SIMPLE LINEAR REGRESSION (supervised model)

> one independent feature and one dependent feature.

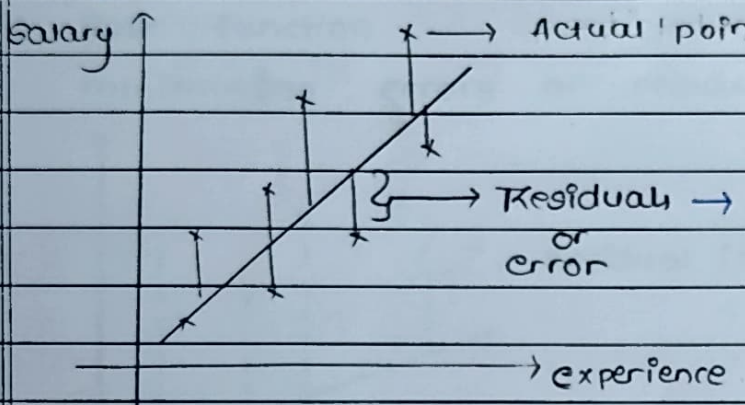
> Eg: Aim to create a model which take the input as height and predict weight.

Dataset : height, weight

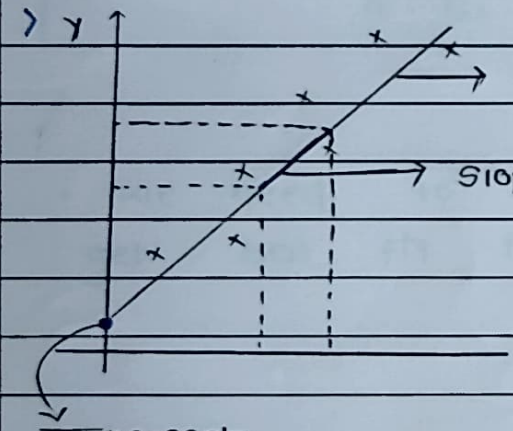
> eg: Aim to create a model which takes no. of rooms and predict the price.

Dataset : no. of rooms, price

> Eg: Aim to predict salary based on experience.



> Based on the training dataset, it find the best fit line in such way that the sum of difference between actual point & predicted point should be minimum



$y = mx + c$

$y = B_0 + B_1 x$ → eq of line

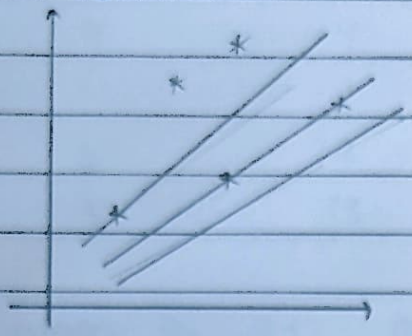
$h_\theta(x) = \theta_0 + \theta_x$

intercept ← → slope

> intercept: When $x=0$ line meeting the y-axis and that point is known as intercept.

> slope: with the unit movement in the x-axis what is the movement in y-axis

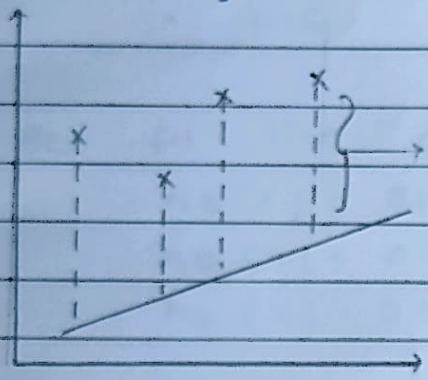
> In order to find best fit line we need to change the values θ_0 and θ_1.



⟶ changing the values of θ_0, θ_1 we will get best fit line

> Cost function

minimizing errors or residuals using equation.



