Session 49 – Simple Linear Regression

→ Simple Linear Regression is a method used to find the relationship between two variables – one independent variable (x) and one dependent variable (y)

→ It tries to find a straight line that best fits the data points.

$$\boxed{\begin{array}{c} y = b_0 + b_1 x \\ b + mx \end{array}}$$

y → Dependent variable (the one we want to predict)
x → Independent variable (the one used for prediction)
b_0 → Intercept (value of y when $x = 0$)
b_1 → Slope (how much y changes when x increases by 1)

Linear Regression
- Simple Linear Regression
- Multiple LR
- Polynomial LR
- Regularization

Ex

i/p	o/p

cgpa	Package (lpa)
6.66	3.5

cgpa	gender	12th marks	state	pack

multiple i/ps

10/0

```
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np

df = pd.read_csv('placement.csv')
df.head()
```

⇒ o/p:

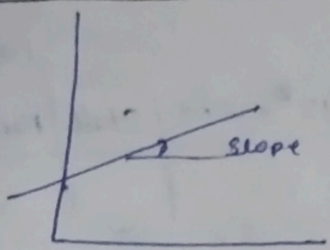
	cgpa	pack
0	6.89	3.2
1	5.12	1.9
2	7.82	3.2
3	7.42	3.6
4	6.94	3.5

```
plt.scatter(df['cgpa'], df['package'])
plt.xlabel('CGPA')
plt.ylable('Package (in lpa)')
```

```
y = df.iloc[:, -1]

from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y,
                                   test_size = 0.2, random_state = 2)

from sklearn.linear_model import LinearRegression
lr = LinearRegression()
lr.fit(X_train, y_train)
lr.predict(X_test.iloc[0].values.reshape(1,1))

plt.scatter(df['cgpa'], df['package'])
plt.plot(X_train, lr.predict(X_train), color='red')
plt.xlabel('CGPA')
plt.ylabel('Package(in lpa)')
```

$\boxed{\# \ y = mx + b}$ m = lr.coef_
b = lr.intercept_

Mathematical Intuition :-



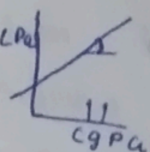
$\boxed{y = mx + b}$

$\boxed{\text{Package} = m \times \text{cgpa} + b}$

here m ⟶ slope ⟶ weightage } how Package depends on cgpa

→ when m ↓ ⇒ less dependent on cgpa

→ when m ↑ ⇒ Package depends more on cgpa



$\boxed{\text{when } b = 0 \quad \text{Package} = m \times \text{cgpa}}$

not good ex

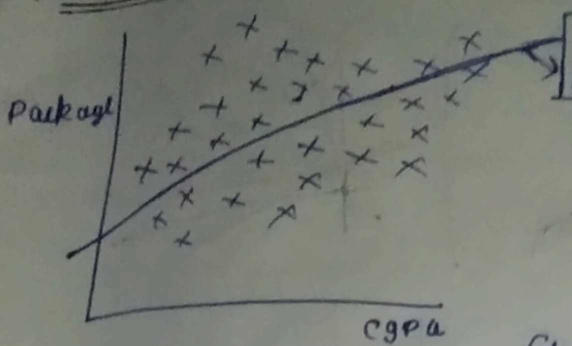
when cgpa = 0, then Package = 0

Ex:- $\boxed{\text{Package} = m \times \text{Exp} + b}$ ⟶ offset
$m \times 0 + 0$

when Exp = 0, then Package = 0

but freshers do get Salary, so here Salary is Package that is $\boxed{\text{Package} = b}$

How to find m and b ?



$\boxed{y = mx + b}$

↑

(m, b)

Closed form ⇒ direct formula $\boxed{OLS}$

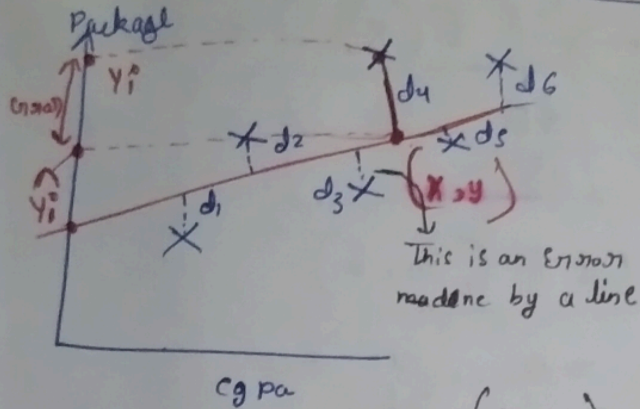
Non closed form $\boxed{\text{Gradient Descent}}$

