

## B.3 Least Squares Regression

### ▶ What you should learn

- Use the sum of squared differences to measure how well a model fits a set of data.
- Find a least squares regression line for a set of data.
- Find a least squares regression parabola for a set of data.

### ▶ Why you should learn it

The method of least squares provides a way of creating mathematical models for a set of data, which can then be analyzed. For instance, in Exercise 9 on page B15, you will find the least squares regression line for the quantity of college textbooks sold in the United States from 2000 to 2003.

In many of the examples and exercises in the text, you have been asked to use the *regression* feature of a graphing utility to find mathematical models for sets of data. The *regression* feature of a graphing utility uses the **method of least squares** to find a mathematical model for a set of data. As a measure of how well a model fits a set of data points

$$\{(x_1, y_1), (x_2, y_2), (x_3, y_3), \dots, (x_n, y_n)\}$$

you can add the squares of the differences between the actual  $y$ -values and the values given by the model to obtain the **sum of the squared differences**. For instance, the table shows the heights  $x$  (in feet) and the diameters  $y$  (in inches) of eight trees. The table also shows the values of a linear model  $y^* = 0.54x - 29.5$  for each  $x$ -value. The sum of squared differences for the model is 51.7.

$x$	70	72	75	76	85	78	77	80
$y$	8.3	10.5	11.0	11.4	12.9	14.0	16.3	18.0
$y^*$	8.3	9.38	11.0	11.54	16.4	12.62	12.08	13.7
$(y - y^*)^2$	0	1.2544	0	0.0196	12.25	1.9044	17.8084	18.49

The model that has the least sum of squared differences is the **least squares regression line** for the data. The least squares regression line for the data in the table is  $y \approx 0.43x - 20.3$ . The sum of squared differences is 43.3.

To find the least squares regression line  $y = ax + b$  for the points  $\{(x_1, y_1), (x_2, y_2), (x_3, y_3), \dots, (x_n, y_n)\}$  algebraically, you need to solve the following system for  $a$  and  $b$ .

$$\begin{cases} nb + \left(\sum_{i=1}^n x_i\right)a = \sum_{i=1}^n y_i \\ \left(\sum_{i=1}^n x_i\right)b + \left(\sum_{i=1}^n x_i^2\right)a = \sum_{i=1}^n x_i y_i \end{cases}$$

In the system,

$$\sum_{i=1}^n x_i = x_1 + x_2 + \dots + x_n$$

$$\sum_{i=1}^n y_i = y_1 + y_2 + \dots + y_n$$

$$\sum_{i=1}^n x_i^2 = x_1^2 + x_2^2 + \dots + x_n^2$$

$$\sum_{i=1}^n x_i y_i = x_1 y_1 + x_2 y_2 + \dots + x_n y_n.$$

**Example 1** ▶ Finding a Least Squares Regression Line

Find the least squares regression line for the points  $(-3, 0)$ ,  $(-1, 1)$ ,  $(0, 2)$ , and  $(2, 3)$ .

**Solution**

Begin by constructing a table like that shown below.

$x$	$y$	$xy$	$x^2$
-3	0	0	9
-1	1	-1	1
0	2	0	0
2	3	6	4
$\sum_{i=1}^n x_i = -2$	$\sum_{i=1}^n y_i = 6$	$\sum_{i=1}^n x_i y_i = 5$	$\sum_{i=1}^n x_i^2 = 14$

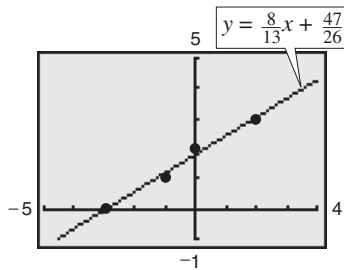


FIGURE B.9

Applying the system for the least squares regression line with  $n = 4$  produces

$$\begin{cases} nb + \left(\sum_{i=1}^n x_i\right)a = \sum_{i=1}^n y_i \\ \left(\sum_{i=1}^n x_i\right)b + \left(\sum_{i=1}^n x_i^2\right)a = \sum_{i=1}^n x_i y_i \end{cases} \quad \longrightarrow \quad \begin{cases} 4b - 2a = 6 \\ -2b + 14a = 5 \end{cases}$$

Solving this system of equations produces  $a = \frac{8}{13}$  and  $b = \frac{47}{26}$ . So, the least squares regression line is  $y = \frac{8}{13}x + \frac{47}{26}$ , as shown in Figure B.9.

