Linear Regression

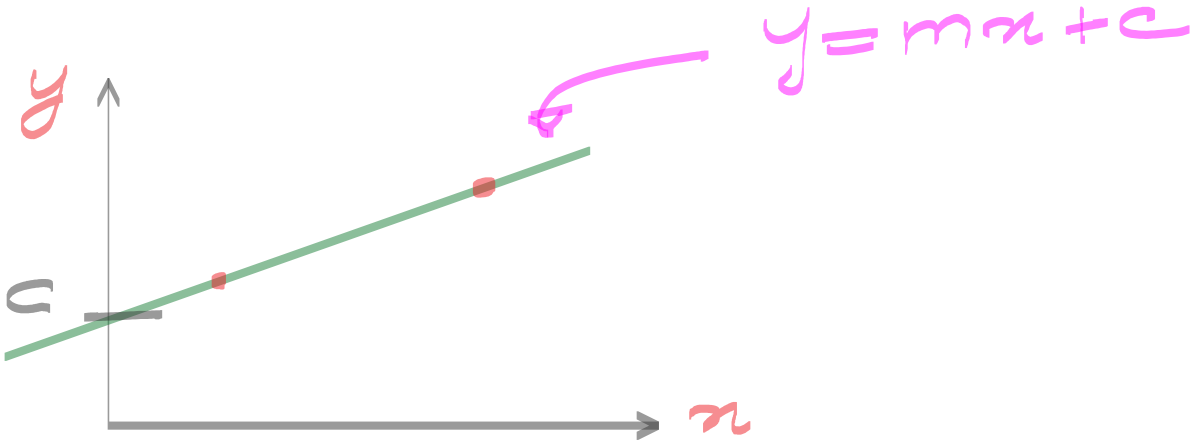
07 December 2022 20:35

* Equation of Line :

