9 COVARIANCE & CORRELATION

Up till now we have only looked at the distribution of a single random variable in isolation. However, often we are interested in studying and quantifying the association or relationship between two or more variables. That's where covariance and correlation come into play, which we discuss in this chapter.

9.1 COVARIANCE

It is a measure of the strength of linear relationship between two random variables.
If we have 'n' observations for two variables X and Y, the covariance is computed as shown:

$$Cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

Where,
X_i = i^{th} value of X
Y_i = i^{th} value of Y
$\bar{X}$ = Mean of X
$\bar{Y}$ = Mean of Y

Covariance can take up any value from -∞ to +∞ and gives the direction of the relationship between two variables. Covariance values are interpreted in the following manner:

- Positive covariance: reveals that two variables tend to move in the same direction
- Negative covariance: indicates that the two variables tend to move in the opposite direction
- Zero covariance: indicates that the two variables are not linearly related to each other

Some Important Points About Covariance

- Variance of any variable is nothing but the covariance of a variable with itself.
- Covariance only measures the linear relationship between two variables. Even if they have zero covariance, it need not mean that they are independent. They could still be related in a non-linear way.

9.2 CORRELATION

One problem with the covariance measure which we covered in the last section is that it can take any value from -∞ to +∞. This makes comparisons between two pairs of variables difficult. The answer to this problem is found in correlation, which is nothing but a standardised version of covariance.

Correlation is also used to measure the strength of the linear relationship between two variables however, due to the way it is defined, it is a dimensionless metric and its value varies from -1 to +1 only. Here, -1 refers to a perfectly negative relationship between two variables while +1 refers to a perfectly positive relationship.

We will look at the Pearson correlation coefficient, one of the ways of calculating correlation. The Pearson correlation coefficient between two variables X and Y is given by the Greek letter rho (ρ), and calculated as shown.

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

where,

$\rho\ (X, Y)$ = Pearson Correlation coefficient between X and Y

$Cov(X, Y)$ = Covariance between X and Y

σ_X = Standard deviation of X

σ_Y = Standard deviation of Y

Some Important Points About Correlation

- A correlation analysis assumes that the relationship between the two variables is linear, otherwise, the results are not useful.
- The Pearson correlation analysis described here assumes that the data are normally distributed; otherwise the Spearman non-parametric analysis would be more appropriate.
- There are inbuilt functions in Excel, R and Python which can be used to compute the covariance and correlation directly.

Correlation and Causation

Causation implies the cause and effect between two or more variables i.e. change in one variable leads to changes in other variables. Correlation and causation are sometimes used interchangeably. However, keep in mind that correlation does not imply causation. Two variables might be correlated to each other but it does not mean that one variable is making the others happen. Correlation could be one of the factors for causation but there may be other factors which lead to the observed phenomenon.

For example, the prices of two stocks tend to move in the same direction and are likely positively correlated, but it is not necessary that change in the price of one leads to change in the price of the other. The prices of both the stocks could move due to some underlying common factors like high GDP growth, etc.

The larger point being made here is that we need to be careful in our interpretation of any observed correlation.

9.3 Application of Correlation in Finance

An important consideration in finance is choosing the right combination of instruments to build a portfolio. Diversification of instruments is one method of portfolio optimization and overall risk can be ideally reduced by choosing stocks having zero or negative correlation with each other.