DegreeLive

B.Com Honours | Semester I | Calicut University

Essential Statistics for Business Analytics

Course Code: COM1MN109 • Module 3 Notes

1. Correlation Analysis: Karl Pearson and Spearman Rank Methods

In business analytics, variables rarely exist in isolation. Understanding the relationship between variables—such as advertising expenditure and sales, or price and quantity demanded—is essential for forecasting and pricing strategies. This module introduces correlation analysis (Karl Pearson's and Spearman's rank methods), bivariate data analysis, and the properties of linear regression models.

Correlation: Meaning and Types

Correlation measures the strength and direction of linear association between two variables, ranging from -1 to +1:

- **Positive Correlation:** Both variables move in the same direction (e.g., height and weight).
- **Negative Correlation:** Variables move in opposite directions (e.g., product price and consumer demand).
- **Zero Correlation:** No relationship exists between variables (e.g., shoe size and intelligence).
- **Linear vs. Non-linear:** Linear correlation forms a straight line when plotted; non-linear forms a curve.

Calculating Correlation Coefficients

Analysts use two main mathematical formulas to calculate correlation:

Karl Pearson's Product-Moment Correlation (r)

Used for quantitative (metric) data, measuring the strength of linear association:

[r = ⬆rac{sum (X - □ar{X})(Y - □ar{Y})}{sqrt{sum (X - □ar{X})^2 sum (Y - □ar{Y})^2}}]

Alternatively: [r = ⬆rac{n sum XY - sum X sum Y}{sqrt{[n sum X^2 - (sum X)^2][n sum Y^2 - (sum Y)^2]}}]

Spearman's Rank Correlation (R)

Used when data is qualitative (ordinal, e.g., ranking products by beauty, intelligence) or when outliers are present:

$$R = 1 - \frac{6 \sum D^2}{n(n^2 - 1)}$$

Where: **D** is the difference between ranks of corresponding pairs of variables, **n** is the number of pairs.

2. Regression Analysis

While correlation only measures association, regression analysis models the mathematical relationship between variables to predict the value of a dependent variable (Y) based on an independent variable (X).

Regression Lines

For bivariate data, there are two distinct lines of regression:

Regression Line of Y on X

Used to estimate/predict Y based on a given value of X.

$$Y - \bar{Y} = b_{yx}(X - \bar{X})$$

Where $\mathbf{b_{yx}}$ is the regression coefficient of Y on X, calculated as:

$$b_{yx} = r \frac{\sigma_y}{\sigma_x} = \frac{n\sum XY - \sum X\sum Y}{n\sum X^2 - (\sum X)^2}$$

Regression Line of X on Y

Used to estimate/predict X based on a given value of Y.

$$X - \bar{X} = b_{xy}(Y - \bar{Y})$$

Where $\mathbf{b_{xy}}$ is the regression coefficient of X on Y, calculated as:

$$b_{xy} = r \frac{\sigma_x}{\sigma_y} = \frac{n\sum XY - \sum X\sum Y}{n\sum Y^2 - (\sum Y)^2}$$

Properties of Regression Coefficients

- **Relationship with r:** The geometric mean of the two regression coefficients equals the correlation coefficient:
  $$r = \pm \sqrt{b_{yx} \times b_{xy}}$$
- **Sign Rule:** Both regression coefficients and the correlation coefficient must have the same algebraic sign (all positive or all negative).
- **Magnitude Limit:** If one regression coefficient is greater than 1, the other must be less than 1 (as $r^2 \leq 1$).

- **Intersection Point:** The two regression lines always intersect at the mean values ((□ar{X}, □ar{Y})).