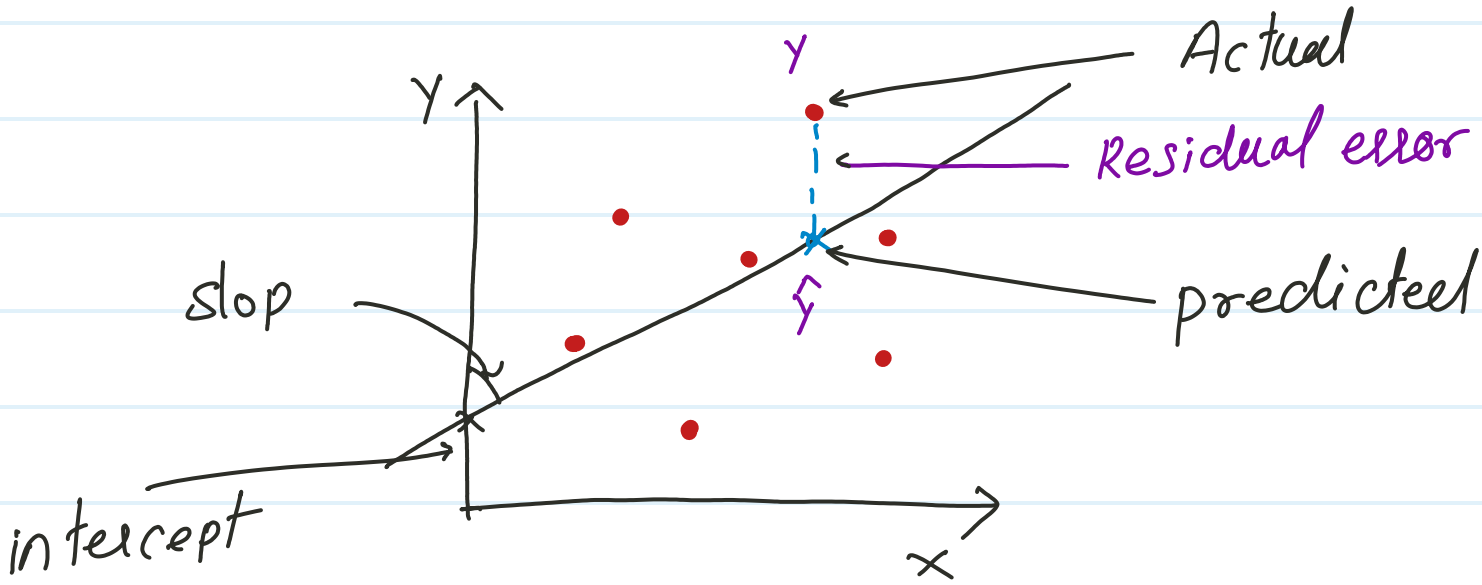
Linear Regression



Line eqn –

$$y = mx + c$$

y = Actual data

$\hat{y}$ = Predicted data

y = Dependent variable

x = Independent variable

m = slop

c = Intercept

$y - \hat{y}$ = Residual error

Base equation

$$y = mx + c$$

OR

$$y = h_\theta(x)$$

OR

Simple linear eqn.

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

$\theta_0 = c$

$\theta_1 = m$

$x_1 =$

OR

multi linear eqn

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_1 x_2 + \text{-------} + \theta_n x_n$$

$$y = J = h_\theta(x)$$

$$J = J(\theta_0, \theta_1)$$

Loss function

y $\hat{y}$

$$J(\theta_0, \theta_1) = (y - \hat{y})^2$$

It only calculate single datapoint

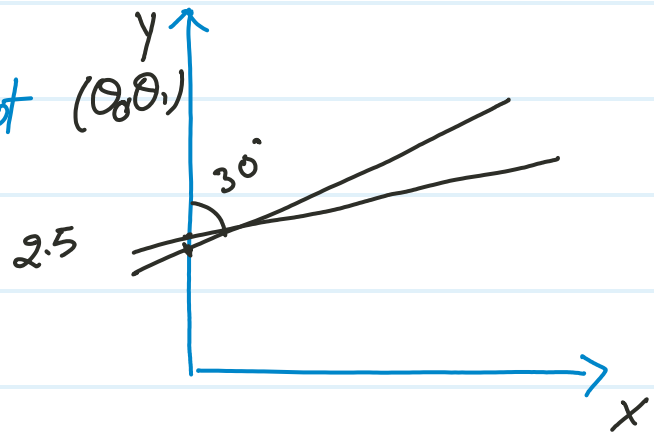
cost function

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} [h_\theta(x) - \hat{y}]^2$$

