Exploratory Statistical Data Analysis With R Software (ESDAR)

Swayam Prabha

Lecture 35

Correlation Coefficient

Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

Slides can be downloaded from

http://home.iitk.ac.in/~shalab/sp

Association of Two Variables

Consider continuous variables

- **Number of hours of study affect the marks obtained is an examination.**
- **Electricity/power consumption increases when the weather temperature increases.**
- **Weight of infants and small children increases as their height increases under normal circumstances.**

Two variables are associated and the nature of variables is continuous.

Correlation Coefficient

How to quantitatively measure the degree of linear relationship?

Use correlation coefficient.

It is based on the concepts of covariance and variance.

Correlation Coefficient

***X*, *Y*   :   Variables measured on contenuous scale.**

***X* and *Y* are linearly related.**

$Y = a + bX$ where a and b are unknown constant values.

Correlation is a statistical tool to study the linear relationship between two variables.

Correlation Coefficient

Two variables are said to be correlated if the change in one variable results in a corresponding change in the other variable.

If two variables deviate in the same direction, i.e., the increase (or decrease) in one variable results in a corresponding increase (or decrease) in the other, the correlation is said to be <u>positive</u> or variables are said to be <u>positively correlated</u>.

Correlation Coefficient

If two variables deviate in the opposite direction, i.e., as one variable increases, the other decreases and vice versa, the correlation is said to be negative or the variables are said to be negatively correlated.

Correlation Coefficient

If one variable changes and the other variable remains constant on average or there is no change in the other variable, the variables are said to be independent or they have no correlation.

Correlation Coefficient

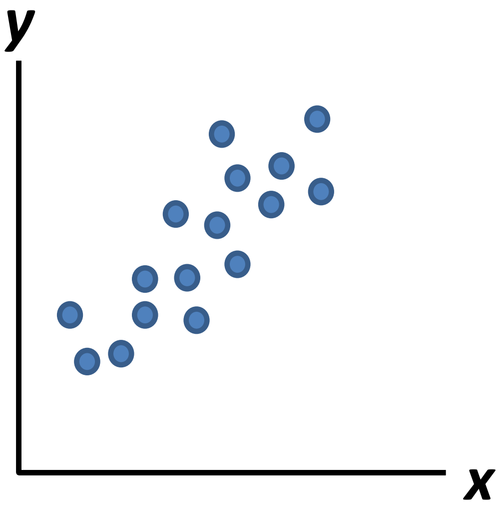


Fig. 1 : Positive correlation

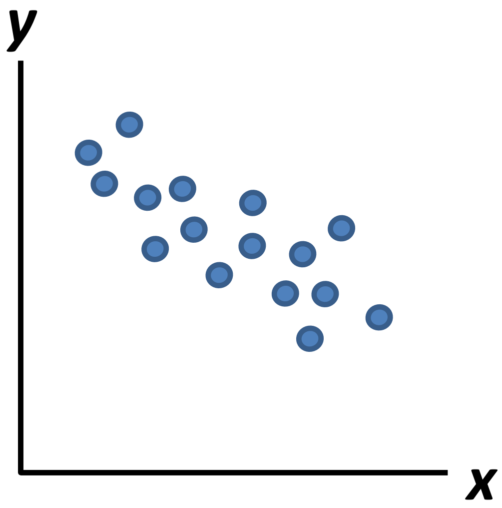


Fig. 2 : Negative correlation

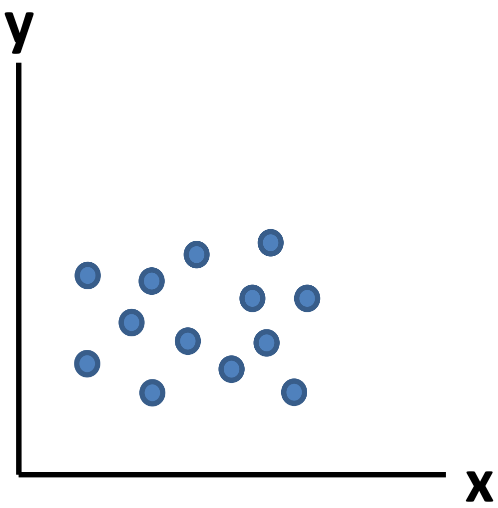


Fig. 3 : No correlation

Correlation Coefficient

Consider following two plots:

