Unit-I
Regression

Bivariate Regression

1. **Definition**: Bivariate regression is a statistical method used to study the relationship between two variables—one dependent (Y) and one independent (X).

2. **Equation**: The linear regression model is represented as:
   $$Y=a+bX$$
   where *a* is the intercept, *b* is the slope
3. **Application**: Used in data analysis to identify trends, make predictions, and understand the impact of one variable on another.

4. **Assumptions**:
   - ❖ A linear relationship exists between X and Y.
   - ❖ The residuals (errors) are normally distributed.
   - ❖ Homoscedasticity (constant variance of residuals).
   - ❖ Independence of observations.
5. **Interpretation**:
   - ❖ A positive *b* indicates a direct relationship.
   - ❖ A negative *b* shows an inverse relationship.
6. **Quality of Fit**: Assessed using metrics like **R^2 (coefficient of determination)** and **p-values** to determine how well the model explains the variability in data.

Multivariate Regression

1. **Definition**: Multivariate regression is an extension of linear regression that models the relationship between one dependent variable (Y) and multiple independent variables (X_1, X_2, X_3, ...).

2. **Equation**: The linear multivariate regression model is represented as:
   $Y=a+b_1X_1+b_2X_2+b_3X_3+\ldots\ldots$
   where *a* is the intercept, $b_1, b_2, b_3,\ldots$ are the coefficients
3. **Application**: Used to analyse how multiple factors influence an outcome and make better predictions by considering various inputs.

4. **Assumptions**:
   - ❖ A linear relationship exists between the dependent and independent variables.
   - ❖ Residuals (errors) are normally distributed.
   - ❖ No multicollinearity (independent variables should not be highly correlated).
   - ❖ Homoscedasticity (constant variance of residuals).
5. **Interpretation**:
   - ❖ Each coefficient ($b_1, b_2, \ldots$) represents the change in Y for a one-unit change in the corresponding X, holding other variables constant.
   - ❖ A positive coefficient shows a direct relationship, while a negative coefficient indicates an inverse relationship.
6. **Quality of Fit**: Evaluated using **Adjusted R^2**, **p-values**, and **Multicollinearity (VIF - Variance Inflation Factor)** to ensure the model is reliable for prediction.

Logistic Regression: Basic Points

1. **Definition**: Logistic regression is a statistical method used for binary or multi-class classification, predicting the probability of an outcome belonging to a specific category. Unlike linear regression, it deals with categorical dependent variables.

2. **Equation**: The logistic regression model is represented as:

$$P(Y) = \frac{e^{(a+b_1X_1+b_2X_2+...+b_nX_n)}}{1+e^{(a+b_1X_1+b_2X_2+...+b_nX_n)}}$$

where $P(Y)$ is the probability of the event occurring, a is the intercept, b_1, b_2, ... are the coefficients, and X_1, X_2, ... are the independent variables.

3. **Application**: Used when the dependent variable is categorical (e.g., Yes/No, Pass/Fail, Fraud/Not Fraud).

Types of Distances in Data Analytics & Their Applications

Distance metrics are essential in machine learning, clustering, and classification to measure similarity or dissimilarity between data points. Below are common distance measures, their formulas, and use cases:

1. Euclidean Distance (Straight-line distance)

Formula:

$$d = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Application:

- Used in **K-Nearest Neighbors (KNN)** and **K-Means clustering** to determine the closest points.
- Ideal when data points are continuous and follow geometric relationships.

2. Manhattan Distance (City-block distance)

Formula:

$$d = \sum_{i=1}^{n}|x_i - y_i|$$

Application:

- Useful in **grid-based path planning** (e.g., movement in chess, robotics).
- Works well when movement is constrained to axes (e.g., warehouse robots following fixed paths).

3. Minkowski Distance (Generalized form of Euclidean & Manhattan)

Formula:

$$d = \left(\sum_{i=1}^{n}|x_i - y_i|^p\right)^{\frac{1}{p}}$$

- When **p = 1**, it becomes Manhattan Distance.
- When **p = 2**, it becomes Euclidean Distance.

Application:

- Allows flexibility in measuring distances based on data structure.