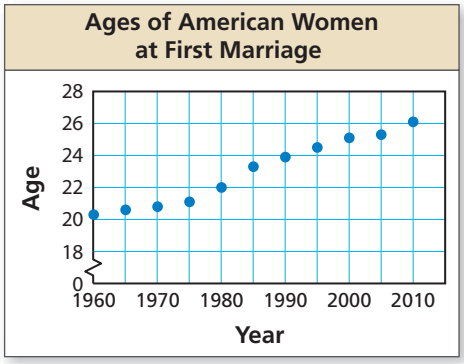
4.5 Analyzing Lines of Fit

Essential Question How can you *analytically* find a line of best fit for a scatter plot?

EXPLORATION 1 Finding a Line of Best Fit

Work with a partner.
The scatter plot shows the median ages of American women at their first marriage for selected years from 1960 through 2010. In Exploration 2 in Section 4.4, you approximated a line of fit graphically. To find the line of *best* fit, you can use a computer, spreadsheet, or graphing calculator that has a *linear regression* feature.



a. The data from the scatter plot is shown in the table. Note that 0, 5, 10, and so on represent the numbers of years since 1960. What does the ordered pair (25, 23.3) represent?

L1	L2	L3
0	20.3	
5	20.6	
10	20.8	
15	21.1	
20	22	
25	23.3	
30	23.9	
35	24.5	
40	25.1	
45	25.3	
50	26.1	

L1(55)=

b. Use the *linear regression* feature to find an equation of the line of best fit. You should obtain results such as those shown below.

```
LinReg
 y=ax+b
 a=.1261818182
 b=19.84545455
 r²=.9738676804
 r=.986847344
```

c. Write an equation of the line of best fit. Compare your result with the equation you obtained in Exploration 2 in Section 4.4.

CONSTRUCTING VIABLE ARGUMENTS
To be proficient in math, you need to reason inductively about data.

Communicate Your Answer

2. How can you *analytically* find a line of best fit for a scatter plot?

3. The data set relates the number of chirps per second for striped ground crickets and the outside temperature in degrees Fahrenheit. Make a scatter plot of the data. Then find an equation of the line of best fit. Use your result to estimate the outside temperature when there are 19 chirps per second.

Chirps per second	20.0	16.0	19.8	18.4	17.1
Temperature (°F)	88.6	71.6	93.3	84.3	80.6

Chirps per second	14.7	15.4	16.2	15.0	14.4
Temperature (°F)	69.7	69.4	83.3	79.6	76.3

4.5 Lesson

What You Will Learn

- Use residuals to determine how well lines of fit model data.
- Use technology to find lines of best fit.
- Distinguish between correlation and causation.

Core Vocabulary

residual, *p. 202*
linear regression, *p. 203*
line of best fit, *p. 203*
correlation coefficient, *p. 203*
interpolation, *p. 205*
extrapolation, *p. 205*
causation, *p. 205*

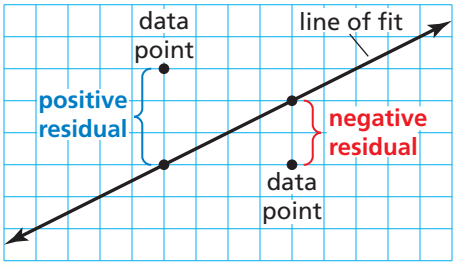
Analyzing Residuals

One way to determine how well a line of fit models a data set is to analyze *residuals*.

Core Concept

Residuals

A **residual** is the difference of the y-value of a data point and the corresponding y-value found using the line of fit. A residual can be positive, negative, or zero.



A scatter plot of the residuals shows how well a model fits a data set. If the model is a good fit, then the absolute values of the residuals are relatively small, and the residual points will be more or less evenly dispersed about the horizontal axis. If the model is not a good fit, then the residual points will form some type of pattern that suggests the data are not linear. Wildly scattered residual points suggest that the data might have no correlation.

EXAMPLE 1 Using Residuals

Week, x	Sales (millions), y
1	\$19
2	\$15
3	\$13
4	\$11
5	\$10
6	\$8
7	\$7
8	\$5

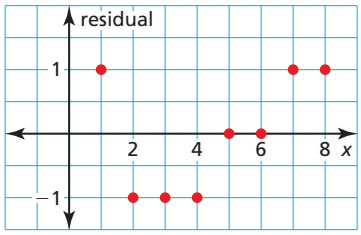
In Example 3 in Section 4.4, the equation $y = -2x + 20$ models the data in the table shown. Is the model a good fit?

