Evaluation metrics for Linear Regression

* Performance

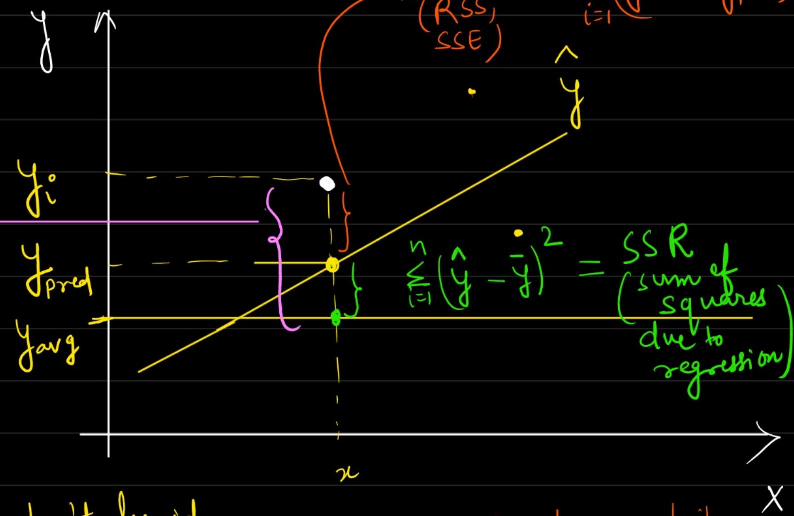
① Rsquare

② adjusted R square

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

TSS

Total sum of squares



Error (RSS, SSE) $= \sum_{i=1}^{n}(y_{act}-y_{pred})^2$

$\sum_{i=1}^{n}(\hat{y}-\bar{y})^2$ = SSR (sum of squares due to regression)

Area of house	Price of hous
1000	50
1100	60
	-
	-
	-

$\bar{y}$ ← avg of y.

→ $\bar{y}$ → if you don't build any model, the average value is the baseline solution.

* Error/RSS/SSE — if 1/n divided then it becomes MSE

Error/RSS ↓ Residual sum of square

SSE ↓ Sum of Squared Error

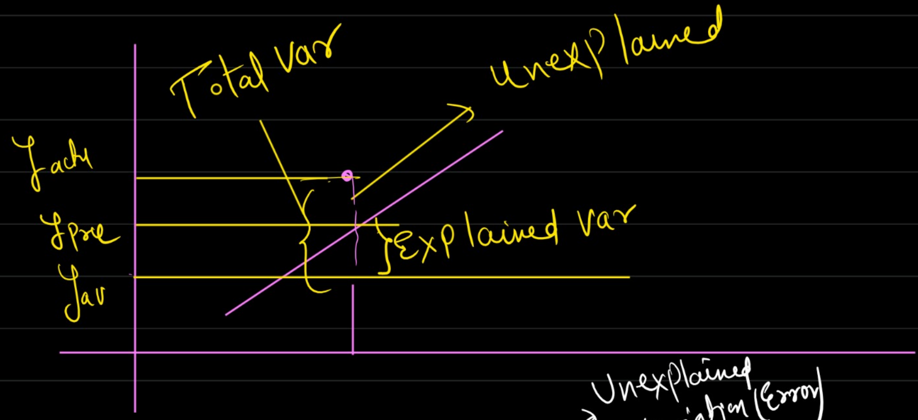
SSR (Sum of Squares due to regression) = Explained variation in y by best fit line $= \sum_{i=1}^{n}(\hat{y}-\bar{y})^2$

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

RSS (Residual Sum of square)

TSS (Total sum of squared Error) $\Rightarrow$ Total Variation in y / difference in y wrt to baseline ($\bar{y}$) $= \sum_{i=1}^{n}(y_i-\bar{y})^2$

TSS (Total variation in y) = Explained Variation (SSR) + Unexplained Variation (RSS, SSE)



$$R_{squared} = 1 - \frac{RSS}{TSS} \quad \text{or} \quad \frac{SSR}{TSS}$$

RSS → Unexplained variation (Error)

SSR → Explaining Variation

TSS → Total Variation.

TSS → Total variation

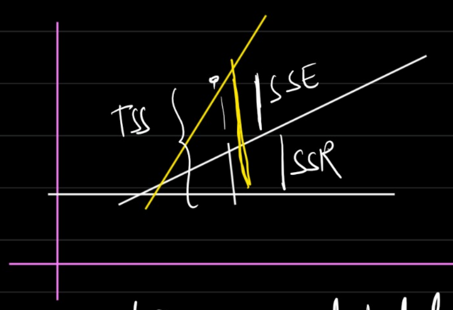
Rsquare = Coefficient of determination ⇒ Out of total variation, SSR is variation explained

Total percentage of variation explained by model.

