Section 5.3
The Least Squares Regression Line

A regression line is a line that describes the relationship between the explanatory variable and the response variable. It is used to make predictions.

The most common mathematical method for fitting the best fit line is the least squares method. The least squares regression line (LSRL) is a regression line that makes the vertical distances of the points in a scatter plot from the line as small as possible.

The least squares line is of the form: \( \hat{y} = bx + a \), where the slope \( b = r \frac{s_y}{s_x} \) and the y-intercept \( a = \bar{y} - b \bar{x} \). In these formulas:

- \( b \) is the slope
- \( a \) is the y-intercept
- \( r \) is the correlation coefficient
- \( s_x \) is the standard deviation of the x-values
- \( s_y \) is the standard deviation of the y-values
- \( \bar{x} \) is the mean of the x-values
- \( \bar{y} \) is the mean of the y-values

Example 1: Use the following statistics to find the equation of the LSRL.
\[
\bar{x} = 3, \quad s_x = 3, \quad \bar{y} = 4, \quad s_y = 5.29, \quad r = 0.189
\]

Example 2: Find the correlation coefficient given: \( b = 0.123, \quad s_x = 5.01, \quad s_y = 1.02 \)

Recall: \( b = r \frac{s_y}{s_x} \)

The correlation and regression line are both sensitive to extreme values.
We can create the LSRL using R. The command is: \( \text{lm}(y \sim x) \)

Example 3: For your science project you collected the following data for a particular bird species.

<table>
<thead>
<tr>
<th>Wing Length (cm)</th>
<th>Age (days)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.3</td>
<td>4</td>
</tr>
<tr>
<td>2.1</td>
<td>5</td>
</tr>
<tr>
<td>3.0</td>
<td>8</td>
</tr>
<tr>
<td>3.1</td>
<td>9</td>
</tr>
<tr>
<td>3.3</td>
<td>10</td>
</tr>
<tr>
<td>3.8</td>
<td>11</td>
</tr>
<tr>
<td>4.0</td>
<td>12</td>
</tr>
<tr>
<td>4.5</td>
<td>14</td>
</tr>
<tr>
<td>5.1</td>
<td>16</td>
</tr>
</tbody>
</table>

a. Find the least squares regression line for this data. Interpret the meaning of the value of the slope of the LSRL.

Command: \( \text{Result: } (\text{Intercept}) \ \text{wing} \)

b. You recently caught a bird of the same species and its wing length measured 3.5 cm. About how old is the bird?

If given an \( x \) far larger or far smaller than the other \( x \) values, the prediction for \( y \) will not be a very good one. This is called extrapolation.
The correlation \( r \), when squared, is called the **coefficient of determination**, \( r^2 \) where \( 0 \leq r^2 \leq 1 \).

The coefficient of determination tells us how certain we can be in making predictions with the line of best fit. It’s the fraction of variation in the \( y \) values that is explained by the regression line and the explanatory variable.

- When \( r^2 \) is close to 1, it implies that the model may be very useful.
- When \( r^2 \) is close to 0, it implies that the model may not be very useful.

**Example 4:** Calculate the coefficient of determination for the problem in Example 3. Interpret your result.

**Command:**

**Answer:**

**Example 5:** One of nature’s patterns connects the percent of adult birds in a colony that return from the previous year and the number of new adults that join the colony. Here are data for 13 colonies of sparrowhawks:

<table>
<thead>
<tr>
<th>Percent Return, ( x )</th>
<th>74</th>
<th>66</th>
<th>81</th>
<th>52</th>
<th>73</th>
<th>62</th>
<th>52</th>
<th>45</th>
<th>62</th>
<th>46</th>
<th>60</th>
<th>46</th>
<th>38</th>
</tr>
</thead>
<tbody>
<tr>
<td>New Adults, ( y )</td>
<td>5</td>
<td>6</td>
<td>8</td>
<td>11</td>
<td>12</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td>18</td>
<td>18</td>
<td>19</td>
<td>20</td>
<td>20</td>
</tr>
</tbody>
</table>

**a.** Make a scatter plot of the data that shows how the number of new adults relates to the percent of returning birds.

**Commands:**
b. Compute the LSRL. Provide an interpretation of the slope of this line.

Command: 

Result: 

Coefficients:

(Intercept) pr

c. Find the correlation coefficient for the relationship. Interpret this number.

Command: 

Answer: 

d. Find the coefficient of determination for the relationship. Interpret this number.

Command: 

Answer: 
Using Mosaic the commands are:

\[
\begin{align*}
\text{lm(Package Name$response variable~Package Name$explanatory variable)} \\
\text{cor(Package Name$response variable,Package Name$explanatory variable)} \\
\text{plot(Package Name$explanatory variable,Package Name$response variable)}
\end{align*}
\]

Example 6: Use the Utilities data from the mosaicData package in RStudio to find the LSRL of kwh (electricity usage) against temp (average temperature (F) for billing period). Describe the correlation? Find the coefficient of determination for the relationship.

Commands:  
  \[
  \text{Coefficients:}
  \]
  \[
  (\text{Intercept}) \quad \text{Utilities$temp}
  \]