Fahad Purohit

SM assignment

PART-A

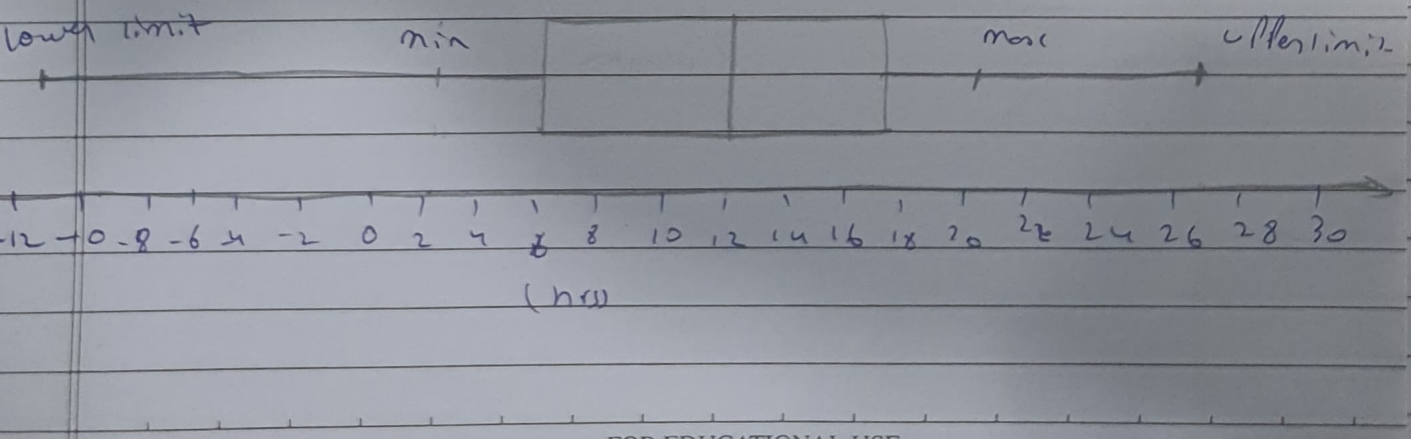
Box - Plot

Q.1] A teacher collects data on the number of hours that her students spend studying per night. The data is as follows →

2, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20.

→

- Minimum = 2
- Maximum = 20
- Median = $\frac{11+12}{2} = 11.5 \approx 12$ $\left(\frac{N/2^{th} + (N/2+1)^{th}}{2}\right)$
- $Q_1 = \frac{N}{4} = \frac{18^{th}}{4} = 6$
- $Q_3 = \frac{3N}{4} = \frac{54^{th}}{4} \approx 17$
- Inter Quartile Range = 17 - 6 = 11
- Lower whisker = $Q_1 - 1.5 \times IQR = 6 - (1.5 \times 11) = -10.5$
  Higher whisker = $Q_3 + 1.5 \times IQR = 11 + (1.5 \times 11) = 27.5$
- Since max and min are written lower and upper whisker, there are no outliers.



Q.2] The following dataset representing the 8 test scores (out of 100): 85, 78, 92, 88, 95, 89, 76, 90, 84, 88.

→

76, 78, 84, 85, 88, 88, 89, 90, 92, 95

→ minimum = 76

maximum = 95

→ median = 88

→ $Q_1 = \frac{N}{4} = 84$

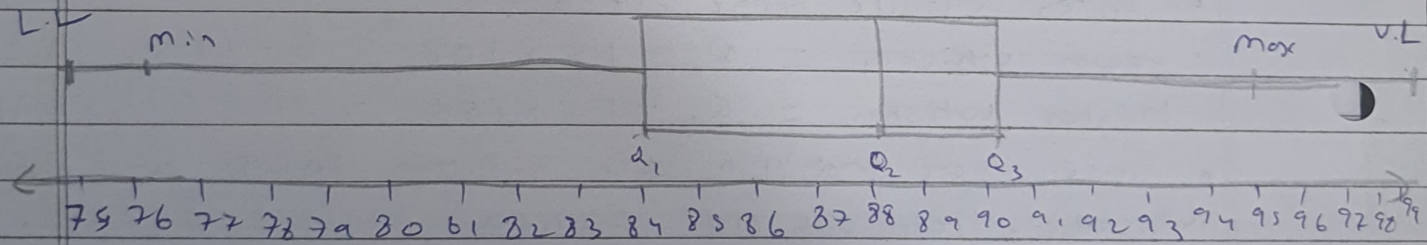
$Q_3 = \frac{3N}{4} = 90$

→ IQR = $Q_3 - Q_1 = 90 - 84 = 6$

→ Lower whisker = $Q_1 - (1.5 \times IQR) = 84 - (1.5 \times 6) = 75$

Upper whisker = $Q_3 + (1.5 \times IQR) = 90 + (1.5 \times 6) = 99$

(Since max and min are whithin the whiskers, there are no outliers)



