explain only what i asked , we will move later towards the linear regression using gradient descent

Below is a **strictly mathematical and technical derivation** of **Multiple Linear Regression using Ordinary Least Squares (OLS)**, following **exactly** the structure you requested.

---

1. Problem Definition

Let

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$$

where

$$x_i \in \mathbb{R}^p, \quad y_i \in \mathbb{R}$$

Define:

- Input matrix:

$$X \in \mathbb{R}^{n \times (p+1)}$$

- Output vector:

$$y \in \mathbb{R}^n$$

- Parameter vector:

$$\beta \in \mathbb{R}^{(p+1)}$$

Assumptions:

$$y = X\beta + \varepsilon$$
$$\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$$
$$\text{rank}(X) = p + 1$$

Objective:

$$\min_{\beta \in \mathbb{R}^{(p+1)}} J(\beta)$$

where

$$J(\beta) = \|y - X\beta\|_2^2$$

---

2. Mathematical Setup

Define:

- Augmented feature matrix:

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

- Parameter vector:

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

- Prediction vector:

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

Residual vector:

$$r = y - X\beta$$

Objective function:

$$J(\beta) = r^\top r = (y - X\beta)^\top (y - X\beta)$$

---

3. Derivation of the Core Equations

3.1 Model Equation

For each observation i:

$$\hat{y}_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$$

Vector form:

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

3.2 Loss Function

Squared error loss:

$$J(\beta) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Matrix expansion:

$$\begin{aligned} J(\beta) &= (y - X\beta)^\top (y - X\beta) \\ &= y^\top y - 2\beta^\top X^\top y + \beta^\top X^\top X\beta \end{aligned}$$

3.3 Partial Derivatives

Gradient of $J(\beta)$:

$$\nabla_\beta J(\beta) = \frac{\partial}{\partial \beta} \left(y^\top y - 2\beta^\top X^\top y + \beta^\top X^\top X\beta \right)$$

Using matrix derivative identities:

$$\frac{\partial}{\partial \beta}(y^\top y) = 0$$

$$\frac{\partial}{\partial \beta}(-2\beta^\top X^\top y) = -2X^\top y$$

$$\frac{\partial}{\partial \beta}(\beta^\top X^\top X\beta) = 2X^\top X\beta$$

Thus:

$$\nabla_\beta J(\beta) = -2X^\top y + 2X^\top X\beta$$

---

4. Solving for Optimal Parameters

Set gradient to zero:

$$\nabla_\beta J(\beta) = 0$$

$$-2X^\top y + 2X^\top X\beta = 0$$

Divide by 2:

$$X^\top X\beta = X^\top y$$

This yields the **normal equations**.

4.1 Closed-Form Solution

If $X^\top X$ is invertible:

$$\boxed{\beta^* = (X^\top X)^{-1} X^\top y}$$

---

5. Final Formula Summary

Objective:

$$\min_\beta \|y - X\beta\|_2^2$$

Gradient:

$$\nabla_\beta J(\beta) = 2X^\top X\beta - 2X^\top y$$

Normal equations:

$$X^\top X\beta = X^\top y$$

Closed-form estimator:

$$\boxed{\beta^* = (X^\top X)^{-1} X^\top y}$$

---

6. Computational Form

Closed-Form Computation

Given:

$$X \in \mathbb{R}^{n \times (p+1)}, \quad y \in \mathbb{R}^n$$

Compute:

1. Gram matrix:

$$G = X^\top X$$

2. Cross-product:

$$c = X^\top y$$

3. Parameter estimate:

$$\beta = G^{-1}c$$

End of derivation.

When you are ready, we can **move next to multiple linear regression using gradient descent**, starting strictly from the same objective and gradient.