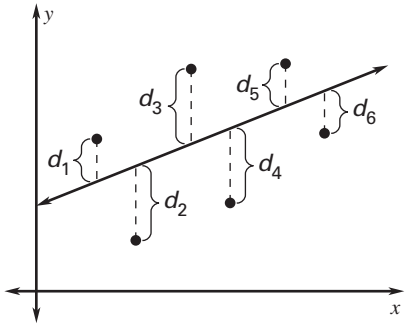
LESSON 30

Linear Regression

GOAL Find equations of lines that model data sets.

The process of fitting a line to a set of data is called **linear regression**. The *least-squares line* is the most common type of linear regression model. Another type of linear regression model is the *median-median line*.

The diagram at the right shows a scatter plot of a set of data and a line approximating the data. When the sum of the squares of the d-values (the vertical distances between the data points and the line) is at a minimum, the line is called a **least-squares line**.



EXAMPLE 1 Using a Least-Squares Line

On cold days, wind makes the air feel colder than it actually is. This effect is called wind chill. The table below shows the wind chill temperature (in degrees Fahrenheit) at various wind speeds (in miles per hour) when the actual air temperature is 15°F. Use a graphing calculator to find and graph an equation of the least-squares line for the data. Then predict the wind chill temperature when the wind speed is 30 miles per hour.

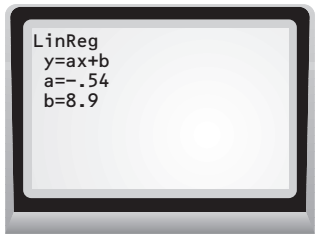
Wind speed, *x*	5	10	15	20	25
Wind chill temperature, *y*	7	3	0	−2	−4

SOLUTION

Step 1 Enter the wind speeds into L1 and the wind chill temperatures into L2.

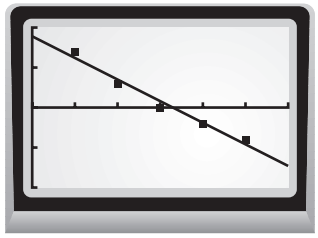
Step 2 Calculate the least-squares line by entering the following keystrokes:

STAT ▶ 4 ENTER



An equation of the least-squares line is $y = -0.54x + 8.9$.

Step 3 Graph the data and the least-squares line.



Step 4 Use the equation of the least-squares line to predict the wind chill temperature when the wind speed is 30 miles per hour.

$y = -0.54x + 8.9$ *Write linear model.*

$y = -0.54(30) + 8.9$ *Substitute 30 for x.*

$y = -7.3$ *Simplify.*

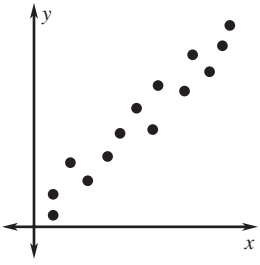
You can predict that when the wind speed is 30 miles per hour, the wind chill temperature will be −7.3°F.

CHECK Example 1

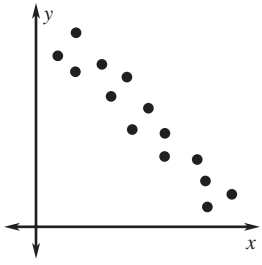
1. The table shows the wind chill temperature (in degrees Fahrenheit) at various wind speeds (in miles per hour) when the actual air temperature is 0°F. Use a graphing calculator to find and graph the equation of the least-squares line for the data. Then predict the wind chill temperature when the wind speed is 10 miles per hour.

Wind speed, *x*	20	25	30	35	40
Wind chill temperature, *y*	−22	−24	−26	−27	−29

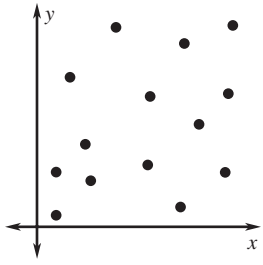
Correlation Coefficient The **correlation coefficient** r for a set of paired data is a measure of how well the least-squares line fits the data. If all of the graphed data pairs lie exactly on a line with a positive slope, the correlation coefficient is 1. If all of the graphed data pairs lie exactly on a line with a negative slope, the correlation coefficient is −1. If the graphed data pairs tend not to lie on any line, the correlation coefficient is close to 0.



***r* is close to 1.**
x and *y* have a strong positive correlation.



***r* is close to −1.**
x and *y* have a strong negative correlation.



***r* is close to 0.**
x and *y* do not seem to have any correlation.

EXAMPLE 2 Finding a Correlation Coefficient

Use a graphing calculator to find the correlation coefficient for the data in Example 1. Then describe the correlation between *x* and *y*.

SOLUTION

Enter the data into L1 and L2. Then calculate the least-squares line.

Calculate the correlation coefficient by entering the following keystrokes:

VARS 5 ▶ ▶ 7 ENTER

The correlation coefficient r is about −0.987. Because r is close to −1, *x* and *y* have a strong negative correlation.

