## 9.11 REGRESSION, CORRELATION

Suppose that we have two data samples of different statistical features of some population

$$x: \ x_1, x_2, \dots, x_n \qquad y: \ y_1, y_2, \dots, y_n$$

We are interested in the following question: Is there any relation between these two features of the population?

**Definition: Covariance**

*Covariance* of x, y is defined by

$$cov(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

**Definition: Pearson's correlation coefficient**

*Pearson's correlation coefficient* is defined by

$$r = \frac{cov(x, y)}{s(x)s(y)}$$

where $s(x)$ is the standard deviation of $x$ and $s(y)$ is the standard deviation of $y$.

Remark:

The coefficient $r$ has a value between $-1$ and $1$.

The value $1$ means that there is a total positive linear correlation between $x$ and $y$, $0$ that there is no linear correlation between $x$ and $y$, and $-1$ means that there is a total negative linear correlation between $x$ and $y$.

*Example 9.36*

For

$$x: \ 3, 3, 4, 5, 5 \qquad y: \ 5, 7, 6, 4, 8$$

$$\bar{x} = 4, \qquad \bar{y} = 6$$

$$cov(x, y) =$$

$$\frac{1}{5}\big((3-4)(5-6) + (3-4)(7-6) + (4-4)(6-6) + (5-4)(4-6) + (5-4)(8-6)\big)$$

$$= \frac{1}{5}(1 - 1 + 0 - 2 + 2) = 0$$

$$r = \frac{0}{s(x)s(y)} = 0.$$


Co-funded by the Erasmus+ Programme of the European Union

*Example 9.37*

For

$$x:\;\; 3,3,4,5,5 \qquad y:\;\; 5,8,6,6,10$$

$$\bar{x} = 4, \qquad \bar{y} = 7$$

$$cov(x,y) =$$

$$\frac{1}{5}\big((3-4)(5-7)+(3-4)(8-7)+(4-4)(6-7)+(5-4)(6-7)$$
$$+(5-4)(10-7)\big) = \frac{1}{5}(2-1+0-1+3) = \frac{3}{5}$$

$$s^2(x) =$$

$$\frac{1}{5}((3-4)^2+(3-4)^2+(4-4)^2+(5-4)^2+(5-4)^2)$$

$$= \frac{1}{5}(1+1+0+1+1) = \frac{4}{5}$$

$$s(x) = \sqrt{\frac{4}{5}} = \frac{2}{\sqrt{5}}.$$

$$s^2(y) =$$

$$\frac{1}{5}((5-7)^2+(8-7)^2+(6-7)^2+(6-7)^2+(10-7)^2)$$

$$= \frac{1}{5}(4+1+1+1+9) = \frac{16}{5}$$

$$s(y) = \sqrt{\frac{16}{5}} = \frac{4}{\sqrt{5}}.$$

$$r = \frac{\frac{3}{5}}{\frac{2}{\sqrt{5}}\cdot\frac{4}{\sqrt{5}}} = \frac{3}{8} = 0.375.$$

### 9.11.1 Linear regression

Suppose that we have two samples:

$$x:\ \ x_1, x_2, \dots, x_n, \qquad y:\ \ y_1, y_2, \dots, y_n.$$

By using *linear regression*, we model a relationship between two variables $x$ and $y$.

One of the variables $(x)$ is called independent (or explanatory) variable. The second variable $(y)$ is called dependent (or response) variable.

The relation between $x$ and $y$ is modelled using a linear function

$$y = ax + b$$

whose unknown parameters $a, b$ are estimated from the data.

To find $a, b$ we use the "least squares" method. This method builds the line which minimizes the squared distance of each point from this line. We call this line a line of best fit.

We must solve the following problem

$$\sum_{i=1}^{n} (y_i - ax_i - b)^2 \to \min, \ \ a, b = ?$$

It is possible to demonstrate that such minimizing problem has always solution $a, b$ given by

$$a = \frac{cov(x, y)}{s^2(x)}, \ \ b = \bar{y} - a\bar{x},$$

where

$$cov(x, y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

is covariance of $x, y$,

$$s^2(x) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

is sample variation of $x$ and

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i\,, \qquad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

are sample means of $x$ and $y$ respectively.


Co-funded by the
Erasmus+ Programme
of the European Union

*Example 9.38*

Consider data

$$x: \; 3, 5, 2, 2, 1, 4, 6, 1 \qquad y: \; 3, 4, 3, 4, 2, 5, 4, 3.$$

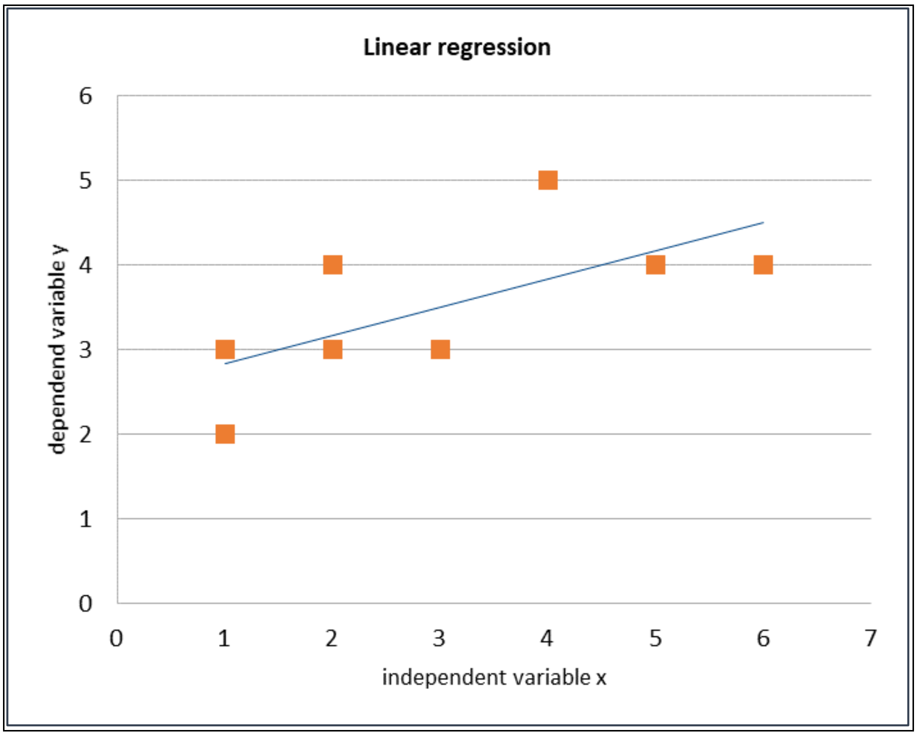


Figure 9.8. *Regression function* $y = \frac{1}{3}x - \frac{5}{2}$.

We have

$$cov(x, y) = 1, \;\; s^2(x) = 3, \;\; \bar{x} = 3, \;\; \bar{y} = \frac{7}{2}$$

$$a = \frac{cov(x, y)}{s^2(x)} = \frac{1}{3}, \qquad b = \bar{y} - a\bar{x} = \frac{7}{2} - \frac{1}{3} \cdot 3 = \frac{5}{2}.$$

A linear regression is $y = \frac{1}{3}x + \frac{5}{2}$. We can see its graph in the figure.

*Exercise 9.8*

Find a linear regression function for data:

$$x: \; 1, 2, 4 \qquad y: \; 0, 2, 1.$$

Solution:

$$y = \frac{3}{14}x + \frac{1}{2}.$$