Outline

The Correlation coefficient
   Definition
   Relationship between two variable
   Scatterdiagram
   The correlation coefficient
   Computing the correlation coefficient
The Correlation coefficient

Definition

The correlation gives a measure of the linear association between two variables. It is a coefficient that does not depend on the units that are used to measure the data and is bounded between -1 and 1. In this class we will consider the following points:

- Scatterplots
- Defining and computing the correlation
- The effect of changing the variables’ units
- The effect of non-linear trends and outliers
Relationship between two variable

Pearson considered the data corresponding to the heights of 1,078 fathers and their son’s at maturity. The plot appears on page 120 of the textbook. A list of these data is difficult to understand, but the relationship between the two variables can be visualized using a scatter diagram, where each pair father-son is represented as a point in a plane. The $x$-coordinate corresponds to the father’s height and the $y$-coordinate to the son’s.

We observe that the taller the father the taller son, as a general tendency. This corresponds to a positive association.

In this example we consider the height of the father as an independent variable and the height of the son as a dependent variable.
Scatterdiagram

The figure shows the scatter diagram of the temperature in 58 locations in Northern California on August 1950. These data are plotted against the elevations of the station where the measurement was recorded. We can see that there is a tendency for locations at high elevations to have lower temperatures compared to those at low elevations.
Correlation and Regression

August Temperatures vs Elevation in Northern California

Degrees

feet

2000 4000 6000 8000
The correlation coefficient

We have seen that mean and standard deviation provide a description of the behavior of a sample for a given variable. When we consider two variables we can calculate the mean and the standard deviation of each of them, but none of those four numbers will give a measure of the association between the two variables.

The correlation coefficient gives a measure of the linear association of two variables

That is, when the scatter plot of the two variables is very close to the straight line we have a correlation that is close to one.

A zero correlation corresponds to a diagram where the data are widely scattered around the line.
The correlation coefficient is usually denoted by $r$ and takes values between -1 and 1.

A negative coefficient means that the data are clustered around lines with a negative slope. That is, as one variable increases, the other one decreases.

The closer $r$ is to 1 the stronger the positive linear association between the variables.

The closer $r$ is to -1 the stronger the negative linear association between the variables.

When $r$ is equal to 1 or -1 there is total linear association between the variables, this implies that all points lie on a line. The slope of the line is positive for $r = 1$ and negative for $r = -1$. 
Computing the correlation coefficient

The procedure to compute the correlation coefficients is the following

1. Convert each variable to standard units

2. Calculate the average of the products

The result is the correlation coefficient. The formula is given by

\[ r = \text{average of } (x \text{ in standard units } \times y \text{ in standard units}) \]
Example  Consider the data

<table>
<thead>
<tr>
<th>$x$</th>
<th>1</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>5</td>
<td>9</td>
<td>7</td>
<td>1</td>
<td>13</td>
</tr>
</tbody>
</table>
Correlation and Regression

Each pair of numbers in the table corresponds to a subject.

1. Convert $x$ to standard units. The average of the $x$-values is 4, the SD is 2.

2. Convert $y$ to standard units. The mean of the $y$-values is 7 and the SD is 4.

3. Compute the products of the standard units of the $x$-values and the $y$-values.

$$0.75 - 0.25 0.00 - 0.75 2.25$$

4. Take the average of the products

$$r = \frac{0.75 - 0.25 + 0.00 - 0.75 + 2.25}{5} = 0.40$$
Chitons: Consider the data in the table

<table>
<thead>
<tr>
<th>Length</th>
<th>10.7</th>
<th>11.0</th>
<th>9.5</th>
<th>11.1</th>
<th>10.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Width</td>
<td>5.8</td>
<td>6.0</td>
<td>5.0</td>
<td>6.0</td>
<td>5.3</td>
</tr>
<tr>
<td>Length</td>
<td>10.7</td>
<td>9.9</td>
<td>10.6</td>
<td>10.0</td>
<td>12.0</td>
</tr>
<tr>
<td>Width</td>
<td>5.8</td>
<td>5.2</td>
<td>5.7</td>
<td>5.3</td>
<td>6.3</td>
</tr>
</tbody>
</table>

These correspond to the lengths and the widths of 10 chitons.

We have that the averages are 10.58 for the lengths and 5.64 for the widths. The standard deviations are 0.6734 and 0.3980 respectively. Thus the transformation to standard units yields.
Then $r = 0.969$ is the average of the products. This implies that there is a strong positive linear association between the length and the width of the chitons.