(marks

Q.3] In a study of how students use mobile Phones, the Phone usage of random sample of 11 student was examined for a Particular week

The total length of calls, 4 mins for 11 students were

17, 23, 35, 36, 51, 53, 54, 55, 60, 72, 110

→ median = 53 → minimum = 17

→ Q_1 = 35 (N/4) → max = 110

→ Q_3 = 60 (3N/4) → IQR = 60 − 35 = 25

→ Lower whisker = $Q_1 - (1.5 \times IQR)$

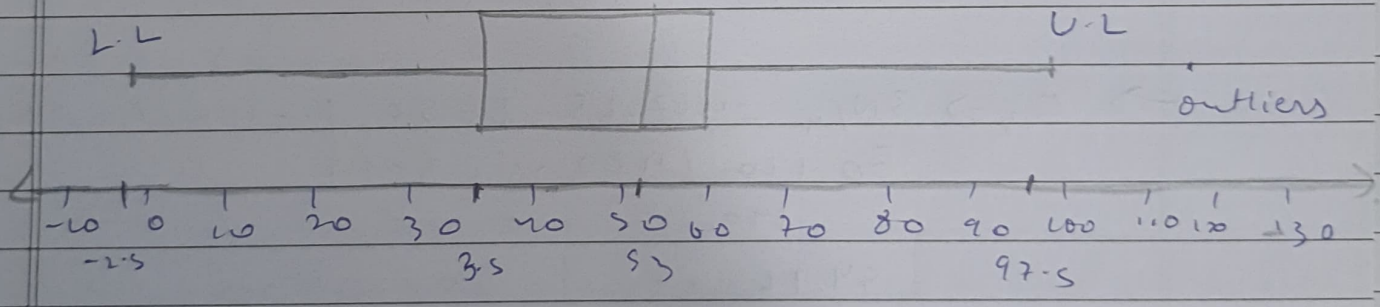
$= 35 - (1.5 \times 25)$

$= -25$

→ upper whisker = $Q_3 + (1.5 \times IQR$

$= 60 + (1.5 \times 25$

$= 97.5$



Qu) There are 543 members in a village, with the following ages:-

Age (yr)	f	cf
10-20	3	3
20-30	61	64
30-40	132	196
40-50	153	349
50-60	140	489
60-70	51	540
70-80	3	543

