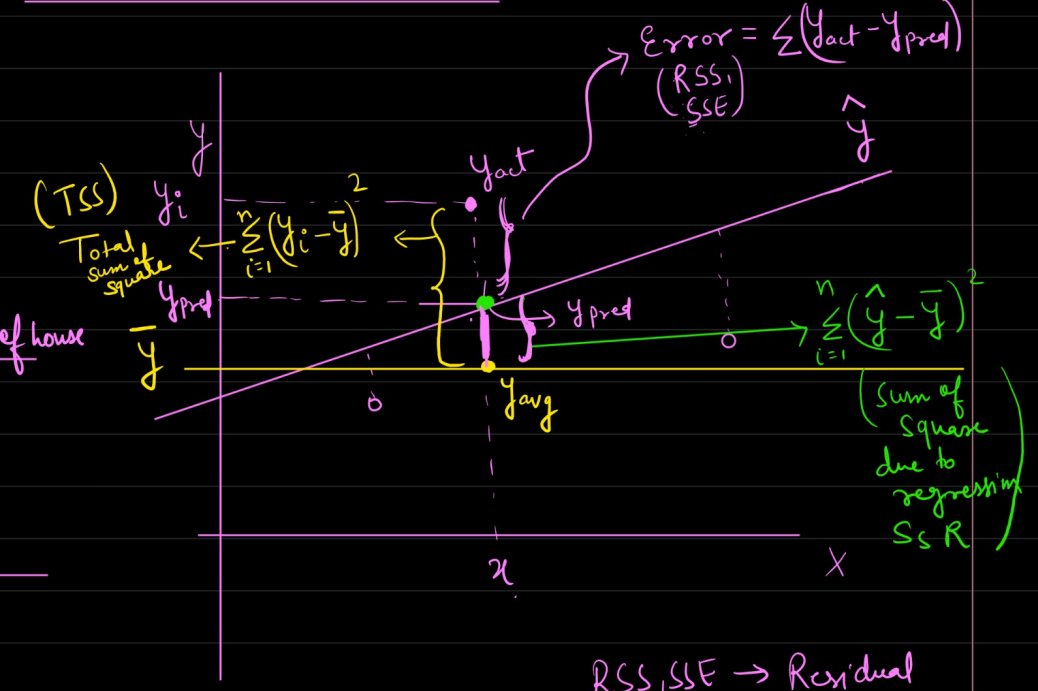
Evaluation metrics for Linear Regression

* Performance

① R square
② adjusted R square

Area of house	Price of house
1000	50
1100	60
=	=
=	=
=	=
	$\bar{y}$



RSS, SSE → Residual Sum of Square, → sum of squared error.

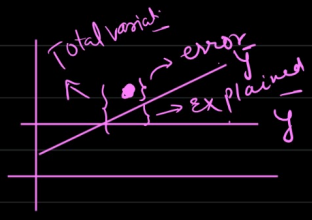
→ $\bar{y}$ becomes your baseline soln

SSR (Sum of Square due to regression) $\Rightarrow$ Explained variation in y by best fit line $\Rightarrow \sum_{i=1}^{n}(\hat{y} - \bar{y})^2$

SSE (Sum of Square Error) / RSS (Residual Sum of Square) $\Rightarrow$ Unexplained variation $\Rightarrow \sum_{i=1}^{n}(y_i - \hat{y})^2$

TSS (Total Sum of Squared) (SST - Sum of Square total) = Total variation in y = Ex Var + Unexplained var.

$= \sum_{i=1}^{n}(y_i - \bar{y})^2$



$$R^2 = 1 - \frac{RSS}{TSS} \quad \text{or} \quad \frac{SSR}{TSS}$$

RSS → Unexplained (error) variation; SSR → Explained variation; TSS → Total variation

R square = Coefficient of determination → Out of (total variation), SSR is variation explained.

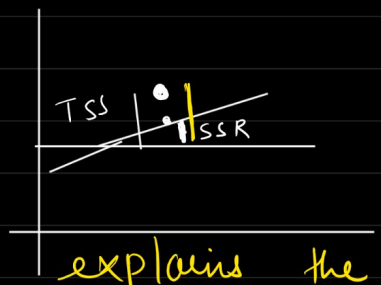
What is total percentage of variation explained by model?

