

2.7 Linear Models and Scatter Plots

Scatter Plots and Correlation

Many real-life situations involve finding relationships between two variables, such as the year and the outstanding household credit market debt. In a typical situation, data is collected and written as a set of ordered pairs. The graph of such a set, called a *scatter plot*, was discussed briefly in Section P.5.

Example 1 Constructing a Scatter Plot



The data in the table shows the outstanding household credit market debt D (in trillions of dollars) from 1998 through 2004. Construct a scatter plot of the data. (Source: Board of Governors of the Federal Reserve System)

Year	Household credit market debt, D (in trillions of dollars)
1998	6.0
1999	6.4
2000	7.0
2001	7.6
2002	8.4
2003	9.2
2004	10.3

Solution

Begin by representing the data with a set of ordered pairs. Let t represent the year, with $t = 8$ corresponding to 1998.

$(8, 6.0), (9, 6.4), (10, 7.0), (11, 7.6), (12, 8.4), (13, 9.2), (14, 10.3)$

Then plot each point in a coordinate plane, as shown in Figure 2.59.

CHECKPOINT Now try Exercise 1.

From the scatter plot in Figure 2.59, it appears that the points describe a relationship that is nearly linear. The relationship is not *exactly* linear because the household credit market debt did not increase by precisely the same amount each year.

A mathematical equation that approximates the relationship between t and D is a *mathematical model*. When developing a mathematical model to describe a set of data, you strive for two (often conflicting) goals—accuracy and simplicity. For the data above, a linear model of the form

$$D = at + b$$

appears to be best. It is simple and relatively accurate.

What you should learn

- Construct scatter plots and interpret correlation.
- Use scatter plots and a graphing utility to find linear models for data.

Why you should learn it

Real-life data often follows a linear pattern. For instance, in Exercise 20 on page 240, you will find a linear model for the winning times in the women's 400-meter freestyle Olympic swimming event.



Nick Wilson/Getty Images

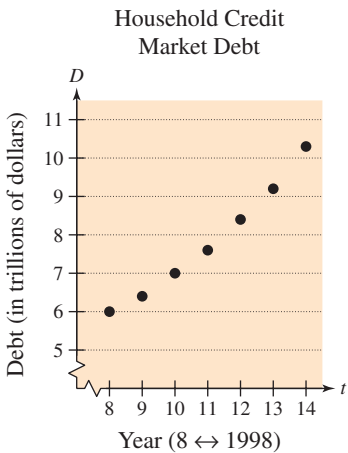
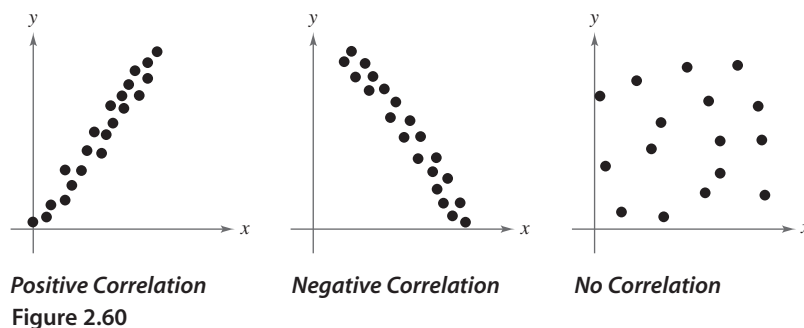


Figure 2.59

Consider a collection of ordered pairs of the form (x, y) . If y tends to increase as x increases, the collection is said to have a **positive correlation**. If y tends to decrease as x increases, the collection is said to have a **negative correlation**. Figure 2.60 shows three examples: one with a positive correlation, one with a negative correlation, and one with no (discernible) correlation.



Example 2 Interpreting Correlation



On a Friday, 22 students in a class were asked to record the numbers of hours they spent studying for a test on Monday and the numbers of hours they spent watching television. The results are shown below. (The first coordinate is the number of hours and the second coordinate is the score obtained on the test.)

