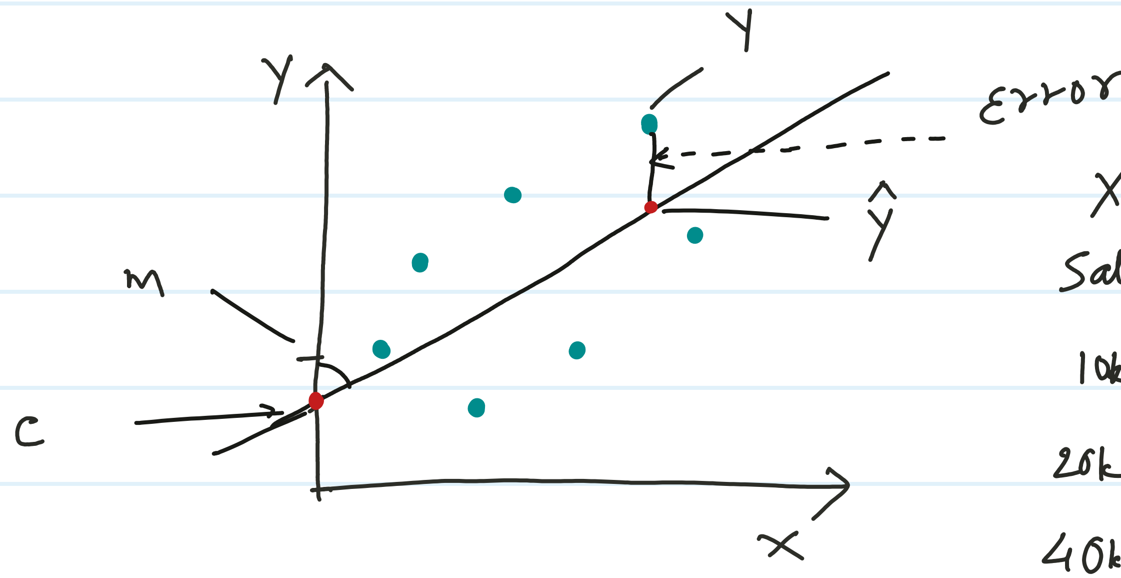
Linear Regression



X	Y
Salary	loan
10k	1L.
20k	2L.
40k	10L.