SOLUTION

Step 1 Calculate the residuals. Organize your results in a table.

Step 2 Use the points (x, residual) to make a scatter plot.

x	y	y-Value from model	Residual
1	19	18	$19 - 18 = 1$
2	15	16	$15 - 16 = -1$
3	13	14	$13 - 14 = -1$
4	11	12	$11 - 12 = -1$
5	10	10	$10 - 10 = 0$
6	8	8	$8 - 8 = 0$
7	7	6	$7 - 6 = 1$
8	5	4	$5 - 4 = 1$



- The points are evenly dispersed about the horizontal axis. So, the equation $y = -2x + 20$ is a good fit.

EXAMPLE 2 Using Residuals

The table shows the ages x and salaries y (in thousands of dollars) of eight employees at a company. The equation $y = 0.2x + 38$ models the data. Is the model a good fit?

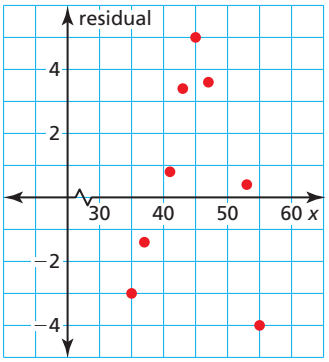
Age, x	35	37	41	43	45	47	53	55
Salary, y	42	44	47	50	52	51	49	45

SOLUTION

Step 1 Calculate the residuals. Organize your results in a table.

Step 2 Use the points (x, residual) to make a scatter plot.

x	y	y-Value from model	Residual
35	42	45.0	$42 - 45.0 = -3.0$
37	44	45.4	$44 - 45.4 = -1.4$
41	47	46.2	$47 - 46.2 = 0.8$
43	50	46.6	$50 - 46.6 = 3.4$
45	52	47.0	$52 - 47.0 = 5.0$
47	51	47.4	$51 - 47.4 = 3.6$
53	49	48.6	$49 - 48.6 = 0.4$
55	45	49.0	$45 - 49.0 = -4.0$



The residual points form a ∩-shaped pattern, which suggests the data are not linear. So, the equation $y = 0.2x + 38$ does not model the data well.

Monitoring Progress

Help in English and Spanish at *BigIdeasMath.com*

1. The table shows the attendances y (in thousands) at an amusement park from 2005 to 2014, where $x = 0$ represents the year 2005. The equation $y = -9.8x + 850$ models the data. Is the model a good fit?

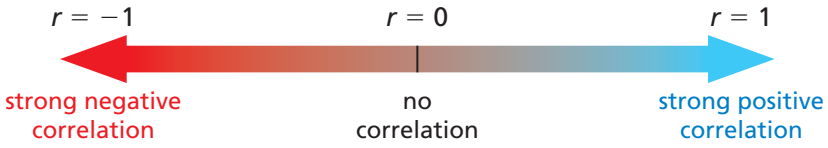
Year, x	0	1	2	3	4	5	6	7	8	9
Attendance, y	850	845	828	798	800	792	785	781	775	760

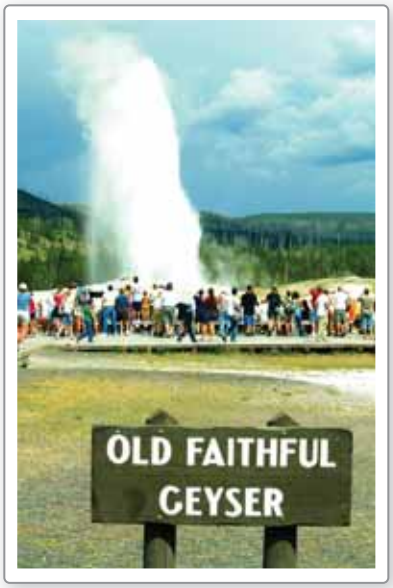
STUDY TIP

You know how to use two points to find an equation of a line of fit. When finding an equation of the line of best fit, every point in the data set is used.

Finding Lines of Best Fit

Graphing calculators use a method called **linear regression** to find a precise line of fit called a **line of best fit**. This line best models a set of data. A calculator often gives a value r, called the **correlation coefficient**. This value tells whether the correlation is positive or negative and how closely the equation models the data. Values of r range from -1 to 1. When r is close to 1 or -1, there is a strong correlation between the variables. As r, gets closer to 0, the correlation becomes weaker.