$$y = mx + c$$

$$y = a + bx$$

$$y = \theta_1 + \theta_2$$



m ⟶ slope of line

c ⟶ intercept

$$y = 5x + 3$$

If $x = 2$, $y = 13$

$x = 5$, $y = 28$

* Types of LR

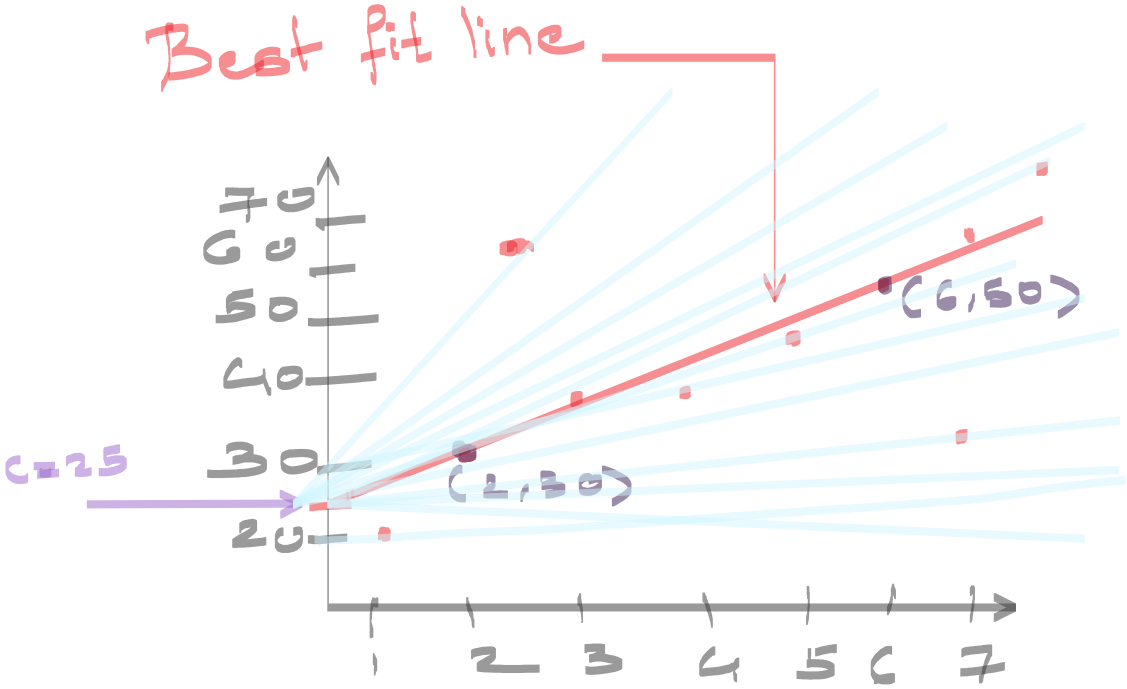
1. Simple LR

1. Simple LR

$$y = mx + c$$

2. Multiple LR

$$y = m_1x_1 + m_2x_2 + \dots + m_nx_n + c$$



(x) Exp	(y) Sal
2	30
3	35
4	35
5	45
6	50
7	60
8	70

$$m = \frac{y_2 - y_1}{x_2 - x_1} = \frac{50 - 30}{6 - 2} = \frac{20}{4} = 5$$

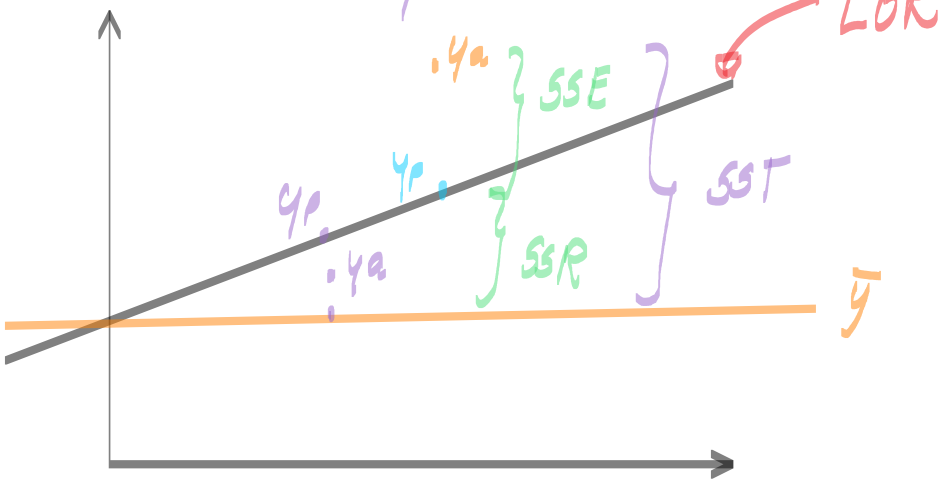
$$y = 5x + 25$$ ← Linear Model

* Blue lines → infinite number of possibilities

Linear Regression

08 December 2022 07:31

SST = SSE + SSR.



MSE → Scale Variant.

Ht.	Temp.
5.	150
5.2	160
5.4	170
6	183
6.1	185

MSE
25-36

MSE
2890

R-squared → R^2 → Scale Invariant

0 to 1 , -ve

1. $R^2 = 1$

$$R^2 = 1 - \frac{SSE}{SST}$$

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

2. $R^2 = 0$ R^2

2. $R^2 = 0$ $\quad R^2 = 1 - 0 = 1$, $\quad SSE = SST$

3. $R^2 = +ve$, $\quad SSE < SST$

$R^2 = 1 - 0.1 = 0.9$

4. $R^2 = -ve$, $\quad SSE > SST$

$R^2 = 1 - 1.4 = -0.4$

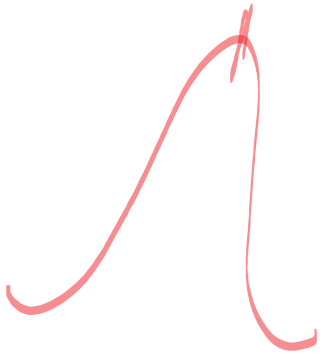
Non-linear.