Study Hours: (0, 40), (1, 41), (2, 51), (3, 58), (3, 49), (4, 48), (4, 64), (5, 55), (5, 69), (5, 58), (5, 75), (6, 68), (6, 63), (6, 93), (7, 84), (7, 67), (8, 90), (8, 76), (9, 95), (9, 72), (9, 85), (10, 98)

TV Hours: (0, 98), (1, 85), (2, 72), (2, 90), (3, 67), (3, 93), (3, 95), (4, 68), (4, 84), (5, 76), (7, 75), (7, 58), (9, 63), (9, 69), (11, 55), (12, 58), (14, 64), (16, 48), (17, 51), (18, 41), (19, 49), (20, 40)

- Construct a scatter plot for each set of data.
- Determine whether the points are positively correlated, are negatively correlated, or have no discernible correlation. What can you conclude?

Solution

- Scatter plots for the two sets of data are shown in Figure 2.61.
- The scatter plot relating study hours and test scores has a positive correlation. This means that the more a student studied, the higher his or her score tended to be. The scatter plot relating television hours and test scores has a negative correlation. This means that the more time a student spent watching television, the lower his or her score tended to be.

CHECKPOINT Now try Exercise 3.

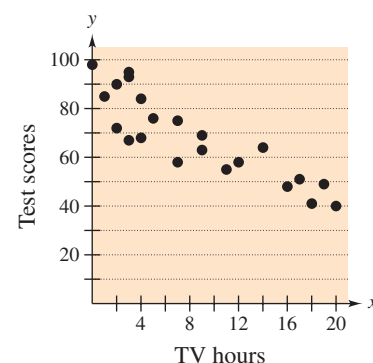
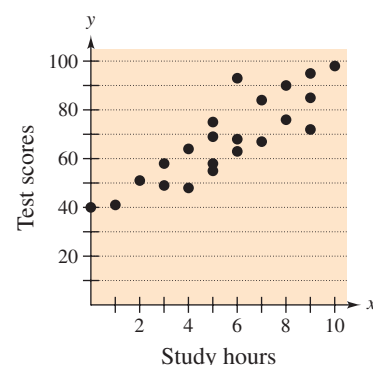


Figure 2.61

Fitting a Line to Data

Finding a linear model to represent the relationship described by a scatter plot is called **fitting a line to data**. You can do this graphically by simply sketching the line that appears to fit the points, finding two points on the line, and then finding the equation of the line that passes through the two points.

Example 3 Fitting a Line to Data

Find a linear model that relates the year to the outstanding household credit market debt. (See Example 1.)

Year	Household credit market debt, D (in trillions of dollars)
1998	6.0
1999	6.4
2000	7.0
2001	7.6
2002	8.4
2003	9.2
2004	10.3

Solution

Let t represent the year, with $t = 8$ corresponding to 1998. After plotting the data in the table, draw the line that you think best represents the data, as shown in Figure 2.62. Two points that lie on this line are $(9, 6.4)$ and $(13, 9.2)$. Using the point-slope form, you can find the equation of the line to be

$$\begin{aligned} D &= 0.7(t - 9) + 6.4 \\ &= 0.7t + 0.1. \end{aligned} \quad \text{Linear model}$$

CHECKPOINT Now try Exercise 11(a) and (b).

Once you have found a model, you can measure how well the model fits the data by comparing the actual values with the values given by the model, as shown in the following table.

	t	8	9	10	11	12	13	14
Actual	D	6.0	6.4	7.0	7.6	8.4	9.2	10.3
Model	D	5.7	6.4	7.1	7.8	8.5	9.2	9.9

The sum of the squares of the differences between the actual values and the model values is the **sum of the squared differences**. The model that has the least sum is the **least squares regression line** for the data. For the model in Example 3, the sum of the squared differences is 0.31. The least squares regression line for the data is

$$D = 0.71t. \quad \text{Best-fitting linear model}$$

Its sum of squared differences is 0.3015. See Appendix C for more on the least squares regression line.

