Like a Glove
Least Squares Regression

LEARNING GOALS
In this lesson, you will:
• Determine and interpret the least squares regression equation for a data set using a formula.
• Use interpolation to make predictions about data.
• Use extrapolation to make predictions about data.

KEY TERMS
• interpolation
• extrapolation
• least squares regression line

How do the nerve cells in your brain communicate with each other? Signals have to be sent all across the brain—from your eyes to your occipital lobe in the back of your brain, from your ears to your temporal lobe, and so on. How does this happen?

In a sense, your nerve cells actually communicate using shapes. When a nerve cell is activated, it releases chemical messengers called neurotransmitters. These messengers have specific shapes, and they fit like keys into the locks on the next cell receiving the message. This message tells the next cell what to do.

And this process happens trillions of times per day!
PROBLEM 1 Music, Anyone?

The table shows the percent of all recorded music sales that came from music stores for the years 1998 through 2004.

<table>
<thead>
<tr>
<th>Year</th>
<th>1998</th>
<th>1999</th>
<th>2000</th>
<th>2001</th>
<th>2002</th>
<th>2003</th>
<th>2004</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percent of Total Sales from Music Stores</td>
<td>50.8</td>
<td>44.5</td>
<td>42.4</td>
<td>39.7</td>
<td>36.8</td>
<td>33.2</td>
<td>32.5</td>
</tr>
</tbody>
</table>

1. Represent the data as ordered pairs with the percent of total sales that came from music stores as a function of time. Let \( x \) represent the number of years since 1998.

2. Use your calculator to construct a scatter plot of the data. Sketch the scatter plot on the coordinate plane. Label the axes.

3. Describe any patterns you see in the data.

4. Use a graphing calculator to calculate the linear regression equation for the data. Round the values to the nearest hundredth.
5. Interpret the equation of the line in terms of the problem situation.

If there is a linear association between the independent and dependent variables of a data set, you can use a linear regression to make predictions within the data set. Using a linear regression to make predictions within the data set is called interpolation.

6. Use your equation to predict the percent of total music sales that came from music stores in the year 2000.

7. Compare the predicted percent in 2000 to the actual percent in 2000.

8. Use your equation to predict the percent of total music sales that came from music stores in 2003.

10. Do you think a prediction made using interpolation will always be close to the actual value? Explain your reasoning.

To make predictions for values of \( x \) that are outside of the data set is called **extrapolation**.

11. Use the equation to predict the percent of total music sales that would come from music stores in:
   a. 2010.
   b. 2020.
   c. 1900.

12. Are these predictions reasonable based on the problem situation?
1. Suppose a data set is composed of the points (1, 3), (–3, –7), and (5, 7) on a coordinate plane.

2. Are these points collinear? How can you tell?

3. Determine the equation of a line passing through the points at (1, 3) and (5, 7). Graph this line on the coordinate plane.

4. Determine and graph the equation of a line passing through the points at:
   a. (–3, –7) and (5, 7).
   b. (–3, –7) and (1, 3).
5. Would you consider any of the three lines you just graphed to be a line that “best fits” the three points? If yes, explain your reasoning. If no, describe where the line of best fit should be drawn.

One method to determine the line of best fit, or linear regression line, is the method of least squares. A least squares regression line is the line of best fit that minimizes the squares of the distances of the points from the line.

For a least squares regression line, ensure the line is written in the form \( y = ax + b \).

To calculate \( a \) and \( b \), use the equations:

\[
 a = \frac{n \sum xy - (\sum x)(\sum y)}{n \sum x^2 - (\sum x)^2} \\
 b = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n \sum x^2 - (\sum x)^2}
\]

where \( x \) represents all \( x \)-values from the data set, \( y \) represents all \( y \)-values from the data set, and \( n \) represents the number of coordinate pairs in the data set.

Let’s use this formula to determine the least squares regression line using these points:

\((-3, -3), (1, 2), \text{ and } (3, 4)\)

Calculate the values of each part of the equation separately. Then put it all together.

Determine the number of coordinate points in the data set.

\( n = 3 \)

Determine the sum of all the \( x \)-values in the data set.

\( \Sigma x = -3 + 1 + 3 = 1 \)

Determine the sum of all the \( y \)-values in the data set.

\( \Sigma y = -3 + 2 + 4 = 3 \)
Determine the sum of the squares of the x-values.

\[ \Sigma x^2 = (-3)^2 + 1^2 + 3^2 = 19 \]

Determine the sum of the products of each coordinate pair.

\[ \Sigma xy = (-3 \cdot -3) + (1 \cdot 2) + (3 \cdot 4) = 23 \]

Determine the square of the sum of the x-values.

\[ (\Sigma x)^2 = 1^2 = 1 \]

Insert each part into the formulas to solve for the values of \( a \) and \( b \).

\[
a = \frac{n \Sigma xy - (\Sigma x)(\Sigma y)}{n \Sigma x^2 - (\Sigma x)^2} = \frac{(3)(23) - (1)(3)}{(3)(19) - 1} = \frac{66}{56} = 1.18
\]

\[
b = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n \Sigma x^2 - (\Sigma x)^2} = \frac{(3)(19) - (1)(23)}{(3)(19) - 1} = \frac{34}{56} = 0.61
\]

6. What is the equation of the line of best fit for the points given in the worked example?
7. Margie calculated the least squares linear regression for the worked example.

\[
\begin{align*}
\hat{a} &= \frac{n \sum xy - (\sum x)(\sum y)}{n \sum x^2 - (\sum x)^2} \\
&= \frac{(3)(23) - (1)(3)}{(3)(19) - 19} \\
&= \frac{66}{38} \\
&= 1.74
\end{align*}
\]

Explain to Margie why her solution is incorrect.

8. Calculate the least squares linear regression using the points from Question 1.
   a. Calculate the sums and values for each part of the equation.
   b. Calculate the values of \(a\) and \(b\).
   c. Write the equation of the line of best fit.
   d. Graph the line on the coordinate plane in Question 1. Does this line “fit” the data better than the others? Explain your reasoning.
The table shown displays the median weekly earnings for U.S. workers according to the number of years of schooling they received.

<table>
<thead>
<tr>
<th>Years of Schooling</th>
<th>Median Weekly Earnings (dollars)</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>444</td>
</tr>
<tr>
<td>12</td>
<td>626</td>
</tr>
<tr>
<td>13</td>
<td>712</td>
</tr>
<tr>
<td>14</td>
<td>767</td>
</tr>
<tr>
<td>16</td>
<td>1038</td>
</tr>
<tr>
<td>18</td>
<td>1272</td>
</tr>
<tr>
<td>22</td>
<td>1510</td>
</tr>
</tbody>
</table>

1. Calculate the equation of the least squares regression line. Define your variables.

2. Interpret the least squares regression equation in terms of this problem situation.
3. Predict the weekly earnings of a worker with 12 years of schooling using the least squares regression equation. How does this compare to the actual earnings?

4. Predict the weekly earnings of a doctor with 25 years of schooling using the least squares regression equation. How does this compare to the actual earnings?

Talk the Talk

1. Why are predictions made by extrapolation more likely to be inaccurate than predictions made by interpolation?

Be prepared to share your solutions and methods.