UNIT 05

Regression and Correlation

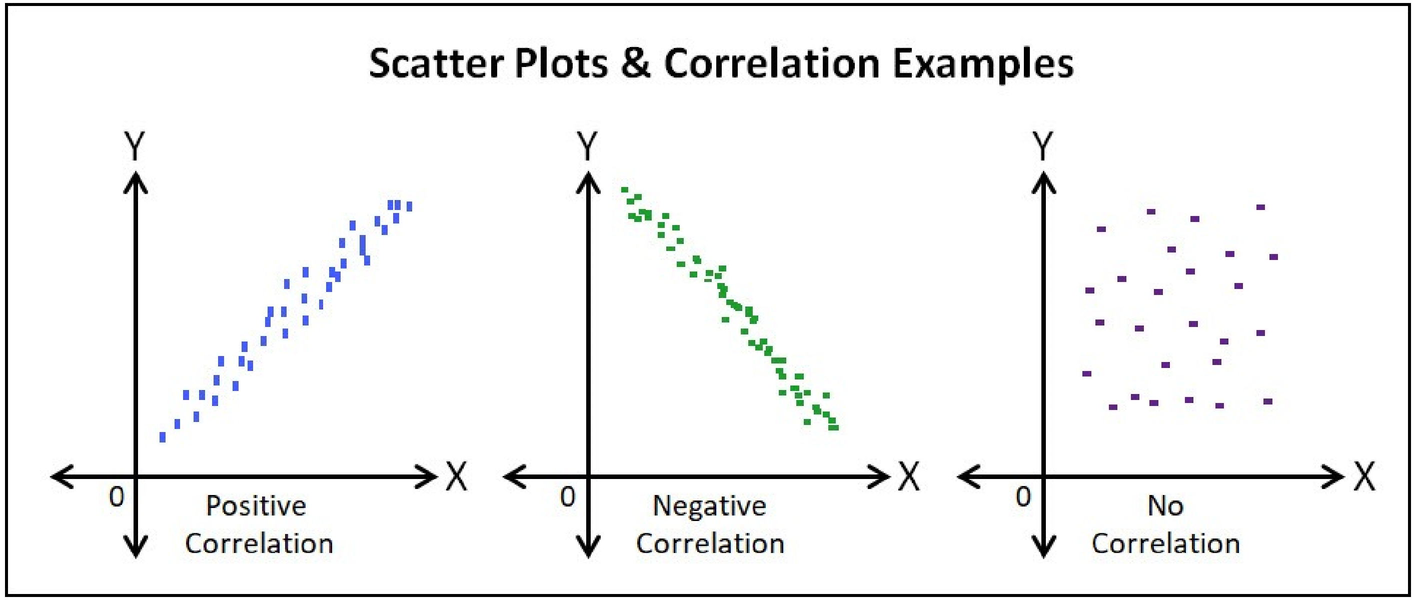
Correlation

Correlation is a statistical measure that describes the extent to which two variables are linearly related. It quantifies the direction and strength of the relationship between these variables.

Regression and Correlation

Scatter diagram/ Scatterplot

Scatter plots or Scatter diagram are the graphs that present the relationship between two variables in a data-set. It represents data points on a two-dimensional plane or on a Cartesian system. The independent variable or attribute is plotted on the X-axis, while the dependent variable is plotted on the Y-axis.



Correlation coefficient

Correlation is a statistical measure that describes the extent to which two variables are linearly related. It quantifies the direction and strength of the relationship between these variables.

The degree of association is measured by a **correlation coefficient**, denoted by ***r***. It is sometimes called Pearson's correlation coefficient. The correlation coefficient is measured on a scale that varies from + 1 through 0 to – 1.

i.e. $-1 \leq r \leq 1$

The correlation coefficient is measured/calculated by:

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Regression and Correlation

- A linear correlation coefficient that is greater than zero indicates a positive relationship.
- A value that is less than zero signifies a negative relationship.
- A value of zero indicates no relationship between the two variables x and y.
- Complete correlation between two variables is expressed by either + 1 or -1.
- When one variable increases as the other increases the correlation is positive
- When one decreases as the other increases it is negative.

Regression and Correlation

Regression Analysis

Regression analysis is a technique of studying the dependence of one variable (called dependent variable) on one or more variables (called explanatory variables) with a view to estimating or predicting the average value of the dependent variable in terms of the known or fixed values of the independent variable.

