Lesson 3-5

Fitting a Line to Data

Vocabulary

linear regression

line of best fit, least-squares line, regression line

deviation

BIG IDEA When data are almost on a line, it is often helpful to approximate the data by a line that minimizes the sum of the squares of the distances from the data points to the corresponding points on the line.

Mental Math

Find the slope of the line.

a. $y = 2.5x + 7$

b. $y = 14$

c. $y - 0.2 = 0.144(x - 0.4)$

d. $16x + 20y = 45$

Square Grabber, which may be provided by your teacher, is a fun game with simple rules. You control a black square. Your job is to capture the other black squares by touching them while avoiding the red squares.

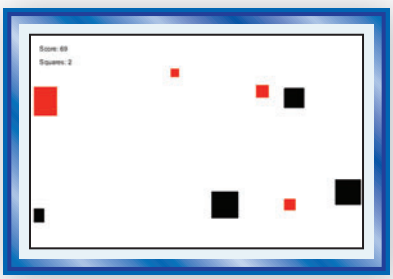
Play the game once. At the end of the game you are told the number of squares you captured and are given a score.

You can write the numbers as an ordered pair (number of squares, score). If you play many times and generate many ordered pairs, you can graph the ordered pairs and use the graph to find a model for the relationship between the number of squares you capture and your score. Is this a simple linear relationship?

Activity 1

MATERIALS internet connection

Step 1 Play Square Grabber 15 times. Record the number of squares you capture and your score for each game in a table like the one below.



Number of Squares (n)	Score (s)
43	855
23	461
12	235

Step 2 Graph your data. Let n = number of squares and s = score. Use n as the independent variable and s as the dependent variable. Does it appear that a line could be a good model for your data?

Step 3 Eyeball a line that comes close to modeling all of the points in your data set. Draw it on your graph.

Step 4 Estimate the coordinates of any two points on your line. (They do not have to be actual data points.) Find an equation for the line through these points.

Step 5 Your equation from Step 4 can be used to estimate your score based on the number of squares you captured. Use your model to estimate your score for capturing 100 squares.

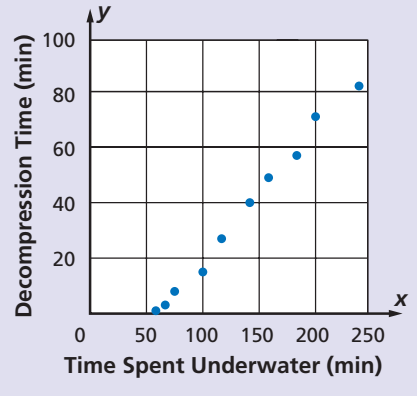
The Regression Line

In Activity 1, you approximated a *line of best fit* for your data by eye. How can you tell which line fits the data the *best*? If you passed your graph of ordered pairs around the room and asked each of your classmates to find a *line of best fit*, you might get as many different lines as you have classmates.

To solve this problem, statisticians have developed a method called **linear regression** that uses all the data points to find the line. A line found by using regression is what people call the **line of best fit**. It is also called the **least-squares line**, or simply the **regression line**. You will learn how regression works in a later course. For now, to find the line of best fit, use a statistics application. For details on how your application works, check the manual or ask your teacher.

Activity 2

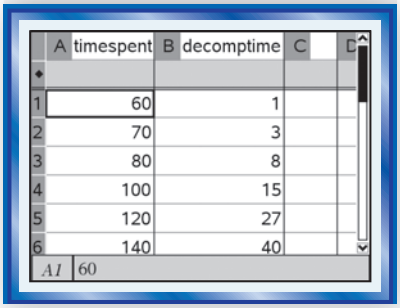
Navy divers who remain underwater for long periods of time cannot come quickly back to the surface due to the high pressure under the water. They must make what are known as *decompression stops* on the way up. If divers skip this procedure they risk a serious medical condition known as *the bends*. The U.S. Navy has created tables that allow divers to know when and for how long they should stop on the way to the surface. The table below gives the decompression time needed (including ascent time) based on how many minutes were spent at a given depth. The points in the table are graphed at the right. Calculate the regression line for the decompression data.