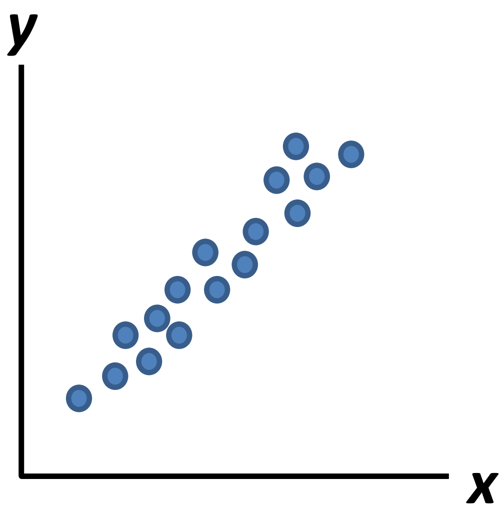


Fig A: strong positive correlation

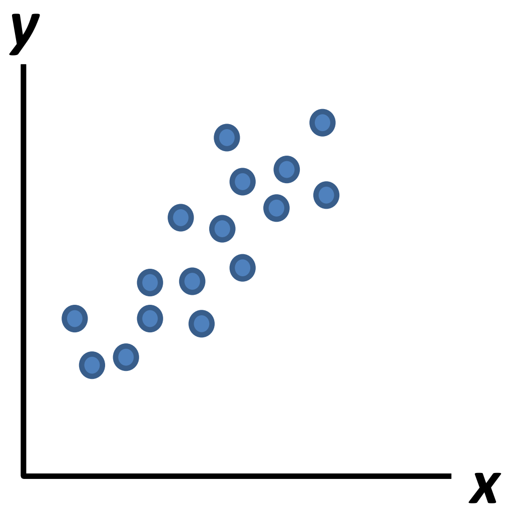


Fig. B: Moderate positive correlation

Covariance

What is covariance?

Recall variance.

When there is only one variable, variation exists.

When there are two variables, beside their individual variations, their co-variation also exists, provided they affect each other.

Covariance

X, Y **: Two variables**

n **pairs of observations are available as** $(x_1,y_1), (x_2,y_2),\ldots, (x_n,y_n)$

The covariance between the variables *X* and *Y* is defined as

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

where

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

Similar definition is available for grouped data in frequency table.

Covariance

R command:

`x,y` : Two data vectors

`cov(x,y)`: covariance between x and y.

Command `cov(x,y)` calculates the covariance with divisor (*n* – 1)

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

Coefficient of Correlation

Also called as Karl Pearson Coefficient of Correlation, Bravis-Pearson Correlation Coefficient or Product Moment Correlation Coefficient

$$r \equiv r(x,y) = \frac{\operatorname{cov}(x,y)}{\sqrt{\operatorname{var}(x).\operatorname{var}(y)}}$$

$$= \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2}\ \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2}}$$

$$= \frac{\sum_{i=1}^{n}x_i y_i - n\,\overline{x}\,\overline{y}}{\sqrt{\left(\sum_{i=1}^{n}x_i^{\,2}-n\overline{x}^2\right)\left(\sum_{i=1}^{n}y_i^{\,2}-n\overline{y}^2\right)}}$$

Coefficient of Correlation

r measures the degree of linear relationship

Limits of r : $-1 \le r \le 1$

$r > 0$: Indicates positive association between X and Y

$\Rightarrow$ X and Y are positively correlated.

$r < 0$: Indicates negative association between X and Y

$\Rightarrow$ X and Y are negatively correlated.

$r = 0$: Indicates no association between X and Y

$\Rightarrow$ X and Y are uncorrelated.

Coefficient of Correlation

Value of *r* has two components – sign and magnitude.

Sign of *r* indicates the nature of association.

+ sign of *r* indicates positive correlation. As one variable increases (or decreases), other variable also increases (or decreases).

– sign of *r* indicates negative correlation. As one variable increases (or decreases), other variable decreases (or increases).

Coefficient of Correlation

Magnitude of r indicates the degree of linear relationship.

