression line for each gender, if appropriate. How might an insurance company use this information?
- (c) Predict the number of accidents for males if there were 8700 thousand licensed drivers. Predict the number of accidents for females if there were 8700 thousand licensed drivers.
25. Mark Twain, in his book *Life on the Mississippi* (1884), makes the following observation:


Therefore, the Mississippi between Cairo and New Orleans was twelve hundred and fifteen miles long one hundred and seventy-six years ago. It was eleven hundred and eighty after the cut-off of 1722. It was one thousand and forty after the American Bend cut-off. It has lost sixty-seven miles since. Consequently its length is only nine hundred and seventy-three miles at present.

Now, if I wanted to be one of those ponderous scientific people, and “let on” to prove what had occurred in the remote past by what had occurred in a given time in the recent past, or what will occur in the far future by what has occurred in late years, what an opportunity is here! Geology never had such a chance, nor such exact data to argue from! Nor “development of species,” either! Glacial epochs are great things, but they are vague—vague. Please observe:

In the space of one hundred and seventy-six years the Lower Mississippi has shortened itself two hundred and

forty-two miles. That is an average of a trifle over one mile and a third per year. Therefore, any calm person, who is not blind or idiotic, can see that in the Old Oolitic Silurian Period, just a million years ago next November, the Lower Mississippi River was upwards of one million three hundred thousand miles long, and stuck out over the Gulf of Mexico like a fishing-rod. And by the same token any person can see that seven hundred and forty-two years from now the Lower Mississippi will be only a mile and three-quarters long, and Cairo and New Orleans will have joined their streets together, and be plodding comfortably along under a single mayor and a mutual board of aldermen. There is something fascinating about science. One gets such wholesale returns of conjecture out of such a trifling investment of fact.

Discuss how this relates to the material in this section.

26. **Regression Applet** Load the regression by eye applet.  Create a scatter diagram with twelve points and a positive linear association. Try to choose the points so that the correlation is about 0.7.

- (a) Draw a line that you believe describes the relation between the two variables well.
- (b) Now click the Show Least-squares Line on the applet. Compare the sum of squared residuals for the line that you draw to the sum of squared residuals for the least-squares regression line. Repeat parts (a) and (b) as often as you like. Does your eyeballed line ever coincide with the least-squares regression line?

Technology Step by Step

Determining the Least-Squares Regression Line

TI-83/84 Plus Use the same steps that were followed to obtain the correlation coefficient.

MINITAB **Step 1:** With the explanatory variable in C1 and the response variable in C2, select the **Stat** menu and highlight **Regression**. Highlight **Regression . . .**

Step 2: Select the explanatory (predictor) and response variables and click OK.

Excel **Step 1:** Be sure the Data Analysis Tool Pak is activated by selecting the **Tools** menu and highlighting **Add-Ins . . .**. Check the box for the Analysis ToolPak and select OK.

Step 2: Enter the explanatory variable in column A and the response variable in column B.

Step 3: Select the **Tools** menu and highlight **Data Analysis . . .**

Step 4: Select the **Regression** option.

Step 5: With the cursor in the Y-range cell, highlight the column that contains the response variable. With the cursor in the X-range cell, highlight the column that contains the explanatory variable. Press OK.

4.3 The Coefficient of Determination

Preparing for This Section Before getting started, review the following:

- Outliers (Section 3.4, pp. 155–156)

Objectives 1 Compute and interpret the coefficient of determination

In Section 4.2, we discussed the procedure for obtaining the least-squares regression line. In this section, we discuss another numerical measure of the strength of relation that exists between two quantitative variables.

1 Compute and Interpret the Coefficient of Determination

Consider the club-head speed versus distance data introduced in Section 4.1. If we were asked to predict the distance of a randomly selected shot, what would be a good guess? Our best guess might be the average distance of all shots taken. Since we don't know this value, we would use the average distance from the sample data given in Table 1, $\bar{y} = 266.75$ yards.

Now suppose we were told this particular shot resulted from a swing with a club-head speed of 103 mph. We could use the least-squares regression line to adjust our guess to $\hat{y} = 3.1661(103) - 55.7964 = 270.3$ yards. Knowing the linear relation that exists between club-head speed and distance allows us to improve our estimate of the distance of the shot. In statistical terms, we say that some of the variation in distance is explained by the linear relation between club-head speed and distance.

The percentage of variation in distance that is explained by the least-squares regression line is called the *coefficient of determination*.

Definition

The **coefficient of determination**, R^2 , measures the percentage of total variation in the response variable that is explained by the least-squares regression line.