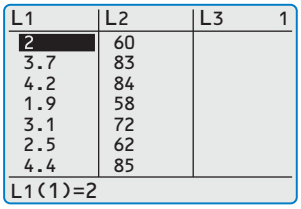
EXAMPLE 3 Finding a Line of Best Fit Using Technology

The table shows the durations x (in minutes) of several eruptions of the geyser Old Faithful and the times y (in minutes) until the next eruption. (a) Use a graphing calculator to find an equation of the line of best fit. Then plot the data and graph the equation in the same viewing window. (b) Identify and interpret the correlation coefficient. (c) Interpret the slope and y-intercept of the line of best fit.

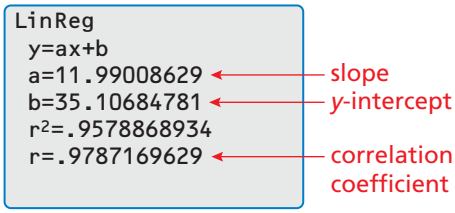
Duration, x	2.0	3.7	4.2	1.9	3.1	2.5	4.4	3.9
Time, y	60	83	84	58	72	62	85	85

SOLUTION

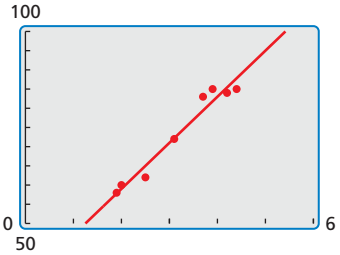
a. Step 1 Enter the data from the table into two lists.



Step 2 Use the *linear regression* feature. The values in the equation can be rounded to obtain $y = 12.0x + 35$.



Step 3 Enter the equation $y = 12.0x + 35$ into the calculator. Then plot the data and graph the equation in the same viewing window.



b. The correlation coefficient is about 0.979. This means that the relationship between the durations and the times until the next eruption has a strong positive correlation and the equation closely models the data, as shown in the graph.

c. The slope of the line is 12. This means the time until the next eruption increases by about 12 minutes for each minute the duration increases. The y-intercept is 35, but it has no meaning in this context because the duration cannot be 0 minutes.

PRECISION

Be sure to analyze the data values to select an appropriate viewing window for your graph.

Monitoring Progress

Help in English and Spanish at *BigIdeasMath.com*

2. Use the data in Monitoring Progress Question 1. (a) Use a graphing calculator to find an equation of the line of best fit. Then plot the data and graph the equation in the same viewing window. (b) Identify and interpret the correlation coefficient. (c) Interpret the slope and y-intercept of the line of best fit.

Using a graph or its equation to *approximate* a value between two known values is called **interpolation**. Using a graph or its equation to *predict* a value outside the range of known values is called **extrapolation**. In general, the farther removed a value is from the known values, the less confidence you can have in the accuracy of the prediction.

STUDY TIP

To approximate or predict an unknown value, you can evaluate the model algebraically or graph the model with a graphing calculator and use the *trace* feature.

EXAMPLE 4 Interpolating and Extrapolating Data

Refer to Example 3. Use the equation of the line of best fit.

a. Approximate the duration before a time of 77 minutes.

b. Predict the time after an eruption lasting 5.0 minutes.

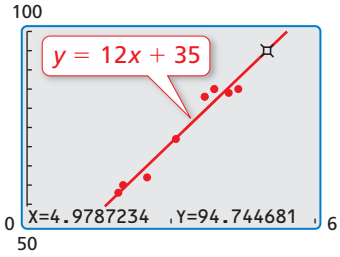
SOLUTION

a. $y = 12.0x + 35$ Write the equation.

$77 = 12.0x + 35$ Substitute 77 for y.

$3.5 = x$ Solve for x.

▶ An eruption lasts about 3.5 minutes before a time of 77 minutes.



b. Use a graphing calculator to graph the equation. Use the *trace* feature to find the value of y when $x \approx 5.0$, as shown.

▶ A time of about 95 minutes will follow an eruption of 5.0 minutes.

Monitoring Progress Help in English and Spanish at *BigIdeasMath.com*

3. Refer to Monitoring Progress Question 2. Use the equation of the line of best fit to predict the attendance at the amusement park in 2017.

Correlation and Causation

When a change in one variable causes a change in another variable, it is called **causation**. Causation produces a strong correlation between the two variables. The converse is *not* true. In other words, correlation does not imply causation.

READING

A causal relationship exists when one variable causes a change in another variable.

EXAMPLE 5 Identifying Correlation and Causation

Tell whether a correlation is likely in the situation. If so, tell whether there is a causal relationship. Explain your reasoning.

a. time spent exercising and the number of calories burned

b. the number of banks and the population of a city

SOLUTION

a. There is a positive correlation and a causal relationship because the more time you spend exercising, the more calories you burn.

b. There may be a positive correlation but no causal relationship. Building more banks will not cause the population to increase.