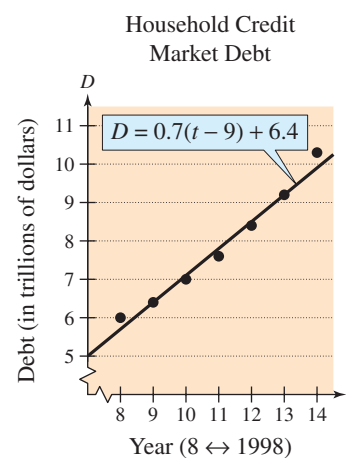


Figure 2.62

STUDY TIP

The model in Example 3 is based on the two data points chosen. If different points are chosen, the model may change somewhat. For instance, if you choose $(8, 6)$ and $(14, 10.3)$, the new model is

$$\begin{aligned} D &= 0.72(t - 8) + 6 \\ &= 0.72t + 0.24. \end{aligned}$$

Example 4 A Mathematical Model

The numbers S (in billions) of shares listed on the New York Stock Exchange for the years 1995 through 2004 are shown in the table. (Source: New York Stock Exchange, Inc.)

Year	Shares, S
1995	154.7
1996	176.9
1997	207.1
1998	239.3
1999	280.9
2000	313.9
2001	341.5
2002	349.9
2003	359.7
2004	380.8

TECHNOLOGY SUPPORT

For instructions on how to use the *regression* feature, see Appendix A; for specific keystrokes, go to this textbook's *Online Study Center*.

- Use the *regression* feature of a graphing utility to find a linear model for the data. Let t represent the year, with $t = 5$ corresponding to 1995.
- How closely does the model represent the data?

Graphical Solution

- Enter the data into the graphing utility's list editor. Then use the *linear regression* feature to obtain the model shown in Figure 2.63. You can approximate the model to be $S = 26.47t + 29.0$.
- You can use a graphing utility to graph the actual data and the model in the same viewing window. In Figure 2.64, it appears that the model is a fairly good fit for the actual data.

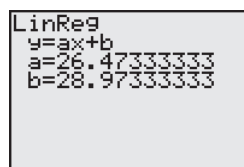


Figure 2.63

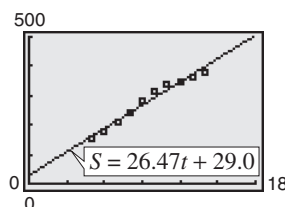


Figure 2.64

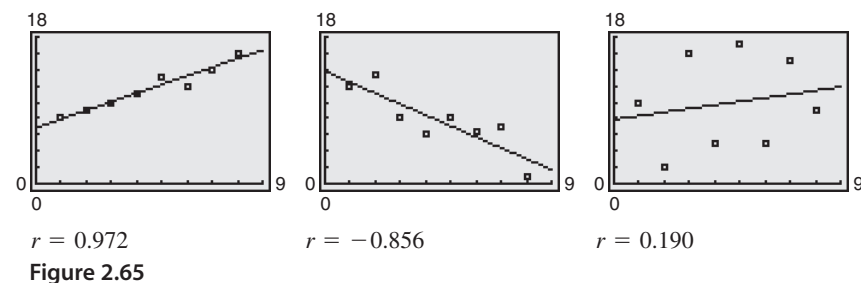
Numerical Solution

- Using the *linear regression* feature of a graphing utility, you can find that a linear model for the data is $S = 26.47t + 29.0$.
- You can see how well the model fits the data by comparing the actual values of S with the values of S given by the model, which are labeled S^* in the table below. From the table, you can see that the model appears to be a good fit for the actual data.

Year	S	S^*
1995	154.7	161.4
1996	176.9	187.8
1997	207.1	214.3
1998	239.3	240.8
1999	280.9	267.2
2000	313.9	293.7
2001	341.5	320.2
2002	349.9	346.6
2003	359.7	373.1
2004	380.8	399.6

CHECKPOINT Now try Exercise 9.

When you use the *regression* feature of a graphing calculator or computer program to find a linear model for data, you will notice that the program may also output an “*r*-value.” For instance, the *r*-value from Example 4 was $r \approx 0.985$. This *r*-value is the **correlation coefficient** of the data and gives a measure of how well the model fits the data. Correlation coefficients vary between -1 and 1 . Basically, the closer $|r|$ is to 1 , the better the points can be described by a line. Three examples are shown in Figure 2.65.