Cost function

$$J(\theta_0, \theta_1) = \frac{1}{n} \sum_{i=1}^{n} \left(h_\theta(x)^{(i)} - y^{(i)} \right)^2$$

$h_\theta(x)^{(i)}$: predicted, $y^{(i)}$: Actual

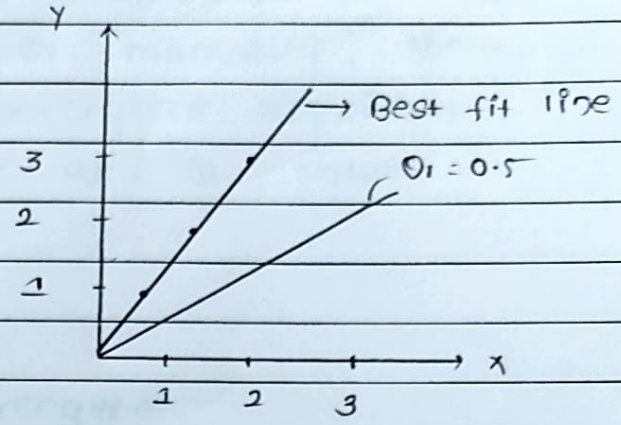
⟶ MEAN SQUARE ERROR

- We need to minimize the cost function to get best fit line

> eg

Training dataset

X	Y
1	1
2	2
3	3



When $\theta_1 = 0.5$, $\theta_0 = 0$

$$J(\theta_1) = \frac{1}{3}\left[(0.5-1)^2 + (1-2)^2 + (1.5-3)^2\right]$$

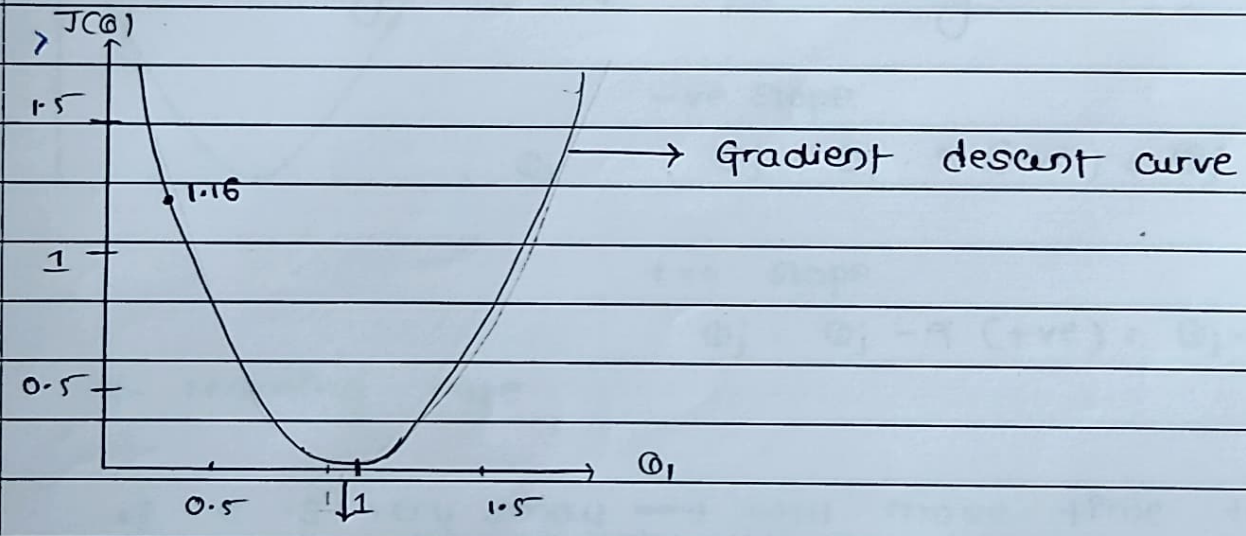
$$= 1.16$$

> Best fit line, $\theta_1 = 0.5$

$h_\theta(x) = 0 + 0.5(1) = 0.5$

$h_\theta(x) = 0 + 0.5(2) = 1$

$h_\theta(x) = 0 + 0.5(3) = 1.5$



minimum error at this point

or Global minima

) We need to reach global minima so we cannot change the value of θ_0, θ_1 manually. there is some mechanism called convergence algorithm that optimize the change of θ_1 value.