$-1 \le r \le 1$

$r = 1 \quad \Rightarrow$ Perfect linear relationship

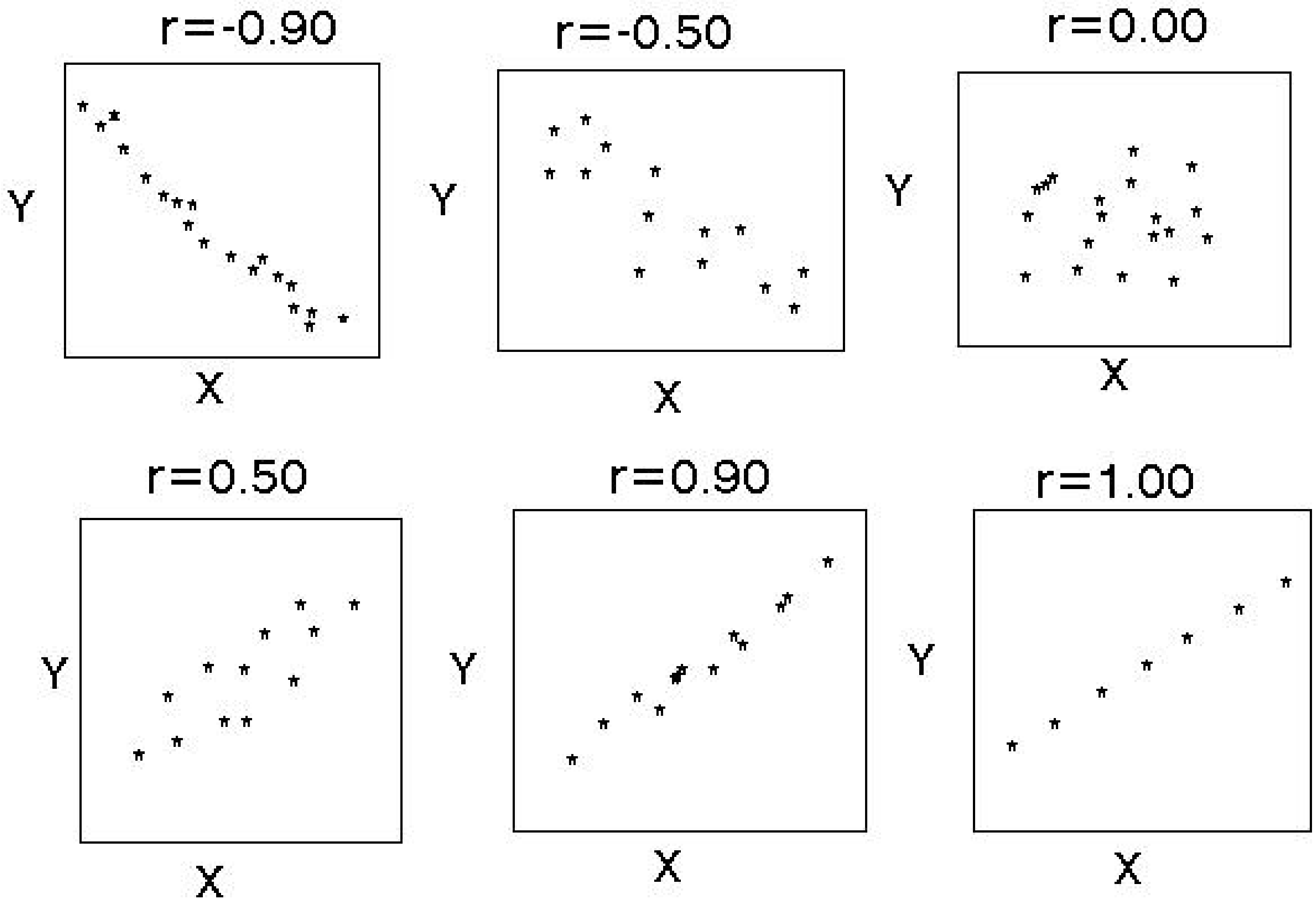
$r = 0 \quad \Rightarrow$ No correlation

Any other value of r between 0 and 1 indicates the degree of linear relationship.

$r = +1 \Rightarrow$ Perfect linear and increasing relationship between X and Y.

$r = -1 \Rightarrow$ Perfect linear and decreasing relationship between X and Y.

Coefficient of Correlation



Coefficient of Correlation

Value of *r* close to zero indicates that

- **the variables are independent**

  or

- **the relationship is nonlinear.**

If relationship between *X* and *Y* is nonlinear, then the degree of linear relationship may be low and *r* is then close to 0 even if the variables are clearly not independent.

So when *X* and *Y* are independent then *r*(*X*, *Y*) = 0 but not conversely true.

Coefficient of Correlation

Correlation coefficient is symmetric

$$r\,(X, Y) = r(Y, X)$$

Example:

Correlation coefficient between height and weight is the same as of the correlation between weight and height.

Coefficient of Correlation

Correlation coefficient is independent of units of measurement of *X* and *Y.*

<u>Example</u>:

- **One person measures height is meters and weight in kilograms. Finds correlation coefficient r_1**
- **Another person measures height in centimeters and weight in grams. Finds correlation coefficient r_2**

Then $r_1 = r_2$