Cost Function in Linear Regression

Linear Regression is a method used to predict values by drawing the best-fit line through the data. When we first create a model, the predictions may not always match the actual data. To understand how well the model is performing we use a **cost function**. This function helps us to measure the difference between the **predicted values** and the **actual data**. In this article, we'll see cost function in linear regression, what it is, how it works and why it's important for improving model accuracy.

Lets understand it with a example, imagine we are building a linear regression model to predict **house prices** based on the **size of the house** (in square feet).

Here's some training data:

Size (sq. ft.)	True Price (in $1000s)
500	50
1000	100
1500	150
2000	200

The linear regression equation is:

$$\hat{y} = w \cdot x$$

where,

- $\hat{y}$ is **predicted house price**
- x is **size of the house (input feature)**
- w is **weight (slope of the line)**

Our goal is to find the weight w that minimizes the difference between the predicted and actual prices. This difference is calculated using the cost function.

Understanding Mean Squared Error (MSE)

A commonly used cost function is **Mean Squared Error (MSE)**. It finds larger errors which helps the model focus on reducing mistakes between predictions and actual values.

The MSE formula is:

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 \Pi$$

Where:

- $J(\theta)$ is the cost function
- m is the number of data points
- $h\theta(x(i))$ is the predicted value for the i-th data point
- $y(i)$ is the actual value.

Let's calculate the MSE for initial predictions with w=0.04. Using this value of w, we can predict the prices for the given house sizes:

Size (x) (sq. ft.)	True Price (y) (in $1000s)	Predicted Price (y^) (in $1000s)
500	50	0.04×500=200.04×500=20
1000	100	0.04×1000=400.04×1000=40
1500	150	0.04×1500=600.04×1500=60
2000	200	0.04×2000=800.04×2000=80

Proceeding with our example Let's calculate the MSE for our predictions. For each data point we square the error to ensure it's positive then sum all the squared errors:

- 500= (30)2=900(30)2=900
- 1000= (60)2=3600(60)2=3600
- 1500= (90)2=8100(90)2=8100
- 2000= (120)2=14400(120)2=14400

Now, sum the squared errors:

900+3600+8100+14400=27000900+3600+8100+14400=27000

Divide by the number of points m=4:

MSE=27000/4=6750