CONVERGENCE ALGORITHM :

Repeat until the convergence

{

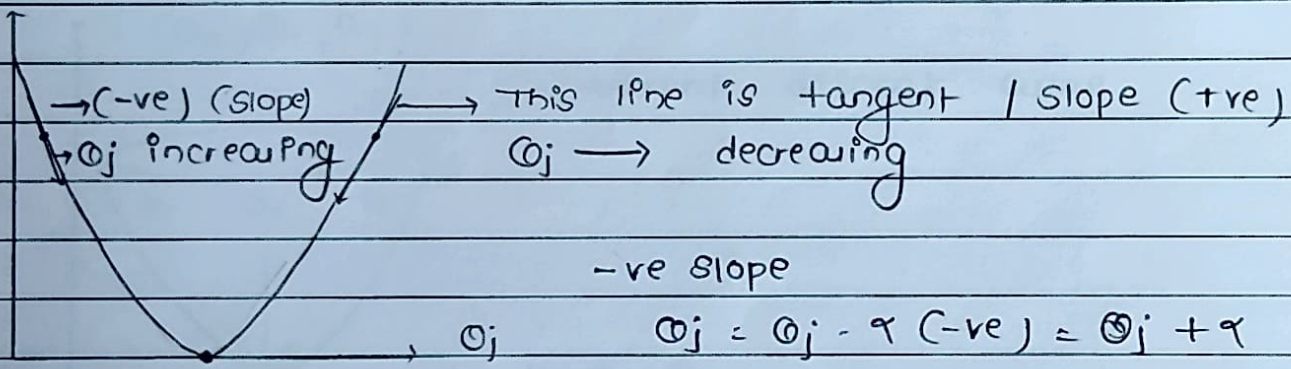
$$\theta_j = \theta_j - \alpha \frac{\partial J(\theta_j)}{\partial \theta_j}$$

slope

learning rate decide the convergence speed

learning rate

$J(\theta_j)$



-ve slope

$\theta_j = \theta_j - \alpha(-ve) = \theta_j + \alpha$

+ve slope

$\theta_j = \theta_j - \alpha(+ve) = \theta_j - \alpha$

α - learning rate

If α is very small → will more time to reach global minima

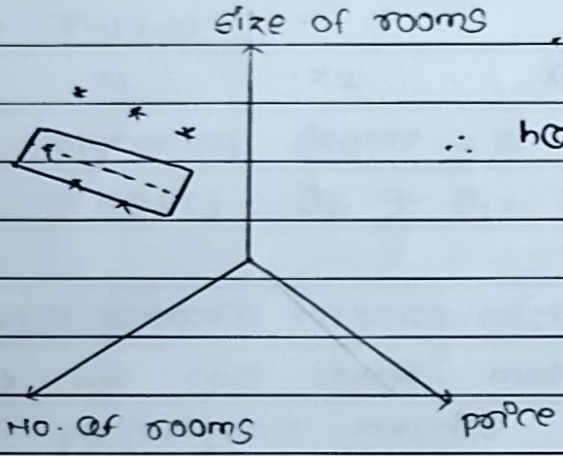
If α is very large → It will jump here & there won't reach global minima.

It should be around → 0.001 for smaller steps.

MULTIPLE LINEAR REGRESSION

- Two or more independent variable and one dependent variable.
- Example:

x_1	x_2	y
No. of rooms	size of rooms	price

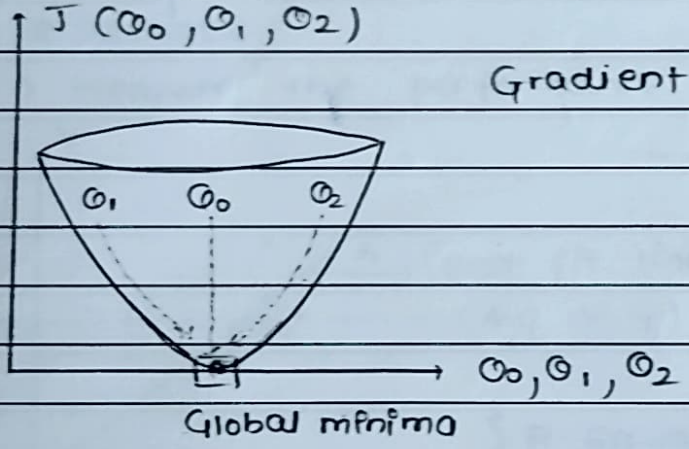


Try to find best fit plane

$\therefore h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$

$\theta_1, \theta_2 \longrightarrow$ slope or coefficient

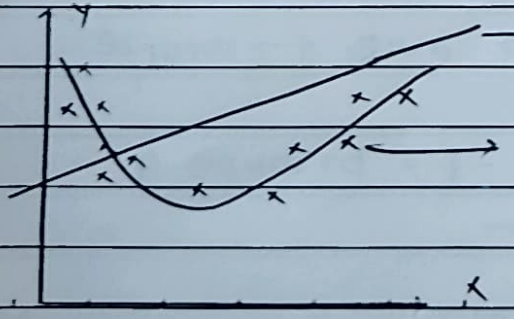
$\theta_0 \longrightarrow$ intercept



Gradient descent curve

POLYNOMIAL REGRESSION

- Non-linear relationship between independent and dependent variable



Simple linear relation

Non-linear relationship

polynomial degree = n

$\{ h_\theta(x) = \theta_0 x_1^0 + \theta_1 x_1^1 + \theta_2 x_1^2 + \dots + \theta_n x_1^n \}$

MULTIPLE POLYNOMIAL REGRESSION

› Multiple independent feature & one dependant feature

› Assumption

- linear relationship
- normaly distributed
- no multi collinearity

› Dataset

x_1 x_2 x_3 $y \longrightarrow$ o/p feature

polynomial degree = 2

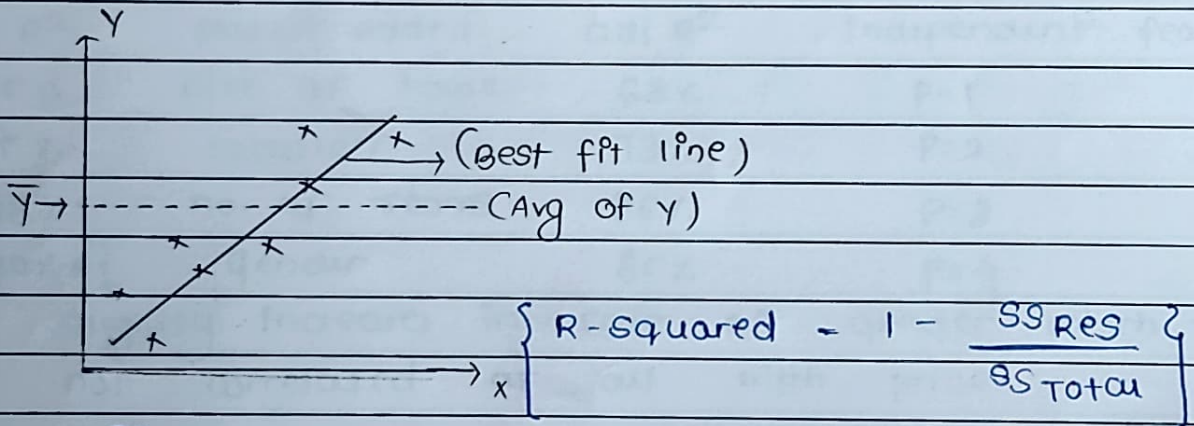
$\therefore h\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2$

PERFORMANCE METRICS USED IN LINEAR REGRESSION

› We can check model is good or not wing performance metrics.

1.) R-SQUARED

› Meaure the performance of the model



$$R\text{-squared} = 1 - \frac{SS_{RES}}{SS_{TOTAL}}$$

$SS_{RES} \longrightarrow$ sum of square residuals $(y_i - \hat{y}_i)^2$

$SS_{TOTAL} \longrightarrow$ sum of square average $(y_i - \bar{y}_i)^2$

$$R\text{-squared} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2}$$

$\sum (y_i - \hat{y}_i)^2 \longrightarrow$ low value

$\sum (y_i - \bar{y}_i)^2 \longrightarrow$ high value

R-Squared ≤ 1

If R-Squared = 0.85 = 85% accuracy
= 0.75 = 75% accuracy

ADJUSTED R-SQUARED

Example: Dataset

x_1	x_2	x_3	y
Size of howe	No. of room	Location	price

> R^2 keep on increasing when we add more independent feature. whether the independent feature is highly correlated or not it will keep on increasing as the number of independent feature increasing.