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

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

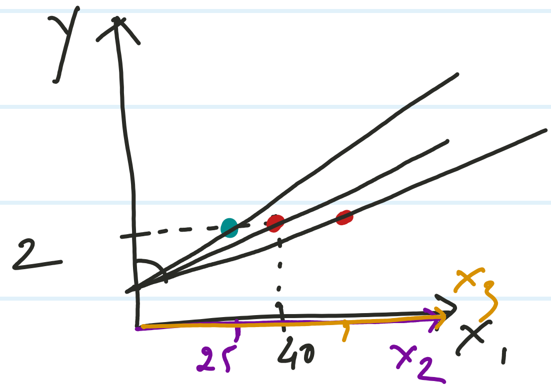
OR

Simple linear eqn. $$h_\theta(x) = \theta_0 + \theta_1 x_1$$

multi linear eqn $$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4$$

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

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



Loss function

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

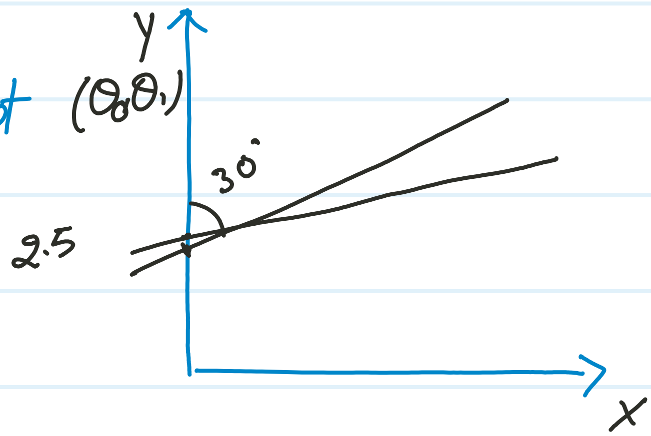
It only calculate single datapoint

cost function

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

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

θ_0 = slop
θ_1 = intercept

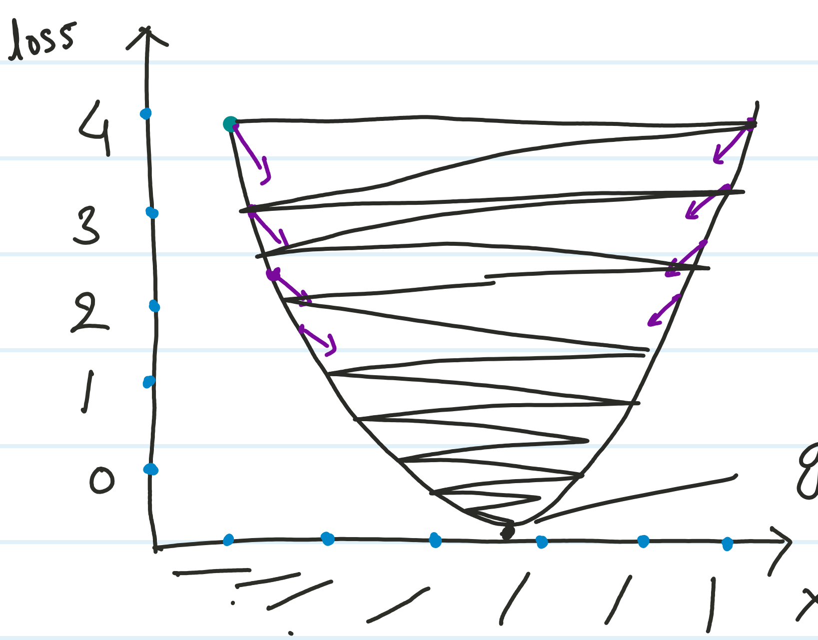


$J_\theta =$

Repeat conversion theorm

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

Residual error



θ_0, θ_1

global minima

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

$$\boxed{\theta_j = \theta_j - \alpha \frac{\partial J(\theta)}{\partial \theta_j}}$$ update eqn^n

for θ_0

$$\frac{\partial J\theta}{\partial \theta_0} = \frac{1}{m} \sum_{i=1}^{m} (\overbrace{h_\theta(x)^i}^{\hat{y}} - y^i) \quad \text{— (1)}$$

for θ_1

$$\frac{\partial J\theta}{\partial \theta_1} = \frac{1}{m} \sum_{i=1}^{m} (\overbrace{h_\theta(x)^i}^{\hat{y}} - y^i) * x^i \quad \text{— (2)}$$

update $\theta_0 = \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x)^i - y^i)$

update $\theta_1 = \theta_1 - \alpha \frac{1}{m} \sum (h_\theta(x)^i - y^i) * x^i$

α is learning rate

(0.01, 0.05, 0.1, 0.25, 0.3, 0.4. - - - -)

(θ_0, θ_1)

✱ Model evaluation or performance metrix

1. MSE (mean squared error)
2. RMSE (Root mean square error)
3. MAE (mean Absolute error)
4. R^2
5. Adj. R^2

2 − 3	(−1) − 1
3 − 4	(−1) − 1
5 − 5	(0) − 0

① MSE

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

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

② RMSE

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

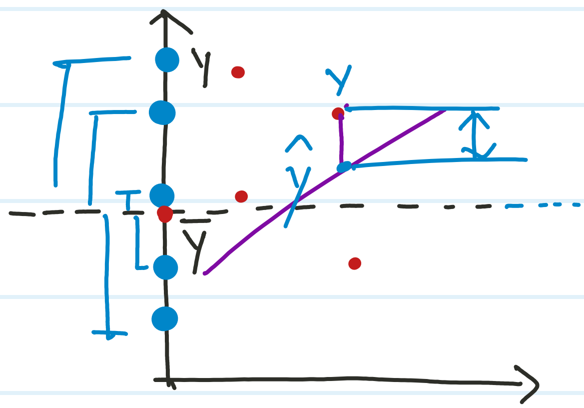
③ MAE

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

lower value better.