→ $N = 543$

→ $Q_2 = l + \frac{h}{f}\left(\frac{N}{2} - c\right)$

$= 40 + \frac{10}{153}(271.5 - 196)$

$= 44.93$

→ $Q_3 \Rightarrow 3N/4 = 407.25,\ l = 50,\ f = 140,\ c = 349$

$= 50 + \frac{10}{140}(407.25 - 349)$

$= 54.16$

→ $Q_1 \Rightarrow N/4 = 135.75,\ l = 30,\ f = 132,\ c = 64$

$= 30 + \frac{10}{132}(135.75 - 64)$

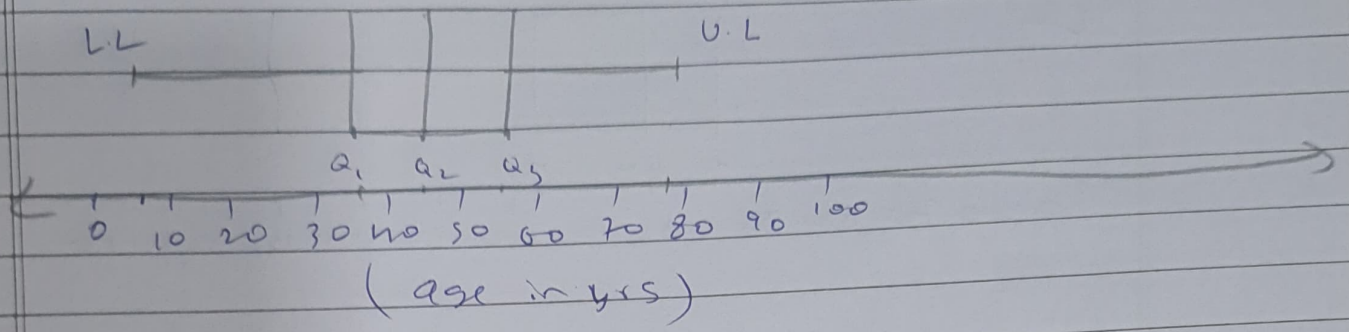
$= 35.43$

→ $IQR = Q_3 - Q_1 = 18.724$

→ $\text{Lower limit} = Q_1 - (1.5 \times IQR) = 7.31$

$\text{Upper limit} = Q_3 + (1.5 \times IQR) = 82.24$

→ No outliers



Q5) A Professor analyses the scores of 2 classes → A & B in maths. Create reflective box Plots and analyze if there are any outliers.

A = 85, 90, 92, 95, 97, 98, 100, 102, 105, 110, 112, 120, 125, 130, 135, 140

B = 75, 78, 80, 82, 85, 88, 92, 95, 96, 98, 100, 105, 108, 112, 115, 150.

→ for class A

• Median = $\left(\frac{\frac{N^{th}}{2} + \frac{N+1}{2}^{th}}{2}\right)$ = $\frac{102+105}{2}$ = 103.5

Q_1 = ~~Q - (1.5 × IQR)~~ $\frac{95+92}{2}$ = 96

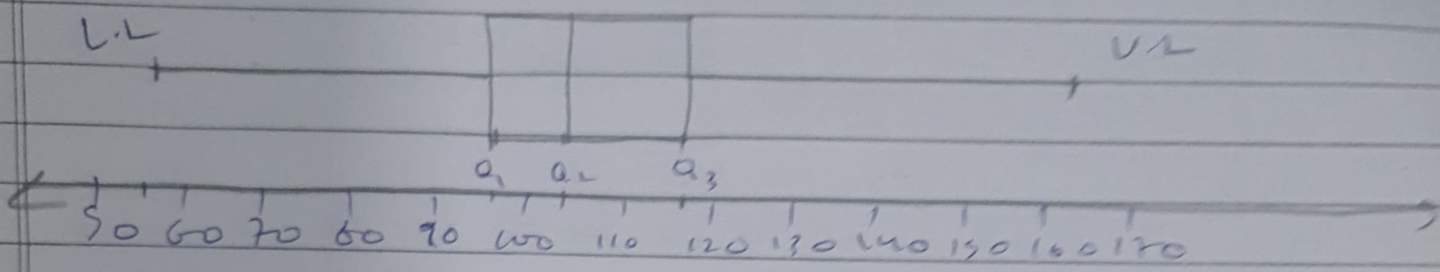
~~56.25~~

Q_2 = $\frac{120+125}{2}$ = 122.5

• IQR = $Q_3 - Q_1$ = 122.5 − 96 = 26.5

• lower limit = Q_1 − (1.5 × IQR)
= 56.25

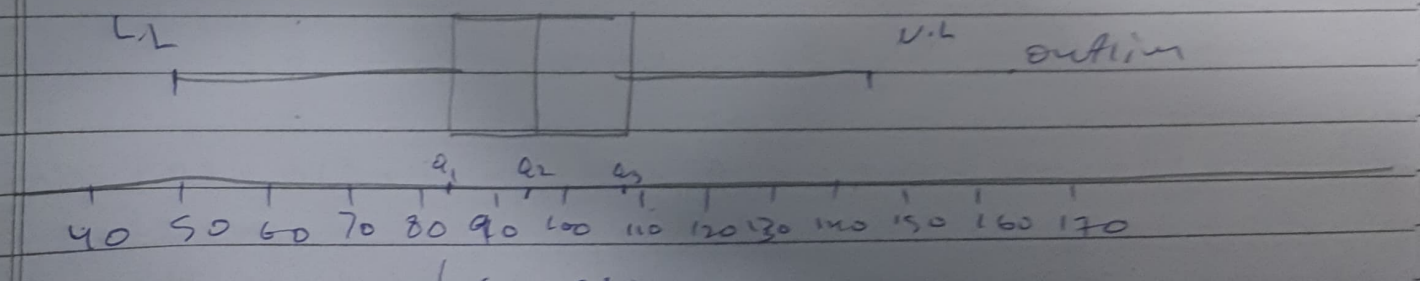
upper limit = Q_3 + (1.5 × 26.5)
= 162.25