>

R^2	feature added	adj R^2	independent feature
65%	Size of howe	63%	P=1
75%	location	73%	P=2
88%	no. of rooms	86%	P=3
90%	Gender	85%	P=4

Slightly increase in case of Gender which is not correlated at all with price.

$$\text{Adjusted } R^2 = 1 - \frac{(1-R^2)(N-1)}{N-P-1}$$

N ⟶ Number of datapoint

P ⟶ Number of independent features

Adjusted R^2 is best metrics to evaluate the model

TYPES OF COST FUNCTION

1.) MSE (Mean square error)

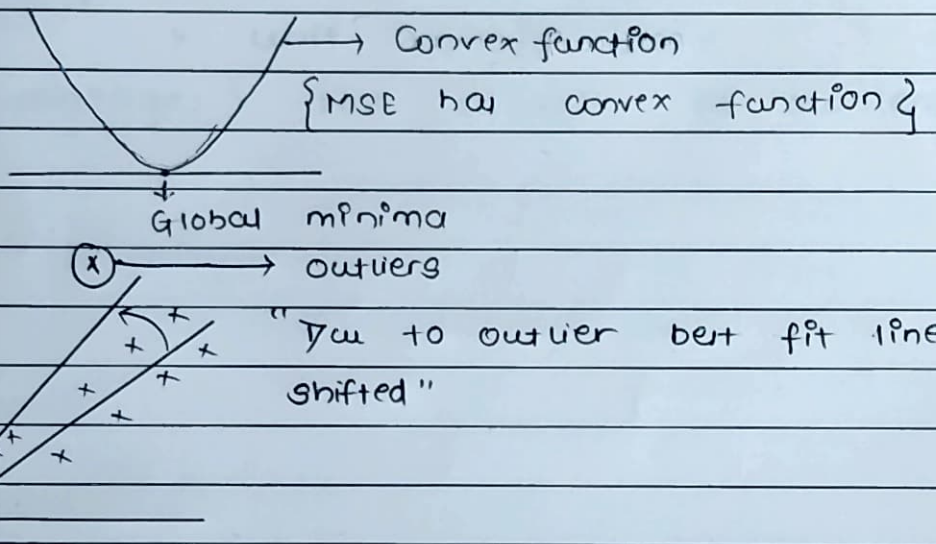
$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$y_i \longrightarrow$ Actual , $\hat{y}_i \longrightarrow$ predicted .

$\therefore \hat{y}_i = \theta_0 + \theta_1 x$

Advantage: > This equation is differentiable.
> It also has one global minima.

Disadvantag: > Not robut to outliers
> not same unit
eg: Rs (INR) $\longrightarrow$ $(INR)^2$



Eg: Experience $\longrightarrow$ independent feature
Salary $\longrightarrow$ dependent feature

$(y-\hat{y})^2$ $(lakhs)^2 \longrightarrow$ unit changing $\longrightarrow$ time complexity increases ↑↑

Ie don't do scaling for dependent feature

TYPES OF COST FUNCTION

1.) MSE (Mean square error)

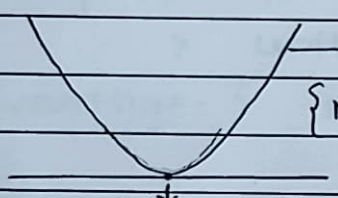
$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$y_i \longrightarrow$ Actual , $\hat{y}_i \longrightarrow$ predicted .

$\therefore \hat{y}_i = \theta_0 + \theta_1 x$

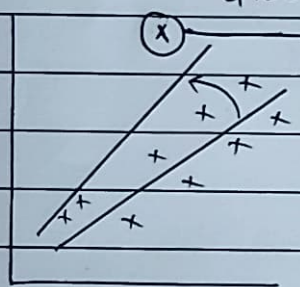
Advantage: > this equation is differentiable.
> It also has one global minima.

Disadvantag: > Not robut to outliers
> not same unit
eg: Rs (INR) $\longrightarrow$ $(INR)^2$



$\longrightarrow$ Convex function
{MSE has convex function}

Global minima



$\longrightarrow$ outliers

"Due to outlier best fit line get shifted"

Eg: Experience $\longrightarrow$ independent feature
salary $\longrightarrow$ dependent feature
$(y-\hat{y})^2$ $(lakhs)^2 \longrightarrow$ unit changing $\longrightarrow$ time complexity increase ↑↑
We don't do scaling for dependent feature

2.) MAE (Mean Absolute Error)

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y - \hat{y}|$$

Advantage: > robust to outliers
> same unit

Disadvantag: > convergence takes time
> optimization is complex

3.) RMSE (Root mean Squared Error)

$$RMSE = \sqrt{MSE}$$

Advantage: > it is differentiable
> unit remain same

Disadvantage: > Not robust to outliers.