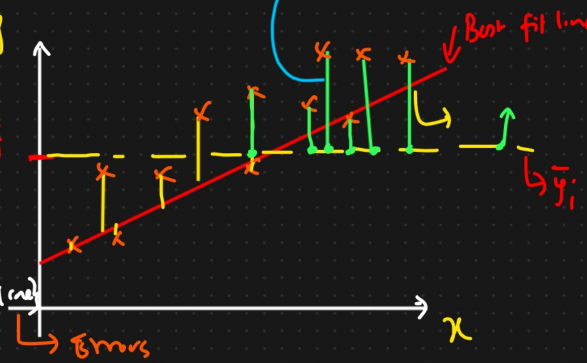
Performance Metrics Used In Linear Regression

① R squared ② Adjusted R Squared

① R Squared



(Mean) Average of y_i

R Squared = 1 − $\frac{SS_{Res} \rightarrow \{\text{Best fit line}\}}{SS_{Total} \rightarrow \{\text{Average of } y_i \text{ line}\} \rightarrow \text{Errors}}$

$(y-\hat{y})$

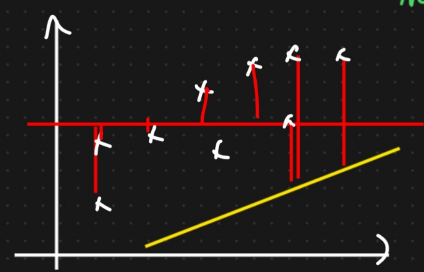
SS_{Res} = Sum of Squares Residual or Errors → Best fit line

SS_{Total} = Sum of Squared Total → Sum of these differences → $\bar{y}_i$

$$\boxed{R\ squared = 1 - \frac{\sum (y_i-\hat{y}_i)^2}{\sum (y_i-\bar{y}_i)^2}} \Rightarrow \frac{\text{Small value}}{\text{Big value}}$$

The best fit line should be atleast better than the line denoting $\bar{y}_i$

"R Squared can be −ve" if this best fit line is worse than average line



$f_1 \quad f_2 \quad f_3 \quad f_4 \rightarrow O/P$

Max R squared = 1

If −ve bad model

R squared ⇒ 0.75 ⇒ 75%

⇒ 0.85 ⇒ 85%

R squared ↑↑↑ inc if features are related to output

{Overfitting Problem}.

$0 < Rsq < 1$ → overfitting

② Adjusted R squared

x_4	x_1	x_2	x_3	y
Gender	Size of the house	No. of bedrooms	$\boxed{\text{Location}}$	→ Price

R squared = 85%

R squared = 90% (when location added)

R squared = 91% when X_4 added

Adjusted R squared = $1 - \dfrac{(1-R^2)(N-1)}{N-p-1}$.

N = No. of data points

p = No. of Independent features

$R^2 = 80\%$ $N = 11$ $P = 2$

Adjusted R squared = $1 - \dfrac{(0.2)(10)}{11-2-1} = 1 - \dfrac{\cancel{2}\,1}{\cancel{8}\,4}$

$= 1 - \frac{1}{4}$

$= 3/4 = 0.75$

when R^2 = 80% then adjusted R^2 = 75%

p=3 R^2 = 85%   adjusted R^2 = 78% a imp feat added

R^2 = 88%   Adjusted R^2 = 76% a useless feature added

adjusted R^2 always less than R^2

⇓

Independent is not

important

adjusted R^2 is better than R^2