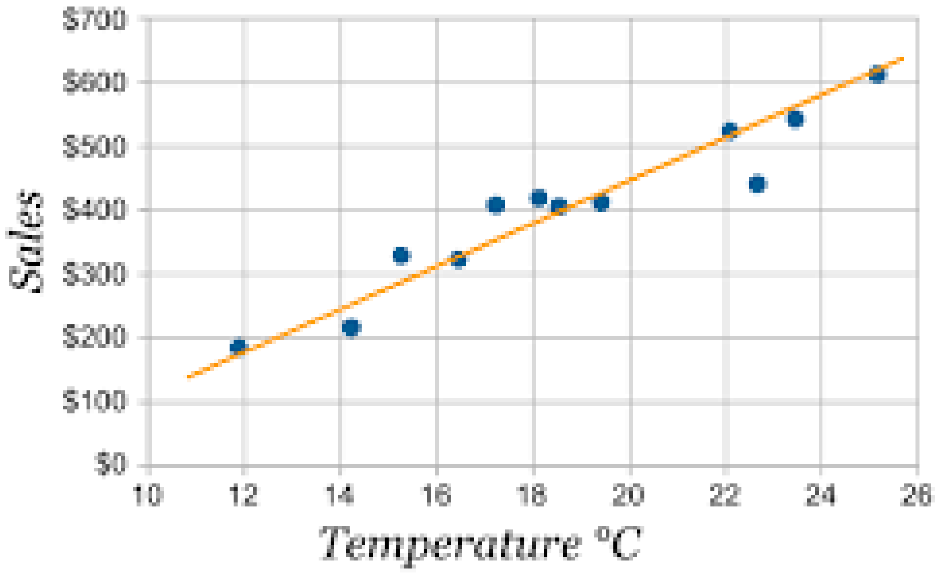
Regression Analysis:

i. Linear Regression
ii. Multiple linear Regression
iii. Non-linear Regression

Regression and Correlation

Line of best fit

Line of best fit refers to a line through a scatter plot of data points that best expresses the relationship between those points.



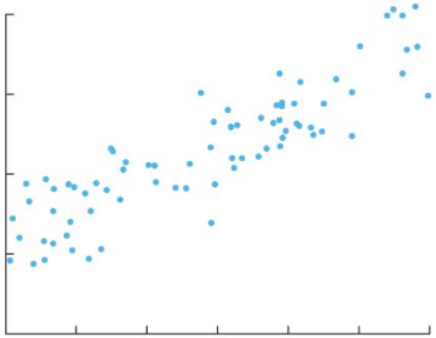
Simple regression line (The least Square method)

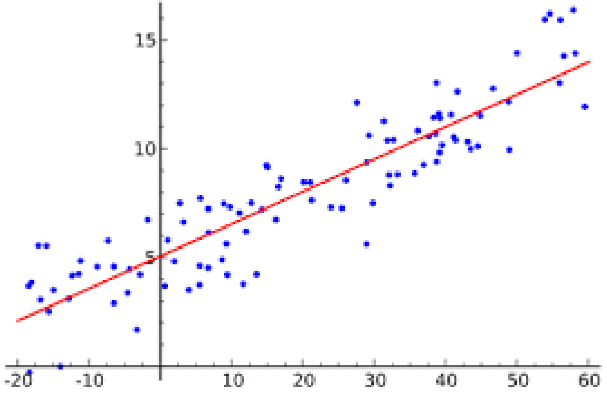
There are many lines that could possibly be drawn on a given scatter diagram. Which one then should we choose as the estimating line? Of course, the best fitted line. The most commonly used method for selecting such a line is the least-squares method and the resulting line is called the least -square lines.