$R^2 = 1 - \frac{SSE}{SST}$ → $(y_a - y_p)^2$

SST → $(y_a - \bar{y})^2$

Exp	Sal
2	30
3	40
4	50
5	60
6	70

y_a, $\bar{y} = 50$



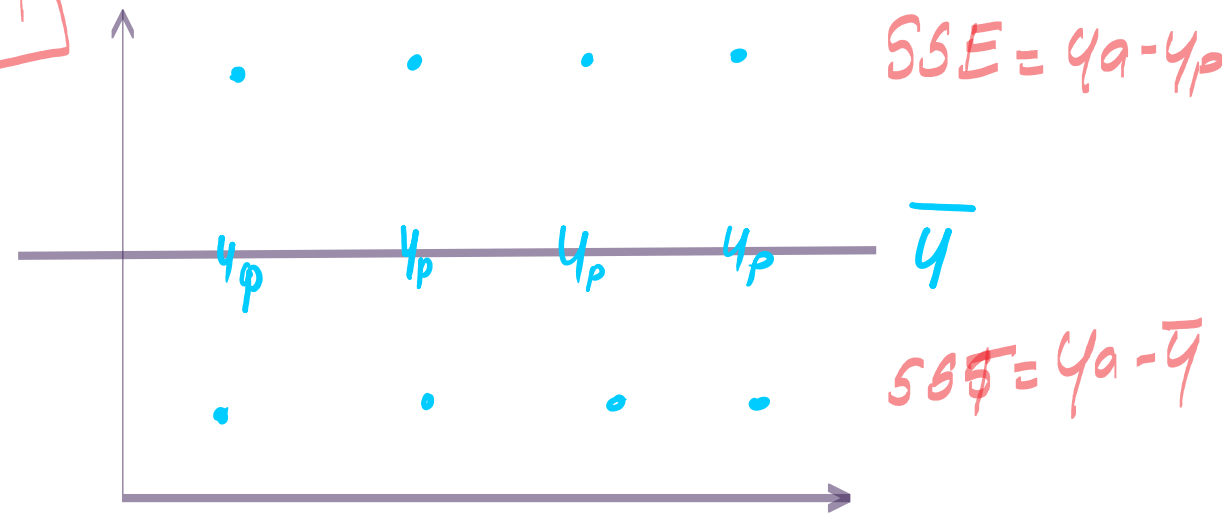
$$R^2 = 1 - \frac{SSE}{SST}$$

$$R^2 = 1 - \frac{\text{Unexplained Variation}}{\text{Total Variation}}$$

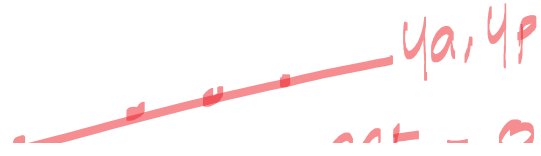
$$R^2 = \frac{SST - SSE}{SST} \rightarrow \frac{\text{Explained Var}}{\text{Total Var}}$$

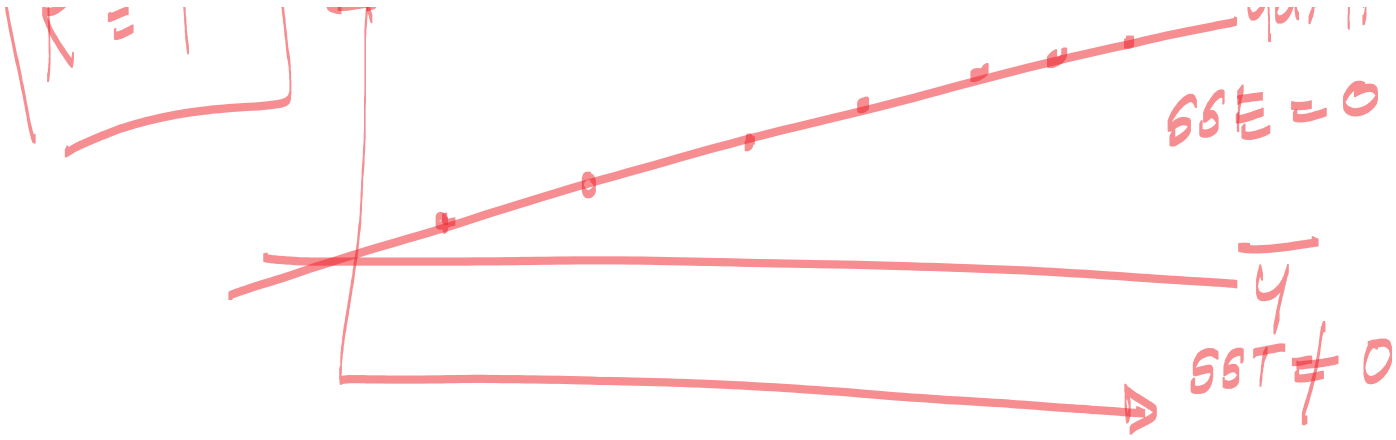
$SSE = 0$

$SSE = SST$



$R^2 = 1$





R-squared

Coeff. of Corr.

$R^2 = R \times R \longrightarrow$ SLR

$R^2 \neq R \times R \longrightarrow$ MLR

$R^2 \longrightarrow$ Goodness of Best fit line (r2_score)

x_1	x_2	x_3	x_4	x_5	y
0.9	0.4	0.8	0.7		-0.5

3 features $\longrightarrow R^2 = 0.85$ } 0.852
(x_1, x_3, x_4)

