Linear Regression

Intercept, Slope

$$y = b + ax$$

$$a = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$ Slope

$$b = \frac{1}{n}\left(\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i\right)$$ Intercept

x	y
-2	-1
1	1
3	2

x	y	xy	x^2
$\Sigma x_i = 2$	$\Sigma y_i = 2$	$\Sigma xy = 9$	$\Sigma x^2 = 14$

$$a = \frac{23}{38} \qquad b = \frac{5}{19}$$

$$y = \frac{5}{19} + \frac{23}{38}x$$

$x = 4$

$$y = \frac{5}{19} + \frac{23}{28} \times 4 = \frac{5+19}{7}$$

$$= 0.26 + 0.82 \times 4$$

$$= 0.26 + 3.29$$

$$= 3.54$$

---

use the linear Reg to find the sales of company

2018 to 2023,

x	y	$t = x - 2018$	yt	t^2
2018	12	0		
2019	19	1		
2020	29	2		
2021	37	3		
2022	45	4		

$$y = 11.4 + 8.4t$$

$$y = 11.4 + 8.4(x - 2018)$$

Multiple Linear Regression:-

if multiple linear Regression Need to Calculate intercept only no. need to calculate Slope

multiple linear Regression formula:-

$\hat{y} = b_0 + b_1 * x_1 + b_2 * x_2$

→ more than one intercept

$b_0 = \bar{y} - b_1 \bar{x}_1 - b_2 \bar{x}_2$

- How to Calculate b?

$$b_1 = \frac{[(\Sigma x_2^2)(\Sigma x_1 y) - (\Sigma x_1 x_2)(\Sigma x_2 y)}{(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2}$$ if there are three variable

$$b_2 = \frac{[(\Sigma x_1^2)(\Sigma x_2 y) - (\Sigma x_1 x_2)(\Sigma x_1 y)}{[(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2]}$$

y	x_1	x_2
140	60	22
155	62	25
159	67	24
179	70	20
192	71	15
200	72	14
212	75	14
215	78	11
M = 181.5	69.38	18.13
S. 1452	555	145

Calculate $x_1^2 = 38767$

$x_2^2 = 2823$

$x_1 y = 101895$

$x_2 y = 25364$

$x_1 y_1 = 9859$

regression sum

$$\Sigma x_1^2 = \frac{\Sigma x_1^2 - (\Sigma x_1)^2}{n} = \frac{38767 - (555)^2}{8} = 263.875$$

$$\Sigma x^2 = \frac{\Sigma x_2^2 - (\Sigma x_2)^2}{n} = 194.88$$

$\Sigma x_2 y = -953.5$

$$\Sigma x_1 y = \frac{\Sigma x_1 y - \Sigma x_1 \Sigma y}{n} = 1162.5$$

$$\Sigma x_1 x_2 = \frac{\Sigma x_1 x_2 - \Sigma x_1 \Sigma x_2}{n} = -200.375$$

Calculate b_0, b_1, b_2

$b_1 = 3.148$

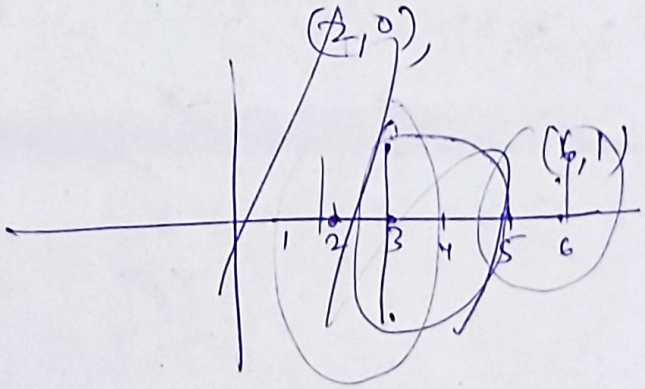
$b_2 = -1.656$

$b_0 = 181.5 - 3.148(69.38) - (-1.656)(18.125)$

$= -6.867$

$b_1 = 3.148 \quad b_2 = -1.656.$

$\hat{y} = -6.867 + 3.148x_1 - 1.656x_2$



Q1 Linear Regression – It stablish a relationship b/w dependent var of one or more Indepent variables using best fit straight line