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

→ The percentage variation in y explained by X is called R square.

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



SSR ≈ TSS, it explains the variation completely

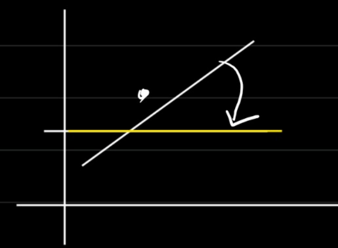
⇒ A perfect model.

$$\frac{TSS}{RSS \approx TSS} = 1$$

→ The maximum value of r square is 1.

→ The minimum value of r square is 0

$$\frac{SSR}{TSS} \qquad SSR \approx 0$$

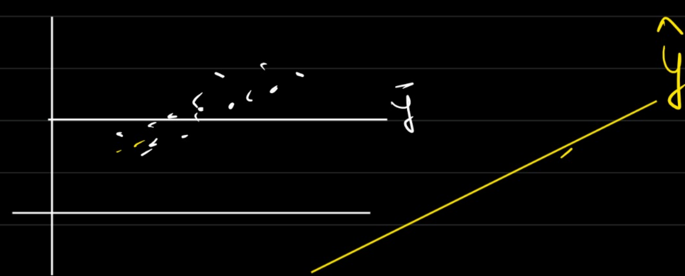


Range of r-square = 0 to 1

Closer to 1 better model will be.

* <u>Can r square be negative?</u>

Yes it can be negative.
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 ✓
$m_2 \rightarrow$,, 0.60

② <u>adjuster r-square</u>

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

x_1 (Area of house) x_2 (No of room) x_3 Parking space y price of house

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

rsquare

$x_1 - y \rightarrow 80\%$

$x_1 x_2 - y \rightarrow 85\%$

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

$x_1 x_2 x_3 \underline{x_4} - y \rightarrow 88.1\%$

$x_1 \quad x_2 \quad x_3 \quad x_4 \quad x_5 \quad \ldots \quad x_n \qquad y$

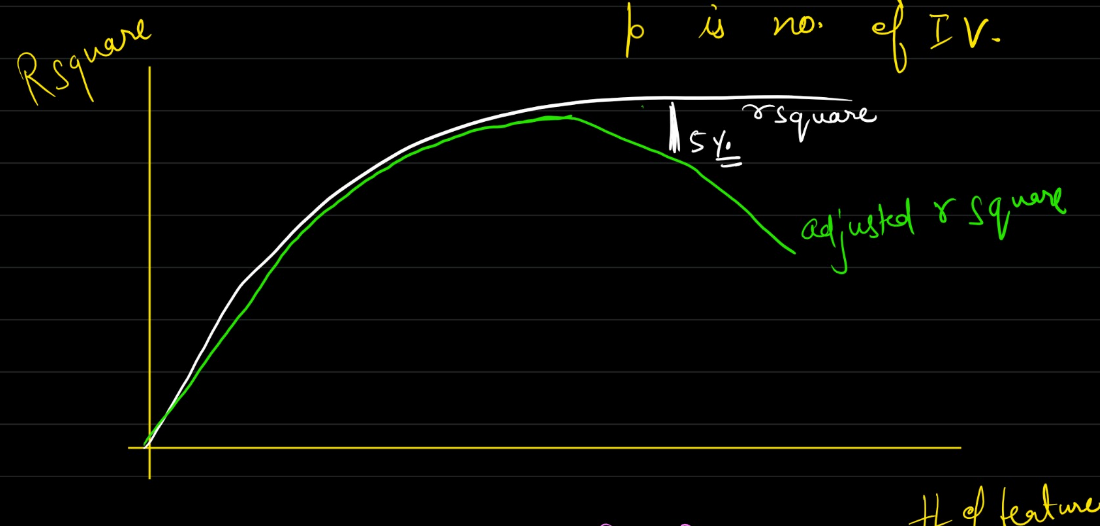
Area, # of room, Parking space, Gender

To Understand if the added feature | IV is contributing to the performance or not, we have adjusted r-square.

Adjusted r-square

$$1 - \frac{(1 - R\text{square})(N-1)}{N - P - 1}$$

where n is no. of dp's
p is no. of IV.



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

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

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

Scen-2

$Rsq = 80$, $N=11$, $p=8$

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

$$= 1 - \frac{2}{2} = 0$$

① Rsquare will always be lesser than adj rsquare. (Adj rsq < Rsquare)

② The difference between Rsq and adj Rsquare should not be more the 5%.

<u>Best practice</u> → Only add features in the model if the difference b/w rsquare & adjusted rsquare is not more than 3-5%.