19/02/21 dec-19

→ M.S.E of reg. model

→ multiple linear regression

→ coefficient of regression

→ non linear reg. models

multiple linear regression : → R squared or coefficient of determination

$$\boxed{y = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n}$$

Linear reg.

x y

$y = b + wx$

multiple linear reg.

$\langle x_1 \quad x_2 \quad x_3 \ldots x_n \rangle$ — y

x_1 → sq. ft. area, x_2 → no. of rooms, x_3 → loc., x_n → flooring

y → price of house.

for 2-variable case :

$$\boxed{y = a + b_1 x_1 + b_2 x_2}$$

$$b_1 = \frac{(\sum x_2^2)(\sum x_1 y) - (\sum x_1 x_2)(\sum x_2 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

$$b_2 = \frac{(\sum x_1^2)(\sum x_2 y) - (\sum x_1 x_2)(\sum x_1 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

$$\boxed{a = \bar{y} - b_1 \bar{x}_1 - b_2 \bar{x}_2}$$

$$S_e = \sqrt{\frac{\sum e_i^2}{n-k-1}}$$

n → no. of observation

k → no. of explanatory variables

Numerical :

X_1	X_2	Y	Y^2
Age	wt	<B pressure>	
—	—	—	—
—	—	—	—
—	—	—	—
—	—		—
	—		$\Sigma =$

$\Sigma y^2 = 29.75$ $\qquad \Sigma x_2^2 = 521.75$

$\Sigma x_1 y = 139.5$ $\qquad \Sigma x_1^2 = 1091.8$

$\Sigma x_2 y = 90.25$ $\qquad \Sigma x_1 x_2 = 515.5$

$\bar{y} = 3.25$

$b_1 = 0.0864$ $\qquad \bar{x}_1 = 51.9$

$b_2 = 0.087$ $\qquad \bar{x}_2 = 32.75$

$$a = \bar{y} - b_1 \bar{x}_1 - b_2 \bar{x}_2$$

$$= 3.25 - 0.086\,(51.9) - 0.087\,(32.75)$$

$$= -4.10$$

$\therefore$ Reg Eqn : $\boxed{-4.10 + 0.086\,x_1 + 0.087\,x_2}$

Ans

calculation of coefficient of regression

→ Coefficient of reg. of Y on X :

$$\boxed{b_{yx} = \frac{n\,\Sigma xy - (\Sigma x)(\Sigma y)}{n\,\Sigma x^2 - (\Sigma x)^2}}$$

reg. eqn. of Y on X

$$\boxed{y - \bar{y} = b_{yx}\,(x - \bar{x})}$$

→ coefficient of reg. of X on Y :

$$\boxed{b_{xy} = \frac{n\,\Sigma xy - (\Sigma x)(\Sigma y)}{n\,\Sigma y^2 - (\Sigma y)^2}}$$

$$\boxed{\text{reg. eqn. of } x \text{ on } y \quad x - \bar{x} = b_{xy}\,(y - \bar{y})}$$

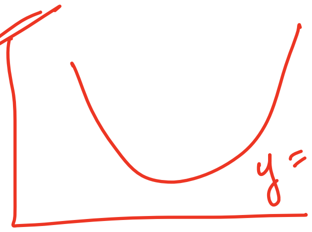
x	y	xy	x^2	y^2
6	9	54	36	81
2	11	22	4	121
10	5	50	100	25
4	8	32	16	64
8	7	56	64	49
$\Sigma = 30$	40	214	220	340

$\boxed{b_{yx} = -0.65}$ $\boxed{b_{xy} = -1.3}$ Ans

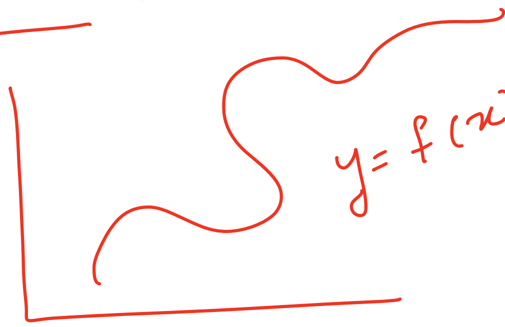
$$\boxed{\begin{aligned} y &= -0.65x + 11.9 \\ x &= -1.3y + 16.4 \end{aligned}}$$