The general form of a straight line is $\boldsymbol{y = a + bx}$, where *a* and *b* are constants to be found

Regression and Correlation

Simple regression line (The least Square method)





$y = a + bx$

$$b = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$\bar{y} = a + b\bar{x}$$

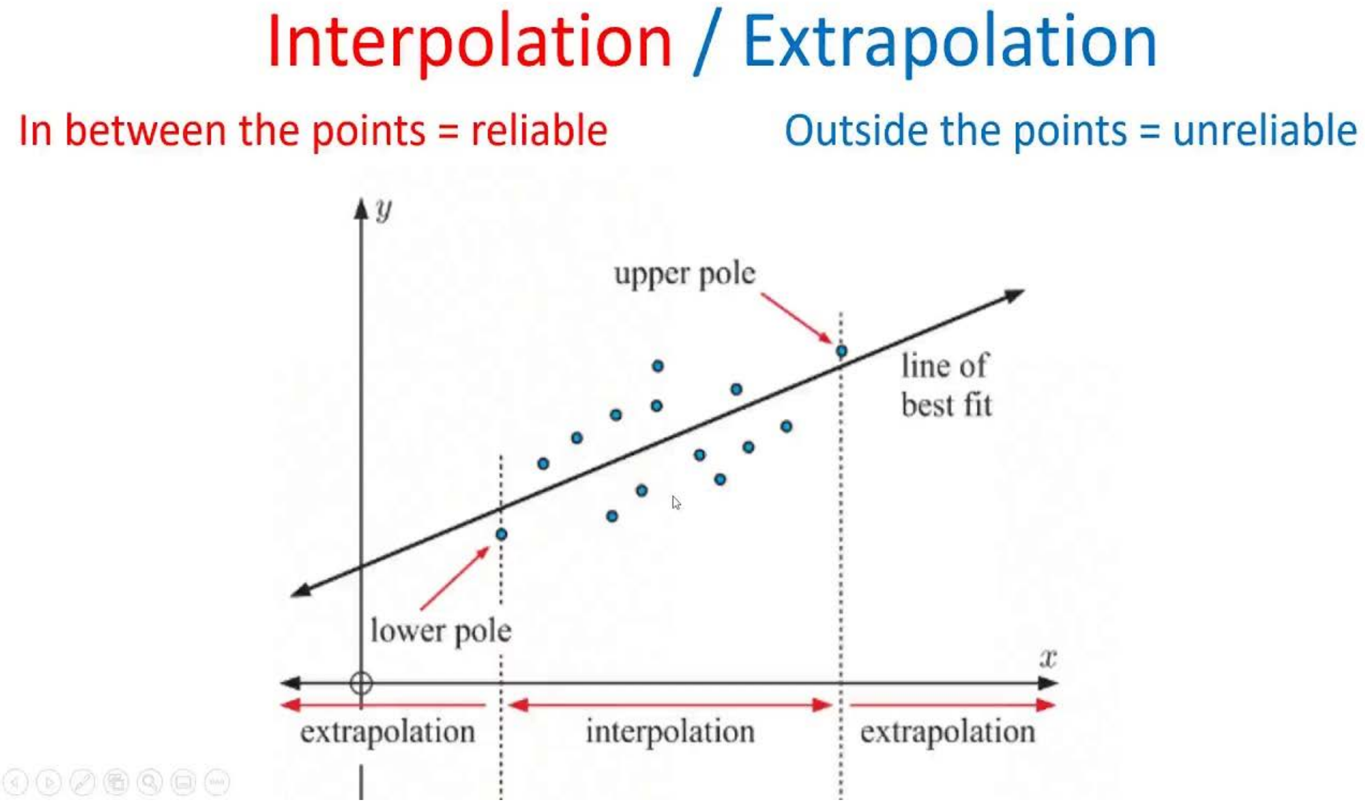
Extrapolation and Interpolation

Extrapolation is an estimation of a value based on extending a known sequence of values or facts beyond the area that is certainly known.

Interpolation is an estimation of a value within two known values in a sequence of values.

