3.1 Correlation

Meaning of Correlation

Correlation is a statistical method used to measure the strength and direction of the relationship between two or more quantitative variables.
It indicates whether an increase (or decrease) in one variable will result in an increase (or decrease) in another variable.

For example:
If the amount of rainfall increases and the crop yield also increases, there is a positive correlation between rainfall and yield.

Correlation does not establish a cause-and-effect relationship — it only tells how closely two variables move together.

Types of Correlation

1. **Positive Correlation:**
   When an increase in one variable leads to an increase in another, or when both decrease together.
   Example: Height and Weight.
2. **Negative Correlation:**
   When an increase in one variable leads to a decrease in another.
   Example: Price and Demand.
3. **Zero Correlation:**
   When there is no relationship between the two variables.
   Example: Shoe size and Intelligence.
4. **Perfect Correlation:**
   - **Perfect Positive Correlation:** When both variables change in the same proportion ($r = +1$).
   - **Perfect Negative Correlation:** When both variables change in opposite proportion ($r = -1$).

Methods of Studying Correlation

1. **Graphical Method**
   The relationship between two variables can be shown graphically using scatter diagrams.
   - The independent variable (X) is plotted on the X-axis.
   - The dependent variable (Y) is plotted on the Y-axis.
   - Each pair (x, y) is represented as a point on the graph.

   The pattern of plotted points shows the type of correlation:

   - Points closely rising upwards → Positive correlation
   - Points falling downwards → Negative correlation

- Random points → Zero correlation

Mathematical Methods

1. Karl Pearson's Coefficient of Correlation

Karl Pearson developed a mathematical formula to measure the degree of linear relationship between two variables.
It is denoted by **r**.

Formula:

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}}$$

Where:

- (x) and (y) are the variables,
- (\bar{x}) and (\bar{y}) are their respective means.

The value of **r** lies between –1 and +1.

- r = +1 → Perfect positive correlation
- r = –1 → Perfect negative correlation
- r = 0 → No correlation

2. Shortcut (or Direct) Formula

If deviations are taken from the mean:

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$$

Where x and y are deviations from the respective means.

3. Probable Error (P.E.)

Probable error helps to determine the reliability of the correlation coefficient.

Formula:

$$P.E. = 0.6745 \times \frac{(1 - r^2)}{\sqrt{N}}$$

Where N = number of pairs of observations.

Interpretation:

- If r < P.E., there is no significant correlation.
- If r > 6 × P.E., the correlation is definitely significant.

4. Spearman's Rank Correlation Coefficient

Used when data are given in ranks (instead of actual values).
It measures the degree of relationship between the rankings of two variables.

Formula:

$$r = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$$

Where:

- d = difference between ranks of each pair
- n = number of pairs

Interpretation:
The closer the value of r is to +1, the higher the similarity between rankings.

5. Covariance

Covariance measures how two variables vary together.
It indicates whether two variables move in the same or opposite directions.

Formula:

$$\text{Cov}(X, Y) = \frac{\sum (x - \bar{x})(y - \bar{y})}{n}$$

Interpretation:

- Positive covariance → variables move together
- Negative covariance → variables move in opposite directions

- Zero covariance → variables are independent

Relationship Between Covariance and Correlation

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

Where:

σX and σY are the standard deviations of X and Y respectively.

Interpretation of Correlation Coefficient

Range of r	Type of Correlation
0.90 to 1.00	Very High Positive
0.70 to 0.89	High Positive
0.40 to 0.69	Moderate Positive
0.20 to 0.39	Low Positive
0.00 to 0.19	Very Low or Negligible
–0.19 to 0.00	Very Low or Negligible Negative
–0.39 to –0.20	Low Negative
–0.69 to –0.40	Moderate Negative
–0.89 to –0.70	High Negative
–1.00 to –0.90	Very High Negative

3.2 Regression Analysis

Meaning of Regression

Regression analysis is a statistical method used to estimate or predict the value of one variable based on the value of another related variable.

If two variables are correlated, we can express their relationship mathematically using a regression equation.

For example:
We can predict a person's weight based on their height, or a company's sales based on its advertising expenditure.

Correlation only measures the degree of relationship,
whereas regression explains the nature and form of this relationship.

Objectives of Regression Analysis

1. To study the functional relationship between two variables.
2. To predict the value of one variable based on another.
3. To estimate the average value of a dependent variable (Y) for given values of an independent variable (X).
4. To make forecasts and analyze trends.
5. To measure the strength and direction of the relationship.

Types of Regression

1. **Simple Regression** – Involves two variables, one dependent and one independent.
   Example: Predicting sales (Y) from advertisement (X).
2. **Multiple Regression** – Involves more than one independent variable.
   Example: Predicting sales (Y) based on advertisement (X_1) and income (X_2).