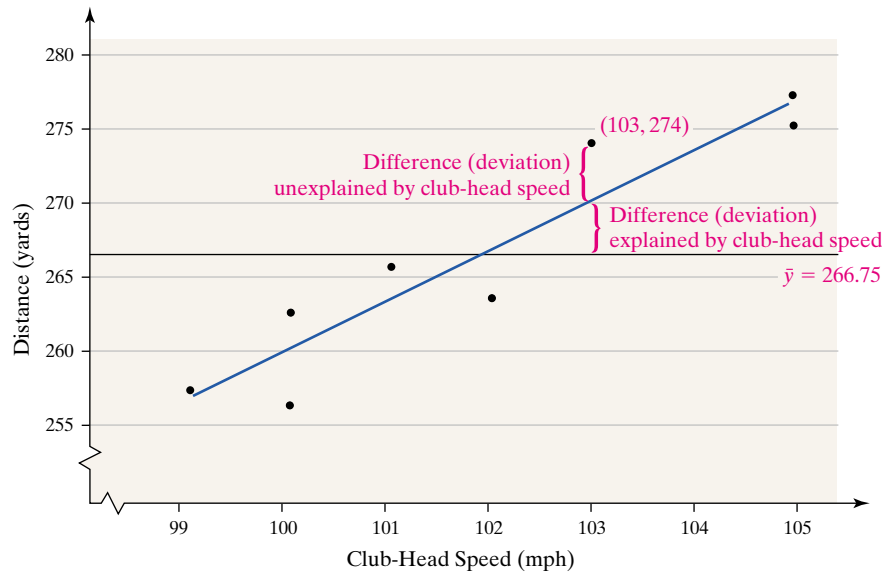
In Other Words

The coefficient of determination is a measure of how well the least-squares regression line describes the relation between the explanatory and response variable. The closer R^2 is to 1, the better the line describes how changes in the explanatory variable affect the value of the response variable.

The coefficient of determination is a number between 0 and 1, inclusive. That is, $0 \leq R^2 \leq 1$. If $R^2 = 0$, the least-squares regression line has no explanatory value. If $R^2 = 1$, the least-squares regression line explains 100% of the variation in the response variable.

Consider Figure 14, where a horizontal line is drawn at $\bar{y} = 266.75$. This value represents the predicted distance of a shot without any knowledge of club-head speed. Armed with the additional information that the club-head speed is 103 miles per hour, we increased our guess to 270.3 yards. The difference between the predicted distance of 266.75 yards and the predicted distance of 270.3 yards is due to the fact that the club-head-speed is 103 miles per hour. In other words, the difference between the prediction of $\hat{y} = 270.3$ and $\bar{y} = 266.75$ is explained by the linear relation between club-head speed and distance. The observed distance when club-head speed is 103 miles per hour is 274 yards (see Table 3 on page 196). The difference between our predicted value, $\hat{y} = 270.3$, and the actual value, $y = 274$, is due to factors (variables) other than the club-head speed and random error. The differences just discussed are called **deviations**.

Figure 14

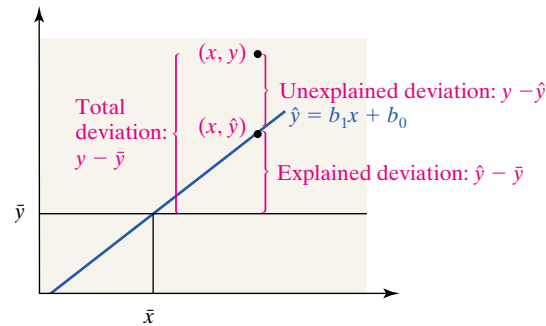


In Other Words

The word *deviations* comes from *deviate*. To *deviate* means “to stray”.

The deviation between the observed value of the response variable, y , and the mean value of the response variable, \bar{y} , is called the **total deviation**, so total deviation = $y - \bar{y}$. The deviation between the predicted value of the response variable, \hat{y} , and the mean value of the response variable, \bar{y} , is called the **explained deviation**, so explained deviation = $\hat{y} - \bar{y}$. Finally, the deviation between the observed value of the response variable, y , and the predicted value of the response variable, \hat{y} , is called the **unexplained deviation**, so unexplained deviation = $y - \hat{y}$. See Figure 15.

Figure 15



From the figure, it should be clear that

$$\text{Total deviation} = \text{unexplained deviation} + \text{explained deviation}$$

or

$$y - \bar{y} = (y - \hat{y}) + (\hat{y} - \bar{y})$$

Although beyond the scope of this text, it can be shown that

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

or

$$\text{Total variation} = \text{unexplained variation} + \text{explained variation}$$

Dividing both sides by total variation, we obtain

$$1 = \frac{\text{unexplained variation}}{\text{total variation}} + \frac{\text{explained variation}}{\text{total variation}}$$

Subtracting $\frac{\text{unexplained variation}}{\text{total variation}}$ from both sides, we obtain

$$R^2 = \frac{\text{explained variation}}{\text{total variation}} = 1 - \frac{\text{unexplained variation}}{\text{total variation}}$$

Unexplained variation is found by summing the squares of the residuals, $\sum \text{residuals}^2$. So the smaller the sum of squared residuals, the smaller the unexplained variation and, therefore, the larger R^2 will be. Therefore, the closer the observed y 's are to the regression line (the predicted y 's), the larger R^2 will be.

The coefficient of determination, R^2 , is the square of the linear correlation coefficient for the least-squares regression model. Written in symbols, $R^2 = (r)^2$.

EXAMPLE 1**Computing the Coefficient of Determination, R^2**

Problem: Compute and interpret the coefficient of determination, R^2 , for the club-head speed versus distance data shown in Table 2.

Approach: To compute R^2 , we square the linear correlation coefficient, r , found in Example 2 from Section 4.1 on page 182.

