ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

DAY – 13 **9 July 2025**

Linear Regression in Machine Learning

Linear regression is a supervised machine learning algorithm used to predict values by finding the best-fitting straight line through the data. It assumes a linear relationship between the input (independent variable) and the output (dependent variable). This means as the input increases or decreases, the output does so at a constant rate.

Example:

If we want to predict a student's exam score based on hours studied:

- Input (Independent Variable): Hours studied
- Output (Dependent Variable): Exam score

As study hours increase, exam scores generally increase too. Linear regression learns this pattern from past data and uses it to make future predictions.

Why Linear Regression Is Important:

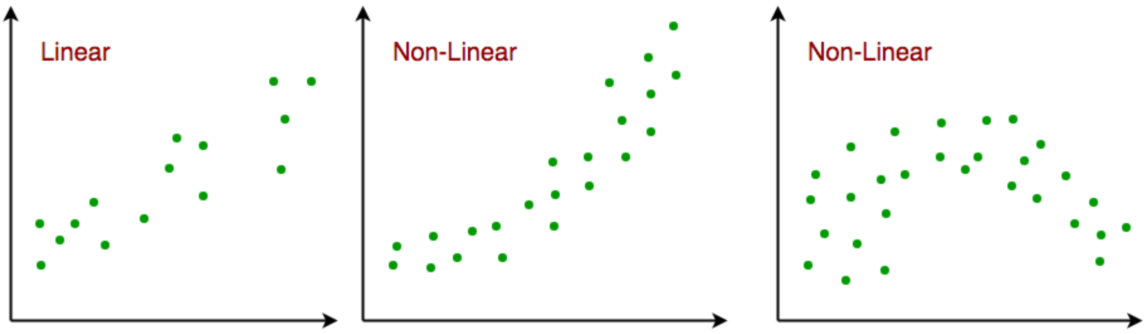
- Simple and Easy to Understand
- Good for Predictions when there’s a clear trend
- Foundation for Advanced Models like logistic regression
- Helps Analyze the relationship between variables

Best Fit Line:

In linear regression, the model finds a straight line (called the **best-fit line**) that represents the relationship between input and output. The line is chosen so that the differences between actual and predicted values are as small as possible.

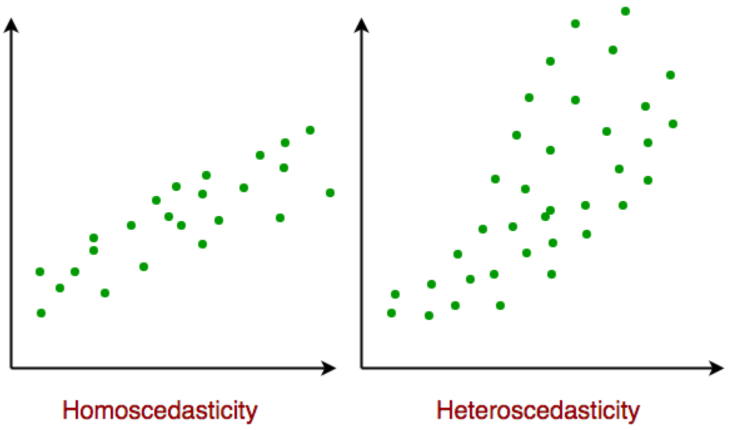
Assumptions of the Linear Regression

1. Linearity: The relationship between inputs (X) and the output (Y) is a straight line.



2. Independence of Errors: The errors in predictions should not affect each other.

3. Constant Variance (Homoscedasticity): The errors should have equal spread across all values of the input. If the spread changes (like fans out or shrinks), it's called heteroscedasticity and it's a problem for the model.



4. Normality of Errors: The errors should follow a normal (bell-shaped) distribution.

5. No Multicollinearity *(*for multiple regression): Input variables shouldn't be too closely related to each other.

6. No Autocorrelation: Errors shouldn't show repeating patterns, especially in time-based data.

7. Additivity: The total effect on Y is just the sum of effects from each X, no mixing or interaction between them.

Types of Linear Regression

Linear regression comes in two main types based on the number of input features (independent variables):

1. Simple Linear Regression

- Used when there is **only one input feature**.

- Assumes a straight-line relationship between the input (x) and output (y).

Formula:

$$\hat{y} = \theta_0 + \theta_1 x$$

Where:

- $\hat{y}$ = predicted value
- x = input (independent variable)
- θ_0 = intercept (value when $x = 0$)
- θ_1 = slope (how much y changes when x increases by 1)

Example: Predicting salary based on years of experience.

2. Multiple Linear Regression

- Used when there are **two or more input features**.
- Still predicts one output, but considers multiple inputs.

Formula:

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$

Where:

- $x_1, x_2, \ldots, x_n$ = input features
- $\theta_1, \theta_2, \ldots, \theta_n$ = corresponding weights
- θ_0 = intercept
- $\hat{y}$ = predicted value

Example: Predicting house price based on size, location, and number of rooms.

How It Works

In both types, the model learns from existing data (X and Y) to find the best-fit line. Once trained, this line (or function) can be used to predict Y for new unseen values of X.