Least square reg. equation

$y = b + ax$ → Independent

dependent

∵ a = slope of line ∵ b = Intercept.

$$a \Rightarrow \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \quad \text{(slope)}$$

$$b \Rightarrow \frac{1}{n}\left(\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i\right) \quad \text{(Intercept)}$$

Ques1 find the Linear Reg best fit line for the basis of given data & draw best fit Reg line for the given Reg eqn.

x	y
-2	-1
1	1
3	2

∵ n = 3

	x	y	xy	x^2
	-2	-1	2	4
	1	1	1	1
	3	2	6	9
Σ	Σx_i 2	Σy_i 2	Σxy 9	Σx^2 14

$$a = \frac{3(9) - (2)(2)}{3((14) - (2)^2)} \rightarrow \frac{23}{38} \quad \frac{15}{30} \Rightarrow \frac{1}{2}$$

$a \rightarrow \frac{23}{38}$

$$b = \frac{1}{3}\left(2 - \frac{23}{38}(2)\right) \rightarrow \frac{1}{3}\left(\frac{76 - 56}{38}\right) \rightarrow \frac{20}{144}$$

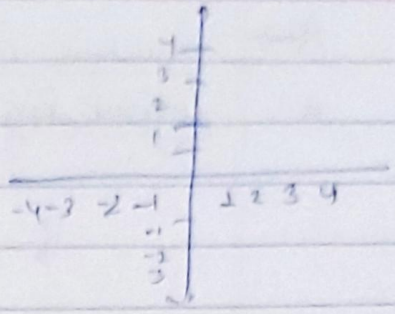
$4y^2 - 4$
36
$+2$
46

$$y = \frac{5}{19} + \frac{23}{38}x$$ → Best fit line Reg eqn.

y val of y at $x=4$ prediction

$$y = \frac{5}{19} + \frac{23}{38}(4) \Rightarrow 51.19$$

$$y = \frac{5}{19} + \frac{92}{38} \rightarrow$$



1) LSRL (LR1)
2) Draw?
3) prediction for Next.

Q Use the Lineau Reg to find the sales of the Company in 2012 & draw the trained (best fit) line for the same

scaling original only data ites $y = b + at \rightarrow y = b + a(x - 2005)$

x (in years)	y (sales) (mill dollers)	$t = x - 2005$	yt	xi^2
original 2005	12	0	0	0
2006	19	1	19	1
2007	29	2	58	4
2008	37	3	111	9
2009	45	4	180	16
	Σy 142	$\Sigma t = 10$	$\Sigma yt = 368$	$\Sigma x^2 = 30$

8.4 $a = \frac{5(368) - 10 \times 142}{5(30) - (10)^2} = \frac{1840 - 1450}{150 - 100} \Rightarrow \frac{390}{50} = 8.4$ (420)

11.4 $b = \frac{1}{5}\left(142 - 8.4(10)\right) \Rightarrow 11.6$ 12.2 13.4

$y = 11.4 + 8.4t$ $y = 11.4 + 8.4(x - 2005)$

for 2012 $t = x - 2005 \rightarrow 2012 - 2005 = 7$

$y = 11.6 + 8.4 * 7$ $y = 11.4 + 8.4(2012 - 2005)$

2012 sales $\rightarrow$ $\boxed{y = 70.4}$

because not multiple indep. value in Linear Reg when x = multiple then

$a = a_1 + a_2 \cdots a_n$

$[y = b + a_1x_1 + a_2x_2 \cdots a_nx_n] \rightarrow y = b + a[x_1 + x_2 + \ldots]$

multiple Regression = multiple Linear Regression (multiple li...

