Lesson 6-7

Fitting a Line to Data

Vocabulary

line of best fit

linear regression

least squares line

BIG IDEA When points lie nearly on a line, it is useful to determine an equation for a line that lies on or comes close to the points.

Mental Math

Give the conversion factor for converting

a. $\frac{\text{inches}}{\text{year}}$ to $\frac{\text{inches}}{\text{month}}$.

b. $\frac{\text{meters}}{\text{pound}}$ to $\frac{\text{millimeters}}{\text{pound}}$.

c. $\frac{\text{feet}}{\text{second}}$ to $\frac{\text{inches}}{\text{minute}}$.

If data points are not all on one line but are close to being linear, you can often use an equation for a line to describe trends in the data. For example, the table and graph below show the life expectancy of people in the United States at birth in ten-year intervals from 1930 to 2000. Notice that the life expectancy has been increasing each decade.

It is natural to wonder what the life expectancy will be in 2010, 2020, or 2050. Of course no one knows, but we can make educated guesses by using algebra.

Notice that the change in life expectancy each decade is not constant. So the points do not lie on the same line. Still, the points seem to be reasonably close to a line. There are different ways of estimating a line that comes close to the data. This is called *fitting a line to the data.*

Year	Life Expectancy (yr)
1930	59.7
1940	62.9
1950	68.2
1960	69.7
1970	70.8
1980	73.7
1990	75.4
2000	77.0

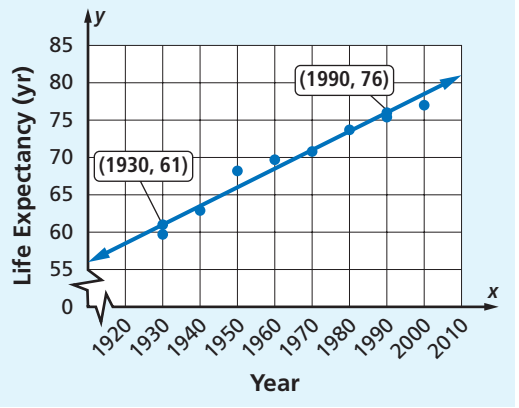
Source: National Center for Health Statistics

Eyeballing a Line of Fit

Activity 1

Step 1 After carefully graphing the data from the table, take a ruler and draw a line that seems close to all the points. This is called "fitting a line by eye" or "eyeballing." One such line is graphed at the right.

Step 2 Find two points on the line. The line we drew happens to not pass through any of the original data points. Our line contains (1930, 61) and (1990, 76).



Step 3 Find an equation for the line through the two points. We follow the algorithm in Lesson 6-6 for finding an equation of a line given two points. First we use these two points to find the slope of the line.

$$\text{slope} = \frac{76 - 61}{1990 - 1930} = \frac{15}{60} = 0.25$$

Now substitute the slope and the coordinates of one of the points into $y = mx + b$ and solve. We use (1930, 61).

$$61 = 0.25 \cdot 1930 + b$$

$$61 = 482.5 + b$$

$$-421.5 = b$$

An equation for the line is $y = 0.25x - 421.5$.

With this method, an estimate for the life expectancy for someone born in 2020 is $0.25 \cdot 2020 - 421.5$, or about 83.5 years.

Eyeballing is a simple method but it has a weakness in that two different people will likely eyeball two different lines.

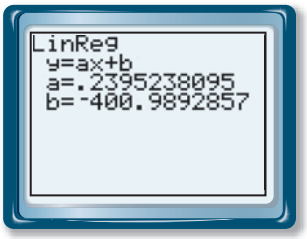
Linear Regression

Most graphing calculators have a feature that will give you what is known as the **line of best fit.** The method they use is called **linear regression.** This is the most common way of finding a line to fit data. In Activity 2, we show only how to use a calculator to find an equation for this line.

Activity 2

Step 1 Enter the eight ordered pairs for the data into two lists, one for the x-coordinate and the other for the corresponding y-coordinate. On some calculators, these lists are called L1 and L2.

Step 2 Have the calculator automatically calculate the line of best fit. One calculator showed the screen at the right. The letter a indicates slope.



Step 3 Round a and b to reasonable accuracy. Here we need four decimal places for a because the x values for the years are so large. Substitute the rounded values for a and b into the equation $y = ax + b$. So, by this method, a line of best fit is $y = 0.2395x - 400.9893$.