Solution: $R^2 = r^2 = 0.9387^2 = 0.8812 = 88.12\%$

Interpretation: 88.12% of the variation in distance is explained by the least-squares regression line, and 11.88% of the variation in distance is explained by other factors.

**CAUTION**

Squaring the linear correlation coefficient to obtain the coefficient of determination works only for the least-squares linear regression model

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

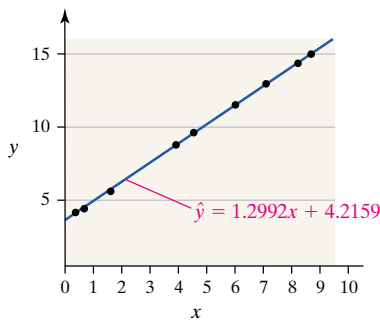
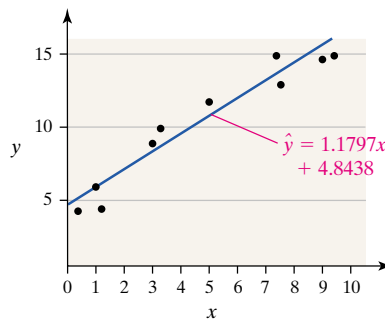
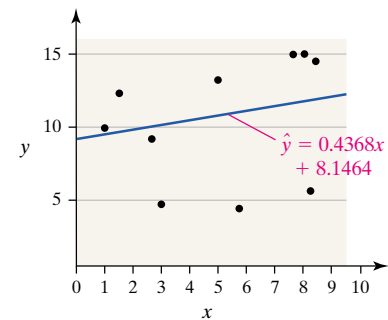
The method does not work in general.

To help reinforce the concept of the coefficient of determination, consider the three data sets in Table 5.

Table 5

Data Set A		Data Set B		Data Set C	
x	y	x	y	x	y
3.6	8.9	3.1	8.9	2.8	8.9
8.3	15.0	9.4	15.0	8.1	15.0
0.5	4.8	1.2	4.8	3.0	4.8
1.4	6.0	1.0	6.0	8.3	6.0
8.2	14.9	9.0	14.9	8.2	14.9
5.9	11.9	5.0	11.9	1.4	11.9
4.3	9.8	3.4	9.8	1.0	9.8
8.3	15.0	7.4	15.0	7.9	15.0
0.3	4.7	0.1	4.7	5.9	4.7
6.8	13.0	7.5	13.0	5.0	13.0

Figure 16(a) represents the scatter diagram of data set A, Figure 16(b) represents the scatter diagram of data set B, and Figure 16(c) represents the scatter diagram of data set C.

Figure 16**(a)****(b)****(c)**

Notice that the y -values in each of the three data sets are the same. The variance of y is 17.49. If we look at the scatter diagram in Figure 16(a), we notice that almost 100% of the variability in y can be explained by the least-squares regression

line, because the data almost lie perfectly on a straight line. In Figure 16(b), a high percentage of the variability in y can be explained by the least-squares regression line because the data have a strong linear relation. Higher x -values are associated with higher y -values. Finally, in Figure 16(c), a low percentage of the variability in y is explained by the least-squares regression line. If x increases, we cannot easily predict the change in y . If we compute the coefficient of determination, R^2 , for the three data sets in Table 5, we obtain the following results:

Coefficient of determination for Data Set A: 99.99%

Coefficient of determination for Data Set B: 94.7%

Coefficient of determination for Data Set C: 9.4%

Notice that, as the explanatory ability of the line decreases, so does the coefficient of determination, R^2 .

EXAMPLE 2

Determining the Coefficient of Determination Using Technology

Problem: Determine the coefficient of determination, R^2 , for the club-head speed versus distance data found in Example 2 from Section 4.1 using a statistical spreadsheet or graphing calculator with advanced statistical features.

Approach: We will use Excel to determine R^2 . The steps for obtaining the coefficient of determination using Excel, MINITAB, and the TI-83/84 Plus graphing calculators are given in the Technology Step by Step on page 215.

Result: Figure 17 shows the results obtained from Excel. The coefficient of determination, R^2 , is highlighted.

Figure 17

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.938695838
R Square	0.881149876
Adjusted R Square	0.861341522
Standard Error	2.882638465
Observations	8

Now Work Problems 3 and 5.

4.3 ASSESS YOUR UNDERSTANDING

Concepts and Vocabulary

- Suppose it is determined that $R^2 = 0.75$ when a linear regression is performed. Interpret this result.
- Explain what is meant by total deviation, explained deviation, and unexplained deviation.

Skill Building

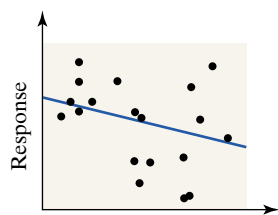
3. Match the coefficient of determination to the scatter diagram. The scales on the horizontal and vertical axis are the same for each scatter diagram.

(a) $R^2 = 0.58$

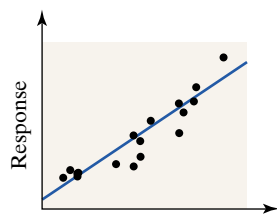
(b) $R^2 = 0.90$

(c) $R^2 = 1$

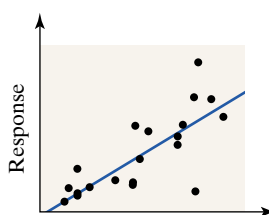
(d) $R^2 = 0.12$



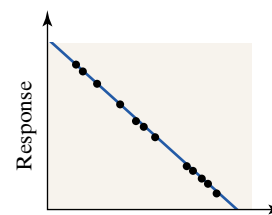
(I)



(II)



(III)



(IV)

4. Use the linear correlation coefficient given to determine the coefficient of determination, R^2 . Interpret each R^2 .

