The Correlation Coefficient

Lelys Bravo de Guenni

April 22nd, 2015
Outline

The Correlation coefficient
  Positive Correlation
  Negative Correlation
  Properties of the Correlation Coefficient
  Non-linear association
  Outliers effect on the correlation
The Correlation coefficient

Positive Correlation

Suppose that the following scatter diagram corresponds to the standard units of the data. The signs of the quadrants are determined by the sign of the products. There are more points in the positive quadrants than in the negative quadrants. So the average is positive yielding a positive correlation.
In this case there are more points in the negative quadrants than in the positive quadrants. So the average is negative, yielding a negative correlation.
Properties of the Correlation Coefficient

Suppose you measure the correlation between the temperature in New York and that in Boston during a month. Do you expect to get the same correlation if you measure it in Celsius than if you measure it in Fahrenheit?

Recall the process of obtaining the correlation. The first step is to convert the samples of both variables to standard units. This implies that the correlation does not depend on units. Thus, no matter if you use Celsius of Fahrenheit, you will get the same correlation.

On the other hand, since the correlation depends on the product of the two variables, it does not matter whether you consider the correlation the temperature in NY and that in Boston or the correlation between the temperature in Boston and that in NY.
Correlation and Regression

- The correlation is not affected when the two variables are interchanged.

- The correlation is not changed if the same number is added to all the values of one of the variables.

- The correlation is not changed if all the values of one of the variables is multiplied by the same positive number. It will change sign if the number is negative.
Non-linear association

Correlation is useful only when measuring the degree of linear association between two variables. That is, how much the values from two variables cluster around a straight line.
The variables in this plot have an obvious non-linear association. Nevertheless, the correlation between them is 0.3. This is because the points are clustered around a curve that has the shape of a parabola and not a straight line.
Outliers effect on the correlation

Similarly, the presence of outliers produces an artificially low correlation between variables that have a high degree of linear association.
The correlation of the data in the plot is 0.944. When the points marked with crosses in red are considered, the correlation drops to 0.75.
Find the correlation coefficient for each of the following data sets:

**Example 1:**

<table>
<thead>
<tr>
<th>$x$</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>2</th>
<th>2</th>
<th>2</th>
<th>3</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>7</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

The average of $x$ is 2, the SD is 1. The average of $y$ is 3 and the SD is 2. Then, the standardized values are
Correlation and Regression

<table>
<thead>
<tr>
<th></th>
<th>$x$</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>$x \times y$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>$y$</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
</tbody>
</table>

and the average of the last column is -0.80.
Example 2:

\[
\begin{array}{c|cccccccccc}
  x & 1 & 1 & 1 & 1 & 2 & 2 & 2 & 3 & 3 & 4 \\
  y & 1 & 2 & 1 & 3 & 1 & 4 & 1 & 2 & 2 & 3 \\
\end{array}
\]

\(x\) has not changed, \(y\) has average equal to 2 and a SD equal to 1. Then, the standardized values are
and the average of the last row is 0.3.

<table>
<thead>
<tr>
<th>$x$</th>
<th>-1</th>
<th>-1</th>
<th>-1</th>
<th>-1</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>1</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>-1</td>
<td>0</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>2</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$x \times y$</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>
Example 3

<table>
<thead>
<tr>
<th>$x$</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>2</th>
<th>2</th>
<th>2</th>
<th>3</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td>6</td>
<td>8</td>
</tr>
</tbody>
</table>

We observe that $y = 2 \times x$ and so there is total linear association between $x$ and $y$, implying that the correlation is 1.
Problem: Guess the correlation coefficients in these scatter diagrams

a

b

c

d
Answers: a) +1.00; b) -0.50; c) +0.85; d) +0.15