Regression Lines

In regression, the relationship between variables is represented by a straight line (if linear regression) known as **a** regression line.

There are two regression lines:

1. Regression Line of Y on X
2. Regression Line of X on Y

Both lines are used to predict one variable from another.

1. Regression Equation of Y on X

This equation is used to estimate the value of Y when X is known.

Formula:
$Y - Y' = byx\ (X - X')$

or

$Y = a + byxX$

Where:

- (byx) = regression coefficient of Y on X
- (a) = Y-intercept (constant term)
- (Y'), (X') = means of Y and X

2. Regression Equation of X on Y

This equation is used to estimate **X** when **Y** is known.

Formula:
X – X' = bxy(Y – Y')

or

X = a + bxyY

Where:

- (bxy) = regression coefficient of X on Y

Regression Coefficients

Regression coefficients measure the change in the dependent variable due to a unit change in the independent variable.
They indicate the slope or steepness of the regression line.

There are two regression coefficients:

1. (byx): Coefficient of Y on X
2. (bxy): Coefficient of X on Y

Formulas

1. **By Covariance and Standard Deviation:**

$$b_{yx} = \frac{\text{Cov}(X,Y)}{\sigma_X^2} \quad \text{and} \quad b_{xy} = \frac{\text{Cov}(X,Y)}{\sigma_Y^2}$$

2. **By Correlation Coefficient:**

$$b_{yx} = r \times \frac{\sigma_Y}{\sigma_X} \quad \text{and} \quad b_{xy} = r \times \frac{\sigma_X}{\sigma_Y}$$

Where:

- (r) = correlation coefficient
- (σX,σY) = standard deviations of X and Y

Properties of Regression Coefficients

1. Both regression coefficients usually have the same sign as the correlation coefficient (r).

2. The geometric mean of (byx) and (bxy) equals the correlation coefficient:

$$r = \sqrt{b_{yx} \times b_{xy}}$$

3. The arithmetic mean of the regression coefficients is greater than or equal to the correlation coefficient.
4. If (r = 0), then both regression coefficients will be zero.
5. If (r = ±1), both regression lines coincide and become identical.

Relationship Between Correlation and Regression

1. Correlation measures the degree of relationship between two variables.
2. Regression measures the nature and extent of the relationship.
3. Correlation is symmetric — it does not distinguish between dependent and independent variables.
4. Regression is asymmetric — one variable is treated as dependent and the other as independent.
5. Correlation coefficient (r) helps in finding regression coefficients (byx) and (bxy).

Example

If the correlation between X and Y is 0.8,
and the standard deviations of X and Y are 5 and 8 respectively:

$$b_{yx} = r \times \frac{\sigma_Y}{\sigma_X}$$

$$b_{yx} = 0.8 \times \frac{8}{5} = 1.28$$

This means for every one-unit increase in X, Y increases by 1.28 units.

Uses of Regression Analysis

1. For prediction and forecasting (e.g., predicting demand or sales).
2. In business and economics for decision-making.
3. To estimate missing data values.
4. To determine cause-and-effect relationships.
5. To study trends and dependencies between variables.

3.3 Difference Between Correlation and Regression

Correlation and regression are closely related statistical tools used to study the relationship between two or more variables.
However, they serve different purposes. Correlation measures the degree or strength of association between variables,
while regression explains the relationship in mathematical form and helps in prediction.

1. Meaning

Correlation measures the extent to which two variables move together.
Regression explains the relationship between dependent and independent variables and helps estimate the value of one variable based on another.

2. Objective

The objective of correlation is to find out the degree of relationship between two variables.
The objective of regression is to predict or estimate the value of one variable based on the known value of another.

3. Type of Relationship

Correlation is a measure of association only — it does not show cause and effect.
Regression shows the cause-and-effect relationship, explaining how one variable affects another.

4. Variables Involved

In correlation, both variables are treated equally — there is no distinction between dependent and independent variables.
In regression, one variable is dependent (Y) and the other is independent (X).

5. Direction of Relationship

Correlation shows only the degree and direction (positive or negative) of a relationship.
Regression shows the nature and magnitude of change in the dependent variable for a given change in the independent variable.

6. Range of Values

The value of correlation (r) always lies between –1 and +1.
The value of regression coefficients is and can be any real number.

7. Purpose

Correlation is used to measure the relationship.
Regression is used to predict or forecast one variable using another.

8. Expression of Result

Correlation is expressed through the correlation coefficient (r).
Regression is expressed through regression equations and coefficients.

9. Number of Equations

Correlation gives only one coefficient (r) to represent the relationship.
Regression provides two equations:

- Regression of Y on X
- Regression of X on Y

10. Graphical Representation

Correlation is shown by a scatter diagram where the closeness of points indicates the degree of correlation.
Regression is represented by regression lines that best fit the data and help in prediction.