Example 5 Finding a Least Squares Regression Line

The following ordered pairs (w, h) represent the shoe sizes w and the heights h (in inches) of 25 men. Use the *regression* feature of a graphing utility to find the least squares regression line for the data.

(10.0, 70.5)	(10.5, 71.0)	(9.5, 69.0)	(11.0, 72.0)	(12.0, 74.0)
(8.5, 67.0)	(9.0, 68.5)	(13.0, 76.0)	(10.5, 71.5)	(10.5, 70.5)
(10.0, 71.0)	(9.5, 70.0)	(10.0, 71.0)	(10.5, 71.0)	(11.0, 71.5)
(12.0, 73.5)	(12.5, 75.0)	(11.0, 72.0)	(9.0, 68.0)	(10.0, 70.0)
(13.0, 75.5)	(10.5, 72.0)	(10.5, 71.0)	(11.0, 73.0)	(8.5, 67.5)

Solution

After entering the data into a graphing utility (see Figure 2.66), you obtain the model shown in Figure 2.67. So, the least squares regression line for the data is

$$h = 1.84w + 51.9.$$

In Figure 2.68, this line is plotted with the data. Note that the plot does not have 25 points because some of the ordered pairs graph as the same point. The correlation coefficient for this model is $r \approx 0.981$, which implies that the model is a good fit for the data.

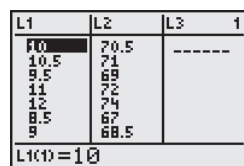


Figure 2.66

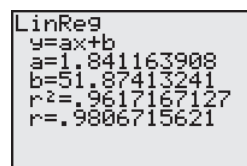


Figure 2.67

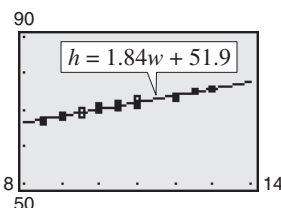


Figure 2.68

CHECKPOINT Now try Exercise 20.

TECHNOLOGY TIP

For some calculators, the *diagnostics on* feature must be selected before the *regression* feature is used in order to see the *r*-value or correlation coefficient. To learn how to use this feature, consult your user's manual.

Emphasize to your students that the correlation coefficient measures the strength of the linear relationship between two variables. Point out that a correlation value near $r = 0$ indicates that there is no linear relationship between the variables, but this does not rule out the possibility of there being some other type of relationship—for instance, a quadratic relationship.

2.7 Exercises

See www.CalcChat.com for worked-out solutions to odd-numbered exercises.

Vocabulary Check

Fill in the blanks.

1. Consider a collection of ordered pairs of the form (x, y) . If y tends to increase as x increases, then the collection is said to have a _____ correlation.
2. Consider a collection of ordered pairs of the form (x, y) . If y tends to decrease as x increases, then the collection is said to have a _____ correlation.
3. The process of finding a linear model for a set of data is called _____.
4. Correlation coefficients vary between _____ and _____.

1. **Sales** The following ordered pairs give the years of experience x for 15 sales representatives and the monthly sales y (in thousands of dollars).

(1.5, 41.7), (1.0, 32.4), (0.3, 19.2), (3.0, 48.4), (4.0, 51.2),
(0.5, 28.5), (2.5, 50.4), (1.8, 35.5), (2.0, 36.0), (1.5, 40.0),
(3.5, 50.3), (4.0, 55.2), (0.5, 29.1), (2.2, 43.2), (2.0, 41.6)

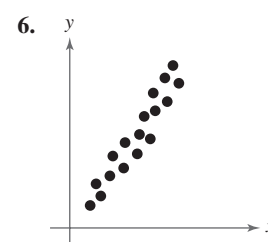
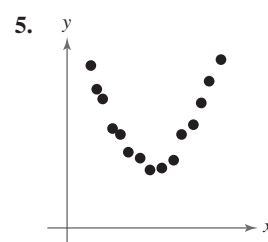
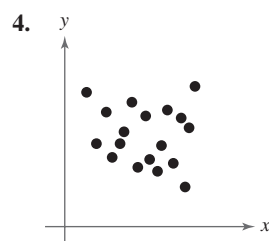
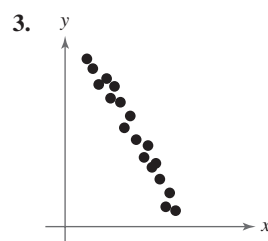
- (a) Create a scatter plot of the data.
- (b) Does the relationship between x and y appear to be approximately linear? Explain.

