Chapter 10: Correlation and Regression

Diana Pell

Section 10.2: Regression

After the scatter plot is drawn and a linear relationship is determined, the next steps are to compute the value of the correlation coefficient and to test the significance of the relationship. If the value of the correlation coefficient is significant, the next step is to determine the equation of the regression line, which is the data’s line of best fit.

In statistics, the equation of the regression line is written as $y' = a + bx$, where $a$ is the $y'$ intercept and $b$ is the slope of the line.

**Formulas for the Regression Line $y' = a + bx$**

$$a = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

where $a$ is the $y'$ intercept and $b$ is the slope of the line.
Exercise 1. Data shown is for car rental companies in the United States for a recent year.

<table>
<thead>
<tr>
<th>Company</th>
<th>Cars (in ten thousands)</th>
<th>Revenue (in billions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>63.0</td>
<td>$7.0</td>
</tr>
<tr>
<td>B</td>
<td>29.0</td>
<td>3.9</td>
</tr>
<tr>
<td>C</td>
<td>20.8</td>
<td>2.1</td>
</tr>
<tr>
<td>D</td>
<td>19.1</td>
<td>2.8</td>
</tr>
<tr>
<td>E</td>
<td>13.4</td>
<td>1.4</td>
</tr>
<tr>
<td>F</td>
<td>8.5</td>
<td>1.5</td>
</tr>
</tbody>
</table>

a) Draw the scatter plot for the variables (in the space next to the table).

b) Compute the value of the correlation coefficient.

\[ \Sigma x = 152.8 \]
\[ \Sigma y = 18.7 \]
\[ \Sigma xy = 682.77 \]
\[ \Sigma x^2 = 5259.26 \]
\[ \Sigma y^2 = 80.67 \]
c) State the hypotheses.

\[ H_0 : \rho = 0 \quad \text{and} \quad H_1 : \rho \neq 0 \]

d) Test the significance of the correlation coefficient at \( \alpha = 0.05 \), using Table I.

\[ n = 6 \quad \text{C.V.} = \pm 0.811 \]

\[ d.f. = n - 2 = 6 - 2 = 4 \]

\[ \alpha = 0.05 \quad \text{Reject the null hyp.} \]

e) Give a brief explanation of the type of relationship. Assume all assumptions have been met.

There is significant linear relationship between the number of car rentals and revenue.

f) Find the equation of the regression line (if possible) and graph the line on the scatter plot of the data.

\[ a = 0.396 \]

\[ b = 0.106 \]

\[ y' = 0.396 + 0.106x \]

\[ x = 15 \quad y' = 1.986 \]

\[ x = 40 \quad y' = 4.636 \]

\[ \text{Plot} \]

g) Use the equation of the regression line to predict the income of a car rental agency that has 200,000 automobiles.

\[ X \text{ values are in } 10,000, \rightarrow \frac{200,000}{10,000} = 20 \rightarrow X = 20 \]

\[ y' = 0.396 + 0.106(20) \]

\[ = 2.516 \]

When a rental agency has 200,000 automobiles, its revenue will be approx. \$2.516 billion.
Exercise 2. Find the equation of the regression line for the data in Exercise 2 Section 10.1 (Absences and Final Grades), and graph the line on the scatter plot.

\[ n = 7 \]
\[ \sum x = 57 \]
\[ \sum y = 511 \]
\[ \sum xy = 3745 \]
\[ \sum x^2 = 579 \]

\[ a = 102.493 \]
\[ b = -3.622 \]

\[ y' = 102.493 - 3.622x \]

Extrapolation, or making predictions beyond the bounds of the data, must be interpreted cautiously. For example, in 1979, some experts predicted that the United States would run out of oil by the year 2003. This prediction was based on the current consumption and known oil reserves at that time. However, since then, the automobile industry has produced many new fuel-efficient vehicles. Also, there are many as yet undiscovered oil fields. Remember that when predictions are made, they are based on present conditions or on the premise that present trends will continue. This assumption may or may not prove true in the future.