Time Spent at a Maximum Depth of 60 feet (min)	60	70	80	100	120	140	160	180	200	240
Decompression Time Needed (min)	1	3	8	15	27	40	49	57	71	82

(continued on next page)

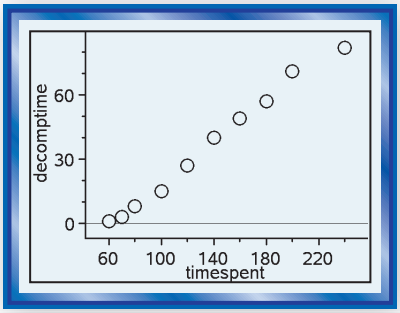
Step 1 Enter the data into columns in your statistics or spreadsheet application. Name the first column *timespent* and the second column *decomptime* to indicate which is the independent and which is the dependent variable.



	A timespent	B decomptime	C	D
1	60	1		
2	70	3		
3	80	8		
4	100	15		
5	120	27		
6	140	40		

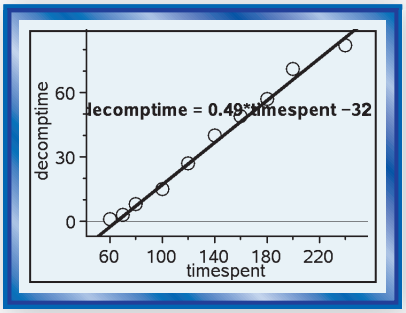
A1 60

Step 2 Create a scatterplot of the data. You may be prompted to choose one column to use for the independent variable data and choose one column to use for the dependent variable data.

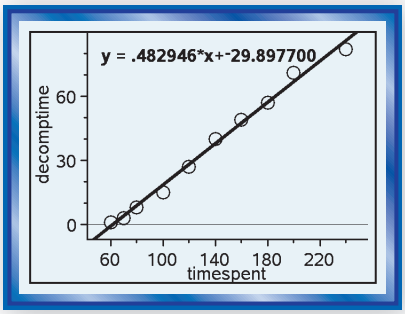


Step 3 If possible, add a movable line to your scatterplot. Adjust the position of the line to eyeball a line of best fit. Record the equation of the equation of your movable line. Also record the predicted values when $x = 80$ and when $x = 220$.

Step 4 Remove your movable line. Choose to show a linear regression line. The application will graph the regression line on the same axes as the data. Record the equation of the regression line. How does your movable line compare to the regression line?



Step 5 Compare your movable line's predictions to those of the regression line. When $x = 220$, by how much do the predicted values differ from each other? When $x = 80$, by how much does the predicted value differ from the value in the table?



You can see from the graphs that the regression line is a reasonable model for the data.

For each value of the independent variable, the difference between the actual value of the dependent variable and the value predicted by the model is called the **deviation**. The line of best fit has the following property: *The sum of the squares of the deviations of its predicted values from the actual values is the least among all possible lines that could model the data*. This is why it is called the least-squares line.

Questions

COVERING THE IDEAS

In 1–3, refer to Activity 1.

1. A student playing Square Grabber recorded a data point of (48, 985). What does this ordered pair mean in this context?
2. a. Find the regression line for the data you collected in Activity 1.
   b. Pick two values of x to compare how well the regression line predicts y-values compared to the line you eyeballed.
3. A person found $y = 19.0x + 35$ to be an equation for the line of best fit for Activity 1.
   a. What is the slope of the line?
   b. What does the slope mean in this situation?
   c. What is the y-intercept of the line?
   d. What does the y-intercept mean in this situation?
   e. Does the x-intercept have a practical meaning in this case?
4. Refer to Activity 2. If a diver spent 130 minutes at a maximum depth of 60 feet, estimate his decompression time using the regression line.

APPLYING THE MATHEMATICS

5. a. Use regression to find an equation of the line through the points (1, 4) and (-2, 8).
   b. Verify your equation in Part a by finding the slope of the line and using the Point-Slope Theorem.
6. The table below gives the total payroll for the Chicago Cubs from 1998 to 2006.

