UTM

UNIVERSITI TEKNOLOGI MALAYSIA

FACULTY OF COMPUTING

SEMESTER 2 2023/2024

SECI1143 – PROBABILITY & STATISTICAL DATA ANALYSIS

SECTION 3

ASSIGNMENT 4

LECTURER: DR. NOR HAIZAN BT MOHAMED RADZI

GROUP MEMBERS

STUDENT NAME	MATRIC NO
JOANNE CHING YIN XUAN	A23CS0227
EVELYN GOH YUAN QI	A23CS0222
CHUA JIA LIN	A23CS0069

1. a) Pearson's product-moment correlation coefficient
Spearman's rho rank correlation coefficient

b)

Num of items (x)	Production cost (y)	xy	x^2	y^2
26	42	1092	676	1764
44	60	2640	1936	3600
53	69	3657	2809	4761
29	47	1363	841	2209
77	91	7007	5929	8281
80	98	7840	6400	9604
20	39	780	400	1521
40	55	2200	1600	3025
67	85	5695	4489	7225
86	104	8944	7396	10816
17	37	629	289	1369
61	77	4697	3721	5929
$\Sigma x = 600$	$\Sigma y = 804$	$\Sigma xy = 46544$	$\Sigma x^2 = 36486$	$\Sigma y^2 = 60104$

$$r = \frac{\Sigma xy - (\Sigma x \Sigma y)/n}{\sqrt{((\Sigma x^2) - (\Sigma x)^2/n)((\Sigma y^2) - (\Sigma y)^2/n)}}$$

$$= \frac{46544 - [(600)(804)]/12}{\sqrt{(36486 - (600)^2/12)(60104 - (804)^2/12)}}$$

$$= \frac{6344}{\sqrt{40446696}}$$

$$= 0.998$$

c) Since $r = 0.998$, there is a relatively strong positive linear relationship between the number of items produced and the production cost.

2. a)

X	Y	XY	X^2	Y^2
0.27	2	0.54	0.073	4
1.41	3	4.23	1.988	9
2.19	3	6.57	4.796	9
2.83	6	16.98	8.009	36
2.19	4	8.76	4.796	16
1.81	2	3.62	3.276	4
0.85	1	0.85	0.723	1
3.05	5	15.25	9.303	25
$\Sigma x = 14.6$	$\Sigma y = 26$	$\Sigma xy = 56.8$	$\Sigma x^2 = 32.964$	$\Sigma y^2 = 104$

$$r = \frac{56.8 - (14.6)(26)/8}{\sqrt{(32.964 - (14.6)^2/8)(104 - (26)^2/8)}}$$

$$= \frac{9.35}{\sqrt{123.2205}}$$

$$= 0.842$$

b) $H_0 : \rho = 0$ (no linear correlation)

$H_1 : \rho \neq 0$ (linear correlation exists)

$$t = \frac{0.84}{\sqrt{\frac{1 - 0.842^2}{8-2}}}$$

$$= 3.814$$

$\alpha = 0.05$, $df = 8 - 2 = 6$

$t_{0.025,6} = 2.447$

∴ Since $3.814 > 2.447$, we reject H_0. There is sufficient evidence at 0.05 significance level to support that there is linear relationship between the weight of plastic usage and size of household.

c) $t_{0.005,6} = 3.707$

∴ Since $3.814 > 3.707$, we reject H_0. The decision in (ii) is still same.

3)

a) Engagement Score and Sentiment Score

b) Number of likes and Number of Shares

c)

Post ID	Rank of Engagement score	Ranks of sentiment score	di	d_i^2
1	3	3	0	0
2	5	4.5	0.5	0.25
3	1	1.5	-0.5	0.25
4	6	6	0	0
5	2	1.5	0.5	0.25
6	4	4.5	-0.5	0.25

$\Sigma d_i^2 = 1$

$$r_s = 1 - \frac{6(1)}{6(6^2-1)} = 0.971$$

d) a.)