	y	$\hat{y}$	
1	3	4	$3 - 4 = -1$
2	6	5	$6 - 5 = 1$
3			
4	7	4	$7 - 4 = 3$

This is cost function to min. error by changing value of θ_0, θ_1

θ_0 = slop
θ_1 = intercept



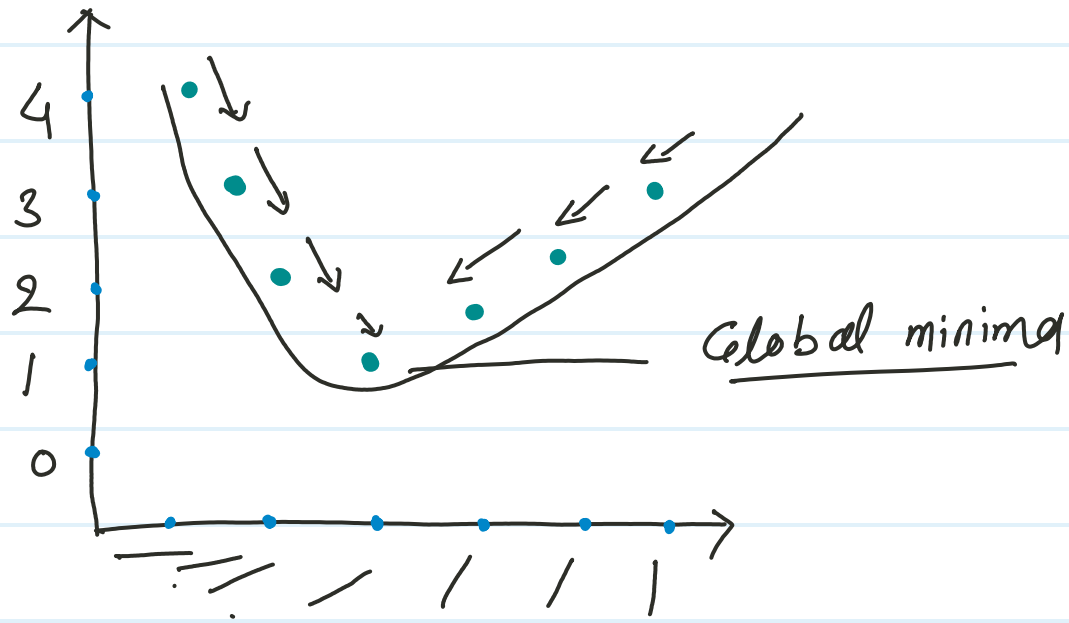
Repeat conversion theorm

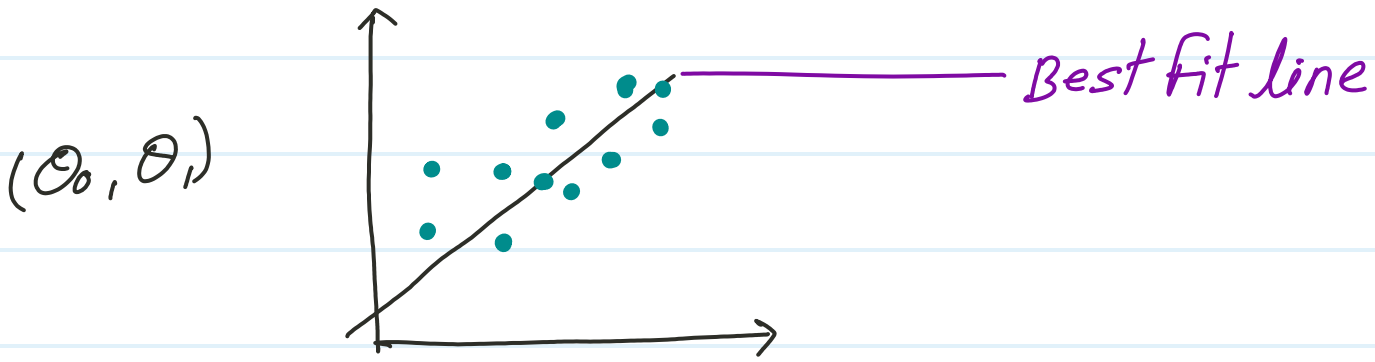
$$J(\theta_1) = J\theta_1 - \alpha \frac{d}{d} \left[J(\theta_1) \right]$$