(a) $r = -0.32$

(b) $r = 0.13$

(c) $r = 0.40$

(d) $r = 0.93$

Applying the Concepts

5. The Other Old Faithful Perhaps you are familiar with the famous Old Faithful geyser in Yellowstone National Park. Another Old Faithful geyser is located in Calistoga in California's Napa Valley. The following data represent the time between eruptions and the length of eruption for 11 randomly selected eruptions.



Time between Eruptions, x	Length of Eruption, y	Time between Eruptions, x	Length of Eruption, y
12.17	1.88	11.70	1.82
11.63	1.77	12.27	1.93
12.03	1.83	11.60	1.77
12.15	1.83	11.72	1.83
11.30	1.70		

Source: Ladonna Hansen, Park Curator

The coefficient of determination is determined to be 83.0%. Interpret this result.

6. Concrete As concrete cures, it gains strength. The following data represent the 7-day and 28-day strength (in pounds per square inch) of a certain type of concrete.



7-Day Strength, x	28-Day Strength, y	7-Day Strength, x	28-Day Strength, y
2300	4070	2480	4120
3390	5220	3380	5020
2430	4640	2660	4890
2890	4620	2620	4190
3330	4850	3340	4630

The coefficient of determination, R^2 , is determined to be 57.5%. Interpret this result.

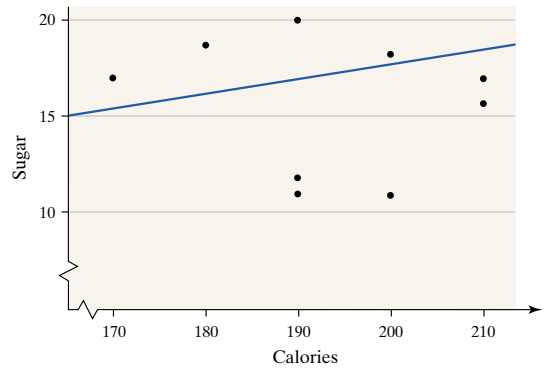
7. Calories versus Sugar The following data represent the number of calories per serving and the number of grams of sugar per serving for a random sample of high-fiber cereals.



Calories, x	Sugar, y	Calories, x	Sugar, y
200	18	210	23
210	23	210	16
170	17	210	17
190	20	190	12
200	18	190	11
180	19	200	11

Source: Consumer Reports

(a) A scatter diagram with the least-squares regression line is shown. The least-squares regression equation is $\hat{y} = 0.0821x + 0.93$. Do you think that calories and sugar content are linearly related? Why?



- (b) The coefficient of determination, R^2 , for these data is 6.8%. Interpret this result. Does this support your conclusion from part (a)? Why or why not?
- (c) Suppose that we add Kellogg's All-Bran cereal, which has 80 calories and 6 grams of sugar per serving, to the data set. Draw a scatter diagram of the data with this cereal included. The coefficient of determination, R^2 , with Kellogg's All-Bran cereal included, is 42.1%. Interpret this result. Why do you think that All-Bran cereal has such a large impact on the value of the coefficient of determination?

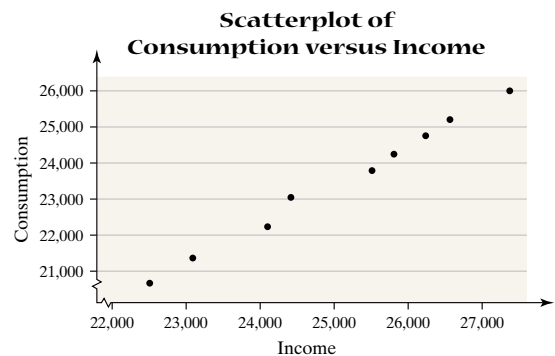
8. Consumption versus Income The following data represent the per capita disposable income (income after taxes) and per capita consumption in constant 2000 dollars in the United States for 1996–2004.



Year	Per Capita Disposable Income, x	Per Capita Consumption, y
1996	22,546	20,835
1997	23,065	21,365
1998	24,131	22,183
1999	24,564	23,050
2000	25,472	23,862
2001	25,698	24,216
2002	26,229	24,715
2003	26,570	25,270
2004	27,240	25,965

Source: Bureau of Economic Analysis

- (a) A scatter diagram is shown. Do you think that per capita disposable income and per capita consumption are linearly related? Why or why not?



- (b) The coefficient of determination, R^2 , for the data is 99.4%. Interpret this result. Does this support your conclusion from part (a)? Why?

Problems 9–12 use the results from Problems 23–26 in Section 4.1 and Problems 17–20 in Section 4.2.