~~Logistic Regression~~

$$a = \frac{a_1 + a_2 \cdots + a_n}{n}$$

$$b = \frac{1}{n}\left(\sum_{i}^{n} y_i - a\sum x_i\right) \quad , \quad b_1 = \frac{1}{n}\sum y_i - a_1 x_1$$

$$\boxed{y = b + ax}$$

x_1	x_2	y	x_1y	x_2y	$y = b + a_1x_1 + a_2x_2$	x_1^2	x_2^2
2	1	4	8	4		4	1
3	2	6	18	12		9	4
1	4	5	5	20		1	16
6	3	8	48	24		36	9
$\Sigma 12$	$\Sigma 10$	$\Sigma 23$	$\Sigma 79$	$\Sigma 60$		50	30

(x_1, y)

$$a_1 = \frac{4(79) - (12 \times 23)}{4(50) - (12)^2} \quad \frac{316 - 276}{56} = \frac{40}{56} \Rightarrow 0.7$$

$$b_1 = \frac{1}{4}(23 - (0.714)12) \Rightarrow \frac{14.432}{4} \Rightarrow 3.6$$

(x_2, y)

$$a_2 = \frac{4(60) - (10)(23)}{4(30) - (10)^2} = \frac{240 - 230}{20} = \frac{10}{20} = \frac{1}{2} = 0.5$$

$$b_2 = \frac{1}{4}(23 - (0.5)(10)) \Rightarrow 4.5$$

$$b = \frac{b_1 + b_2}{2} \Rightarrow \frac{3.6 + 4.5}{2} \Rightarrow 4.05$$

12.679

$$\boxed{y = 4.05 + (0.714)x_1 + 0.5x_2}$$

$y = b + a_1x_1 + a_2x_2$

$y = 4.05 + 3.57 + 3.5 \Rightarrow 11.12$

when $x_1 = 5$

$x_2 = 7$

∵ b intercept same line ||

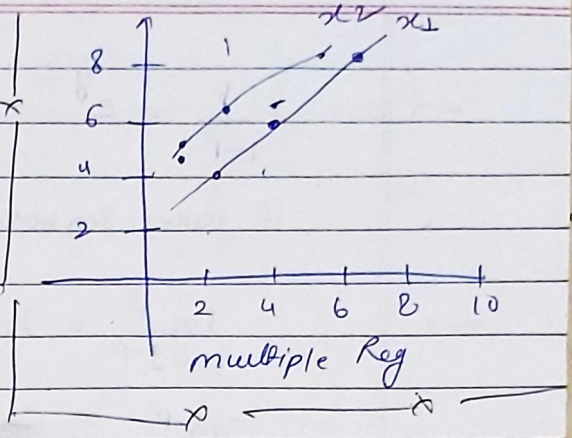
b of intercept diff then line

20-1
20-2
40 - final
20 IA

Color UK Colour classmate

Date
Page 10

★ Logistic Regression: when dependent (Y) vari is in form of yes/No, 0/1 or T/F then we use Logistic Regression



⇒ It is used to find the Prob. of event success & event failure. we should use logistic Regression. when the dependent vari is binary or boolean (0,1), (T/F), (Y,N)
Binary Boolean.

∵ function → logit() → used for evaluating any Logistic Regression.

→ $y = b_0 + b_1 x$ —(1) Linear regression.

P

$P(y) = b_0 + b_1 x \rightarrow P = b_0 + b_1 x$ ∵ $0 \le P \le 1$

0 → $0 \le P$ $e^x \rightarrow +ve$ $e^{-x} \Rightarrow \frac{1}{e^x} \rightarrow +ve$ exponention.

Let $P = e^{b_0 + b_1 x}$ —(2)

P → $P \le 1$ $P = \frac{e^{b_0 + b_1 x}}{1 + e^{b_0 + b_1 x}}$ —(3) ∵ $b_0 + b_1 x = y$

success $P = \frac{e^y}{1 + e^y}$ → (3) $P + q = 1$

$p+q=1$ ∵

$q = 1 - P \rightarrow 1 - \frac{e^y}{e^y + 1} \rightarrow \frac{1}{e^y + 1} = q$ —(4)

eqn (3) divide by (4)

$\frac{P}{q} = \frac{P}{1-P} \Rightarrow \frac{\frac{e^y}{1+e^y}}{\frac{1}{e^y+1}} \Rightarrow e^y$

odd ratio

Classification.