SST | SSR

$$r\text{square} = \frac{SSR}{TSS} \times 100 =$$

→ The percentage variation in y explained by X is called as R-Square.

$$r\text{square} \Rightarrow r^2 \Rightarrow R^2 = \frac{SSR}{TSS}$$



SSR ≈ TSS, if it explains the variation completely.

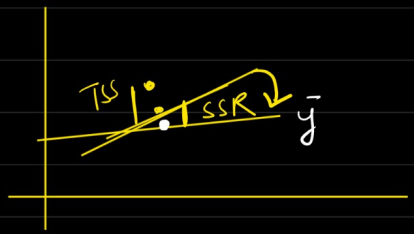
$$\frac{TSS}{TSS} \Rightarrow 1$$

The maximum value of r square is 1

SSR ≈ 0

$$\frac{SSR}{TSS} = 0$$

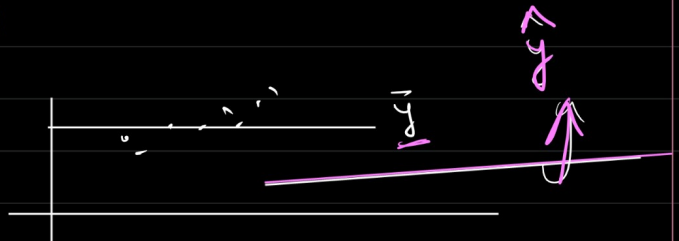
r Square min value = 0



r square (0,1)

Can r square be negative?

→ yes if the best fit



line is very very far from $\bar{y}$ (baseline model) which is not possible given the nature of algorithm.

$m_1 \rightarrow$ Rsquare 0.80 ✓

$m_2 \rightarrow$ Rsquare 0.60

② adjusted R-square

$r^2 \rightarrow$ % age explained variance in y due to x.

x_1 (Area of house)	x_2 (No of rooms)	x_3 Parking space	y Price of house

r square

$x_1 - y \rightarrow 80\%$

$x_1 x_2 - y \rightarrow 85\%$

$x_1 x_2 x_3 - y \rightarrow 88\%$

$x_1, x_2 \ldots x_{10} \rightarrow 89\%$

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

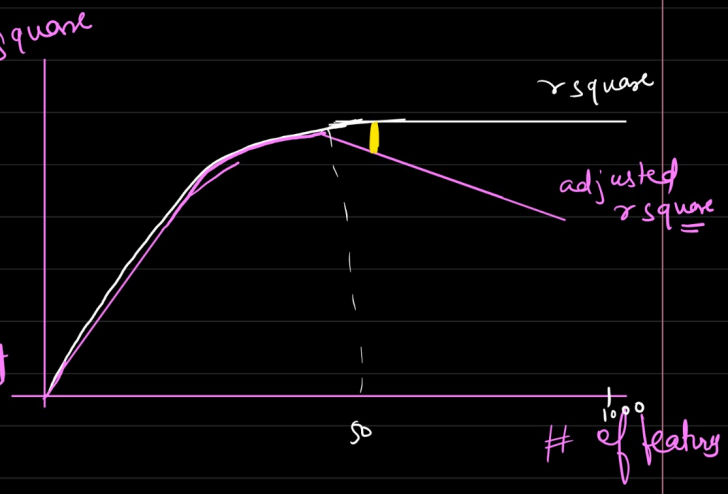
x_1 (Area of house)	x_2 (No of rooms)	x_3 (Parking space)	x_4 (Gender)	$x_5 \ldots x_{10}$	y Price of house

Adjusted R-square

$$= 1 - \frac{(1-R^2)(N-1)}{N-p-1}$$

$N \rightarrow$ No of dp's

$p =$ No of independent features.



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

$$\text{adj Rsquar} = 1 - \frac{(1-0.8)(11-1)}{11-2-1}$$

$$= 1 - \frac{0.2 \times 10}{8} \Rightarrow 0.75$$

Scen-2 $R^2 = 80\%$, $N = 11$, $p = 8$

$$\text{adj rsquare} = 1 - \frac{(0.2)(11-1)}{11-8-1}$$

$$= 1 - \frac{2}{2}$$

$$= 0$$

Points

→ Adj r square < R square

* Only add features in the model if the difference between Rsquare & adjusted Rsquare is not more than 3 – 5 %.