Difference Table

Basis	Correlation	Regression
Meaning	Measures the degree of relationship between two variables.	Explains the relationship and predicts the value of one variable based on another.
Objective	To find the degree of relationship.	To estimate or predict dependent variable.
Dependency	No dependent or independent variable.	One variable is dependent, the other is independent.
Range	Lies between –1 and +1.	Can take any real value.
Equations	Only one coefficient (r).	Two equations (Y on X and X on Y).
Graphical Representation	Shown by scatter diagram.	Shown by regression line.
Purpose	Measures association.	Predicts outcomes.

3.4 Regression Equations

Meaning

Regression equations are mathematical expressions that show the relationship between two correlated variables.
They allow us to estimate or predict the value of one variable (dependent) based on the value of another (independent).

If two variables X and Y are correlated, there are two regression equations:

1. Regression equation of Y on X
2. Regression equation of X on Y

Both equations are straight lines when the relationship is linear.

Regression Equation of Y on X

The regression line of Y on X is used to estimate Y when X is known.

Equation:

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

or

$$Y = a + b_{yx}X$$

Where:

- (byx) = regression coefficient of Y on X
- (a) = intercept (constant)
- (Y') = mean of Y
- (X') = mean of X

Regression Equation of X on Y

This equation is used to estimate X when Y is known.

Equation:

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

or

$$X = a + b_{xy}Y$$

Where:

- (bxy) = regression coefficient of X on Y

Finding the Constants a and b

The constants (a) and (b) in the regression equations can be found using the method of least squares, which minimizes the sum of the squares of deviations.

For the regression of Y on X:

$$\begin{cases} \sum Y = Na + b\sum X \\ \sum XY = a\sum X + b\sum X^2 \end{cases}$$

For the regression of X on Y:

$$\begin{cases} \sum X = Na + b\sum Y \\ \sum XY = a\sum Y + b\sum Y^2 \end{cases}$$

Solving these two simultaneous equations gives the values of a and b.

Formulas for Regression Coefficients

Regression coefficients can be calculated using different methods:

1. **Using Covariance:**

$$b_{yx} = \frac{\text{Cov}(X,Y)}{\sigma_X^2} \quad \text{and} \quad b_{xy} = \frac{\text{Cov}(X,Y)}{\sigma_Y^2}$$

2. **Using Correlation:**

$$b_{yx} = r \times \frac{\sigma_Y}{\sigma_X} \quad \text{and} \quad b_{xy} = r \times \frac{\sigma_X}{\sigma_Y}$$

Where:

- (r) = correlation coefficient
- (σX,σY) = standard deviations of X and Y

Properties of Regression Equations

1. Both regression lines pass through the point of means (X', Y').
2. There are two lines unless the correlation is perfect (r = ±1).
3. If correlation is perfect, both regression lines coincide.
4. The regression line of Y on X is used to estimate Y for a given X, and vice versa.
5. The angle between the two regression lines depends on the value of r.

Example

Given the following data:

X	Y
1	2
2	4
3	5
4	4
5	5

To find:

1. The regression equation of Y on X.
2. The regression equation of X on Y.

Solution:

Step 1: Calculate the means

X' = 3, Y' = 4

Step 2: Calculate deviations and products

X	Y	$(X-\bar{X})$	$(Y-\bar{Y})$	$(X-\bar{X})(Y-\bar{Y})$	$(X-\bar{X})^2$	$(Y-\bar{Y})^2$
1	2	−2	−2	4	4	4
2	4	−1	0	0	1	0
3	5	0	1	0	0	1
4	4	1	0	0	1	0
5	5	2	1	2	4	1

$$\sum(X-\bar{X})(Y-\bar{Y}) = 6, \quad \sum(X-\bar{X})^2 = 10, \quad \sum(Y-\bar{Y})^2 = 6$$

Step 3: Find regression coefficients

$$b_{yx} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(X-\bar{X})^2} = \frac{6}{10} = 0.6$$

$$b_{xy} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(Y-\bar{Y})^2} = \frac{6}{6} = 1.0$$

Step 4: Regression equations

- Regression equation of Y on X:

  Y - 4 = 0.6 (X - 3)
  Y = 0.6X + 2.2
- Regression equation of X on Y:

  X - 3 = 1.0 (Y - 4)
  X = Y - 1

Interpretation

The equation (Y = 0.6X + 2.2) means that for every one-unit increase in X, Y increases by 0.6 units on average.
Similarly, the equation (X = Y - 1) can be used to estimate X when Y is known.