There are no outliers in Class A

→ For Class B

- median $= \frac{95+96}{2} = 95.5$
- $Q_1 = \frac{82+85}{2} = 83.5$
- $Q_3 = \frac{105+108}{2} = 106.5$
- IQR $= Q_3 - Q_1 = 106.5 - 83.5 = 23$
- Lower limit $= Q_1 - (1.5 \times IQR) = 83.5 - (1.5 \times 23)$
  $= 47$
- upper limit $= Q_3 + (1.5 \times IQR) = 106.5 + (1.5 \times 23)$
  $= 141$



∴ In Class B, there is 1 outlier = 150

Rahul Purohit

Assignment - B

Q.1] Check whether a linear regression model is applicable for the following data

x : 60 70 80 85 95 $= 390$

y : 70 65 70 95 85 $= 385$

$\hat{y}$ = 65.41 71.80 70.78 81.5 87.94 $= 384.73$

→

residual $(y - \hat{y})$: 4.59 −6.8 −0.78 13.5 −2.94

$(y - \hat{y})^2$: 21.0681 46.7856 0.5584 182.25 8.6436

$\sum (y - \hat{y})^2 = 327.3057$

xy : 4200 4550 5600 8075 8075 $= 30500$

$$SS_{xy} = 30500 - \frac{390 \times 385}{5}$$

$$= -470$$

$$SS_{xx} = 31500 - 30420$$

$$= 730$$

$$b_1 = \frac{SS_{xy}}{SS_{xx}} = -0.6438$$

$$b_0 = \frac{385}{5} - \frac{(-0.6438)(390)}{5}$$

$$= 127.2164$$

$$y = 127.2164 - 0.6438x$$

$$Se = \sqrt{\frac{327.3057}{5-2-1}} = \sqrt{\frac{327.3057}{2}}$$

$$= 12.7927$$

Q2] The equation of a regression line is, the data is as follows; Solve for the residual and graph a residual plot. Do the data seem to violate any assumptions of regression?

→

x	y	x^2	xy
57	42	3249	2679
11	38	121	418
12	32	144	384
19	24	361	456
25	22	625	550
$\Sigma x = 124$	153	$\Sigma x^2 = 4500$	$\Sigma xy = 4487$
	(Σy)		

→ $n = 5$

→ $SS_{xy} = \Sigma xy - \frac{\Sigma x \Sigma y}{n} = 4487 - \frac{(124 \times 163)}{5}$

$= 444.6$

→ $S_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 4500 - \frac{124^2}{5}$

→ $b_1 = \frac{S_{xy}}{S_{xx}} = \frac{444.6}{1424.8}$

$= 0.3120$

→ $b_0 = \frac{\Sigma y}{n} - b \frac{\Sigma x}{n} = \frac{163}{5} - \left(0.3120 \times \frac{124}{5}\right)$

$= 24.8624$

→ $\hat{y} = b_0 + b_1 x$

2 24.8624 + 0.3120x

→

y	$\hat{y}$	$y - \hat{y}$
47	42.6464	4.3536
38	28.2944	9.7056
32	28.6064	3.3936
24	30.7904	−6.7904
22	32.6624	−10.6624
		$\Sigma(y - \hat{y}) = 0$

→ from the graph we infer that the residuals are scattered randomly around the origin
∴ The model is an appropriate linear regression model

→ No, the data doesn't seem to be violating any assumptions of regression, the residuals are randomly distributed as they should be and the sum of residuals comes out to be zero

Rahul Purohit

1

Assignment - 3 Case Studies

0.1]

(i) response variable : US ticket Sales

explanatory variables : x_0 = budget

x_1 = stars

x_2 = runtime

(ii) $\hat{y} = -22.989025 + 1.134423 x_0 + 24.972358 x_1 - 0.402295 x_2$

(iii) x_0 : US ticket sales is positively correlated to budget

x_1 : US ticket sales is positively correlated to stars

x_2 : US ticket sales is negatively correlated to run time (minutes)

(iv) $x_0 = \$32$ millions, $x_1 = 3$ stars, $x_2 = 100$ minutes

$\therefore \hat{y} = -22.989025 + 1.134423(32) + 24.972358(3) - 0.402295(100)$

$\therefore \hat{y} = 47.899256$

(v)

F - test	T - test
H_0 : model is not significant P value > LOS	H_0 : x_0 is not significant P-value > LOS
H_1 : model is significant P value < LOS	H_1 : x_0 is significant P-value < LOS

(vi) f-test : P-value = < 0.0001 < 0.05

∴ model is significant

(vii) ∵ t-test : Pvalue of run times > 0.05

i.e $0.113 > 0.05$

∴ we will drop runtime from the model and rerun the model

(viii) $R^2 = 0.4737$

adj $R^2 = 0.4603$

∴ R^2 is more closer to 1 than adj R^2.