- 9. Height versus Head Circumference** Use the results from **NW** Problem 23 in Section 4.1 and Problem 17 in Section 4.2 to
- compute the coefficient of determination, R^2 .
 - interpret the coefficient of determination.
- 10. Gestation Period versus Life Expectancy** Use the results from Problem 24 in Section 4.1 and Problem 18 in Section 4.2 to
- compute the coefficient of determination, R^2 .
 - interpret the coefficient of determination.
- 11. Weight of a Car versus Miles per Gallon** Use the results from Problem 25 in Section 4.1 and Problem 19 in Section 4.2 to
- compute the coefficient of determination, R^2 .
 - interpret the coefficient of determination.
- 12. Bone Length** Use the results from Problem 26 in Section 4.1 and Problem 20 in Section 4.2 to
- compute the coefficient of determination, R^2 .
 - interpret the coefficient of determination.
- 13. Weight of a Car versus Miles per Gallon** Suppose we add the Dodge Viper to the data in Problem 19 in Section 4.2. A Dodge Viper weighs 3425 pounds and gets 11 miles per gallon. Compute the coefficient of determination of the expanded data set. What effect does the addition of the Viper to the data set have on R^2 ?
- 14. Gestation Period versus Life Expectancy** Suppose we add humans to the data in Problem 18 in Section 4.2. Humans have a gestation period of 268 days and a life expectancy of 76.5 years. Compute the coefficient of determination of the expanded data set. What effect does the addition of humans to the data set have on R^2 ?

Consumer Reports® Fit to Drink

The taste, color, and clarity of the water coming out of home faucets have long concerned consumers. Recent reports of lead and parasite contamination have made unappetizing water a health, as well as an esthetic, concern. Water companies are struggling to contain cryptosporidium, a parasite that has caused outbreaks of illness that may be fatal to people with a weakened immune system. Even chlorination, which has rid drinking water of infectious organisms that once killed people by the thousands, is under suspicion as an indirect cause of miscarriages and cancer.

Concerns about water quality and taste have made home filtering increasingly popular. To find out how well they work, technicians at Consumer Reports tested 14 models to determine how well they filtered contaminants and whether they could improve the taste of our cabbage-soup testing mixture.

To test chloroform and lead removal, we added concentrated amounts of both to our water, along with calcium nitrate to increase water hardness. Every few days we analyzed the water to measure chloroform and lead content. The following table contains the lead measurements for one of the models tested.

No. Gallons Processed	% Lead Removed
25	85
26	87
73	86
75	88
123	90
126	87
175	92
177	94

- Construct a scatter diagram of the data using % Lead Removed as the response variable.
- Does the relationship between No. Gallons Processed and % Lead Removed appear to be linear? If not, describe the relationship.
- Calculate the linear correlation coefficient between No. Gallons Processed and % Lead Removed. Based on the scatter diagram constructed in part (a) and your answer to part (b), is this measure useful? What is R^2 ? Interpret R^2 .
- Fit a linear regression model to these data.
- Using statistical software or a graphing calculator with advanced statistical features, fit a quadratic model to these data. What is R^2 ? Which model appears to fit the data better?
- Given the nature of the variables being measured, describe the type of curve you would expect to show the true relationship between these variables (linear, quadratic, exponential, S-shaped). Support your position.

Note to Readers: In many cases, our test protocol and analytical methods are more complicated than described in these examples. The data and discussions have been modified to make the material more appropriate for the audience.

© by Consumers Union of U.S., Inc., Yonkers, NY 10703-1057, a nonprofit organization. Reprinted with permission.

Technology Step by Step**Determining R^2**

TI-83/84 Plus	Use the same steps that were followed to obtain the correlation coefficient to obtain R^2 . Diagnostics must be on.
MINITAB	This is provided in the standard regression output.
Excel	This is provided in the standard regression output.

CHAPTER 4 Review

Summary

In this chapter, we introduced techniques that allow us to describe the relation between two quantitative variables. The first step in identifying the type of relation that might exist is to draw a scatter diagram. The explanatory variable is plotted on the horizontal axis and the corresponding response variable on the vertical axis. The scatter diagram can be used to discover whether the relation between the explanatory and the response variables is linear. In addition, for linear relations, we can judge whether the linear relation shows positive or negative association.

A numerical measure for the strength of linear relation between two quantitative variables is the linear correlation coefficient. It is a number between -1 and 1 , inclusive. Values of the correlation coefficient near -1 are indicative of a negative linear relation between the two variables. Values of the correlation coefficient near $+1$ indicate a positive linear relation between the two variables. If the correlation coefficient is near 0 , then there is little *linear* relation between the two variables.

Once a linear relation between the two variables has been discovered, we describe the relation by finding the least-squares regression line. This line best describes the linear relation between the explanatory and the response variables. We can use the least-squares regression line to predict a value of the response variable for a given value of the explanatory variable.

The coefficient of determination, R^2 , measures the percent of variation in the response variable that is explained by the least-squares regression line. It is a measure between 0 and 1 inclusive. The closer R^2 is to 1 , the more explanatory value the line has.

One item worth mentioning again is that a researcher should never claim causation between two variables in a study unless the data are experimental. Observational data allow us to say that two variables might be associated, but we cannot claim causation.

