Machine Learning

Week 3 • Class 1

Introduction to Linear Regression

Class Objective

By the end of this class, students will be able to:

- Understand what regression is
- Explain the idea of linear regression
- Identify dependent and independent variables
- Derive the linear regression equation
- Perform manual calculations for simple linear regression
- Interpret slope and intercept meaningfully

What Is Regression?

Regression is a supervised learning technique where:

- Output variable is **continuous**
- The goal is to **predict a numerical value**

Examples:

- Predict house price
- Predict salary from experience
- Predict fuel consumption from engine size

Why Linear Regression?

Linear regression is used when:

- Relationship between variables is approximately linear
- We want a simple, interpretable model
- Data follows a straight-line trend

It is:

- Easy to understand
- Easy to compute
- Widely used as a baseline model

Real-World Uses of Linear Regression

- Market research & customer surveys
- Automobile engine performance analysis
- Pricing of goods and services
- Astronomy and scientific measurements
- Economics and finance forecasting

Simple Linear Regression

Simple Linear Regression has:

- One **independent variable (X)**
- One **dependent variable (Y)**

General form:

$$Y = mX + c$$

Where:

- m = slope
- c = intercept

Understanding the Variables

- **Independent variable (X)**
  → input, predictor, feature
- **Dependent variable (Y)**
  → output, response, target

Example:

- X = hours studied
- Y = exam score

Graphical Interpretation

- X axis → Independent variable
- Y axis → Dependent variable
- Best-fit straight line represents the model
- **Objective:** minimize error between actual and predicted values

Problem Statement (Example)

Consider the following data:

X (Hours Studied)	Y (Marks)
1	2
2	4
3	5
4	4
5	6

We want to find the best fit line.

Step 1: Formula for Slope (m)

$$m = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum X^2 - (\sum X)^2}$$

Where:

- n = number of observations

Step 2: Compute Required Values

X	Y	X^2	XY
1	2	1	2
2	4	4	8
3	5	9	15
4	4	16	16
5	6	25	30

Summations

- $\sum X = 15$
- $\sum Y = 21$
- $\sum X^2 = 55$
- $\sum XY = 71$
- $n = 5$

Step 3: Calculate Slope (m)

$$m = \frac{5(71) - (15)(21)}{5(55) - (15)^2}$$

$$m = \frac{355 - 315}{275 - 225}$$

$$m = \frac{40}{50} = 0.8$$

Step 4: Formula for Intercept (c)

$$c = \frac{\sum Y - m \sum X}{n}$$

Substitute values:

$$c = \frac{21 - (0.8 \times 15)}{5}$$

$$c = \frac{21 - 12}{5} = \frac{9}{5} = 1.8$$

Final Regression Equation

$$Y = 0.8X + 1.8$$

This is the best fit line for the given data.

Interpretation of the Model

- **Slope (0.8)**
  → For every 1 unit increase in X, Y increases by 0.8 units
- **Intercept (1.8)**
  → Expected value of Y when X = 0

Making Predictions

If a student studies for 6 hours:

$$Y = 0.8(6) + 1.8 = 6.6$$

Predicted marks ≈ 6.6

Error in Linear Regression

Error (Residual):

$$\text{Error} = Y_{\text{actual}} - Y_{\text{predicted}}$$

Goal of linear regression:

- Minimize total error
- Best fit line minimizes sum of squared errors

Limitations of Linear Regression

- Assumes linear relationship
- Sensitive to outliers
- Cannot model complex curves
- Poor performance if data is non-linear

What Comes Next?

Next classes will cover:

- Multiple Linear Regression
- Polynomial Regression
- Matrix methods
- Error metrics (MSE, RMSE)

In-Class Exercise

1. Use the given formula to compute regression for a new dataset
2. Plot the data points and regression line
3. Predict output for unseen values

Practice Problem

A company wants to predict **Sales (Y)** based on **Advertising Budget (X)** in thousands of dollars.

X (Budget)	Y (Sales)
2	3
3	5
4	6
5	8
6	9

Tasks:

1. Calculate the regression equation
2. Predict sales when budget = 7