$$\frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x)^i - y^i \right)^2$$

Multiple Linear Regression

House pricing

No. of Rooms | Size | location | ?

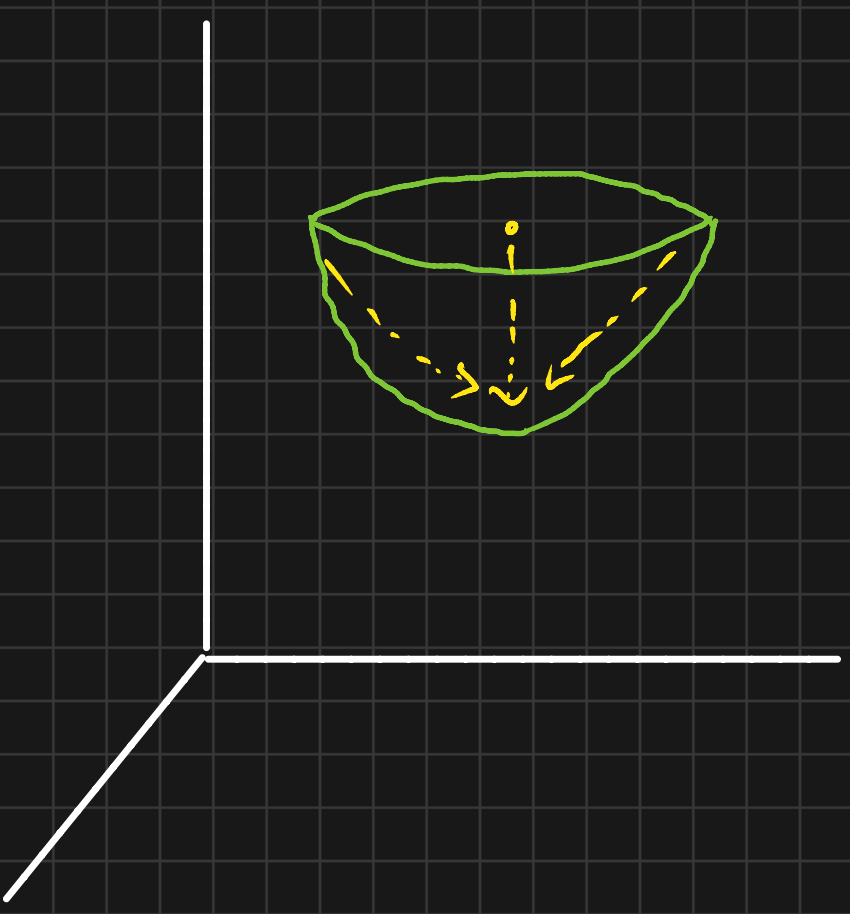
independent feature

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3$$

→ multiple linear regression

$\theta_1, \theta_2, \theta_3$ are Cofficient

θ_0 is intercept

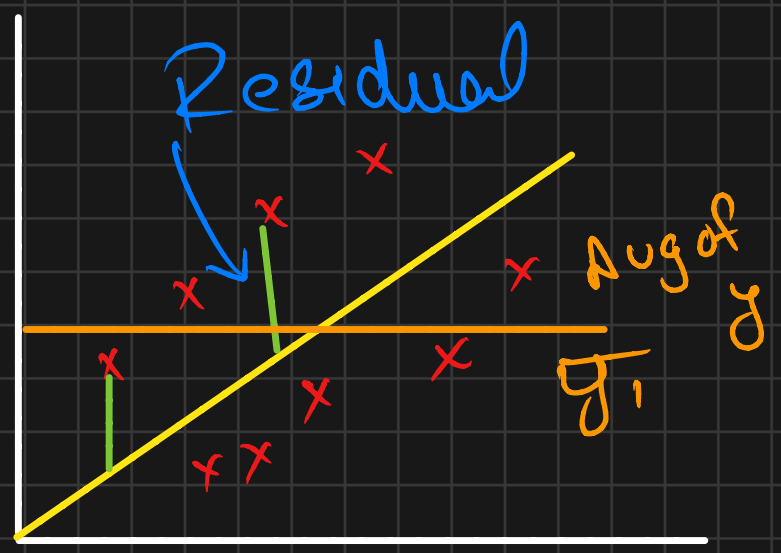


Performance Matrix

① R squared ② Adjusted R squared

$$R_{sq} = 1 - \frac{SS_{residual}}{SS_{total}}$$

$$R_{sq} = 1 - \frac{\Sigma (y - \hat{y}_i)^2}{\Sigma (y_i - \bar{y}_i)^2}$$



$$1 - \frac{\text{Smaller no.}}{\text{Bigger no.}} \approx 1$$

- Value more near to 1 the model performs that well

② Adjusted R square

Size → Price $(R)^2 \propto$ Size $\propto$ Price

+ve Correlation

$R_{sq} = .75$

Size → no. of Room → Price ⇒ Size $\propto$ no. of room $\propto$ Price

$R_{sq} = .80$ (80%)

