Chapter 4: Correlation

Dr. Abbas Rammal

Bachelor's degree in Mathematics
Option: DATA SCIENCE
October 2023

Plan

1. Introduction
2. Scatter Plot
3. Correlation
4. Estimation of the correlation coefficient
5. Correlation coefficient and regression
6. Test on the correlation coefficient
7. Correlation matrix

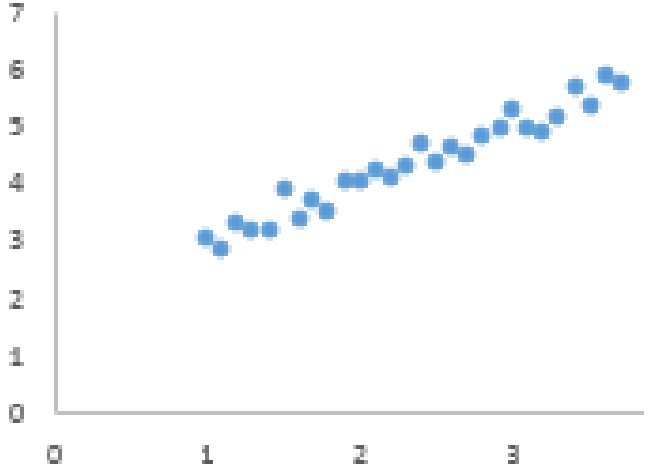
Introduction

- Correlation is a numerical index measuring the degree of connection or the intensity of the relationship between two variables.
- **Objective:** The objective of this chapter is to present the coefficient of correlation and to give the hypothesis tests on this coefficient

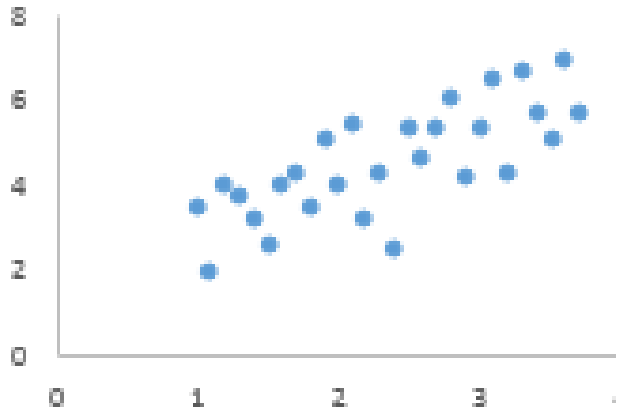
Scatter plot

- The simplest method for observing the relationship between Y and to represent the couples (x_i, y_i), for all the observations, in a two-dimensional graph.
- This graph is called the scatter plot.
- From the cloud of points, we can observe the type of relationship existing between X and Y