Q2)

(i) Response variable = netflix subscribers

Explanatory variables : x_0 = Monthly Price

x_1 = no of movies

x_2 = Price of rival

(ii) $\hat{y} = 2905.9639 + 111.213003 x_0 + 10.087515 x_1$

$- 605.173348 x_2$

(iii) x_0 : monthly Price is positively correlated to netflix subscribers

x_1 = no of movies is positively correlated to netflix subscribers

x_2 = Price of rival is negatively correlated to netflix subscribers

(iv) $x_0 = 0 \quad x_1 = 4250, \quad x_2 = 7$

∴ $\hat{y} = 2905.9639 + (111.213003)(0) + (10.087515)(4250)$

∴ $-605.173348(7) = 42479.84246$

(v) f test	T-test
H_0 : model is not significant P-value > LOS	H_0 : x_0 is not significant P-value > LOS
H_1 : model is significant P value < LOS	H_1 : model is significant P-value < LOS

(vi) f-test : P-value $= 4.30416 \times 10^{-5} < 0.05$

∴ model is significant

(vii) $R^2 = 0.88921293$

adj $R^2 = 0.8855976$

R^2 is more closer to 1 than adj R^2

(viii)

∴ t-test : P-value : $x_0 = 0.9040293 > 0.05$

$x_2 = 0.1997113 > 0.05$

∴ we will drop monthly Prices (x_0)

Q.10)

(i) response variable : y

Explanatory variable : $x_0 = x_1$

$x_1 = x_2$

$x_2 = x_4$

$x_3 = x_7$

(ii) $y = -55.9 + 0.0105 x_1 - 0.107 x_2 + 0.579 x_3 - 0.870 x_4$

(iv) $x_1 = 9, \; x_2 = 11, \; x_3 = 1, \; x_4 = 80$

$\therefore y = -55.9 + 0.0105(9) - 0.107(11) + 0.579(1) - 0.870(80)$

$= -117.0035$

(iii) x_1 is positively correlated to y

x_2 is negatively correlated to y

x_3 is positively correlated to y

x_4 is negatively coveld to y

(iv) F-test	T-test
H_0 : the model is not significa	$H_0 : x_1$ is not significant
P-value > LOS	P-value > los
H_1 : the model is signifi	$H_1 : x_1$ is significant
P-value < LOS	P-value < LOS

(v) $R^2 = 80.2\%$

adj. $R^2 = 70.7\%$

R^2 is more closer to 100% than adj R^2

(vii) f-test : P-value = 0.000 < 0.05

∴ model is Significant

viii) t-test = P value : x_1 = 0.619 > 0.05

∴ we will drop x_1 from the model

Q.ii) Predictors : 2 (x_1 and x_2)

$\hat{y} = 203.3432 + 1.1151 x_1 - 2.2115 x_2$

Strength :

∴ R^2 = 0.663 ; adj R^2 = 0.636

∴ R^2 is more closer to 1 than adj. R^2

∴ model is strong

∴ P-value : x_1 = 0.448 > 0.05

∴ x_1 is not significant

x_2 = 0.0006 < 0.05

∴ x_2 is significant

∴ co-eff : x_1 = 1.1151

x_2 = -2.2115

∴ x_2 has stronger negative correlation than that of the x_1 with comparitively weaker positive correlation

Q.12]

(i) response variable : y
exploratory variable : x_1 = college
x_1 = age - years
x_3 = income - k
x_4 = gender

(ii) $y = 1.94528 0922 + 8.683514224 x_4 + 0.06011530$
$- x_2 + 0.8164345 32 x_3 + 8.683514224 x_4$

(iii) all x_1, x_2, x_3 and x_4 are positively correlated to y
∴ college, age (years), income (k) and gender are positively correlated to the amount of customer purchase.