Monitoring Progress Help in English and Spanish at *BigIdeasMath.com*

4. Is there a correlation between time spent playing video games and grade point average? If so, is there a causal relationship? Explain your reasoning.

4.5 Exercises

Dynamic Solutions available at *BigIdeasMath.com*

Vocabulary and Core Concept Check

1. **VOCABULARY** When is a residual positive? When is it negative?
2. **WRITING** Explain how you can use residuals to determine how well a line of fit models a data set.
3. **VOCABULARY** Compare interpolation and extrapolation.
4. **WHICH ONE DOESN'T BELONG?** Which correlation coefficient does *not* belong with the other three? Explain your reasoning.

$r = -0.98$ $r = 0.96$ $r = -0.09$ $r = 0.97$

Monitoring Progress and Modeling with Mathematics

In Exercises 5–8, use residuals to determine whether the model is a good fit for the data in the table. Explain. *(See Examples 1 and 2.)*

5. $y = 4x - 5$

x	−4	−3	−2	−1	0	1	2	3	4
y	−18	−13	−10	−7	−2	0	6	10	15

6. $y = 6x + 4$

x	1	2	3	4	5	6	7	8	9
y	13	14	23	26	31	42	45	52	62

7. $y = -1.3x + 1$

x	−8	−6	−4	−2	0	2	4	6	8
y	9	10	5	8	−1	1	−4	−12	−7

8. $y = -0.5x - 2$

x	4	6	8	10	12	14	16	18	20
y	−1	−3	−6	−8	−10	−10	−10	−9	−9

9. **ANALYZING RESIDUALS** The table shows the growth y (in inches) of an elk's antlers during week x. The equation $y = -0.7x + 6.8$ models the data. Is the model a good fit? Explain.

Week, x	1	2	3	4	5
Growth, y	6.0	5.5	4.7	3.9	3.3

10. **ANALYZING RESIDUALS** The table shows the approximate numbers y (in thousands) of movie tickets sold from January to June for a theater. In the table, $x = 1$ represents January. The equation $y = 1.3x + 27$ models the data. Is the model a good fit? Explain.

Month, x	Ticket sales, y
1	27
2	28
3	36
4	28
5	32
6	35

In Exercises 11–14, use a graphing calculator to find an equation of the line of best fit for the data. Identify and interpret the correlation coefficient.

11.

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

12.

x	−4	−2	0	2	4	6	8	10
y	17	7	8	1	5	−2	2	−8

13.

x	−15	−10	−5	0	5	10	15	20
y	−4	2	7	16	22	30	37	43

14.

x	5	6	7	8	9	10	11	12
y	12	−2	8	3	−1	−4	6	0

ERROR ANALYSIS **In Exercises 15 and 16, describe and correct the error in interpreting the graphing calculator display.**

```
LinReg
 y=ax+b
 a=-4.47
 b=23.16
 r²=.9989451055
 r=-.9994724136
```

15. ✗ An equation of the line of best fit is $y = 23.16x - 4.47$.

16. ✗ The data have a strong positive correlation.

17. **MODELING WITH MATHEMATICS** The table shows the total numbers y of people who reported an earthquake x minutes after it ended. *(See Example 3.)*

Minutes, x	People, y
1	10
2	100
3	400
4	900
5	1400
6	1800
7	2100

a. Use a graphing calculator to find an equation of the line of best fit. Then plot the data and graph the equation in the same viewing window.

b. Identify and interpret the correlation coefficient.

c. Interpret the slope and y-intercept of the line of best fit.

18. **MODELING WITH MATHEMATICS** The table shows the numbers y of people who volunteer at an animal shelter on each day x.

Day, x	1	2	3	4	5	6	7	8
People, y	9	5	13	11	10	11	19	12

a. Use a graphing calculator to find an equation of the line of best fit. Then plot the data and graph the equation in the same viewing window.

b. Identify and interpret the correlation coefficient.

c. Interpret the slope and y-intercept of the line of best fit.

19. **MODELING WITH MATHEMATICS** The table shows the mileages x (in thousands of miles) and the selling prices y (in thousands of dollars) of several used automobiles of the same year and model. *(See Example 4.)*

Mileage, x	22	14	18	30	8	24
Price, y	16	17	17	14	18	15

a. Use a graphing calculator to find an equation of the line of best fit.

b. Identify and interpret the correlation coefficient.

c. Interpret the slope and y-intercept of the line of best fit.

d. Approximate the mileage of an automobile that costs $15,500.

e. Predict the price of an automobile with 6000 miles.