α is learning rate 0.05, 0.25, 0.1, 0.01

$$J(\theta_0) = \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left[h\theta(x^i) - y^i \right]$$

Residual error





✱ Model evaluation or performance metrix

(i) MSE (mean squared error)
(ii) RMSE (Root mean square error)
(iii) MAE (mean Absolute error)
(iv) R^2
(v) Adj. R^2

$2-3 \quad (-1)-1$
$3-4 \quad (-1)-1$
$5-5 \quad (0)-0$

(i) MSE

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

$\frac{2}{3} \Rightarrow 0.5$

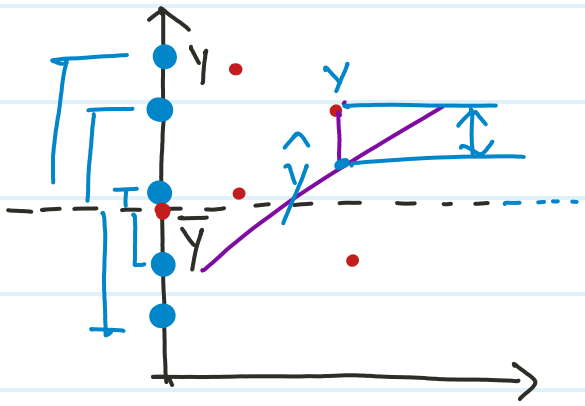
② RMSE

$$RMSE = \frac{1}{n}\sqrt{\sum_{i=1}^{n}(y-\hat{y})^2}$$

③ MAE

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

lower value better.