$$\boxed{b = \bar{y} - m\bar{x}}$$

$y \Rightarrow$ Pack, $x \rightarrow$ cgpa

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

$\bar{x}, \bar{y} \Rightarrow$ mean



$$E = d_1 + d_2 + d_3 + \dots d_n$$

$$E = d_1^2 + d_2^2 + d_3^2 + \dots d_n^2$$

$$\boxed{E = \sum_{i=1}^{n} d_i^2}$$ Error function

find (m, b) values, which should reduce this Equation error

Resons for taking (Squares) on d

1. Mod $\Rightarrow |d_1| + |d_2| + |d_3| + \dots$

mod } R1 ⇒ Since line is Passing b/w points, Some points may be +ve or −ve, So to make all Points +ve

Square } R2 ⇒ if mod is used, at some point or at origin its not differentiable. but by using square, we can easily differentiate

$$d_i = (y_i - \hat{y}_i)$$

⇒ y_i ⇒ actual y value } Pack

$\hat{y}_i$ ⇒ Predicted y value } Pred. Pack

$$\boxed{E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \Rightarrow (m, b)$$

} so we need find the value of m, b such that this Eq gives min error

$$\boxed{\hat{y}_i = mx_i + b}$$

$$\boxed{E(m, b) = \sum_{i=1}^{n} (y_i - mx_i - b)^2}$$

→ $y = f(x)$

when x changes y also changes here to depending on 1 x variable

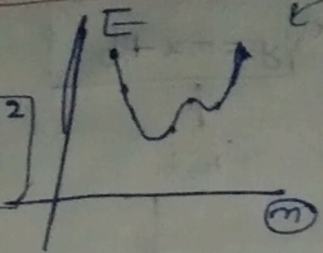
in this case ⇒ $E(m, b)$

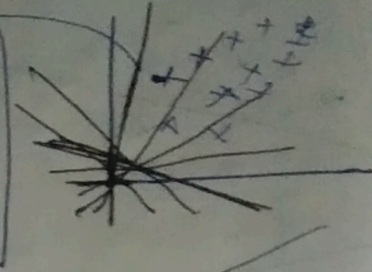
error depending on 2 variable m & b

$$\boxed{E(m,b) = \sum_{i=1}^{n} (y_i - mx_i - b)^2} \Rightarrow \text{minimum}$$

when $\boxed{b=0}$

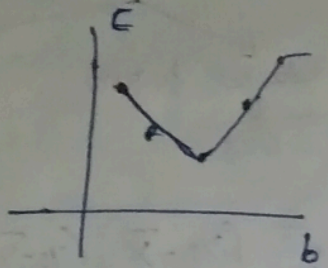
$$\boxed{E(m) = \sum_{i=1}^{n} (y_i - mx_i)^2}$$

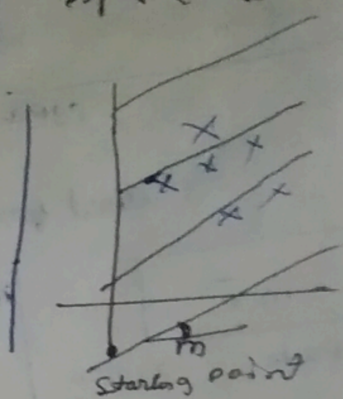




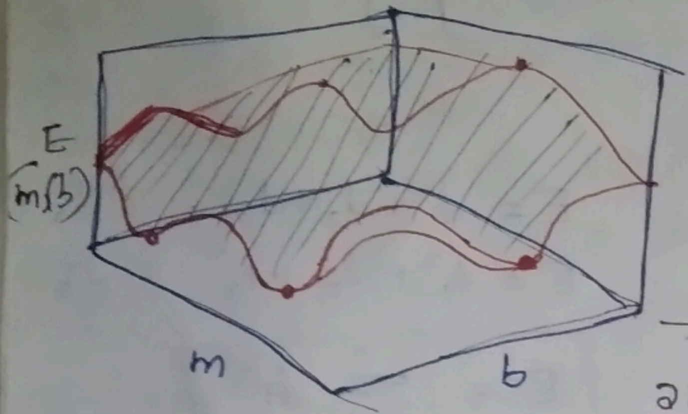
when $\boxed{m=1}$

$$\boxed{E(b) = \sum_{i=1}^{n} (y_i - x_i - b)^2}$$





3D



$\rightarrow$ if $E(x)$

$$\frac{dE}{dx} = 0$$

$\rightarrow$ here $E(m,b)$

$$\frac{\partial E}{\partial m} = 0, \quad \frac{\partial E}{\partial b} = 0$$

$(m,b) \Rightarrow$ we will get equations

$$\frac{\partial E}{\partial b} = \frac{\partial}{\partial b} \sum_{i=1}^{n} (y_i - mx_i - b)^2 = 0$$