Using linear regression, an estimate for the life expectancy for someone born in 2020 is $0.2395 \cdot 2020 - 400.9893$, or about 82.8 years. This is a little lower than what was predicted by eyeballing.

Both an eyeballed line and the line of best fit can be considered as models of life expectancy in the United States from 1930 to 2000. Recall that the difference between the actual amount and the amount predicted by a model is called the deviation.

The table below shows the actual life expectancy, the expectancies predicted by these models, and the deviations for each model. It also shows the predicted life expectancies for 2010 and 2020.

Year	Life Expectancy	Eyeball a Line of Fit ($y = 0.25x - 421.5$)	Line of Best Fit ($y = 0.2395x - 400.9893$)	Eyeball Deviation	Best Fit Deviation
1930	59.7	61.0	61.2	1.3	1.5
1940	62.9	63.5	63.6	0.6	0.7
1950	68.2	66.0	66.0	−2.2	−2.2
1960	69.7	68.5	68.4	−1.2	−1.3
1970	70.8	71.0	70.8	0.2	0
1980	73.7	73.5	73.2	−0.2	−0.5
1990	75.4	76.0	75.6	0.6	0.2
2000	77.0	78.5	78.0	1.5	1.0
2010		81.0	80.4		
2020		83.5	82.8		

The line of best fit has the following property: The sum of the squares of the deviations of its values from the actual values is the least of all lines. For this reason, it is called the **least squares line.**

GUIDED

Example

Show that the sum of the squares of the eyeball line deviations for the life expectancies is greater than the sum of the squares of the best fit deviations.

Solution For the eyeball line, the sum of the squares of the deviations is $1.3^2 + 0.6^2 + (-2.2)^2 + (-1.2)^2 + 0.2^2 + (-0.2)^2 + 0.6^2 + 1.5^2 =$ ___?___.

For the least squares line, the sum of the squares of the deviations is ___?___.

Questions

COVERING THE IDEAS

1. If three people were to use the indicated method for fitting a line to a particular set of data, would their answers necessarily be the same? (Assume they made no errors in calculations.)
   a. eyeballing
   b. linear regression

In 2 and 3, use the table at the right of life expectancies for people in the United States.

Year	All	Females	Males
1930	59.7	61.6	58.1
1940	62.9	65.2	60.8
1950	68.2	71.1	65.6
1960	69.7	73.1	66.6
1970	70.8	74.7	67.1
1980	73.7	77.4	70.0
1990	75.4	78.8	71.8
2000	77.0	79.7	74.3

Source: U.S. Census Bureau

2. a. Construct a scatterplot of the ordered pairs (year, female life expectancy).
   b. Eyeball a line to fit the data and find its equation.
   c. What female life expectancies does your equation predict for the years 1930–2000?
   d. Calculate the sum of the squares of the deviations of the predicted values in Part c from the actual values.
   e. Use your equation to predict the female life expectancy in the U.S. in the year 2020.
3. Follow the directions for Question 2, but use linear regression.
4. Why is the line found using linear regression called the *least squares* line?

APPLYING THE MATHEMATICS

In 5 and 6, the table shows women's 800-meter freestyle swimming long course (50-meter pool) world records between 1971 and 1978.

Person and Country	Year	Time (min)
Shane Gould, Australia	1971	8.97
Keena Rothhammer, USA	1972	8.88
Novella Calligaris, Italy	1973	8.87
Jo Ann Harshbarger, USA	1974	8.79
Jennifer Turrall, Australia	1975	8.72
Petra Thumer, East Germany	1976	8.67
Petra Thumer, East Germany	1977	8.58
Tracey Wickham, Australia	1978	8.40

Source: USA Swimming

In her 11-year career, Evans won 25 of 27 major international races at the 400-meter freestyle and 22 of 23 at the 800-meter freestyle.

Source: United States Olympic Committee

5. a. Construct a scatterplot of the ordered pairs (year, time).
   b. Use linear regression to predict the world record in 1989.
   c. In 1989, Janet Evans set the most recent world record in the long course women's 800-meter. Her time was 8 minutes, 16.22 seconds, or about 8.27 minutes. Calculate the deviation from the linear regression prediction.
   d. The first women's 800-meter freestyle world record was in 1919. The record was set by Gertrude Ederle in a time of 13.32 minutes. This time deviates from the linear regression equation's predicted time by how much?
   e. Is the linear regression line a good model for predicting world record times in the women's 800-meter freestyle before or after the 1970s? Explain why or why not.