CHECK Example 2

2. Use a graphing calculator to find the correlation coefficient for the data in Exercise 1 above. Then describe the correlation between *x* and *y*.

Median-Median Line The **median-median line** is a type of linear regression model that uses *summary points* calculated using medians. In this way, the median-median line is less influenced by outliers than the least-squares line.

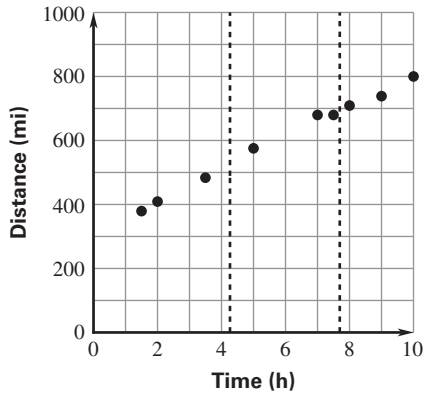
EXAMPLE 3 Finding a Median-Median Line

During one day of a road trip, Jamie leaves her hotel and begins driving. The table shows the time she has been traveling and her total distance from home. Fit a median-median line to the data. Interpret the slope and *y*-intercept.

Time, *x* (hours)	1.5	2	3.5	5	7	7.5	8	9	10
Distance, *y* (miles)	380	408	482	574	679	679	708	737	801

SOLUTION

Step 1 Make a scatter plot of the data. Draw two vertical lines that divide the data into three groups of equal (or nearly equal) size.



Step 2 Find a *summary point* for each of the three groups. The summary point for each group has coordinates (median *x*-value, median *y*-value).

The summary points are (2, 408), (7, 679), and (9, 737).

Step 3 Find the slope of the line that passes through the first and third summary points.

$$m = \frac{737 - 408}{9 - 2} = \frac{329}{7} = 47$$

Step 4 Find the point with the coordinates (mean of *x*-values of summary points, mean of *y*-values of summary points).

$$\left(\frac{2 + 7 + 9}{3}, \frac{408 + 679 + 737}{3}\right) = (6, 608)$$

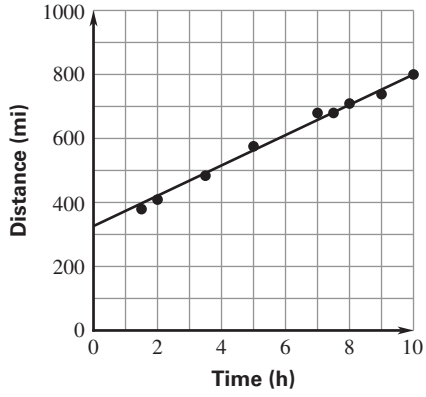
Step 5 Write an equation of the median-median line using the slope in Step 3 and the point in Step 4.

$y - y_1 = m(x - x_1)$ *Point-slope form*

$y - 608 = 47(x - 6)$ *Substitute.*

$y = 47x + 326$ *Solve for y.*

The equation of the median-median line is $y = 47x + 326$. The slope is Jamie's approximate average speed, 47 mi/h. The *y*-intercept is her approximate starting distance from home, 326 miles.



EXAMPLE 4 Using a Calculator to Find a Median-Median Line

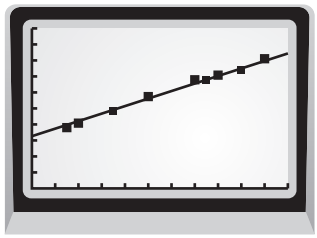
Use a graphing calculator to graph the median-median line and scatter plot for the data in Example 3.

SOLUTION

Step 1 Enter the data into lists.

Step 2 Press STAT. From the CALC menu, choose Med-Med. The median-median line has the form $y = ax + b$, where, in this case, $a = 47$ and $b = 326$.

Step 3 Press Y = and enter $47x + 326$ for Y1. Set up Plot1 to display the scatter plot. Press GRAPH.



CHECK Examples 3 and 4

A hot air balloon begins a slow vertical descent. The table shows the height of the hot air balloon over time. Use the data in the table for Exercises 3 and 4.

Time (min)	1	2	3	4	5	6	7	8	9
Height (ft)	370	340	305	270	250	230	185	160	135

3. Fit a median-median line to the data. Interpret the slope and y-intercept.

4. Use a graphing calculator to graph the median-median line and scatter plot for the data.

EXERCISES

In Exercises 1 and 2, use a graphing calculator to find and graph an equation of the least-squares line, and calculate the correlation coefficient. Then make the indicated prediction.

1. The table below shows the calories burned per hour by a 130 pound person while running at various speeds (in miles per hour). Predict the calories burned per hour by a 130 pound person running at a speed of 9.5 miles per hour.

Speed, *x*	5	5.2	6	6.7	7	7.5	8	8.6	9	10	10.9
Calories burned, *y*	472	531	590	649	679	738	797	826	885	944	1062

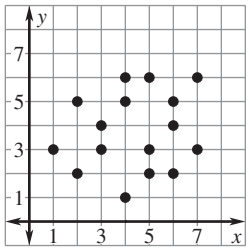
2. The table below shows the unit price (in dollars per dozen) of miniature soccer balls at a novelty store for several different size purchases. Predict the unit price of miniature soccer balls when 30 dozen balls are purchased.

