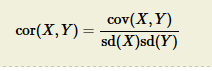
**Correlation**

**Introduction:**

Correlation refers to the extent to which two variables have a linear relationship with each other. It is a statistical technique that can show whether and how strongly variables are related. It is a scaled version of covariance and values ranges from -1 to +1.

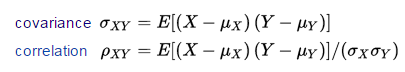
It can be calculated as



**Difference between Covariance and Correlation:**

In [probability theory](https://en.wikipedia.org/wiki/Probability_theory) and [statistics](https://en.wikipedia.org/wiki/Statistics), the mathematical concepts of covariance and correlation are very similar. Both describe the degree to which two [random variables](https://en.wikipedia.org/wiki/Random_variable) or [sets](https://en.wikipedia.org/wiki/Set_(mathematics)) of random variables tend to deviate from their [expected values](https://en.wikipedia.org/wiki/Expected_value) in similar ways.

If X and Y are two [random variables](https://en.wikipedia.org/wiki/Random_variables), with [means](https://en.wikipedia.org/wiki/Mean) μX and μY, and standard deviations σX and σY, respectively, then their covariance and correlation are as follows:

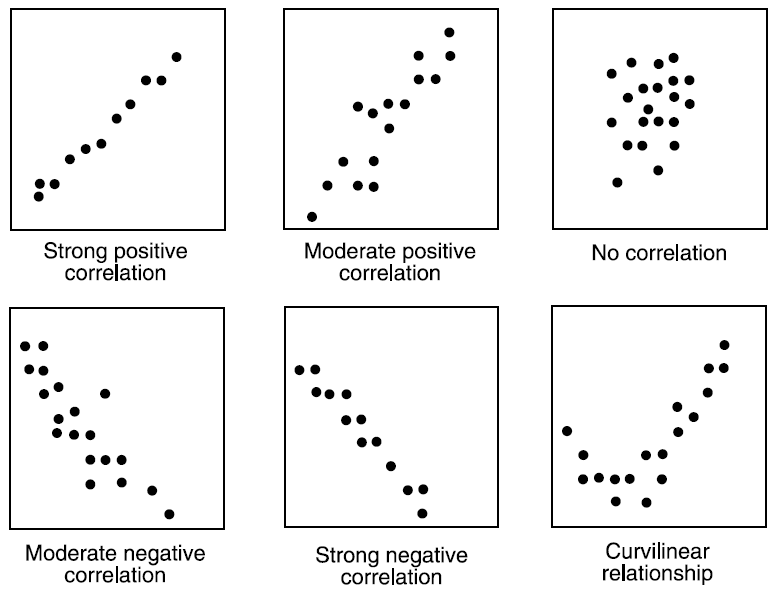


|  |
| --- |
|  |

Where E is the expected value operator. Notably, correlation is [dimensionless](https://en.wikipedia.org/wiki/Dimensionless) while covariance is in units obtained by multiplying the units of the two variables. The correlation of a variable with itself is always 1 (except in the [degenerate case](https://en.wikipedia.org/wiki/Degenerate_case) where the two variances are zero, in which case the correlation does not exist).

**Ways to detect Correlation between variables:**

1. Graphical Method: While doing bi-variate analysis between two continuous variables, we should look at scatter plot. It is a nifty way to find out the relationship between two variables. The pattern of scatter plot indicates the relationship between variables. The relationship can be linear or non-linear.

[](https://www.analyticsvidhya.com/wp-content/uploads/2015/02/Data_exploration_4.png)

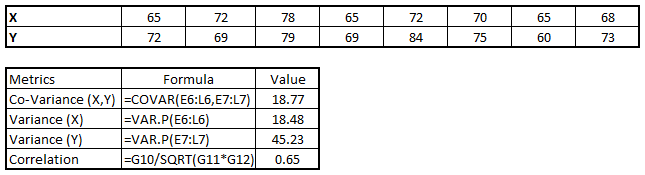
Scatter plot shows the relationship between two variable but does not indicates the strength of relationship amongst them. To find the strength of the relationship, we use statistical technique.

1. Non-graphical method: Build the correlation matrix to understand the strength between variables. Correlation varies between -1 and +1.
   1. -1: Perfect negative linear correlation
   2. +1: Perfect positive linear correlation
   3. 0: No correlation

**Ideal assumptions:**

1. High Correlation between dependent and independent variable.
2. Less correlation between independent variables.

Generally, if the correlation between the two independent variables are high (>= 0.8) then we drop one independent variable otherwise it may lead to multi collinearity problem. Various tools have function or functionality to identify correlation between variables. In Excel, function CORREL() is used to return the correlation between two variables and SAS uses procedure PROC CORR to identify the correlation. These function returns Pearson Correlation value to identify the relationship between two variables: Let us calculate using excel.

[](https://www.analyticsvidhya.com/wp-content/uploads/2015/02/Data_exploration_51.png)

In above example, we have positive relationship (0.65) between two variables X and Y.

**Interview Questions:**

1. How would you explain the difference between correlation and covariance?
2. What’s the difference between correlation and simple linear regression?
3. Can single outlier decrease or increase the correlation with a big magnitude?
4. Does correlation and dependency mean the same thing? In simple words if two events have correlation of zero, does this convey they are not dependent and vice-versa?
5. Difference between correlation and auto correlation?