**CHECKPOINT**

Now try Exercise 5.

The least squares regression parabola  $y = ax^2 + bx + c$  for the points

$$\{(x_1, y_1), (x_2, y_2), (x_3, y_3), \dots, (x_n, y_n)\}$$

is obtained in a similar manner by solving the following system of three equations in three unknowns for  $a$ ,  $b$ , and  $c$ .

$$\begin{cases} nc + \left(\sum_{i=1}^n x_i\right)b + \left(\sum_{i=1}^n x_i^2\right)a = \sum_{i=1}^n y_i \\ \left(\sum_{i=1}^n x_i\right)c + \left(\sum_{i=1}^n x_i^2\right)b + \left(\sum_{i=1}^n x_i^3\right)a = \sum_{i=1}^n x_i y_i \\ \left(\sum_{i=1}^n x_i^2\right)c + \left(\sum_{i=1}^n x_i^3\right)b + \left(\sum_{i=1}^n x_i^4\right)a = \sum_{i=1}^n x_i^2 y_i \end{cases}$$

Fortunately, graphing utilities have built-in least squares *regression* features.

## B.3 Exercises

### VOCABULARY CHECK: Fill in the blanks.

1. A graphing utility uses the \_\_\_\_\_ of \_\_\_\_\_ to find a mathematical model for a set of data.
2. The \_\_\_\_\_ of the \_\_\_\_\_ measures how well a model fits a set of data points.
3. The \_\_\_\_\_ line for a set of data is the linear model that has least sum of squared differences.

In Exercises 1–4, you are given a set of data points and a linear model for the data. Find the sum of squared differences for the given linear model.

1.  $(-3, -1), (-1, 0), (0, 2), (2, 3), (4, 4)$   
 $y = 0.5x + 0.5$
2.  $(0, 2), (1, 1), (2, 2), (3, 4), (5, 6)$   
 $y = 0.8x + 2$
3.  $(-2, 6), (-1, 4), (0, 2), (1, 1), (2, 1)$   
 $y = -1.7x + 2.7$
4.  $(0, 7), (2, 5), (3, 2), (4, 3), (6, 0)$   
 $y = -1.2x + 7$

In Exercises 5–8, find the least squares regression line for the points. Verify your answer with a graphing utility.

5.  $(-4, 1), (-3, 3), (-2, 4), (-1, 6)$
6.  $(0, -1), (2, 0), (4, 3), (6, 5)$
7.  $(-3, 1), (-1, 2), (1, 2), (4, 3)$
8.  $(0, -1), (2, 1), (3, 2), (5, 3)$
9. **Book Sales** The quantity of college textbooks sold in the United States from 2000 to 2003 are represented by the ordered pairs  $(x, y)$ , where  $x$  represents the year, with  $x = 0$  corresponding to 2000 and  $y$  represents the quantity of books sold (in millions). Find the least squares regression line for the data. What is the sum of squared differences? (Source: [Book Industry Study Group, Inc.](#))  
 $(0, 83), (1, 86), (2, 88), (3, 89)$

10. **Cell Phone Calls** The average lengths of a cell phone call from 2000 to 2003 are represented by the ordered pairs  $(x, y)$ , where  $x$  represents the year, with  $x = 0$  corresponding to 2000 and  $y$  represents the average length of a call (in minutes). Find the least squares regression line for the data. What is the sum of squared differences? (Source: [Cellular Telecommunications & Internet Association](#))  
 $(0, 2.56), (1, 2.74), (2, 2.73), (3, 2.87)$

In Exercises 11–14, find the least squares regression parabola for the points. Verify your answer with a graphing utility.

11.  $(0, 0), (2, 4), (4, 2)$
12.  $(-2, 6), (-1, 2), (1, 3)$
13.  $(-1, 4), (0, 2), (1, 0), (3, 4)$
14.  $(-3, -1), (-1, 2), (1, 2), (3, 0)$