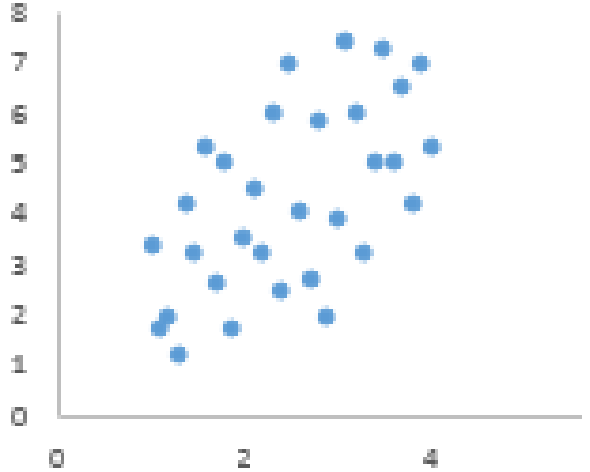
Strong linear correlation



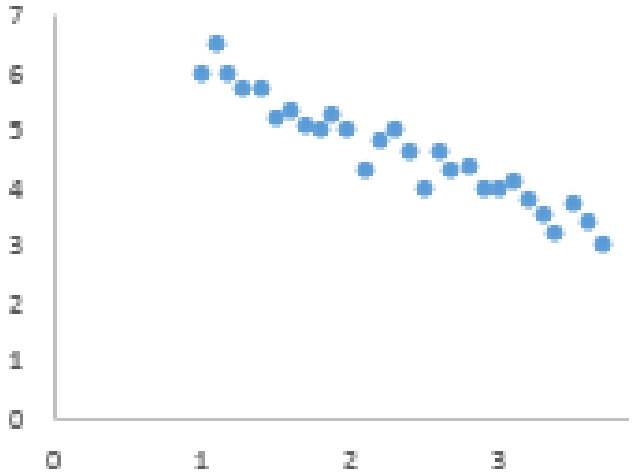
Mean linear correlation



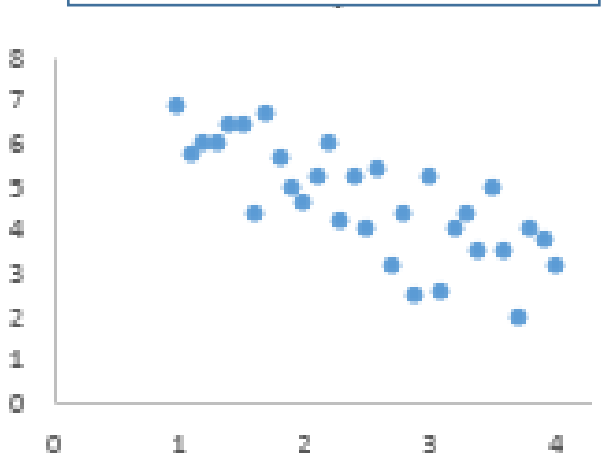
Weak linear correlation



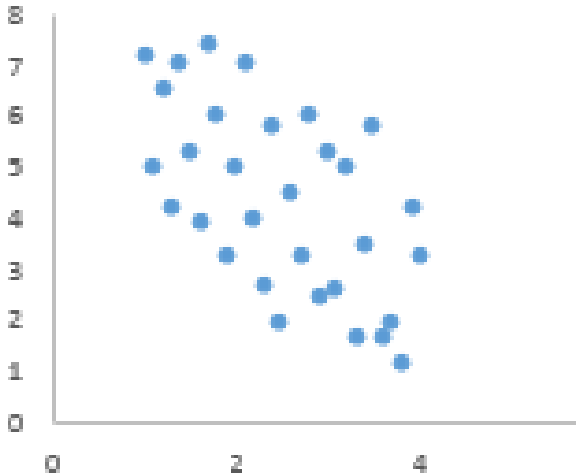
Strong linear correlation



Mean linear correlation



Weak linear correlation



Correlation

- Examining the scatter plot makes it possible to detect if there is a particular structure of association between the two variables. Two questions arise:

1. Can we quantify the strength of this association?
2. If the association between the two variables actually reflects a statistical dependence between Y and X, how can we define a representation of this particular relationship?

   Answer: In the case where the relationship between the two variables is linear, we use the linear correlation coefficient to answer the first question. Then, we introduce the regression line to answer the second question.

Linear Correlation Coefficient

- **Covariance**:
- We call covariance of a double statistical series (X, Y) where the variables X and Y are quantitative, the quantity denoted Cov(X, Y) defined by:

$$Cov(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

- **Linear correlation coefficient**:
- The “linear” correlation coefficient of two vectors X and Y, denoted ρ_{XY}, is defined by:

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sqrt{\mathbb{V}(X)\mathbb{V}(Y)}}$$

- This coefficient measures the degree of linear relationship between the vectors X and Y.

Properties of the coefficient ρ

- The linear correlation coefficient and the covariance take the same sign since the standard deviations are positive.

$$-1 \leq \rho \leq 1$$

- The linear correlation coefficient is independent of the units in which the variables are expressed. Let $Z = a + bX$, $b > 0$ we have:

$$\rho(X, Y) = \rho(Z, Y)$$

- The linear correlation coefficient is independent of the order in which the two variables are expressed.