Accuracy Matrix

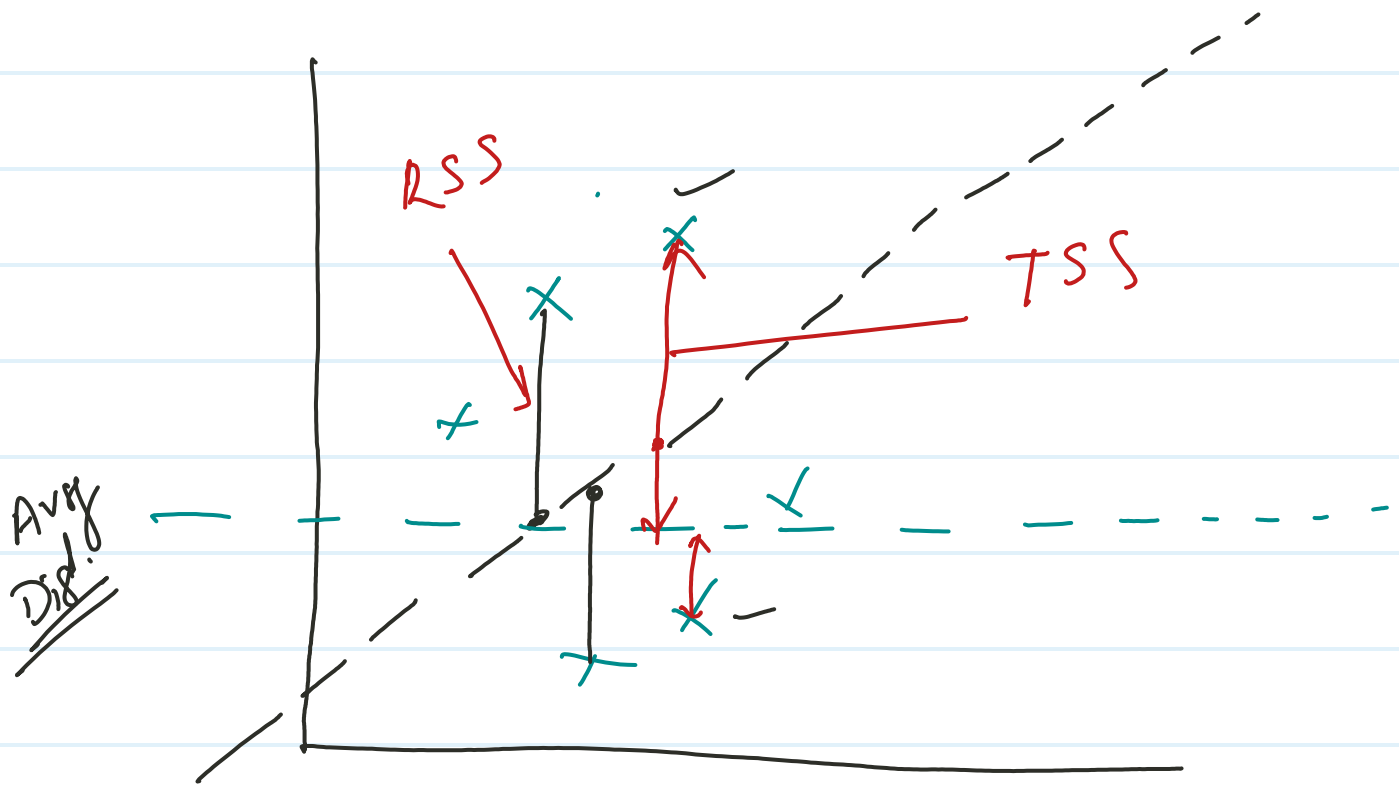
$$R^2 = 1 - \frac{RSS}{TSS}$$

R^2 = coeff. of determination

RSS = sum of square of residual

RSS = Distance b/w y and $\hat{y}$

TSS = Distance b/w y and $\bar{y}$



$$RSS = \sum (y - \hat{y})^2$$

$$TSS = \sum (y - \bar{y})^2$$

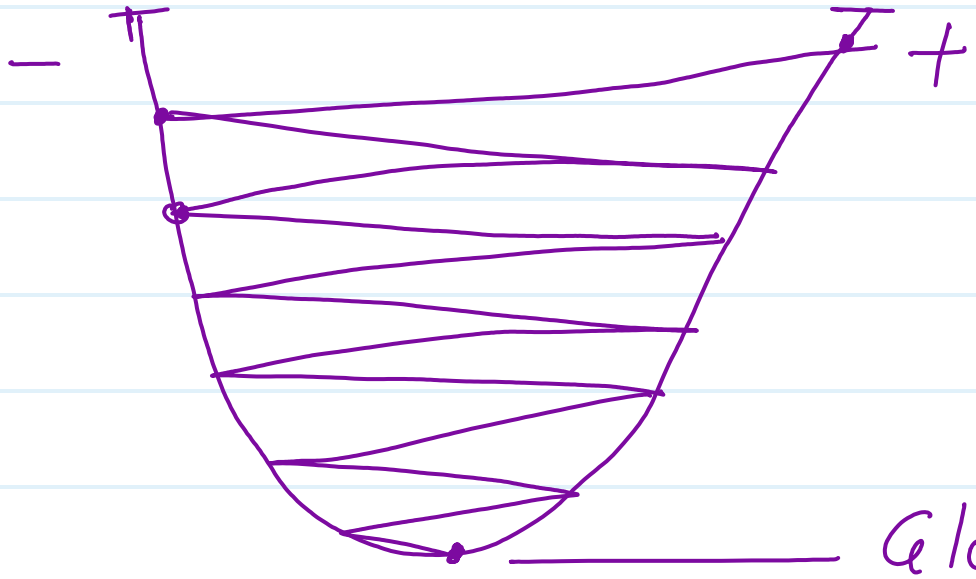
TSS Avg distance

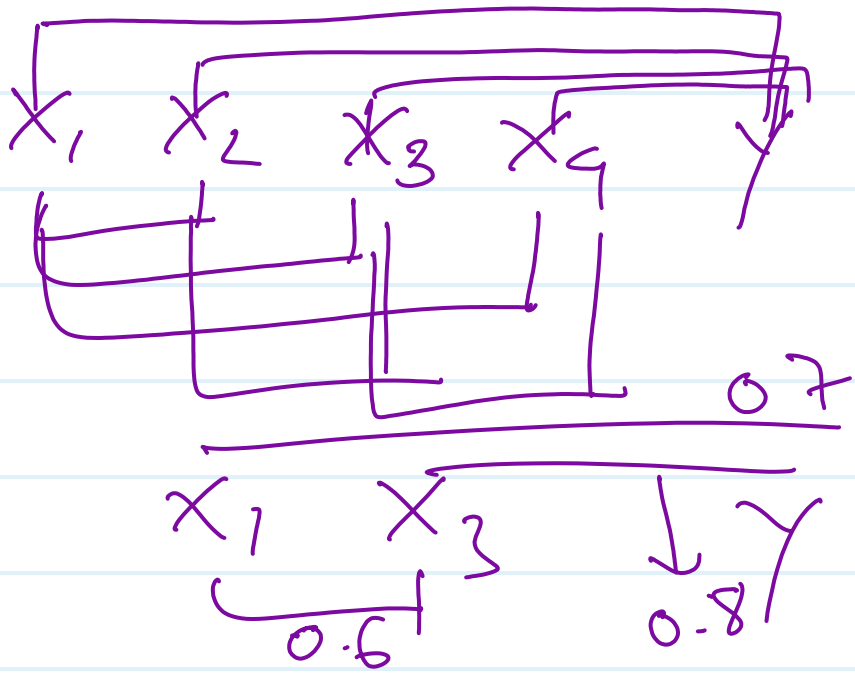
② Adj. R^2

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

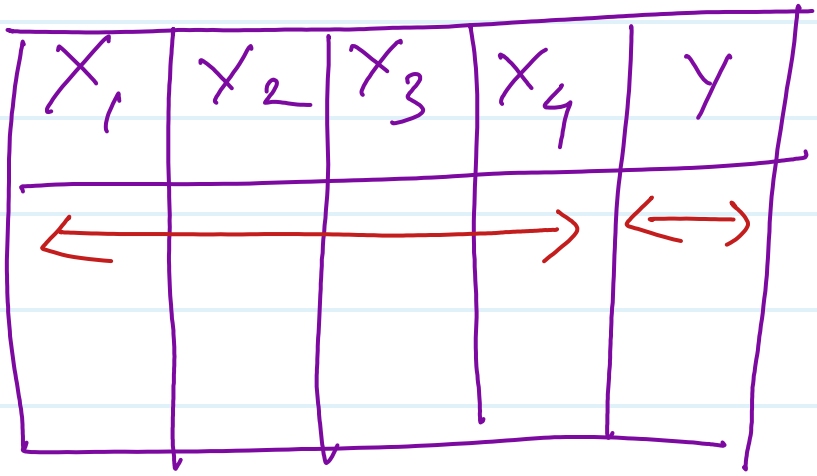
N = no. of datapoint in our dataset

P = no. of independ variable (x_1 x_2 x_3 ----)





✱ To find multi co-linearity



$X_1\ X_2$ $X_2\ X_3$

$X_1\ X_3$ $X_2\ X_4$

$X_1\ X_4$ $X_3\ X_4$

$X_1\ X_3\ X_4\ Y$

✱

VIF (variance inflation factors)

$$VIF = \frac{1}{1-R^2}$$

$X_1 - 3$

~~$X_2 - 6$~~

$X_3 - 4$

$X_4 - 5$

VIF = start 1 and it has no limit

IF 1 or less than 5 so no. multicolinearity

If > 5 so there will be co-linearity blw inde. Feature.

over fitting :-

low biased
high variance

under fitting

high biased
low variance

Best fitting

low biased
low variance