Chapter 9: Correlation and Regression

9.1 Correlation

In this section, we aim to answer the question: Is there a relationship between A and B?

Is there a relationship between the number of employee training hours and the number of on-the-job accidents? Is there a relationship between the number of hours a person sleeps and their reaction time? Is there a relationship between the number of hours a student spends studying for a calculus test and the student’s score on that calculus test?

Definition: a correlation is a relationship between two variables.

Typically, we take $x$ to be the independent variable. We take $y$ to be the dependent variable. Data is represented by a collection of ordered pairs $(x, y)$.

Mathematically, the strength and direction of a linear relationship between two variables is represented by the correlation coefficient. Suppose that there are $n$ ordered pairs $(x, y)$ that make up a sample from a population. The correlation coefficient $r$ is given by:

$$r = \frac{n \sum (xy) - (\sum x)(\sum y)}{\sqrt{n \sum x^2 - (\sum x)^2} \sqrt{n \sum y^2 - (\sum y)^2}}$$

This will always be a number between -1 and 1 (inclusive).

- If $r$ is close to 1, we say that the variables are positively correlated. This means there is likely a strong linear relationship between the two variables, with a positive slope.

- If $r$ is close to -1, we say that the variables are negatively correlated. This means there is likely a strong linear relationship between the two variables, with a negative slope.

- If $r$ is close to 0, we say that the variables are not correlated. This means that there is likely no linear relationship between the two variables, however, the variables may still be related in some other way.
The correlation coefficient of the population is denoted by $\rho$ – and is usually unknown.

Example 1: The time $x$ in years that an employee spent at a company and the employee’s hourly pay, $y$, for 5 employees are listed in the table below. Calculate and interpret the correlation coefficient $r$. Include a plot of the data in your discussion.

<table>
<thead>
<tr>
<th>$x$</th>
<th>$y$</th>
<th>$x^2$</th>
<th>$y^2$</th>
<th>$xy$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>21</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>35</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>38</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sum x$ =</td>
<td>$\sum y$ =</td>
<td>$\sum x^2$ =</td>
<td>$\sum y^2$ =</td>
<td>$\sum xy$ =</td>
</tr>
</tbody>
</table>

Hint: Calculate the numerator:

$$n \sum (xy) - (\sum x) (\sum y) =$$

Then calculate the denominator:

$$\sqrt{n \sum x^2 - (\sum x)^2} \sqrt{n \sum y^2 - (\sum y)^2} =$$

Now, divide to get $r =$

Interpret this result:

Example 2: The table below shows the number of absences, $x$, in a Calculus course and the final exam grade, $y$, for 7 students. Find the correlation coefficient and interpret your result.

<table>
<thead>
<tr>
<th>$x$</th>
<th>1</th>
<th>0</th>
<th>2</th>
<th>6</th>
<th>4</th>
<th>3</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>95</td>
<td>90</td>
<td>90</td>
<td>55</td>
<td>70</td>
<td>80</td>
<td>85</td>
</tr>
</tbody>
</table>

You may use the facts that (double check this for practice)

$$\sum x = 19, \quad \sum y = 565, \quad \sum x^2 = 75, \quad \sum y^2 = 46,775, \quad \sum xy = 1,380.$$
Example 3: The table below shows the height, \( x \), in inches and the pulse rate, \( y \), per minute, for 9 people. Find the correlation coefficient and interpret your result.

<table>
<thead>
<tr>
<th>( x )</th>
<th>68</th>
<th>72</th>
<th>65</th>
<th>70</th>
<th>62</th>
<th>75</th>
<th>78</th>
<th>64</th>
<th>68</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y )</td>
<td>90</td>
<td>85</td>
<td>88</td>
<td>100</td>
<td>105</td>
<td>98</td>
<td>70</td>
<td>65</td>
<td>72</td>
</tr>
</tbody>
</table>

You may use the facts that (double check this for practice)

\[
\sum x = 622, \quad \sum y = 773, \quad \sum x^2 = 43,206, \quad \sum y^2 = 68,007, \quad \sum xy = 53,336.
\]

Example 4: The table below shows the number of absences, \( x \), in a Calculus course and the final exam grade, \( y \), for 7 students. Find the correlation coefficient and interpret your result.

<table>
<thead>
<tr>
<th>( x )</th>
<th>1</th>
<th>0</th>
<th>2</th>
<th>6</th>
<th>4</th>
<th>3</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y )</td>
<td>85</td>
<td>80</td>
<td>70</td>
<td>55</td>
<td>90</td>
<td>90</td>
<td>95</td>
</tr>
</tbody>
</table>

Interpreting the Correlation Between Two Variables:

Suppose that you find a strong positive or negative correlation between two variables. Is there a cause-and-effect relationship between these variables?

- There could be a direct cause-and-effect relationship: that is, \( x \) causes \( y \).
- There could be a reverse cause-and-effect relationship: that is, \( y \) causes \( x \).
- There could be a third (or fourth? or more?) variable that leads to the relationship between \( x \) and \( y \).
- The “relationship” between \( x \) and \( y \) may just be a coincidence.
9.2 Linear Regression

If there is a “significant” linear correlation between two variables, the next step is to find the equation of a line that “best” fits the data. Such an equation can be used for prediction: given a new \( x \)-value, this equation can predict the \( y \)-value that is consistent with the information known about the data. This predicted \( y \)-value will be denoted by \( \hat{y} \). The line represented by such an equation is called the linear regression line.

The equation for a line is

\[
\hat{y} = mx + b,
\]

where \( m \) is the slope of the line and \( b \) is the \( y \)-intercept (the \( y \)-value for which \( x \) is 0).

In general, the regression line, will not pass through each data point. For each data point, there is an error: the difference between the \( y \)-value from the data and the \( y \)-value on the line, \( \hat{y} \). By definition, this linear regression line is such that the sum of the squares of the errors is the least possible. It turns out, given a set of data, there is only one such line. The slope \( m \) and \( y \)-intercept \( b \) are given by

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

**Examples:** Find the equation of the regression line for each of the two examples and two practice problems in section 9.1.

**Example 1:**
First, find the slope \( m \). Start by determining the numerator:

\[
n \sum xy - (\sum x)(\sum y)
\]

Next, find the denominator:

\[
n \sum (x^2) - (\sum x)^2
\]

Divide to obtain \( m = \)

Now, find the \( y \)-intercept: \( b = \frac{\sum y}{n} - m \frac{\sum x}{n} = \)

**Additional Questions:** Use the equations to (Ex 1) predict the hourly pay rate of an employee who has worked for 20 years, and (Ex 2) predict the test score for a student with 5 absences.