↓

Size no. of Room Location Price

$R_{sq} = .85$ (85%)

↓

Size no. of Room Location Gender Price

$R_{sq} = .87$ (87%)

⇒ R_{sq} value increased but Gender & Price doesnot relate with each other

⇒ <u>adjusted R sq</u>

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

N = No. of data point

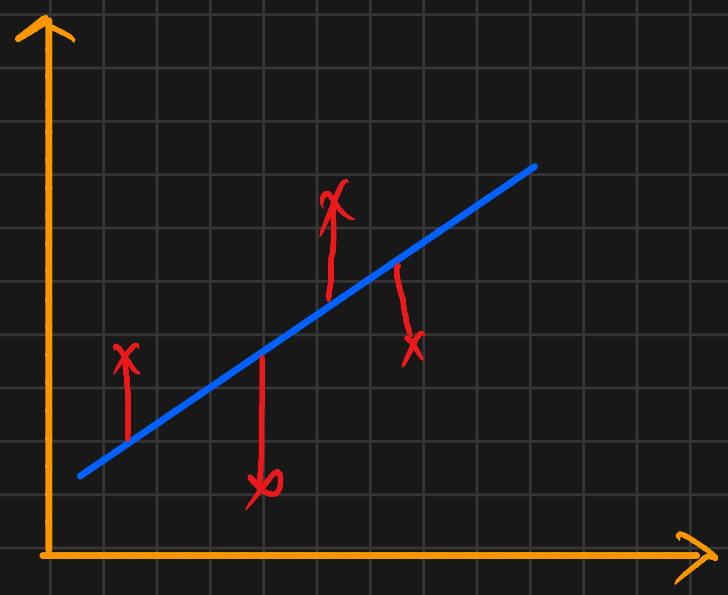
P = No. of Independent feature

⇒ MSE, MAE, RMSE

Exp Salary

= =

= =

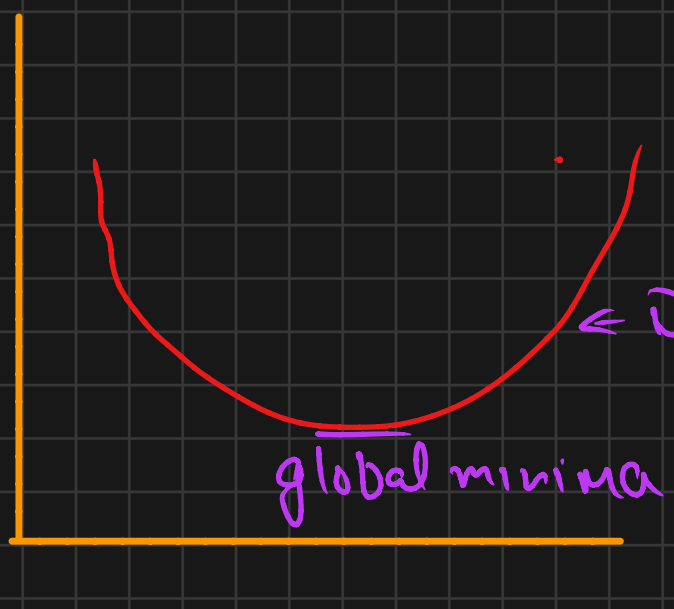


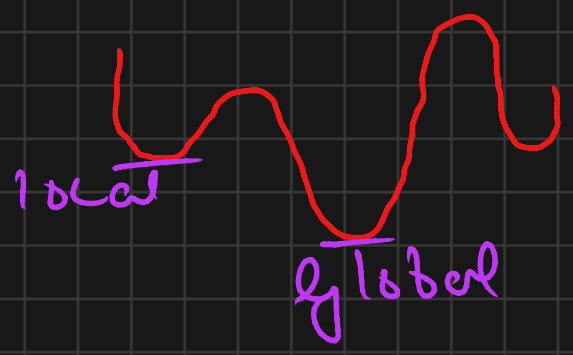
Mean Square Error

$$\sum_{i=1}^{n} \frac{(y - \hat{y})^2}{n} = \text{cost fxn} \downarrow$$

advantage

1) it is diffrentiable
2) It has one local & one global minima



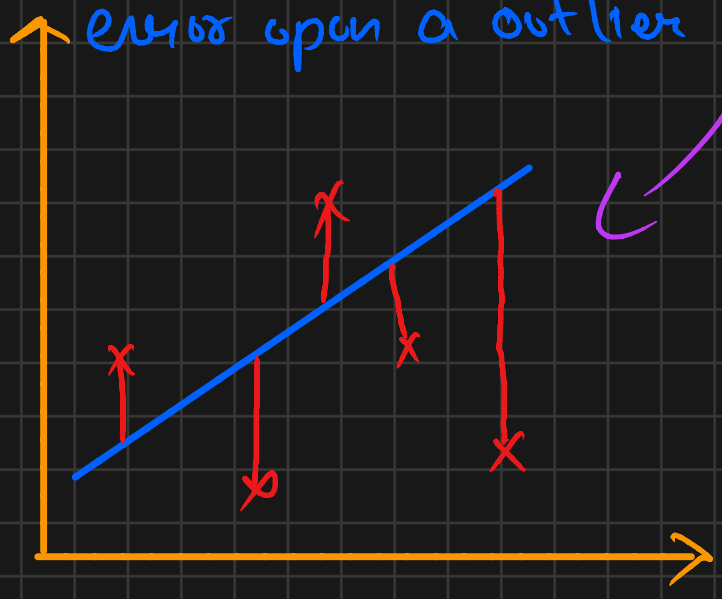


Disadvantage

1) It is not Robust

it will increase error upon a outlier

if error ↑ the best fit line will be changed



if no global it will Converge at local

x (lakhs) y $(y_i - \bar{y}_i)$ (lakhs)^2 (Units changed)

Mean Absolute Error

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

Advantage

- Robust to outliers
- $|y - \hat{y}_i|$ will be in same units

Disadvantage

- Not differentiable (we need sub gradient & convergence take more time)
- Time Consuming