4 features $\longrightarrow R^2 = 0.86$ }

$(x_1\ x_2\ x_3\ x_4)$

Adjusted R-squared $\rightarrow \bar{R}^2$

$$\bar{R}^2 = 1 - \frac{(1-R^2)(n-1)}{n-p-1}$$

n = No of samples (rows)

p = No of Predictors

$R^2 \geq \bar{R}^2$

		R^2	$\bar{R}^2$
1	$R = 0.3$	0.85	0.85
		0.86	0.84
2	$R = 0.9$	0.87	0.86

$$R^2 = 1 - \frac{SSE}{SST}$$

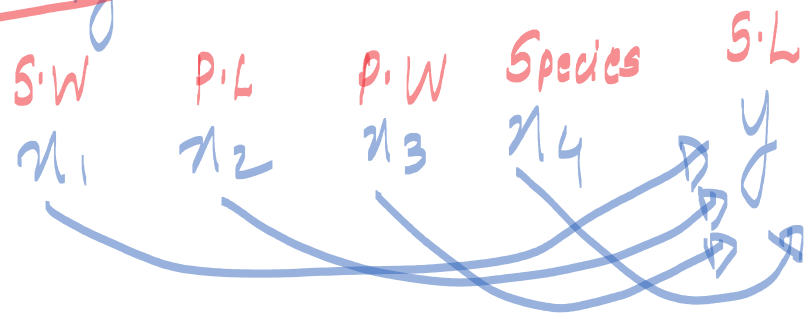
$$\bar{R}^2 = 1 - \frac{(1-R^2)(n-1)}{n-p-1} \quad \begin{matrix} \leftarrow 999 \\ \leftarrow 994 \end{matrix}$$

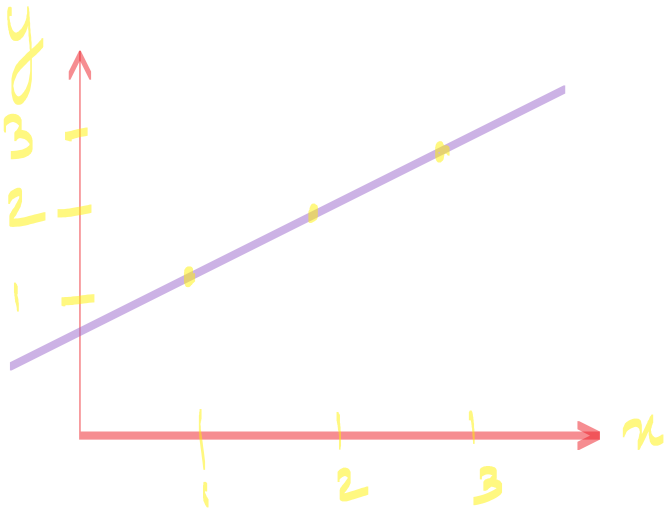
$n = 1000$, $p = 5$

$\boxed{n > p}$

~~5 rows~~
6 columns X

* ~~Linearity~~.





R-value = 1

150 rows

df → x_1 x_2 x_3 x_4 y

Tr / Test ← x → Df → x_1 x_2 x_3 x_4

Tr / Test ← y → Series → y

Train → x_train (120 rows), y-train (120 rows)

Test → x_test (30 rows), y-test (30 rows)

x → x_1 x_2

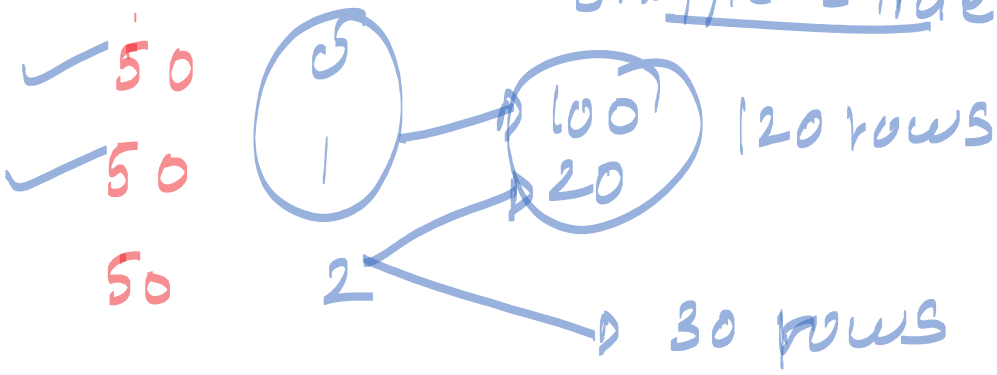
y → y

Tr – 1
2
3
Tr – 4
Tr – 5

Tr - 5

- Train → 3 row → x_train, 2, 4, 5 / y_train, 2, 4, 5
- Test → 2 row → x_test 1, 3 / y_test 1, 3

shuffle = True



100 row → 80 ← 0 → Train

20 ← 1 → Test

* What is Data Science ?

Train — evaluate

4 lines → 3 correct 75%
1 incorrect 25%

* What is KNN ?

Test — Evaluate unseen

4 line → 2 correct 50%

* what is RMN

Test — 4 line → 2 covred 50%

Evaluate

unseen

	Area	Bed	Bath	Location	Price
Train	1000	3	2	Baner	2cr
	700	2	1	Wakad	80
	800	1	1	Bale.	75
Test Unseen	650	2	2	Baner	90L

$y = mx + c$