$$\frac{P}{1-P} = e^{y}$$

if take log both sides

$$\log\frac{P}{1-P} = \log(ey)$$

$$\log\frac{P}{1-P} = y \rightarrow \boxed{\log\frac{P}{1-P} = b_0 + b_1 x}$$

(logit function)

Logistic Regression eqn

Ques Consider the Data Set Given below of cricket players which contains the no of wickets taken in the last match played by them & on the basis of the wicket taken in the previous match by each player, the player is either selected (1) or not selected (0) in the team for the next match to be played. Find the prob of selection of a player who has taken 2 wickets in the previous match that he played. Assume the value of b_0 & $b_1 = 1$

Last match (past data) no of wicket taken	Result (Selected 1 / Not selected 0)
0	0
3	1
1	0
2	1
4	1

13/9/19

Q find the prob of the selection of the player who has taken 5 wickets in the previous match that he played assume value of $b_0 = 1$.

$$\boxed{\log\frac{P}{1-P} = x}$$ when $x = 2$ $P = ?$

Likelyhood Prop → Binary → Regression on features → classification.

$$\frac{P}{1-P} = e^{b_0 + b_1 x} \rightarrow \frac{P}{1-P} = e^{1+1x} \rightarrow \frac{P}{1-P} = e^3$$

$$\frac{P}{1-P} = 20.08$$

$$P = 20.08 - 20.08P$$

$$P + 20.08P = 20.08$$

$$21.08P = \frac{20.08}{21.08} \Rightarrow 0.952 \simeq 1 \text{ selected}$$

★ $\frac{P}{1-P} = e^{1+1(5)}$

$P = (1-P)403.428$

$P = \frac{403.428}{404.428}$

$\boxed{P = 0.99} \Rightarrow 1 \simeq 1$ selected

→ after logistic we use classification like 0,1 (0) 3,2,4,5 (1) ✓

For multiple Independent variable

$$\log \frac{P}{1+P} = b_0 + b_1 x_1 + b_2 x_2 ---- + b_n x x_n$$

★ Data is inconsistant, → use Logistic regression

★ If Data is large → use classification

13/9/19

→ multiple Logistic Regression/

sample ques last 2 match. selected (1) & selected (0)

Q find the prob of selection of a player who has taken zero wkt. in first match & 4 wkt in 2nd match. Assume value of b0, b1, b2 - - - bn = 1

$x_1 = 0$

$x_2 = 4$

$\log \frac{P}{1-P} = b_0 + b_1 x_1 + b_2 x_2$

x_1	x_2	f
0	1	0
1	0	0
3	2	1
4	3	1
2	4	1

13/09/19

Logistic Regression

∴ Universal or single dataset given then we will find training set & Testtest init

$$\log \frac{P}{1-P} = b_0 + b_1 x_1 + b_2 x_2$$

$$= 1 + 1 \times 0 + 1 \times 4$$

$$\boxed{\log \frac{P}{1-P} = 5}$$

$$\frac{P}{1-P} = e^5 \rightarrow \frac{148.413}{149.413} \Rightarrow \boxed{1 = P}$$

$$\frac{P}{1-P} = b_0 + b_1 x_1 + b_2 x_2$$

$$1 + 1.0 + 1.1$$

$$\frac{P}{1-P} = e^2$$

$$P = \frac{7.389}{} = 0.98$$

$$P = 7.38 (1-P)$$

$$P = 7.38 - 7.38 \Rightarrow 0.87 \rightarrow \text{not decided.}$$

∴ Logistic Regression → value if like 0.87 not necessary true

★ Ridge & Lasso Regression — 2 Regularisation $\overset{L1}{\underset{L2}{\leftrightarrow}}$

→ Ridge & Lasso Regression are powerful techniques generally used for creating more accurate least square regression line or more accurate value for dependent variable.

$y = b_0 + ax + Af$ — adjustment factor / error factor.

(L1) Lasso Ridge (L2)

→ Ridge & Lasso regression depends on the value of α (alpha) = level of significance. If the value of level of significance is given in the historical or quantitative data (data set) then we use the concept of Ridge & Lasso Regression.