20. **MODELING WITH MATHEMATICS** The table shows the lengths x and costs y of several sailboats.

Length (feet), x	Cost (thousands of dollars), y
27	94
18	56
25	58
32	123
18	60
26	87
36	145

a. Use a graphing calculator to find an equation of the line of best fit.

b. Identify and interpret the correlation coefficient.

c. Interpret the slope and y-intercept of the line of best fit.

d. Approximate the cost of a sailboat that is 20 feet long.

e. Predict the length of a sailboat that costs $147,000.

In Exercises 21–24, tell whether a correlation is likely in the situation. If so, tell whether there is a causal relationship. Explain your reasoning. *(See Example 5.)*

21. the amount of time spent talking on a cell phone and the remaining battery life

22. the height of a toddler and the size of the toddler's vocabulary

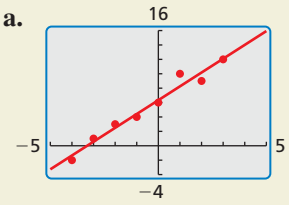
23. the number of hats you own and the size of your head

24. the weight of a dog and the length of its tail

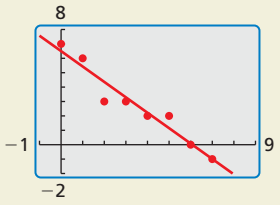
25. OPEN-ENDED Describe a data set that has a strong correlation but does not have a causal relationship.

26. HOW DO YOU SEE IT? Match each graph with its correlation coefficient. Explain your reasoning.

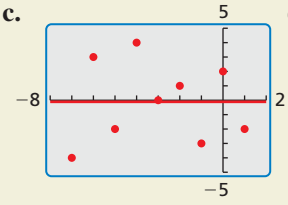
a.



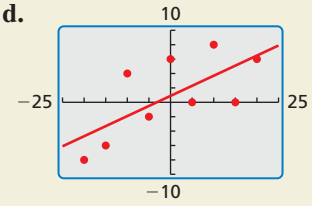
b.



c.



d.



A. $r = 0$ **B.** $r = 0.98$

C. $r = -0.97$ **D.** $r = 0.69$

27. ANALYZING RELATIONSHIPS The table shows the grade point averages y of several students and the numbers x of hours they spend watching television each week.

Hours, x	Grade point average, y
10	3.0
5	3.4
3	3.5
12	2.7
20	2.1
15	2.8
8	3.0
4	3.7
16	2.5

a. Use a graphing calculator to find an equation of the line of best fit. Identify and interpret the correlation coefficient.

b. Interpret the slope and y-intercept of the line of best fit.

c. Another student watches about 14 hours of television each week. Approximate the student's grade point average.

d. Do you think there is a causal relationship between time spent watching television and grade point average? Explain.

28. MAKING AN ARGUMENT A student spends 2 hours watching television each week and has a grade point average of 2.4. Your friend says including this information in the data set in Exercise 27 will weaken the correlation. Is your friend correct? Explain.

29. USING MODELS Refer to Exercise 17.

a. Predict the total numbers of people who reported an earthquake 9 minutes and 15 minutes after it ended.

b. The table shows the actual data. Describe the accuracy of your extrapolations in part (a).

Minutes, x	9	15
People, y	2750	3200

30. THOUGHT PROVOKING A data set consists of the numbers x of people at Beach 1 and the numbers y of people at Beach 2 recorded daily for 1 week. Sketch a possible graph of the data set. Describe the situation shown in the graph and give a possible correlation coefficient. Determine whether there is a causal relationship. Explain.

31. COMPARING METHODS The table shows the numbers y (in billions) of text messages sent each year in a five-year period, where $x = 1$ represents the first year in the five-year period.

Year, x	1	2	3	4	5
Text messages (billions), y	241	601	1360	1806	2206

a. Use a graphing calculator to find an equation of the line of best fit. Identify and interpret the correlation coefficient.

b. Is there a causal relationship? Explain your reasoning.

c. Calculate the residuals. Then make a scatter plot of the residuals and interpret the results.

d. Compare the methods you used in parts (a) and (c) to determine whether the model is a good fit. Which method do you prefer? Explain.

Maintaining Mathematical Proficiency

Reviewing what you learned in previous grades and lessons

Determine whether the table represents a *linear* or *nonlinear* function. Explain. *(Section 3.2)*

32.

x	5	6	7	8
y	−4	4	−4	4

33.

x	2	4	6	8
y	13	8	3	−2