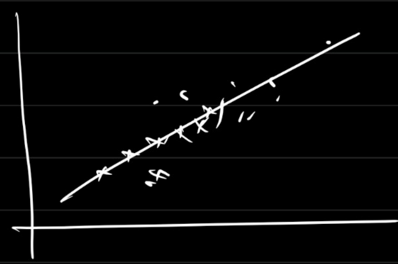


9:30

$$CF = \sum_{i=1}^{n} (y_{act} - y_{pred})^2$$

$$\sum_{i=1}^{n} (y_{act} - m x_i - c)^2$$

$Cf = |y_{act} - y_{pred}|$

Closed form solution → OLS

$$c = \bar{y} - m\bar{x} \quad , \quad m = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x - \bar{x})^2}$$

iterative soln. → $y = mx + c$

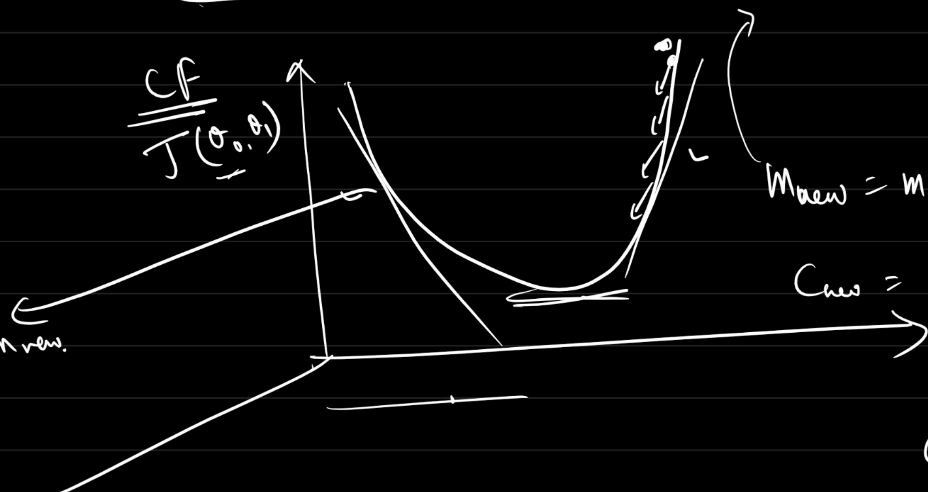
iterative soln

$m_{new} < m_{old}$

$J(w,b)$

$J(m,c)$

$\frac{CF}{J(\theta_0, \theta_1)}$



$$m_{new} = m_{old} - \eta \frac{\partial CF}{\partial m_{old}}$$

$$c_{new} = c_{old} - \eta \frac{\partial CF}{\partial c_{old}}$$

$m_{old} > m_{new}$

$c_{new} < c_{old}$

Agenda

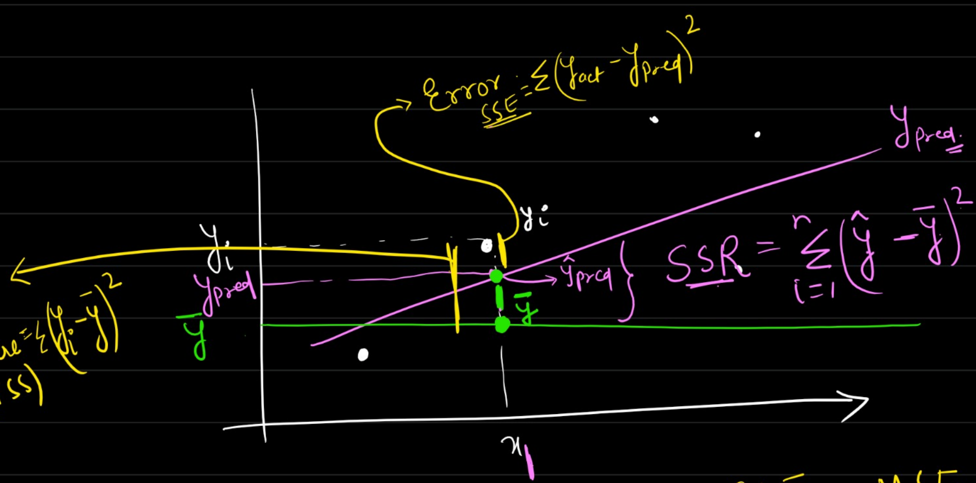
1. SLR implementations
2. MLR "
3. Evaluation metrics → Rsq, adj Rsq, MAE, RMSE, MSE.
4. Polynomial regression.