6. Add the 1919 and 1989 world record times from Question 5 to the table. How does the linear regression equation change? Do you think it is more or less accurate for predicting world records in the women's 800-meter freestyle?

7. Refer to the data below.

City	Latitude (°North)	January Mean Low Temperature (°F)
Lagos, Nigeria	6	74
San Juan, Puerto Rico	18	70
Calcutta, India	23	55
Cairo, Egypt	30	47
Tokyo, Japan	35	31
Rome, Italy	42	39
Belgrade, Serbia	45	28
London, England	52	35
Copenhagen, Denmark	56	29
Moscow, Russia	56	9

Source: infoplease.com

a. Make a scatterplot showing a point for each city.

b. Use linear regression to find an equation.

c. **Fill in the Blank** As you go one degree north, the January low temperature tends to ___?___.

d. Which city's January mean low temperature deviates most from that predicted by the equation?

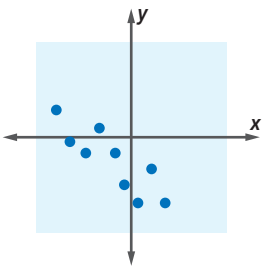
e. Predict the January mean low temperature for the North Pole.

f. The January mean low temperature for Acapulco, Mexico, which is at 17° north latitude, deviates from the equation by +8.3°F. Find the actual January mean low temperature in Acapulco.

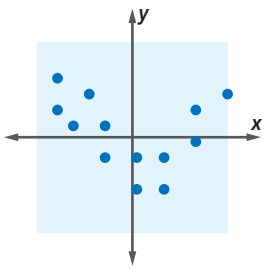
The domes and minarets of the Sultan Hassan Madrasa stand on the eastern edge of Cairo.

In 8–11, tell whether fitting a line to the data points would be appropriate.

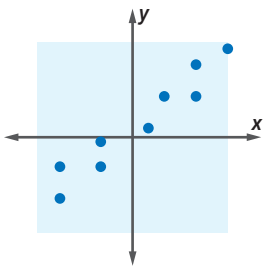
8.



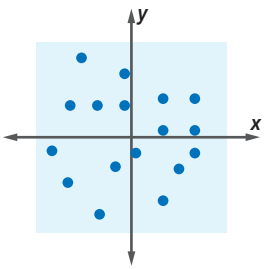
9.



10.



11.



REVIEW

12. Find an equation for the line with y-intercept -2 and x-intercept 4. **(Lesson 6-6)**

13. **a.** Find an equation for the line through $(6, 4)$ with a slope of $\frac{5}{3}$.
    b. Give the coordinates of one other point on this line. **(Lesson 6-5)**

14. **a.** Give the slope and y-intercept of the line $4x - 9y = 3$.
    b. Graph the line. **(Lesson 6-4)**

In 15 and 16, find the slope and y-intercept of the line. (Lesson 6-4)

15. $y = -\frac{x}{6}$

16. $y = 4$

17. **a.** Draw a line that has no y-intercept.
    b. What is the slope of your line? **(Lesson 6-4)**

18. The Tigers scored 15 points in the first 8 minutes of a basketball game. At that rate, how many points would the team score in a 48-minute game? **(Lesson 5-8)**

19. In a game of chess, each player begins the game with 8 pawns, 2 knights, 2 bishops, 2 rooks, 1 queen, and 1 king. What is the ratio of bishops to the total number of pieces? **(Lesson 5-5)**

20. Solve for y. **(Lesson 4-4)**
    a. $y - 3 = 6y + 2$ **b.** $y - a = by + c$

21. Describe all solutions to the following sentences. **(Lesson 2-8)**
    a. Six times a number is zero.
    b. Zero times a number is zero.
    c. Zero times a number is six.

22. If a is the cost of an adult's ticket to an amusement park and c is the cost of a child's ticket, what is the cost of tickets for 2 adults and 5 children? **(Lesson 1-2)**

Prior to the founding of the World Chess Federation in 1924, chess had existed as a sport played at a competitive level for centuries.

Source: World Chess Federation

EXPLORATION

23. The method of least squares was independently discovered by two mathematicians, the German Carl Friedrich Gauss and the Frenchman Adrien-Marie Legendre, in the years 1795–1809. Research to find the problem that led Gauss to this discovery.