Ans

Non-linear regression model:



$y = ax^2 + bx + c$



$y = f(x)$

→ If there exist a non linear relationship between x and y, straight line linear regression won't give accurate result.

$$\boxed{Y = w_0 + w_1 x + w_2 x^2 + w_3 x^3 \ldots\ldots}$$

f_1 x_1	f_2 x_2	y
		C1
		C2

concept of linearly separable data

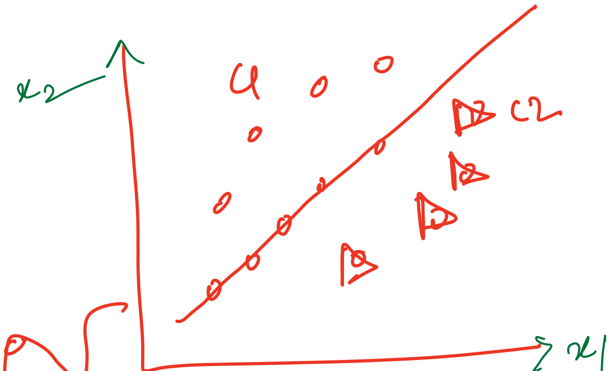


fig1

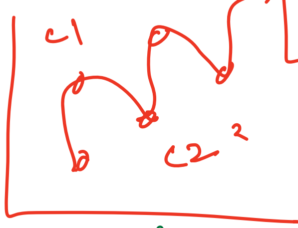
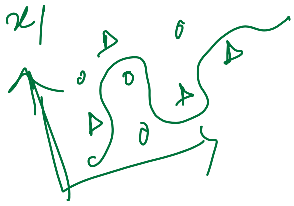
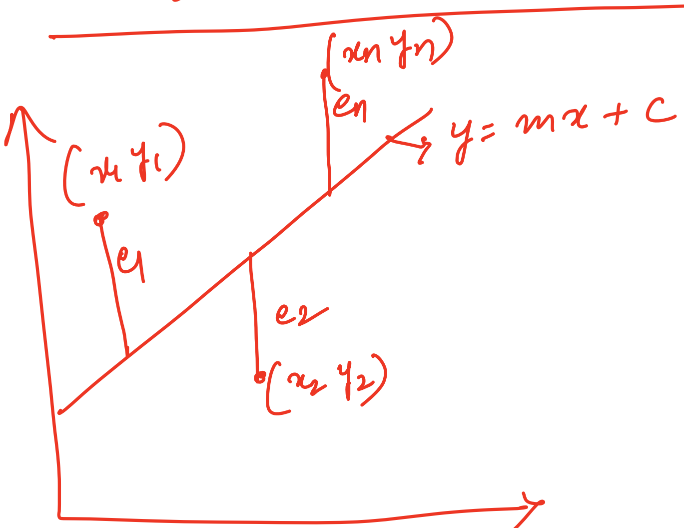


fig2



→ R^2 squared or coefficient of determination :



$$\boxed{SE_{line} = (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \cdots (y_n - \hat{y}_n)^2}$$

squared error from line

what % of total variation in y is described by variation in x ?

$$\boxed{\text{Total variation in } Y = (y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2 = SE_{\bar{y}}}$$

[sq. error from mean of y]

How much % of total variation is not described by the regression line ?

$$= \frac{SE_{line}}{SE_{\bar{y}}}$$

* chk it

$$\therefore \boxed{R^2 = 1 - \frac{SE_{line}}{SE_{\bar{y}}}}$$

→ what % of total variation is described by variation in x

Coefficient of determination

→ If SE_{line} is small, then line is a good fit, and R^2 is close to 01.

→ If SE_{line} is large, then R^2 is close to 0, which indicates line is not a good fit.

→ R^2 is the percent-age of variation of the dependent variable explained by the regression

$$\boxed{R^2 = 1 - \frac{\sum e_i^2}{\sum (Y_i - \bar{Y})^2}}$$

→ R^2 value lies between 0 and 1.

→ The above eqn. indicates that when residuals are small, R^2 will be close to 1, but when they are large, R^2 will be close to 0.

→ R^2 measures the goodness of a linear fit. The better the R^2 fit is, the closer R^2 is to 1.

→ In simple linear regression, R^2 is the square of correlation between the dependent variable and the explanatory variable.