2. **Quiz Scores** The following ordered pairs give the scores on two consecutive 15-point quizzes for a class of 18 students.

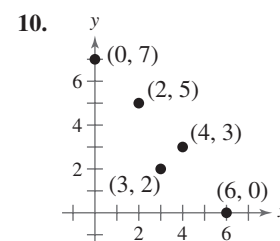
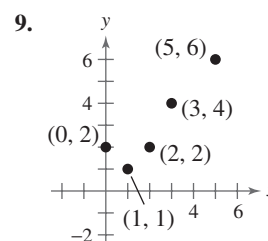
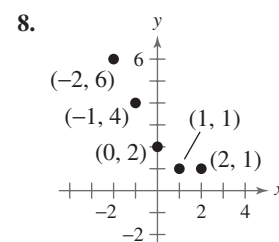
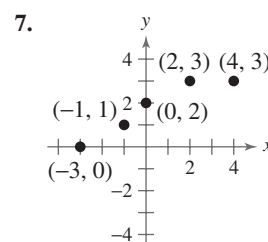
(7, 13), (9, 7), (14, 14), (15, 15), (10, 15), (9, 7),
(14, 11), (14, 15), (8, 10), (9, 10), (15, 9), (10, 11),
(11, 14), (7, 14), (11, 10), (14, 11), (10, 15), (9, 6)

- (a) Create a scatter plot of the data.
- (b) Does the relationship between consecutive quiz scores appear to be approximately linear? If not, give some possible explanations.

In Exercises 3–6, the scatter plots of sets of data are shown. Determine whether there is positive correlation, negative correlation, or no discernible correlation between the variables.




In Exercises 7–10, (a) for the data points given, draw a line of best fit through two of the points and find the equation of the line through the points, (b) use the *regression* feature of a graphing utility to find a linear model for the data, and to identify the correlation coefficient, (c) graph the data points and the lines obtained in parts (a) and (b) in the same viewing window, and (d) comment on the validity of both models. To print an enlarged copy of the graph, go to the website www.mathgraphs.com.




238 Chapter 2 Solving Equations and Inequalities

- 11. Hooke's Law** Hooke's Law states that the force F required to compress or stretch a spring (within its elastic limits) is proportional to the distance d that the spring is compressed or stretched from its original length. That is, $F = kd$, where k is the measure of the stiffness of the spring and is called the *spring constant*. The table shows the elongation d in centimeters of a spring when a force of F kilograms is applied.



Force, F	Elongation, d
20	1.4
40	2.5
60	4.0
80	5.3
100	6.6


- Sketch a scatter plot of the data.
 - Find the equation of the line that seems to best fit the data.
 - Use the *regression* feature of a graphing utility to find a linear model for the data. Compare this model with the model from part (b).
 - Use the model from part (c) to estimate the elongation of the spring when a force of 55 kilograms is applied.
- 12. Cell Phones** The average lengths L of cellular phone calls in minutes from 1999 to 2004 are shown in the table. (Source: Cellular Telecommunications & Internet Association)



Year	Average length, L (in minutes)
1999	2.38
2000	2.56
2001	2.74
2002	2.73
2003	2.87
2004	3.05


- Use a graphing utility to create a scatter plot of the data, with $t = 9$ corresponding to 1999.
- Use the *regression* feature of a graphing utility to find a linear model for the data. Let t represent the year, with $t = 9$ corresponding to 1999.
- Use a graphing utility to plot the data and graph the model in the same viewing window. Is the model a good fit? Explain.
- Use the model to predict the average lengths of cellular phone calls for the years 2010 and 2015. Do your answers seem reasonable? Explain.