Regression and Correlation

- **Extrapolation and Interpolation**



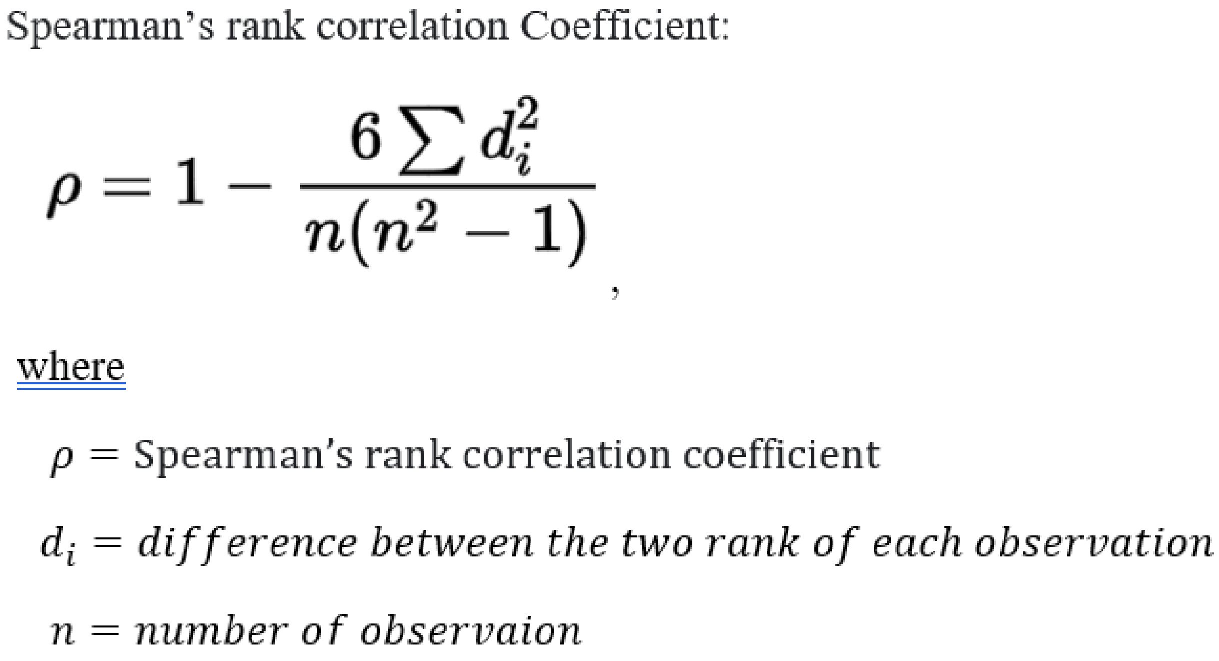
Interpolation / Extrapolation

In between the points = reliable

Outside the points = unreliable

Rank Correlation

A **rank correlation coefficient** measures the degree of similarity between two rankings and can be used to assess the significance of the relation between them. Spearman's correlation coefficient, (ρ) measures the strength and direction of association between two ranked variables.



Spearman's rank correlation Coefficient:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)},$$

where

ρ = Spearman's rank correlation coefficient

$d_i = difference\ between\ the\ two\ rank\ of\ each\ observation$

$n = number\ of\ observaion$

Rank coefficient of correlation value lies between –1 and +1. Symbolically, –1≤ρ≤+1

R_X	R_Y	$d = R_X - R_Y$	d^2
1	1	0	0
2	4	-2	4
3	2	1	1
4	5	-1	1
5	3	2	4
6	9	-3	9
7	7	0	0
8	10	-2	4
9	6	3	9
10	8	2	4
			$\sum d^2 = 36$

The rank correlation is given by

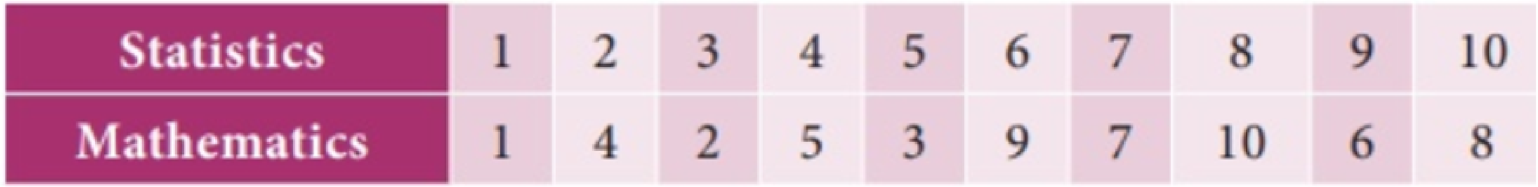
$$\rho = 1 - \frac{6\Sigma d^2}{N(N^2-1)} = 1 - \frac{6(36)}{10(10^2-1)}$$

$$= 1-0.218$$

$$\therefore \qquad \rho = 0.782$$

Example:

The following are the ranks obtained by 10 students in Statistics and Mathematics



Statistics	1	2	3	4	5	6	7	8	9	10
Mathematics	1	4	2	5	3	9	7	10	6	8

Find the rank correlation coefficient.