$$= \sum \frac{\partial}{\partial b} (y_i - mx_i - b)^2 = 0$$

$$\Rightarrow \sum -2(y_i - mx_i - b) = 0$$

$$\Rightarrow \sum (y_i - mx_i - b) = 0$$

$$\Rightarrow \left[\sum y_i - \sum mx_i - \sum b = 0\right] \div N$$

$$\Rightarrow \sum \frac{y_i}{n} - \sum \frac{mx_i}{n} - \sum \frac{b}{n} = 0$$

$$\Rightarrow \bar{y} - m\bar{x} - \frac{\cancel{n}b}{\cancel{n}} = 0 \qquad \frac{b+b+b\ldots+b = nb}{n}$$

$$\Rightarrow \bar{y} - m\bar{x} = b$$

$$\boxed{b = \bar{y} - m\bar{x}} \rightarrow (1)$$

$$\frac{\partial E}{\partial m} = \sum_{i=1}^{n} \frac{\partial}{\partial m}\left(y_i - mx_i - \bar{y} + m\bar{x}\right)^2 = 0 \quad \text{diff}$$

$$\Rightarrow \sum_{i=n}^{n} 2\left(y_i - mx_i - \bar{y} + m\bar{x}\right)\left(-x_i + \bar{x}\right) = 0$$

$$\Rightarrow \sum_{i=1}^{n} -2\left(y_i - mx_i - \bar{y} - m\bar{x}\right)\left(x_i - \bar{x}\right) = 0 \quad \div 2$$

$$\Rightarrow \sum_{i=1}^{n} \left(y_i - mx_i - \bar{y} - m\bar{x}\right)\left(x_i - \bar{x}\right) = 0$$

$$\Rightarrow \Sigma\left[\left(y_i - \bar{y}\right) - m\left(x_i - \bar{x}\right)\right]\left(x_i - \bar{x}\right) = 0$$

$$\Rightarrow \Sigma\left[\left(y_i - \bar{y}\right)\left(x_i - \bar{x}\right) - m\left(x_i - \bar{x}\right)^2\right] = 0$$

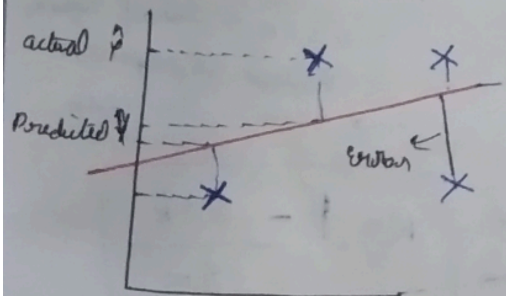
$$\Rightarrow \Sigma\left[\left(y_i - \bar{y}\right)\left(x_i - \bar{x}\right)\right] = m\left(x_i - \bar{x}\right)^2$$

$$\Rightarrow \boxed{m = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}}$$

Regression Metrics

i) MAE ii) MSE iii) RMSE iv) R2 Score v) Adjusted R2 Score

i) MAE [Mean absolute Error]:-



$$\frac{|y_1 - \hat{y}_1| + |y_2 - \hat{y}_2| + \dots |y_n - \hat{y}_n|}{n}$$

(mean

$$\boxed{mae = \sum_{i=1}^{n} |y_i - \hat{y}_i|}$$

advantages

i) Same unit of mae & $\hat{y}$ has same unit Ex:-

Cgpa	lpa
8	4.6

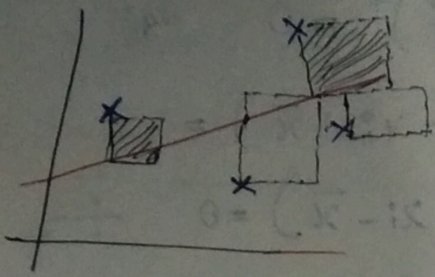
⇒ here mae gives error in lpa, so we can identify the loss deviations easily.

ii) Robust outliers

disadvantage :- mod function

⇒ because of mod function, differentiab at point zero is not possible, so not able to perform the optimization properly.

ii) MSE [Mean Squared Error] :-



$$\boxed{mse = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}}$$

advantage
→ can be used as a loss function because it is differentiable at all the point

disadvantages :-
y - (lpa) then mse ⇒ $(lpa)^2$
Error gives in a Squared ~~of~~.