Dozens purchased, *x*	1	6	12	18	24
Unit price, *y*	7.20	6.50	6.00	5.40	4.80

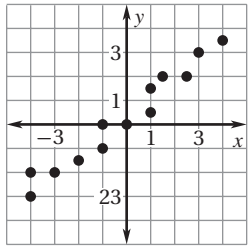
Match the correlation coefficient with the appropriate scatter plot. Explain your reasoning.

3. 0.977 **4.** −0.635 **5.** 0.170 **6.** −0.921

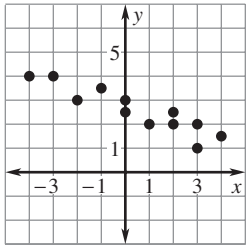
A.



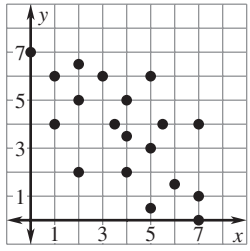
B.



C.



D.



7. The table shows the height of a tree over time. Fit a median-median line to the data. Interpret the slope and *y*-intercept.

Years since 1999, *x*	1	2	3	4	5	6	7	8	9
Height, *y* (ft)	6	9	12	12	15	16	17	21	23

8. You collect data on six books in a series of cookbooks. Use a graphing calculator to graph the data and fit a median-median line to the data. Interpret the slope and *y*-intercept.

Number of pages, *x*	164	182	224	256	288	320
Weight, *y* (oz)	10.5	11.2	12.8	14.4	15.2	16.8

9. The table shows the number of people and the additional weight that can be safely added for several elevators. Fit a median-median line to the data. Use a graphing calculator to check your answer. Interpret the slope and *y*-intercept.

Number of people, *x*	2	4	5	7	7	10	12	12	15
Additional weight, *y* (lb)	2200	1900	1650	1170	1250	920	600	550	100

10. In Example 3 on page 147, suppose Jamie drives for 9 hours as shown but then takes a short airplane flight. Replace the data point (10, 801) with the outlier (10, 1120).

a. Use a graphing calculator to fit a median-median line for the new data set.

b. Use a graphing calculator to fit a least-squares line for both the original data set and the new data set.

c. Discuss the effect the outlier has with respect to the median-median line and the least-squares line.

d. Discuss the effect the outlier has with respect to the correlation coefficient of the least-squares line.

11. Jacob downloads eight files from the internet. The table below shows the file sizes (in megabytes) and the download times (in seconds).

File size, *x* (MB)	1	2	2.5	3	3.5	4	5	6
Download time, *y* (sec)	12	26	32	36	44	48	62	70

a. Find the equation of the least-squares line and the correlation coefficient.

b. Jacob buys a new computer with a faster internet connection. He is able to download each file in half as much time. Find the equation of the least-squares line and the correlation coefficient for the new data.

c. How does the slope of the original least-squares line compare to the slope of the new least-squares line?

d. Compare the correlation coefficients.

12. The table below shows the men's winning long jump distances (in inches) in the Olympics over several years.

Years since 1968, *x*	0	4	8	12	16	20	24	28	32
Distance, *y* (in.)	350.5	324.5	328.75	336.25	336.25	343.25	341.5	334.75	336.75

a. Use the median-median line to predict the men's winning long jump distance in the 2012 Olympics.

b. Do you think you could use the equation of the median-median line to make a reasonable prediction of the men's winning long jump distance in *any* future year? Why or why not?

13. A restaurant has a banquet room that can be rented for special events. The table shows the number of employees who prepared the banquet room for different events and the time it took them to prepare the room. Use a graphing calculator to find and graph the least-squares line and the median-median line for the set of data. Which one is a better model? Explain your reasoning.

Number of employees, *x*	1	3	3	4	4	5	6	7	7	8	9	10
Time to prepare banquet room, *y* (hours)	8	12	7	7	6	5	4	4	3	3	2	1

In Exercises 14–16, use the following information. For Exercises 14 and 15, explain whether you expect a strong correlation and whether there is causation.

Causation refers to a cause-and-effect relationship between variables; that is, an increase or decrease in one variable that directly causes a corresponding increase or decrease in another variable. Variables may have a strong positive or negative correlation without necessarily having a cause-and-effect relationship.

14. Suppose you collect data on the height of a child on January 1 of each year during a 10-year period. On those same days, you also collect data on the average price of a movie ticket.

15. Suppose you collect data on the amount of gasoline sold in your town during 12 months and the monthly tax revenue from selling the gasoline.

16. In some cases, two variables may not have a direct cause-and-effect relationship, but both variables may be influenced by a third variable (or *lurking variable*). Consider a study that shows that as sales of iced tea increase, the number of cases of heatstroke increases. Describe the correlation and explain whether there is causation. Then explain how a third variable might influence the relationship cited in the study.