ML Pipeline

1. Read the data set
2. Prepare the data.
3. X, Y
4. train-test
5. Scaling
6. Model training
7. Model Evaluation.

Evaluation metrics

1. Rsquare
2. adjusted r-square
3. MSE
4. RMSE
5. MAE



x	y
	2
	3
	4
	5
	$\bar{y}$

$$\frac{SSE}{n} = MSE$$

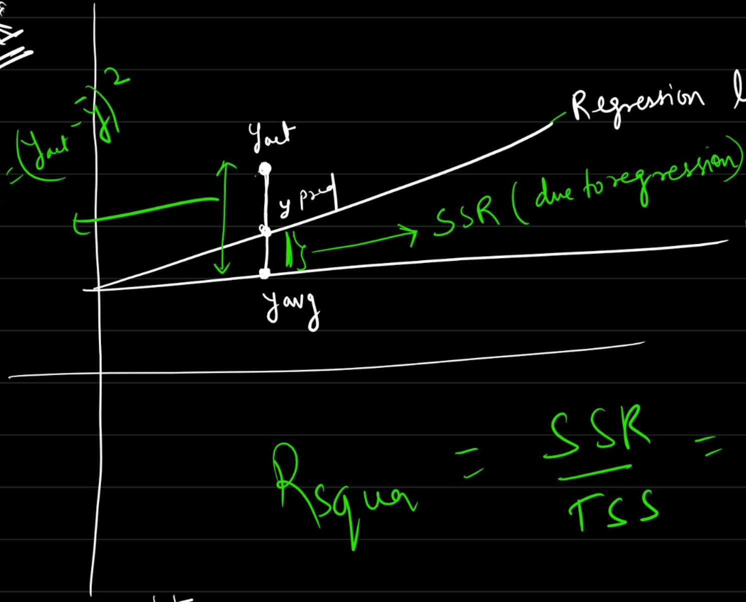
SSR — Sum of square due to regression.

$TSS = SSE + SSR$

$$Rsquare = 1 - \frac{SSE}{TSS} \text{ or } \frac{SSR}{TSS}$$

⇓

(Coefficient of determination) → Out of total Error, SSR is the variation explained by linear regression line



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

SSR (due to regression) $\Rightarrow (Y_{avg} - Y_{pred})^2 \rightarrow$ Can I say this much of variation in Y is explained by regression line

$$R_{squar} = \frac{SSR}{TSS}$$ = out of total variation, SSR is variation explained.

H/w

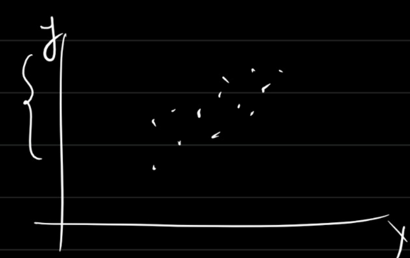
Rsquar
Explained var
&
Unexplained variation

$$r_{squar} = \frac{SSR}{TSS} \times 100$$

→ The percentage variation in Y explained by X.