(iv) $x_1 = 5$, $x_2 = 10$, $x_3 = 12$

∴ $y = 1.94528 0922 + 8.039875895 (5) + 0.0601153 01 (10)$
$+ 0.8164345 39 (12) + 8.6835 4224 x_4$

$= 51.62362794 + 8.683514224 x_4$

(v) F-test	T-test
H_0: model is not significant	H_0: x_i is not significant
p-value > 0.05	p-value > 0.05
H_1: model is significant	H_1: x_i is significant
p-value < 0.05	p-value < 0.05

(vi) $R^2 = 0.974610686$, adj $R^2 = 0.954310036$
R^2 is more closer to 1 than adj R^2

(vii) p-value : model = 0.00352771 < 0.05
∴ model is significant.

(viii) P-value : college = 0.0226144843 > 0.05
age (yrs) = 0.7238696882 > 0.05
income (k) = 0.00032543 < 0.05
gender = 0.0560344432 > 0.05
∴ we will drop college x_1, age (yrs) x_2 and gender (x_4) from the model and rerun the model

Q.3
(i) response variable = concentration of chlorine
explanatory variables = temperature, extraction, CO, Mg, depth, pH, salinity, DO, bicarbonate, Ca, Cl^-, silicate, NH_4, NO_2^-, NO_3, PO_4^{3-}

(ii)
$y = -0.436 + 0.012x_1 - 0.02x_2 - 0.108x_3 - 0.015x_4 +$
$0.003x_5 + 0.013x_6 + 0x_7 + 0x_8 + 0x_9 + 0x_{10} + 0x_{11}$
$- 0.003x_{12} + 0x_{13} + 0x_{14}$

$\therefore y = -0.436 + 0.012x_1 - 0.02x_2 - 0.108x_3 - 0.015x_4$
$+ 0.003x_5 + 0.013x_6 - 0.003x_{12}$

(iv) $x_1 = 9, x_2 = 11, x_3 = 1, x_4 = 20$

$\therefore y = -0.436 + 0.012(9) - 0.02(11) - 0.108 - 0.015(20)$
$+ 0.003x_5 + 0.003x_{12}$
$= -1.156 + 0.003x_5 + 0.013x_6 - 0.003x_{12}$

(x) (v)

F-test	T-test
H_0: model is not significant & P-value > los	H_0: x_4 is not significant P-value > los
H_1: model is significant P-value < los	H_0: x_4 is significant P-value < los

(vi) $R^2 = 0.82327601$; adj $R^2 = 0.54863046$

R^2 is more closer to 1 than adj R^2

(vii) P-value : model $= 0.05139974 > 0.05$

∴ model is slightly insignificant

(viii) P-value : temperature $= 0.2527 > 0.5$

Extinction coef $= 0.303 > 0.5$

DO $= 0.253 > 0.05$

bicarbonate $= 0.34 > 0.05$

ammonium-N $= 0.455 > 0.5$

Nitrate-N $= 0.378 > 0.05$

Phosphate-P $= 0.714 > 0.05$

∴ we will drop temperature, extinction coef, DO, bicarbonate.

0.14) Predictors : 3 (x variable 1, X variable 2, x variable 3)
observations : 1 (4)
$y = 657.053 + 9.5710 x_1 - 0.4169 x_2 - 3.7252 x_3$

Strength : Model

- $R^2 = 0.710$; adj $R^2 = 0.630$
R^2 is more closer to 1 than adj R^2
∴ model is strong.

Model : P-value = 0.007 < 0.05
∴ model is strong

Significant : Predictors

x_1 : P-value = 0.0037 < 0.05
x_2 : P-value = 0.2122 > 0.05
x_3 : P-value = 0.0349 < 0.05
∴ x_1 and x_3 are significant

0.15) Strength : Model

- Model : P-value = 0.021 < 0.05
∴ Model is strong and significant

∴ $R^2 = 51.5\%$: adj $R^2 = 40.4\%$
R^2 is closer to 1 than adj
∴ Model is moderately sig..

Sleigh : Predictors

- x_1 : P-value = 0.05 < 0.05
- x_2 : P-value = 0.805 > 0.05
- x_3 : P-value = 0.230 > 0.05

∴ x_1 is the only significant Predictor

The model doesn't appear to be fit the data well.

Recommend: Drop x_2 and x_3 from the model and re run the model

Q.16) Predictors = 2(x_1 and x_{1*1})

$y = 203.3232$

Note: Some data and repetitive questions from Q.14 from PDF and Q.11 from this sheet

Q.17)

Note: Some data and repetitive questions from Q.8 of QB pdf and Q15 from next sheet.