Year	2006	2005	2004	2003	2002	2001	2000	1999	1998
Payroll (millions of dollars)	94.4	87.0	90.6	79.9	75.7	64.5	62.1	55.4	49.4

Source: http://content.usatoday.com/sports/baseball/salaries/teamresults.aspx?team=17

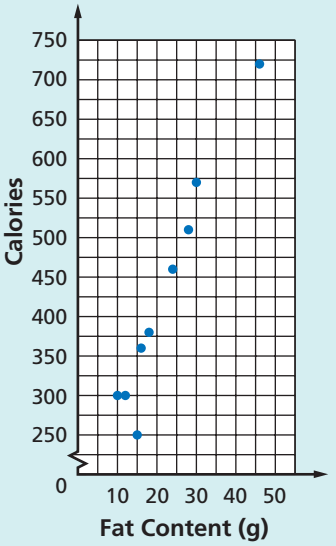
   a. Draw a scatterplot and eyeball a line of best fit to the data.
   b. Which data point has the greatest deviation from your line?
   c. Find an equation for the regression line for the data.
   d. Which data point has the greatest deviation from the regression line?
   e. Use the regression line and your line to predict the 2007 Chicago Cubs payroll.
   f. Which gives the closer prediction to the actual value of 99.7 million dollars, your eyeballed line or the regression line?

7. Recall the data at the right from the beginning of the chapter on the average body mass index b for a male of age a. Verify that $b = 0.07a + 24.9$ is an equation for the line of best fit to the data.

Age	BMI
25	26.6
35	27.5
45	28.4
55	28.7

8. The table below shows nutritional information on various items from a fast-food menu. Included for each item are the number of calories and the fat content. A scatterplot of the data is shown.

Item	Calories	Fat Content (g)
chicken pieces	250	15
Asian salad	300	10
cheeseburger	300	12
fish fillet	380	18
chicken sandwich	360	16
large French fries	570	30
big hamburger	460	24
big cheeseburger	510	28
big breakfast	720	46



   a. Describe the general relationship between calories and fat content.
   b. What are the independent and dependent variables?
   c. Find an equation of the regression line for this data set.
   d. What is the slope of the regression line, and what does it mean in the context of the problem? What is the unit for the slope?
   e. What is the y-intercept of the regression line, and what does it mean in context of the problem? Is the value practical in this situation?
   f. A salad with chicken is not on this menu. It contains 320 calories. Use the regression line to estimate the number of grams of fat in the salad.

9. Add a large vanilla milkshake with 740 calories and 18 grams of fat to the menu in Question 8. Recalculate the regression line.

REVIEW

10. **a.** Find an equation for the line that passes through the points $(2, 5)$ and $\left(-\frac{4}{3}, 9\right)$.

b. Graph the line from Part a. **(Lesson 3-4)**

11. **a.** Write an equation of a line with slope $-\frac{5}{2}$ and y-intercept 3.

b. Write an equation of a line parallel to the line in Part a that passes through the point $(-1, -2)$. **(Lessons 3-4, 3-1)**

12. Consider the formula for the volume of a sphere, $V = \frac{4}{3}\pi r^3$. If the radius of a sphere is divided by three, how many times smaller is the volume of the resulting sphere? **(Lesson 2-3)**

13. Consider the sequence $H_n = (-1)^n(n + 5)$, for integers $n \geq 1$. **(Lesson 1-8)**

a. Write the first four terms.

b. Without explicitly calculating, is the 25th term positive or negative?

14. Jayla's cake company charged \$25 per cake, and each week she paid \$50 for supplies and the upkeep of her equipment. She found that $I = 25c - 50$ is an explicit formula for her income I based on the number c of cakes that she sold. **(Lesson 1-5)**

a. Use a graphing utility to generate a table showing the number of cakes sold and Jayla's income.

b. Graph the first six data values from your table on coordinate axes.

The world's most expensive chocolate cake was decorated with diamonds and sold for over \$8.3 million.

15. Given that $b: x \rightarrow \frac{3x + 5}{7 - 8x}$ evaluate **(Lesson 1-3)**

a. $b(2)$. **b.** $b(0)$.

c. $b(a)$. **d.** $b(-2) + b(-4)$.

EXPLORATION

16. When you calculate a regression equation, some calculators and software include an extra statistic r called *correlation* along with the slope and intercept. Find out what correlation means and what it tells you about the regression line. Go back and examine r for the data sets in Questions 6–8 and see if you can interpret its value in the context of the problems.