Root Mean Squared Error

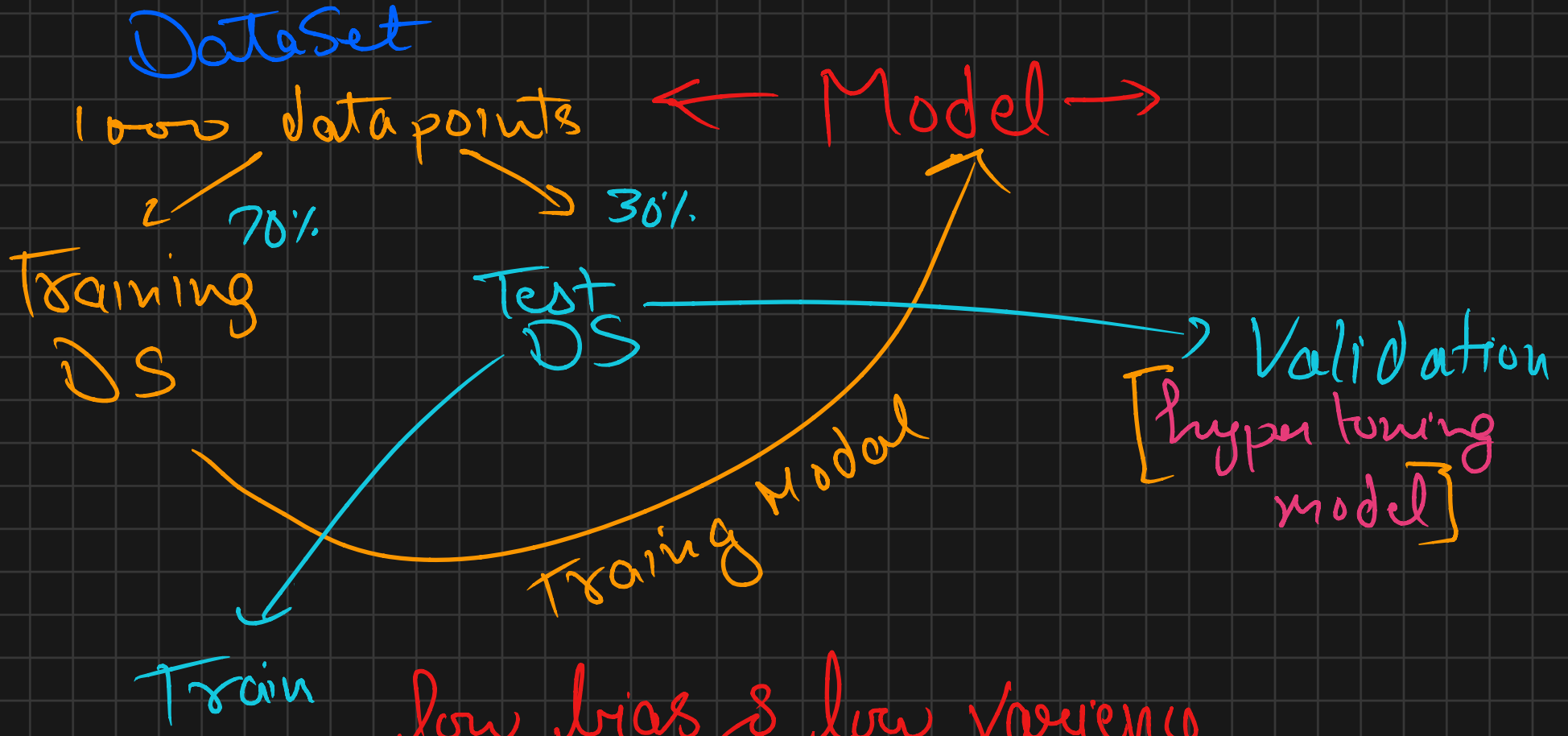
Advantage

① Same unit

② Differentiable

⇒ We should consider all these will Creating Regression Model

Overfitting & Underfitting



low bias & low variance

Train DS good accuracy

Test DS good accuracy

⇒ low bias high variance

Good accuracy.

⇓

modal is overfitting

Bias Variance tradeoff

Train accuracy low

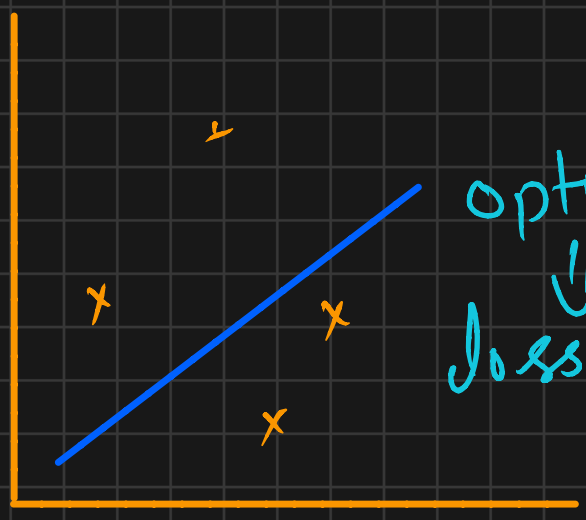
Test accuracy is low

⇓

modal is underfitting

⇒ high bias high variance

Linear Regression Using OLS



optimization
⇓
loss fxn

OLS = formula & Calculate
Reduce error $h_\theta(x) = \beta_0 + \beta_1 x_1$

$$S(\beta_0, \beta_1) =$$

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_1)^2$$

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = \frac{2}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)(0-1-0)$$

$$\frac{-2}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_1) \leftarrow$$

①

$$\frac{\partial S}{\partial \beta_1}(\beta_0, \beta_1) = \frac{2}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_1)(-x_1) = 0$$

②

$$\frac{-2}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_1)(x_1) = 0$$

Eq ①

$$\frac{-2}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_1) = 0$$

$$-\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_1) = 0$$

$$\boxed{\beta_0 = \bar{y} - \beta_1 \bar{x}}$$

Eq (2)

$$\frac{-2}{n} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)(x_i) = 0$$

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)(x_i) = 0$$

$$\sum_{i=1}^{n} x_i y_i - \beta_0 \sum_{i=1}^{n} (x_i) - \beta_1 \sum_{i=1}^{n} (x_i)^2 = 0$$

$$\beta_1 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})}$$