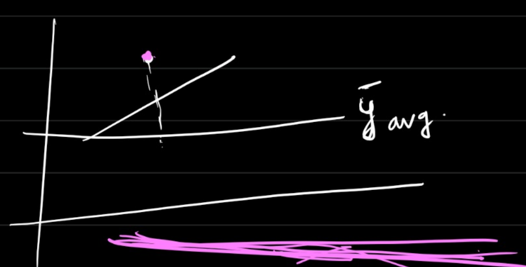
→ Rsquar ↑ Your model will be better.

$R_{sq} \uparrow$



Rsquare ⇒ Variation in Y explained ⇒ X is Capturing Y Well → Error (SSE) will be low ⇒ better the model will be.

Can r square be negative



$1 - \frac{SSE}{TSS}$ +ve

$1 - 1.2$ -0.2

(2) Adjusted rsquare

r^2 = %age explained variance in y due to x.

Predict Price of house. $x_1 - x_2 - x_3 - - x_{100}$

x_1 (Area of house) x_2 (no. of rooms) x_3 (gender) y price of house.

r square - ↑

$\frac{SSR}{TSS}$

$x_1 - y$ — rsquare – 80%

$x_1, x_2 - y$ — rsquare – 85%

$x_1\, x_2\, x_3 - y$ ⟶ rsquare – 85.5%

✓ As we add more features r square will improve / or remain as it is (constant)

x_1, x_2 - - - x_{100}

Adjusted r square

↓

It penalizes r square as we add new features.

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

N – no of dp
p – no of features.

Scen-1 $R^2 = 80, N = 11, p = 2$

$$\text{adj rsqu} = 1 - \frac{(1-0.8)(11-1)}{11-2-1} = 1 - \frac{0.2 \times 10}{8} = 0.75$$

$R^2 = 80, N = 11, p = 8$

$$\text{adj rsq} = 1 - \frac{(0.2)(10)}{11-8-1} = 1 - \frac{2}{2} = 0$$

* adj r sq < r square
* Only add features in the model if the difference b/w r square & adj r square is not more than 3-5 %

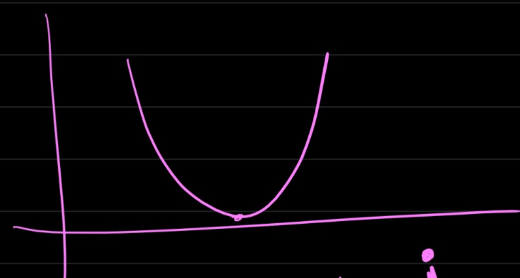
③ **MSE**

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_{act} - y_{pred}\right)^2$$

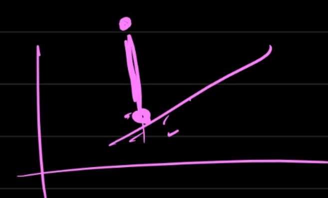
→ Error is quantified
→ Lower the MSE better the model will be.

* MSE measures the average of squared difference

* **Advantage**
→ it is differentiable
→ Emphasis on large error



$$\left(y_{act} - y_{pred}\right)^2$$



* **Disadvantage**
→ Not robust to outliers
→ It is not in same unit as Y

$$\begin{cases} y_{act} = 3\,m \\ y_{pred} = 6\,m \end{cases}$$

Error $= (3-6)^2$
$= 9\,m^2$

④ Mean absolute error

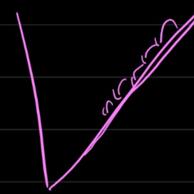
$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_{act} - y_{pred}\right|$$

Advantage
→ Less sensitive to outlier

→ more interpretable
→ It is of same unit

$|x|$ ⇒ 

Disadvantage
→ Not differentiable at $x=0$
→ Time consuming



Lower the MAE Better will be the model.

⑤ RMSE (Root mean squared error)

$$RMSE = \sqrt{MSE}$$

$$= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{act} - y_{pred})^2}$$

* Advantage
→ Same unit
→ Differentiable
→ less sensitive to outlier.