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

Interpretation of $\boldsymbol{\rho}$

1) If $\boldsymbol{\rho = 1}$, X and Y are said to be perfectly positively correlated.
2) If $\boldsymbol{\rho = -1}$, X and Y are said to be perfectly negatively correlated.
3) If $\boldsymbol{\rho = 0}$, X and Y are not linearly correlated.
4) If $|\boldsymbol{\rho}| \to 0$ X and Y are weakly correlated.
5) If $|\boldsymbol{\rho}| \to 1$, X and Y are strongly correlated and we can consider the existence of a linear relationship between the two variables.

Disadvantages of ρ

- The correlation coefficient has some disadvantages:
- There may be a strong non-linear relationship between two variables X and Y without their correlation coefficient being high.
- $\rho = 0$ does not mean that X and Y are independent, but that their relationship is not linear.
- On the other hand, if X and Y are independent then $\rho(X, Y) = 0$.

Remarks:

- The correlation coefficient does not distinguish between the variable explained and the explanatory variables.
- A measure of correlation is not a measure of causality.
- It does not tell us whether it is the first variable which influences the second or the opposite or even if the relationship between these two variables is due to the joint influence of a third variable.
- Correlation measures the intensity of the connection between variables, while regression analyzes the relationship of one variable in relation to one or more others.

Estimation of the correlation coefficient

- From a sample of size n: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$we estimates the correlation coefficient $\boldsymbol{\rho}(\boldsymbol{X}, \boldsymbol{Y})$ by r_{xy} defined as follows:

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

Example

- We consider the following data:

x_i	-4	-3	-2	-1	0	1	2	3	4

- Let $Y = 2X + 3$ and $Z = X^2$. Find the different coefficients of correlations between these variables.

Correlation coefficient and regression

- The Correlation Coefficient is closely linked to the coefficient of determination of a simple regression and to the estimation of the slope of the regression line.

- Link between $\mathrm{r_{xy}}$ and β_1:

$$\hat{\beta}_1 = r_{xy} . \frac{s_y}{s_x}$$

- Link between $\mathrm{r_{xy}}$ and R^2:

$$R^2 = r_{xy}^2$$

Test on the correlation coefficient

- Let X and Y be two random variables. We notice:

  $\mathrm{R_{xy}}$ is the test statistic for the correlation coefficient $\boldsymbol{\rho}(\boldsymbol{X}, \boldsymbol{Y})$. $\mathrm{r_{xy}}$ is a realization of the random variable $\mathrm{R_{xy}}$

- Let the hypotheses be:

$$\mathcal{H}_0 : \rho(X, Y) = 0 \quad \mathcal{H}_1 : \rho(X, Y) \neq 0$$

- Under H_0, the test statistic

$$T = \frac{R_{xy}\sqrt{n-2}}{\sqrt{1 - R_{xy}^2}} \rightsquigarrow t(n-2)$$

- The rejection region is:

$$RR :]-\infty, -t_{(1-\alpha/2,(n-2))}[\cup]t_{(1-\alpha/2,(n-2))}, +\infty[$$

- If $T_{obs} = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}} \in RR$ then we reject $H_0 \Rightarrow$ There exists a linear relationship between X and Y

Example

- The following table gives the number of vehicles in circulation x_i and the number of traffic accidents y_i for 6 cantons:

Véhicules x_i	Accidents y_i
132 981	3 335
112 172	2 224
109 543	1 758
108 064	1 661
103 324	1 941
78 795	2 391

- In this example, there is not a linear relationship between the number of vehicles in circulation (X) and the number of traffic accidents (Y)

Partial Fisher Test

- When we have more than two variables (p variables), we can calculate the $\frac{p(p-1)}{2}$ correlation coefficients between these variables taken two by two.
- We can form the correlation matrix.

$$\begin{pmatrix} 1 & r_{\{x_1x_2\}} & r_{\{x_1x_3\}} & \cdots & r_{\{x_1x_p\}} \\ r_{\{x_1x_2\}} & 1 & r_{\{x_2x_3\}} & \cdots & r_{\{x_2x_p\}} \\ r_{\{x_1x_3\}} & r_{\{x_2x_3\}} & 1 & \cdots & r_{\{x_3x_p\}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{\{x_1x_p\}} & r_{\{x_2x_p\}} & r_{\{x_3x_p\}} & \cdots & 1 \end{pmatrix}$$