→ Not robust to outliers :- when the outliers are very big or very small. The outliers may be impact on the o/p badly strongly

iii) RMSE [Root mean Squared Error]

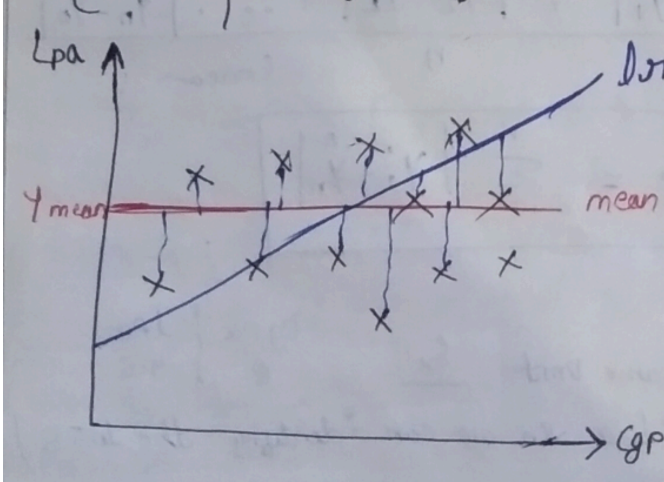
$$\Rightarrow \sqrt{mse} \Rightarrow \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}}$$

advantage
rmse - lpa } Same unit
y - lpa }
→ Easy interpretation

disadvange
→ not robust to outliers

iv) R2 Score :- [Coefficient of determiniation]

[cgpa | package (lpa)]



} in best case

} in worst case if i/p cgpa is not given.

→ SSR = Sum of Square Error in regression
→ SSm - in mean

$$\boxed{R_2 = 1 - \frac{SSR}{SSm}}$$

$$R_2 = 1 - \frac{\left[\sum_{i=1}^{n} (y_i - \hat{y}_i)^2\right]_{lr}}{\left[\sum_{i=1}^{n} (y_i - \hat{y}_i)^2\right]_{m}}$$

1

i) when this is $= 1$, $R_2 = 0$
{ comparing mean vs best fit line, when lr line meets the mean line, $R_2 = 0$, in this case, the i/p cgpa is not used properly then this is not a good model

ii) when R2 = 1,

$R_2 = 1 - \boxed{\frac{SS_R}{SSm}} \Rightarrow 1$ — 0 = SSR

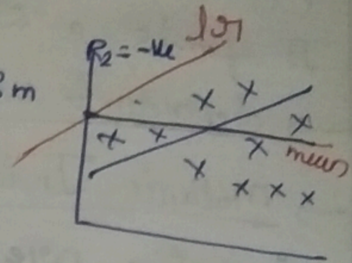
→ when regression line (l_R) is not making any mistake, means the line is passing through all the points (Perfect line)

iii) when R_2 = −Ve

$R_2 = 1 - \boxed{\frac{SSR}{SSm}} \Rightarrow > 1$

→ when lr is worst than mean, then there is a chance of getting −ve R2

when SSR > SSm



→ when highlly non linear data is there

$\boxed{R_2 = 0.80}$ ⇒ dataset: cgpa | lpa ⇒ i/p col: $\boxed{cgpa}$ Explains 80% of Variance in lpa

$\boxed{R_2 = 0.80}$ ⇒ cgpa | iq | lpa ⇒ i/p col: $\boxed{cgpa}$ $\boxed{iq}$ Explains 80% of variance in lpa

→ R_2 Score Explains the Variance of o/p based on the no. of i/P cols

→ Ex
cgpa | lpa ⇒ 0.80
cgpa | iq | lpa ⇒ 0.86 → it may give
when ⇒ cgpa | iq | temp | lpa ⇒ 0.86 / 0.87 / 0.9 } here temp i/p col is an irrelevant column

⇓

in this case R_2 Score ↓ should get down, but here R_2 ↑

to solve this

v) Adjusted R2 Score:

[1 − ↓] = ↑

R_2 → normal RS score

$$R^2_{adj} = 1 - \left[\frac{(1-R_2)(n-1)}{(n-1-k)}\right]$$

large n → no. of rows

little k ⇒ independents colums / i/p

k=1, k=2, k=3

irrelevant col (temp)

$\boxed{\text{Adj } R_2}$ ↓ [k will increase as i/p col increase]

iq (relevant column

$\boxed{R^2_{adj} \uparrow}$

→ when k increases,
→ the denomi (n−1−k) will ↓
→ R_2 may ↑ or remains constant