Linear Regression and the Normal Equation

1. Key Concepts

To understand the derivation of the Normal Equation, the following mathematical foundations are required:

- **Vectors & Matrices**: Data representation formats.
- **Matrix Multiplication**: The process of combining features and weights.
- **Transpose & Inverse**: Essential operations for isolating variables in matrix equations.
- **Column Space**: The subspace spanned by the feature vectors.
- **Least Squares Approximation**: The method of finding the "best fit" by minimizing the sum of squared differences.

2. Problem Statement

Given a dataset where:

- $X \in \mathbb{R}^{n \times m}$ (Input Matrix with n samples and m features)
- $y \in \mathbb{R}^{n \times 1}$ (Target Vector)

Goal: Find the parameter vector β such that the prediction $\hat{y} = X\beta$ minimizes the prediction error.

3. The Normal Equation

In Machine Learning, the Normal Equation provides the closed-form solution for Linear Regression. It allows us to find the optimal parameters directly by minimizing the Mean Squared Error (MSE)[1].

The Derivation

We aim to minimize the Squared Euclidean Norm:

$$\min_{\beta} \|y - X\beta\|^2$$

Properties of this function:

- It is a quadratic form.
- It is convex.
- It has a single global minimum.

The Loss Function $J(\beta)$:

$$J(\beta) = (y - X\beta)^T (y - X\beta)$$

Expanding the expression:

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

Optimization:

By taking the gradient with respect to β and setting it to zero ($\nabla_\beta J = 0$):

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

Simplifying:

$$X^T X\beta = X^T y$$

The Final Equation:

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

This is the **Normal Equation** — the closed-form solution for the optimal regression parameters.

4. Geometric Interpretation

"Linear Regression coefficients define the orthogonal projection of y onto the column space of X."[1]

Key geometric insights:

- The target vector y lies in $\mathbb{R}^n$.
- The prediction $X\beta$ lies in the Column Space of X ($\mathrm{Col}(X)$).
- The residual vector (the error) is perpendicular (orthogonal) to $\mathrm{Col}(X)$.

This interpretation reveals that Linear Regression finds the best approximation of the target vector within the subspace spanned by the feature vectors.

5. Why Use the Normal Equation?

The Normal Equation offers several significant advantages in machine learning:

- **Closed-form Solution**: Provides an analytical result for regression parameters without approximation.
- **No Iteration**: Unlike Gradient Descent, it calculates the minimum in one step without needing multiple loops or epochs.
- **No Hyperparameters**: You do not need to choose or tune a learning rate, which simplifies model development.
- **Efficiency**: There is no need for convergence checks or early stopping criteria.
- **Uniqueness**: It guarantees a unique solution for the coefficients as long as $X^T X$ is invertible.

When to Use the Normal Equation

The Normal Equation is particularly suitable for:

- Small to medium-sized datasets where matrix inversion is computationally feasible.
- Situations where quick solutions are needed without extensive tuning.
- Cases where interpretability and directness are prioritized.

However, for very large datasets with many features, Gradient Descent or other iterative optimization methods may be more efficient due to the computational cost of matrix inversion.