Multiple Linear Regression – Mathematical Formulation (From Scratch)

---

1. Introduction:

- Multiple Linear Regression (MLR) is an extension of **Simple Linear Regression** where we have **more than one independent variable**.
- Example: Predicting **house prices** using multiple features like:
    1. Size of the house (x_1)
    2. Number of rooms (x_2)
    3. Location score (x_3)
    4. Age of property (x_4)

The general equation becomes:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$$

---

2. Types of Linear Regression:

- **Simple Linear Regression:** One input feature → line in 2D.
- **Multiple Linear Regression:** Multiple features → plane or hyperplane in higher dimensions.
- Still assumes a **linear relationship** between independent variables (features) and dependent variable (target).

---

3. Mathematical Formulation:

We express the problem using **matrix notation**, making it easier to handle many features.

Step 1: Define Data in Matrix Form:

For n data points and pp features:

- **Feature matrix (X):**

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

(1st column = 1's for intercept term)

- **Coefficient vector (β):**

$$\beta = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_p \end{bmatrix}$$

- **Target vector (Y):**

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Step 2: Hypothesis (Prediction):

$$\hat{Y} = X\beta$$

Where:

- $\hat{Y}$ = predicted values
- $X\beta$ = linear combination of inputs and weights

Step 3: Cost Function (Error):

We want coefficients β that minimize:

$$J(\beta) = \text{SSE} = (Y - X\beta)^T(Y - X\beta)$$

This is the **Sum of Squared Errors** in matrix form.

Step 4: Find Optimal Solution:

To minimize $J(\beta)$, we take its derivative with respect to β and set to zero.

$$\frac{\partial J}{\partial \beta} = -2X^T(Y - X\beta) = 0$$

Solving this:

$$X^TY = X^TX\beta$$

$$\beta = (X^TX)^{-1}X^TY$$

This is known as the **Normal Equation**, giving the **optimal weights** (coefficients) for multiple linear regression.

---

4. Why Gradient Descent?

Although we have a **direct solution** (Normal Equation), sometimes:

- X^TX is **non-invertible** (features are highly correlated → multicollinearity).
- Dataset is **very large**, making matrix inversion computationally expensive ($O(p^3)$).
- In these cases, **Gradient Descent** is preferred:
  - Iteratively updates weights β to minimize error.
  - More scalable for **big data** and **high-dimensional problems**.

Key Takeaways:

- Multiple Linear Regression finds a **hyperplane** that best fits the data.
- Formula in vectorized form:

$$\hat{Y} = X\beta$$

- Optimal solution (Normal Equation):

$$\beta = (X^T X)^{-1} X^T Y$$

- **Gradient Descent** is an alternative to solve when direct inversion is not feasible.