- 13. Sports** The mean salaries S (in thousands of dollars) for professional football players in the United States from 2000 to 2004 are shown in the table. (Source: National Collegiate Athletic Assn.)



Year	Mean salary, S (in thousands of dollars)
2000	787
2001	986
2002	1180
2003	1259
2004	1331


- Use a graphing utility to create a scatter plot of the data, with $t = 0$ corresponding to 2000.
 - Use the *regression* feature of a graphing utility to find a linear model for the data. Let t represent the year, with $t = 0$ corresponding to 2000.
 - Use a graphing utility to plot the data and graph the model in the same viewing window. Is the model a good fit? Explain.
 - Use the model to predict the mean salaries for professional football players in 2005 and 2010. Do the results seem reasonable? Explain.
 - What is the slope of your model? What does it tell you about the mean salaries of professional football players?
- 14. Teacher's Salaries** The mean salaries S (in thousands of dollars) of public school teachers in the United States from 1999 to 2004 are shown in the table. (Source: Educational Research Service)



Year	Mean salary, S (in thousands of dollars)
1999	41.4
2000	42.2
2001	43.7
2002	43.8
2003	45.0
2004	45.6


- Use a graphing utility to create a scatter plot of the data, with $t = 9$ corresponding to 1999.
- Use the *regression* feature of a graphing utility to find a linear model for the data. Let t represent the year, with $t = 9$ corresponding to 1999.
- Use a graphing utility to plot the data and graph the model in the same viewing window. Is the model a good fit? Explain.
- Use the model to predict the mean salaries for teachers in 2005 and 2010. Do the results seem reasonable? Explain.

- 15. Cable Television** The average monthly cable television bills C (in dollars) for a basic plan from 1990 to 2004 are shown in the table. (Source: Kagan Research, LLC)



Year	Monthly bill, C (in dollars)
1990	16.78
1991	18.10
1992	19.08
1993	19.39
1994	21.62
1995	23.07
1996	24.41
1997	26.48
1998	27.81
1999	28.92
2000	30.37
2001	32.87
2002	34.71
2003	36.59
2004	38.23


- (a) Use a graphing utility to create a scatter plot of the data, with $t = 0$ corresponding to 1990.
- (b) Use the *regression* feature of a graphing utility to find a linear model for the data and to identify the correlation coefficient. Let t represent the year, with $t = 0$ corresponding to 1990.
- (c) Graph the model with the data in the same viewing window.
- (d) Is the model a good fit for the data? Explain.
- (e) Use the model to predict the average monthly cable bills for the years 2005 and 2010.
- (f) Do you believe the model would be accurate to predict the average monthly cable bills for future years? Explain.
- 16. State Population** The projected populations P (in thousands) for selected years for New Jersey based on the 2000 census are shown in the table. (Source: U.S. Census Bureau)



Year	Population, P (in thousands)
2005	8745
2010	9018
2015	9256
2020	9462
2025	9637
2030	9802


- (a) Use a graphing utility to create a scatter plot of the data, with $t = 5$ corresponding to 2005.
- (b) Use the *regression* feature of a graphing utility to find a linear model for the data. Let t represent the year, with $t = 5$ corresponding to 2005.
- (c) Use a graphing utility to plot the data and graph the model in the same viewing window. Is the model a good fit? Explain.
- (d) Use the model to predict the population of New Jersey in 2050. Does the result seem reasonable? Explain.

- 17. State Population** The projected populations P (in thousands) for selected years for Wyoming based on the 2000 census are shown in the table. (Source: U.S. Census Bureau)



Year	Population, P (in thousands)
2005	507
2010	520
2015	528
2020	531
2025	529
2030	523

- (a) Use a graphing utility to create a scatter plot of the data, with $t = 5$ corresponding to 2005.
- (b) Use the *regression* feature of a graphing utility to find a linear model for the data. Let t represent the year, with $t = 5$ corresponding to 2005.
- (c) Use a graphing utility to plot the data and graph the model in the same viewing window. Is the model a good fit? Explain.
- (d) Use the model to predict the population of Wyoming in 2050. Does the result seem reasonable? Explain.
- 18. Advertising and Sales** The table shows the advertising expenditures x and sales volumes y for a company for seven randomly selected months. Both are measured in thousands of dollars.




