Unit 2 Lesson 3: 11 C – D Linear Regression

Regression is the method of fitting a line to a set of data and then finding the equation of the line. The line is often called the model. The regression line is often called the line of best fit.

There are several ways to fit a straight line to a data set. Two of them are:

1. The line of best fit “by eye”.

Step 1: Find \( \bar{x} \) and \( \bar{y} \).

Step 2: Graph \((\bar{x}, \bar{y})\) and draw a line through it which fits the trend of the data (approximately the same number of points should be above the line as below the line).

Step 3: Find the linear equation of the line.

Ex 1: Look at the scatter plot for the data below. Find the mean of Judge A’s and Judge B’s scores, plot on the scatter plot and then draw a line of best fit. Write an equation for your line.

<table>
<thead>
<tr>
<th>Competitor</th>
<th>P</th>
<th>Q</th>
<th>R</th>
<th>S</th>
<th>T</th>
<th>U</th>
<th>V</th>
<th>W</th>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>Judge A</td>
<td>5</td>
<td>6.5</td>
<td>8</td>
<td>9</td>
<td>4</td>
<td>2.5</td>
<td>7</td>
<td>5</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>Judge B</td>
<td>6</td>
<td>7</td>
<td>8.5</td>
<td>9</td>
<td>5</td>
<td>4</td>
<td>7.5</td>
<td>5</td>
<td>7</td>
<td>4.5</td>
</tr>
</tbody>
</table>

| Judge A    | | | | | | | | | | |
| Judge B    | | | | | | | | | | |

\( \bar{x} = 5.6 \)
\( \bar{y} = 6.35 \)

\( y = 4.5x + 1.8 \)
2. A formula exists for finding the least squares regression line for \( y \) on \( x \). It is:

\[
y - \bar{y} = s_y \frac{s_x}{S_x^2} (x - \bar{x})
\]

Remember from last class:

\[
S_{xy} = \frac{\sum (x - \bar{x})(y - \bar{y})}{n}
\]

\[
S_x = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}
\]

\[
\text{covariance}
\]

\[
\text{stddev of } x
\]

with simplifying algebraically:

\[
\frac{s_{xy}}{S_x^2} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\left(\sqrt{\sum (x - \bar{x})^2}\right)^2}
\]

**Ex 2:** Use the formulae for calculating \( m \) and \( c \) for the line of best fit through \((1, 3), (3, 5), \text{ and } (5, 6)\)

\[
y = mx + c
\]

<table>
<thead>
<tr>
<th>( x )</th>
<th>( y )</th>
<th>( x - \bar{x} )</th>
<th>( y - \bar{y} )</th>
<th>( (x - \bar{x})(y - \bar{y}) )</th>
<th>( (x - \bar{x})^2 )</th>
<th>( (x - \bar{x})(y - \bar{y}) \frac{1}{(x - \bar{x})^2} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>-2</td>
<td>-1.67</td>
<td>4</td>
<td>+3.33</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>0</td>
<td>0.333</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>2</td>
<td>1.333</td>
<td>4</td>
<td>2.67</td>
<td></td>
</tr>
</tbody>
</table>

\[
\bar{x} = 3, \quad \bar{y} = 14/3
\]

\[
\Sigma = 8, \quad \Sigma = 6
\]

The regression equation is used for estimating values.

It is usually reliable when data is close.

It can be unreliable when data is spread out.

**Ex 3:** Chef Jeff was trying out a new recipe for pizza. The data in the table shows how popular the new pizza was:

<table>
<thead>
<tr>
<th>Day number</th>
<th>1</th>
<th>3</th>
<th>5</th>
<th>7</th>
<th>9</th>
<th>11</th>
<th>13</th>
<th>15</th>
<th>17</th>
<th>19</th>
</tr>
</thead>
<tbody>
<tr>
<td># of pizzas sold</td>
<td>1</td>
<td>15</td>
<td>19</td>
<td>24</td>
<td>30</td>
<td>37</td>
<td>43</td>
<td>45</td>
<td>53</td>
<td>60</td>
</tr>
</tbody>
</table>

a. Use your GDC to calculate the regression line.

\[
y = 3x + 2.67
\]

b. Use your line to predict how many pizzas will be sold on day 8 and day 14.

\[
\begin{align*}
\text{Day 8} &= 3(8) + 2.67 = 26.67 \text{ pizzas} \\
\text{Day 14} &= 3(14) + 2.67 = 44.67 \text{ pizzas}
\end{align*}
\]
Making predictions: Consider the scatter plot and its line of best fit:

If we use the equation of the least squares regression line to predict weight values for height values in between the smallest and largest value points, we say we are interpolating (in between the poles). If we predict weight values outside the smallest and largest height values we say we are extrapolating (outside the poles). The accuracy of an interpolation depends on the linearity of the trend. We can judge this by calculating the correlation coefficient. The accuracy of an extrapolation depends on two factors: linearity and assumption that the trend will continue.

**TAKE CARE WHEN EXTRAPOLATING!**

**Ex 4:** The table below shows the sales for Hancock’s Electronics established late 1998.

<table>
<thead>
<tr>
<th>Year</th>
<th>1999</th>
<th>2000</th>
<th>2001</th>
<th>2002</th>
<th>2003</th>
<th>2004</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sales ($ x 10,000)</td>
<td>5</td>
<td>9</td>
<td>14</td>
<td>18</td>
<td>21</td>
<td>27</td>
</tr>
</tbody>
</table>

Let $t$ be the time in years from 1998 and $s$ be sales in $10,000$.

a) Find $r$. $r = .997$

b) Find the equation of the line of best fit using the GDC. $y = 4.29x + .667$

c) Predict the sales figures for the year 2006, giving your answers to the nearest $10,000$.

Comment on the reasonableness of your prediction.

$4.29(8) + .667 = 34.987 \approx 35 \times 10,000$

$\$350,000 This is relatively reasonable because correlation is very strong but it is extrapolated.