Formulas

Correlation Coefficient

$$r = \frac{\sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)}{n - 1}$$

Coefficient of Determination, R^2

$$\begin{aligned} R^2 &= \frac{\text{variation explained by explanatory variable}}{\text{total variation}} \\ &= 1 - \frac{\text{unexplained variation}}{\text{total variation}} \\ &= r^2 \text{ for the least-squares regression model } \hat{y} = b_1x + b_0 \end{aligned}$$

Equation of the Least-Squares Regression Line

The equation of the least-squares regression line is given by

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

where

\hat{y} is the predicted value of the response variable,

$b_1 = r \cdot \frac{s_y}{s_x}$ is the slope of the least-squares regression line, and

$b_0 = \bar{y} - b_1\bar{x}$ is the intercept of the least-squares regression line.

Vocabulary

Bivariate data (p. 176)
 Response variable (p. 177)
 Explanatory variable (p. 177)
 Predictor variable (p. 177)
 Lurking variable (p. 177)
 Scatter diagram (p. 177)
 Positively associated (p. 179)

Negatively associated (p. 179)
 Linear correlation coefficient (p. 180)
 Correlation matrix (p. 183)
 Residuals (p. 197)
 Least-squares regression line (p. 198)
 Slope (p. 198)
 y-intercept (p. 198)

Outside the scope of the model (p. 201)
 Coefficient of determination (p. 209)
 Total deviation (p. 210)
 Explained deviation (p. 210)
 Unexplained deviation (p. 210)

Objectives

Section	You should be able to . . .	Example	Review Exercises
4.1	1 Draw and interpret scatter diagrams (p. 177)	1, 3	1(a)–4(a), 9(a), 10(a), 15(a)
	2 Understand the properties of the linear correlation coefficient (p. 179)		18
	3 Compute and interpret the linear correlation coefficient (p. 182)	2, 3	1(b)–4(b), 15(b)
	4 Determine whether there is a linear relation between two variables (p. 185)	4	1(c)–4(c)
4.2	1 Find the least-squares regression line and use the line to make predictions (p. 197)	2, 3	5(a)–8(a), 9(d)–10(d)
	2 Interpret the slope and y -intercept of the least-squares regression line (p. 201)	Page 218	5(c)–8(c)
	3 Compute the sum of squared residuals (p. 202)	4	9(g), 10(g)
4.3	1 Compute and interpret the coefficient of determination (p. 209)	1, 2	11–14

Review Exercises

- 1. Engine Displacement versus Fuel Economy** The following data represent the size of a car's engine (in liters) versus its miles per gallon in the city for various 2005 domestic automobiles.



Car	Engine Displacement (liters), x	City MPG, y	Car	Engine Displacement (liters), x	City MPG, y
Buick Century	3.1	20	Ford Crown Victoria	4.6	18
Buick LeSabre	3.8	20	Ford Focus	2.0	26
Cadillac DeVille	4.6	18	Ford Mustang	3.8	20
Chevrolet Cavalier	2.2	25	Mercury Sable	3.0	19
Chevrolet Impala	3.8	21	Pontiac Grand Am	3.4	20
Chevrolet Malibu	2.2	24	Pontiac Sunfire	2.2	24
Chrysler Sebring Sedan	2.7	22	Saturn Ion	2.2	26
Dodge Magnum	3.5	21			

Source: www.roadandtrack.com

- (a) Draw a scatter diagram treating engine displacement as the explanatory variable and miles per gallon as the response variable.
- (b) Compute the linear correlation coefficient between engine displacement and miles per gallon.
- (c) Based on the scatter diagram and the linear correlation coefficient, comment on the type of relation that appears to exist between the two variables.

- 2. Temperature versus Cricket Chirps** Crickets make a chirping noise by sliding their wings rapidly over each other. Perhaps you have noticed that the number of chirps seems to increase with the temperature. The following data list the temperature (in degrees Fahrenheit) and the number of chirps per second for the striped ground cricket.



Temperature, x	Chirps per Second, y	Temperature, x	Chirps per Second, y
88.6	20.0	71.6	16.0
93.3	19.8	84.3	18.4
80.6	17.1	75.2	15.5
69.7	14.7	82.0	17.1
69.4	15.4	83.3	16.2
79.6	15.0	82.6	17.2
80.6	16.0	83.5	17.0
76.3	14.4		

Source: Pierce, George W. *The Songs of Insects*. Cambridge, MA: Harvard University Press, 1949, pp. 12–21

- (a) Draw a scatter diagram treating temperature as the explanatory variable and chirps per second as the response variable.
- (b) Compute the linear correlation coefficient between temperature and chirps per second.
- (c) Based upon the scatter diagram and the linear correlation coefficient, comment on the type of relation that appears to exist between the two variables.

- 3. Apartments** The following data represent the square footage and rents for apartments in the Borough of Queens and Nassau County, New York.



Queens (New York City)		Nassau County (Long Island)	
Square Footage, x	Rent Per Month, y	Square Footage, x	Rent Per Month, y
500	650	1100	1875
588	1215	588	1075
1000	2000	1250	1775
688	1655	556	1050
825	1250	825	1300
460	1805	743	1475
1259	2700	660	1315
650	1200	975	1400
560	1250	1429	1900
1073	2350	800	1650
1452	3300	1906	4625
1305	3100	1077	1395

Source: apartments.com

- On the same graph, draw a scatter diagram for both Queens and Nassau County apartments, treating square footage as the explanatory variable. Use a different plotting symbol for each group.
- Compute the linear correlation coefficient between square footage and rent for each location.
- Given the scatter diagram and the linear correlation coefficient, comment on the type of relation that appears to exist between the two variables for each group.
- Does location appear to be a factor in rent?