3.5 Properties and Uses of Regression Lines

Properties of Regression Lines

1. **Both Regression Lines Pass Through the Mean Point ($\bar{X}$, $\bar{Y}$):**
   The two regression lines always intersect at the point representing the mean values of both variables.
   This point is called the point of means and it is common to both lines.

2. **There Are Two Regression Lines:**
   - The regression line of Y on X is used to estimate the value of Y when X is known.
   - The regression line of X on Y is used to estimate the value of X when Y is known.

   Both lines generally intersect at the point ($\bar{X}$, $\bar{Y}$).
   If the correlation is perfect (r = ±1), both lines coincide and become identical.

3. **Angle Between the Two Regression Lines Depends on the Correlation Coefficient:**
   - When the correlation is perfect (r = ±1), both lines coincide, and the angle between them is 0°.

- When there is no correlation (r = 0), the lines are perpendicular to each other, forming a 90° angle.
- As correlation increases, the lines come closer together.

4. **Regression Coefficients Have the Same Sign as the Correlation Coefficient:**
   If the correlation between X and Y is positive, both regression coefficients (b_{xy} and b_{yx}) are positive.
   If the correlation is negative, both regression coefficients are negative.
   Thus, regression lines always slope in the same direction as the correlation.

5. **Regression Coefficients Are Independent of Change of Origin But Not of Scale:**
   - If a constant value is added to or subtracted from the variables, the regression coefficients remain unchanged.
   - But if the variables are multiplied or divided by a constant, the regression coefficients are affected.

6. **The Arithmetic Mean of Regression Coefficients Is Always Greater Than or Equal to the Correlation Coefficient:**
   Mathematically,
   $$r = \sqrt{b_{yx} \times b_{xy}}$$
   This shows that the correlation coefficient is the geometric mean of the two regression coefficients.

7. **Regression Lines Are Least Squares Lines:**
   Each regression line is fitted using the least squares method, which minimizes the sum of the squares of deviations between actual and estimated values.
   This ensures the best possible fit for prediction.

8. **Regression Line of Y on X Gives Minimum Error in Estimating Y:**
   The line of Y on X minimizes the total squared vertical distances between actual Y values and estimated Y values from the line.
   Similarly, the regression line of X on Y minimizes horizontal errors.

Uses of Regression Analysis

1. **Prediction and Forecasting:**
   Regression helps in predicting the value of one variable based on another known variable.
   For example, future sales can be predicted based on past advertisement expenditure.

2. **Estimation of Relationships:**
   It helps to estimate the nature and strength of relationships between dependent and independent variables in a dataset.

3. **Business and Economics:**
   Used for forecasting demand, cost estimation, price determination, and trend analysis in economics and management.

4. **Research and Experiments:**
   Helps in determining the relationship between different factors in scientific and social studies.

5. **Data Analysis and Decision Making:**
   Regression provides a quantitative base for decision-making in planning, marketing, finance, and operations.

6. **Trend Study:**
   Regression lines can be used to study and project long-term trends, such as population growth, sales patterns, or production rates.

3.6 Limitations of Correlation and Regression Analysis

Correlation and regression analysis are powerful tools in statistics used to study and predict relationships between variables.
However, they also have certain limitations that must be considered while interpreting results. Misuse or over-reliance on these methods can lead to incorrect conclusions.

1. Correlation Does Not Imply Causation

A high degree of correlation between two variables does not necessarily mean that one variable causes changes in the other.
The relationship may be coincidental or may result from the effect of a third variable.

Example:
There may be a high correlation between ice cream sales and drowning cases, but one does not cause the other.
Both are related to temperature (a third factor).

2. Limited to Linear Relationships

Both correlation and regression measure only linear relationships between variables.
If the relationship is non-linear (curved), these methods do not give accurate results.

3. Affected by Extreme Values (Outliers)

A few extreme observations or outliers can significantly distort the results of correlation and regression analysis.
Hence, data should always be checked for abnormal or extreme values before applying these techniques.

4. Based on Assumptions

Regression analysis assumes:

- A constant relationship between variables,
- No measurement errors, and
- Normal distribution of variables.

If these assumptions are not satisfied, the conclusions may become unreliable.

5. Applicable Only to Quantitative Data

Correlation and regression require numerical or quantitative data.
They cannot be used for qualitative variables such as gender, religion, or satisfaction level unless these are expressed numerically.

6. Limited Scope of Prediction

Regression equations are valid only within the range of observed data.
Predicting values beyond the available range (extrapolation) may lead to incorrect or unrealistic estimates.

7. Sensitive to Sample Size

Both correlation and regression results are affected by the number of observations.
A small sample size may give misleading results, while a large sample provides more reliability.

8. Misinterpretation Risk

Without proper understanding, correlation and regression can be easily misinterpreted.
A high correlation may be wrongly assumed as a cause-and-effect relationship, leading to false conclusions in research and decision-making.