Month	Advertising expenditures, x	Sales volume, y
1	2.4	202
2	1.6	184
3	2.0	220
4	2.6	240
5	1.4	180
6	1.6	164
7	2.0	186

240 Chapter 2 Solving Equations and Inequalities

- Use the *regression* feature of a graphing utility to find a linear model for the data and to identify the correlation coefficient.
- Use a graphing utility to plot the data and graph the model in the same viewing window.
- Interpret the slope of the model in the context of the problem.
- Use the model to estimate sales for advertising expenditures of \$1500.

19. **Number of Stores** The table shows the numbers T of Target stores from 1997 to 2006. (Source: Target Corp.)



Year	Number of stores, T
1997	1130
1998	1182
1999	1243
2000	1307
2001	1381
2002	1475
2003	1553
2004	1308
2005	1400
2006	1505

- Use the *regression* feature of a graphing utility to find a linear model for the data and to identify the correlation coefficient. Let t represent the year, with $t = 7$ corresponding to 1997.
- Use a graphing utility to plot the data and graph the model in the same viewing window.
- Interpret the slope of the model in the context of the problem.
- Use the model to find the year in which the number of Target stores will exceed 1800.
- Create a table showing the actual values of T and the values of T given by the model. How closely does the model fit the data?

20. **Sports** The following ordered pairs (t, T) represent the Olympic year t and the winning time T (in minutes) in the women's 400-meter freestyle swimming event. (Source: *The World Almanac 2005*)

(1948, 5.30)	(1968, 4.53)	(1988, 4.06)
(1952, 5.20)	(1972, 4.32)	(1992, 4.12)
(1956, 4.91)	(1976, 4.16)	(1996, 4.12)
(1960, 4.84)	(1980, 4.15)	(2000, 4.10)
(1964, 4.72)	(1984, 4.12)	(2004, 4.09)

- Use the *regression* feature of a graphing utility to find a linear model for the data. Let t represent the year, with $t = 0$ corresponding to 1950.

- What information is given by the sign of the slope of the model?
- Use a graphing utility to plot the data and graph the model in the same viewing window.
- Create a table showing the actual values of y and the values of y given by the model. How closely does the model fit the data?
- Can the model be used to predict the winning times in the future? Explain.

Synthesis

True or False? In Exercises 21 and 22, determine whether the statement is true or false. Justify your answer.

- A linear regression model with a positive correlation will have a slope that is greater than 0.
- If the correlation coefficient for a linear regression model is close to -1 , the regression line cannot be used to describe the data.
- Writing** A linear mathematical model for predicting prize winnings at a race is based on data for 3 years. Write a paragraph discussing the potential accuracy or inaccuracy of such a model.
- Research Project** Use your school's library, the Internet, or some other reference source to locate data that you think describes a linear relationship. Create a scatter plot of the data and find the least squares regression line that represents the points. Interpret the slope and y -intercept in the context of the data. Write a summary of your findings.

Skills Review

In Exercises 25–28, evaluate the function at each value of the independent variable and simplify.

- $f(x) = 2x^2 - 3x + 5$
 - $f(-1)$
 - $f(w + 2)$
- $g(x) = 5x^2 - 6x + 1$
 - $g(-2)$
 - $g(z - 2)$
- $h(x) = \begin{cases} 1 - x^2, & x \leq 0 \\ 2x + 3, & x > 0 \end{cases}$
 - $h(1)$
 - $h(0)$
- $k(x) = \begin{cases} 5 - 2x, & x < -1 \\ x^2 + 4, & x \geq -1 \end{cases}$
 - $k(-3)$
 - $k(-1)$

In Exercises 29–34, solve the equation algebraically. Check your solution graphically.

- $6x + 1 = -9x - 8$
- $8x^2 - 10x - 3 = 0$
- $2x^2 - 7x + 4 = 0$
- $3(x - 3) = 7x + 2$
- $10x^2 - 23x - 5 = 0$
- $2x^2 - 8x + 5 = 0$