- 4. Boys versus Girls** The following data represent the height (in inches) of boys and girls between the ages of 2 and 10 years.



Age	Boy Height, x	Girl Height, y	Age	Boy Height, x	Girl Height, y
2	36.1	39.0	6	49.8	43.7
2	34.2	38.6	7	43.2	50.5
2	31.1	33.6	7	47.9	47.7
3	36.3	41.3	8	51.4	44.0
3	39.5	40.9	8	48.3	62.1
4	41.5	43.2	8	50.9	44.8
4	38.6	39.8	9	52.2	50.9
5	45.6	50.5	9	51.3	55.6
5	44.8	38.3	10	55.6	61.4
5	44.6	43.9	10	59.5	50.8

Source: National Center for Health Statistics

- On the same graph, draw a scatter diagram for both boys and girls, treating age as the explanatory variable. Use a different plotting symbol for each gender.
 - Compute the linear correlation coefficient between age and height for each gender.
 - Based on the scatter diagram and the linear correlation coefficient, comment on the type of relation that appears to exist between the age and height for each gender.
 - Does gender appear to be a factor in determining height?
- 5.** Using the data and results from Problem 1, do the following:
- Find the least-squares regression line, treating engine displacement as the explanatory variable.
 - Draw the least-squares regression line on the scatter diagram.
 - Interpret the slope and y -intercept, if appropriate.
 - Predict the miles per gallon of a Ford Mustang whose engine displacement is 3.8 liters.
 - Compute the residual of the prediction found in part (d).
 - Is the miles per gallon above or below average for a Ford Mustang?
- 6.** Using the data and results from Problem 2, do the following:
- Find the least-squares regression line, treating temperature as the explanatory variable and chirps per second as the response variable.
 - Draw the least-squares regression line on the scatter diagram.
 - Interpret the slope and y -intercept, if appropriate.
 - Predict the chirps per second if it is 83.3°F .
 - Compute the residual of the prediction found in part (d).
 - Were chirps per second above or below average at 83.3°F ?
- 7.** Using the Queens data and results from Problem 3, do the following:
- Find the least-squares regression line, treating square footage as the explanatory variable.
 - Draw the least-squares regression line on the scatter diagram.
 - Interpret the slope and y -intercept, if appropriate.
 - Predict the rent of an 825-square-foot apartment.
 - Compute the residual of the prediction found in part (d).
 - Is this apartment's rent above or below average?
- 8.** Using the Boy Height data and results from Problem 4, do the following:
- Find the least-squares regression line, treating age as the explanatory variable and height as the response variable.
 - Draw the least-squares regression line on the scatter diagram.
 - Interpret the slope and y -intercept, if appropriate.
 - Predict the height of a 6-year-old boy.
 - Compute the residual of the prediction found in part (d).
 - Is this boy's height above or below average?

In Problems 9 and 10, do the following:

- (a) Draw a scatter diagram treating x as the explanatory variable and y as the response variable.
- (b) Select two points from the scatter diagram, and find the equation of the line containing the points selected.
- (c) Graph the line found in part (b) on the scatter diagram.
- (d) Determine the least-squares regression line.
- (e) Graph the least-squares regression line on the scatter diagram.
- (f) Compute the sum of the squared residuals for the line found in part (b).
- (g) Compute the sum of the squared residuals for the least-squares regression line found in part (d).
- (h) Comment on the fit of the line found in part (b) versus the least-squares regression line found in part (d).

9.

x	3	4	6	7	9
y	2.1	4.2	7.2	8.1	10.6

10.

x	10	14	17	18	21
y	105	94	82	76	63

- 11. Use the results from Problems 1 and 5 to compute and interpret R^2 .
- 12. Use the results from Problems 2 and 6 to compute and interpret R^2 .
- 13. Use Queens data and the results from Problems 3 and 7 to compute and interpret R^2 .
- 14. Use the results from Problems 4 and 8 to compute and interpret R^2 .
- 15. **200-Meter Dash** The following data represent the gold medal times, in seconds, for men and women in the 200-meter dash at the summer Olympics from 1948 to 2004.



Year	Time (Men)	Time (Women)	Year	Time (Men)	Time (Women)
1948	21.10	24.40	1980	20.19	22.03
1952	20.70	23.70	1984	19.80	21.81
1956	20.60	23.40	1988	19.75	21.34
1960	20.50	24.00	1992	20.01	21.81
1964	20.30	23.00	1996	19.32	22.12
1968	19.80	22.50	2000	20.09	21.84
1972	20.00	22.40	2004	19.79	22.05
1976	20.23	22.37			