Accuracy matrix

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

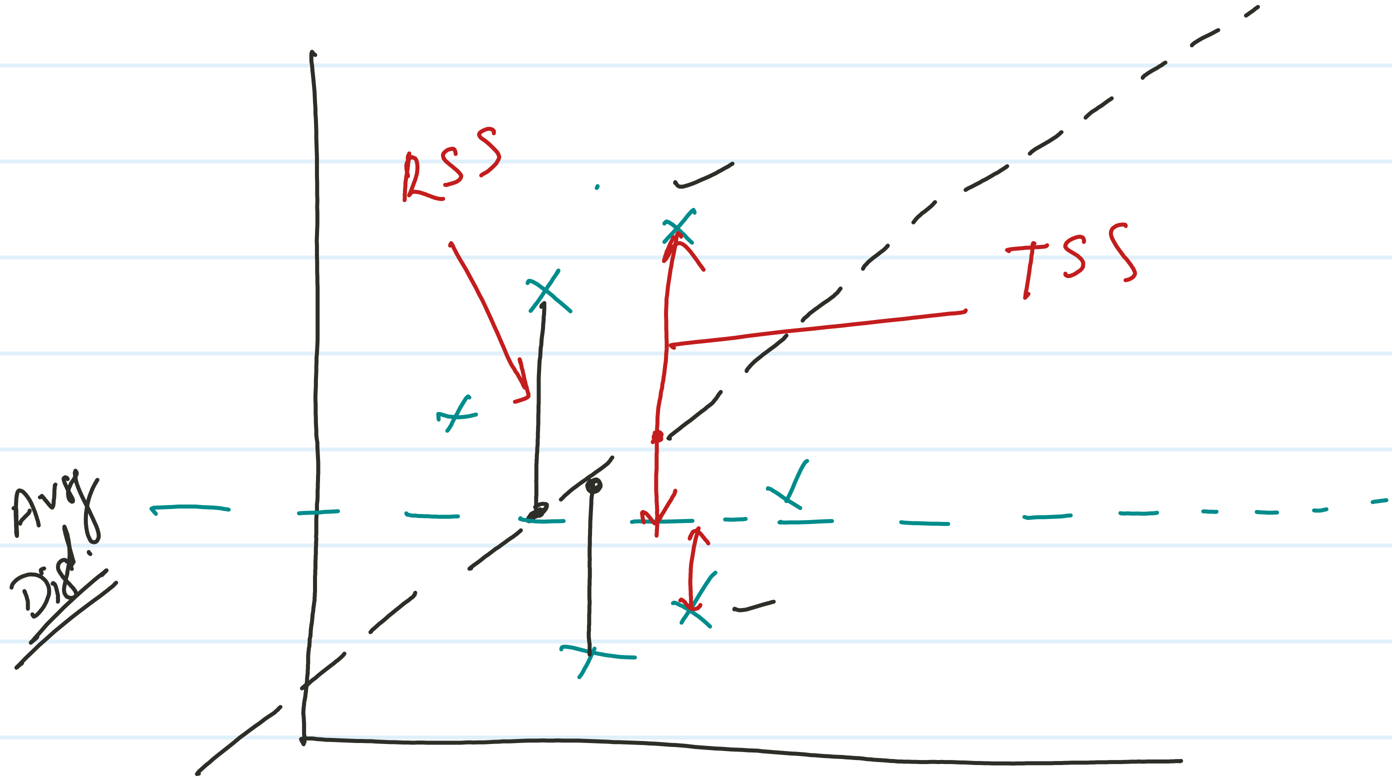


R^2 = coeff. of determination

RSS = sum of square of residual

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

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



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

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

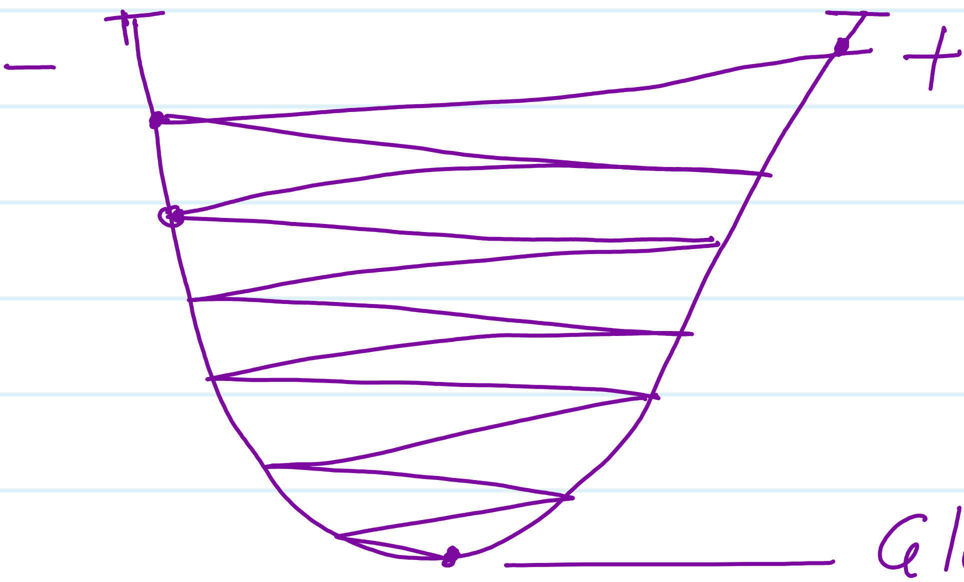
TSS Avg distance

② Adj. R^2

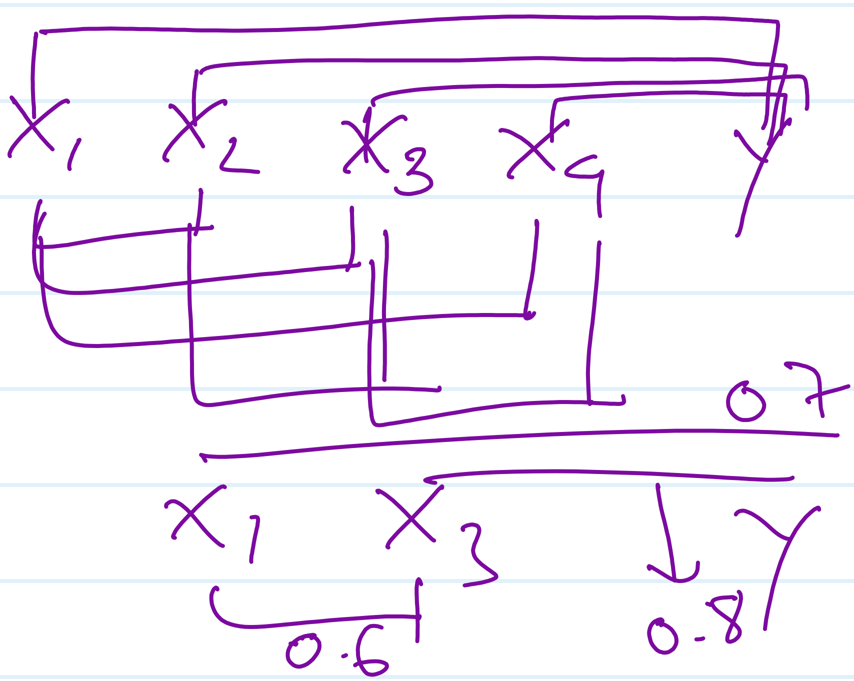
$$\text{Adj. } R^2 = 1 - \frac{(1-R^2)(N-1)}{N-P-1}$$

N = no. of datapoint in our dataset

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



Global minima



To find multi co-linearity

X_1	X_2	X_3	X_4	Y

$X_1\ X_2$ $\quad$ $X_2\ X_3$

$X_1\ X_3$ $\quad$ $X_2\ X_4$

$X_1\ X_4$ $\quad$ $X_3\ X_4$

$\boxed{X_1\ X_3\ X_4\ \ Y}$

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 b/w inde. Feature.

over fitting :-

low biased
high variance

under fitting

high biased
low variance

Best fitting

low biased
low varianced