Source: www.factmonster.com

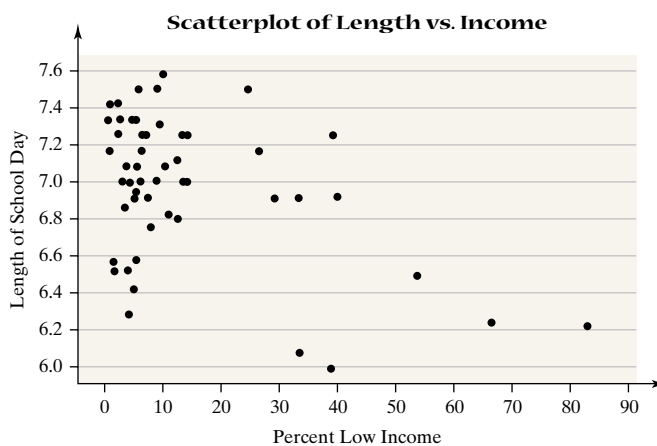
- (a) Draw a scatter diagram of the data using time for men as the explanatory variable and time for women as the response variable.
- (b) Compute the correlation coefficient for the data.
- (c) Based on your results from parts (a) and (b), what type of relation appears to exist between the gold medal time for men and the gold medal time for women in the 200-meter dash? Do you believe that the gold medal time for men causes the gold medal time for women?

16. **Wine and Your Heart** The health benefits of moderate wine consumption are well documented. Researchers wanted to determine if alcohol consumption is positively related to heart-rate variability (HRV) in women with coronary heart disease (CHD). The purpose of the study was to shed some doubt on the heart-health benefits of wine. The researchers

surveyed female patients who have recently been released from the hospital after successful heart procedures such as bypass surgery or angioplasty. A questionnaire evaluated self-reported consumption of individual alcoholic beverage types: beer, wine, and spirits. Other characteristics, such as age, body mass index, smoking habits, history of diabetes, menopausal status, and educational status, were also assessed. The researchers found that wine intake was associated with increased HRV. Based on this study, can we conclude that increased wine consumption in women with recent heart procedures causes an increase in heart-rate variability? Why?

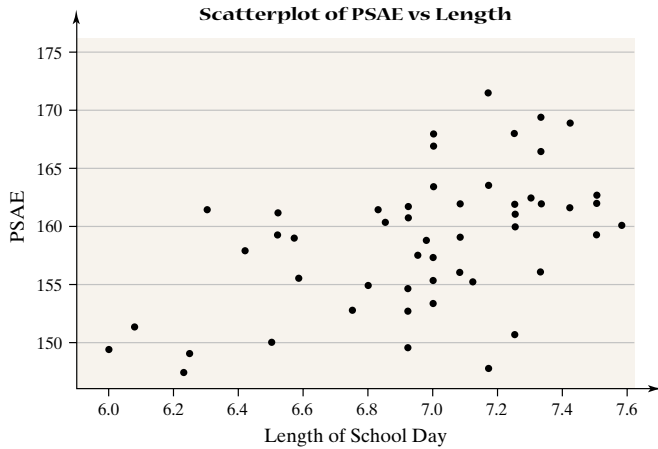
17. **Analyzing a Newspaper Article** In a newspaper article written in the *Chicago Tribune* on September 29, 2002, it was claimed that poorer school districts have shorter school days.

- (a) The following scatter diagram was drawn using the data supplied in the article. In this scatter diagram, the response variable is length of the school day and the explanatory variable is percent of the population that is low income. The correlation between length and income is -0.461 . Do you think that the scatter diagram and correlation coefficient support the position of the article?

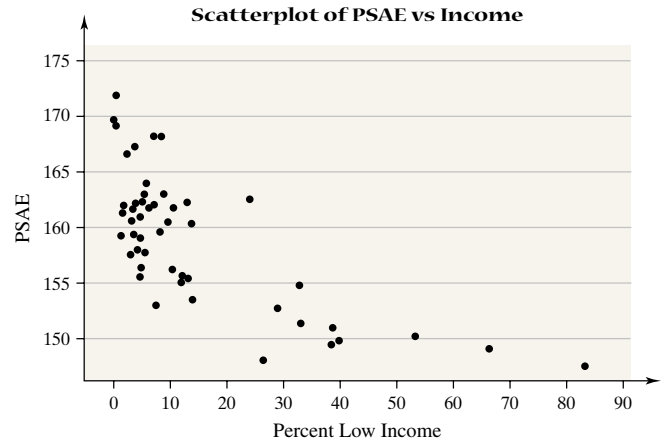


- (b) The least-squares regression line between length, y , and income, x , is $\hat{y} = -0.0102x + 7.11$. Interpret the slope of this regression line. Does it make sense to interpret the intercept? If so, interpret the intercept.
- (c) Predict the length of the school day for a district in which 20% of the population is low income by letting $x = 20$.

(d) This same article included average Prairie State Achievement Examination (PSAE) scores for each district. The article implied that shorter school days result in lower PSAE scores. The correlation between PSAE score and length of school day is 0.517. A scatter diagram treating PSAE as the response variable is shown below. Do you believe that a longer school day is positively associated with a higher PSAE score?



(e) The correlation between percentage of the population that is low income and PSAE score is -0.720 . A scatter diagram treating PSAE score as the response variable is shown below.



Do you believe that percentage of the population that is low income is negatively associated with PSAE score?

(f) Can you think of any lurking variables that are playing a role in this study?

18. List the seven properties of the linear correlation coefficient.

THE CHAPTER 4 CASE STUDY IS LOCATED ON THE CD THAT ACCOMPANIES THIS TEXT.