x likes	y comments	xy	x^2	y^2
150	20	3000	22500	400
100	10	1000	10000	100
200	25	5000	40000	625
80	8	640	6400	64
170	22	3740	28900	484
120	15	1800	14400	225
$\Sigma = 820$	$\Sigma = 100$	$\Sigma = 15180$	$\Sigma = 122200$	$\Sigma = 1898$

$$r = \frac{15180 - (820)(100)/6}{\sqrt{[(122200)-(820)^2/6][(1898)-(100)^2/6]}} = \frac{1513.33}{\sqrt{(10133.33)(231.33)}} = 0.988$$

b)

x likes	y share	xy	x^2	y^2
150	30	4500	22500	900
100	20	2000	10000	400
200	40	8000	40000	1600
80	15	1200	6400	225
170	35	5950	28900	1225
120	25	3000	14400	625
$\Sigma = 820$	$\Sigma = 165$	$\Sigma = 24650$	$\Sigma = 122200$	$\Sigma = 4975$

$$r = \frac{24650 - (820)(165)/6}{\sqrt{[(122200) - (820)^2/6][(4975) - (165)^2/6]}} = \frac{2100}{\sqrt{(10133.33)(437.5)}} = 0.997$$

c) $r = 1 - \frac{6(1)}{6(6^2-1)}$

$= 0.971$

∴ strongest positive correlations is between Likes and share There is no negative correlations which $r = 0.997$ that closer to 1 and indicate a stronger positive linear relationship.

Next, likes and Comments, $r = 0.9884$ > Engagement Score and Sentiment Score, $r = 0.971$

There are no negative correlations since all r are positive

e) H_0 : There is no correlation between 'Engagement Score' and 'Post length'

H_1 : There is correlation between 'Engagement Score' and 'Post length'

x Engagement	y Post length	xy	x^2	y^2
85	200	17000	7225	40000
70	100	7000	4900	10000
90	250	22500	8100	62500
60	150	9000	3600	22500
88	220	19360	7744	48400
75	180	13500	5625	32400
$\Sigma=468$	$\Sigma=1100$	$\Sigma=88360$	$\Sigma=37194$	$\Sigma=215800$

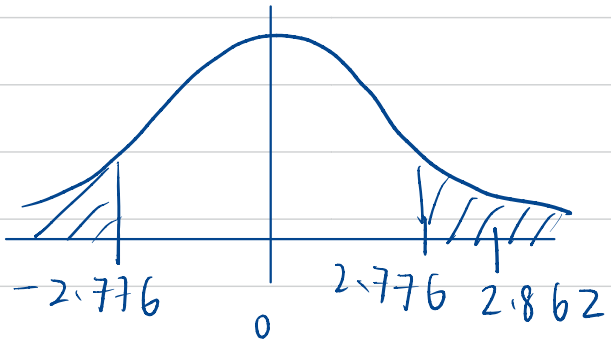
$$r=\frac{88360-(468)(1100)/6}{\sqrt{[(37194)-(468)^2/6][(215800)-(1100)^2/6]}}=\frac{2560}{\sqrt{(690)(14133.33)}}$$

$$=0.8197$$

$$t=\frac{0.8197}{\sqrt{\frac{1-(0.8197)^2}{6-2}}}=2.862$$

$\alpha=0.05$, $df=6-2=4$

$t_{0.025,4}=2.776$



$\therefore$ $2.862 > 2.776$. Thus we reject H_0.

There is sufficient evidence at 5% level of significance claim that there is correlation between 'Engagement Score' and 'Post length'

Ref: Date:

4.

x	y	xy	x^2	$\hat{y}$
95	85	8075	9025	88.76
85	95	8075	7225	82.66
80	70	5600	6400	79.61
70	75	5250	4900	73.51
65	70	4550	4225	70.46
$\Sigma x = 395$	$\Sigma y = 395$	$\Sigma xy = 31550$	$\Sigma x^2 = 31775$	

a. $\Sigma x = 395$

$\Sigma y = 395$

$\Sigma xy = 31550$

$\Sigma x^2 = 31775$

b. $b_1 = \dfrac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}$

$= \dfrac{31550 - \frac{(395)(395)}{5}}{31775 - \frac{(395)^2}{5}}$

$= 0.61$

$\bar{x} = \bar{y} = \frac{395}{5} = 79$

$b_0 = \bar{y} - b_1 \bar{x}$

$= 79 - 0.61(79)$

$= 30.81$

c. $\hat{y} = 30.81 + 0.61x$

d. Scatter plot and regression line with y-intercept of 30.81 and slope of 0.61.

e. $\hat{y} = 30.81 + 0.61(60)$

$= 67.41$

f. $SSR = \Sigma(\hat{y} - \bar{y})^2$

$= (88.76-79)^2 + (82.66-79)^2 + (79.61-79)^2 + (73.51-79)^2 + (70.46-79)^2$

$= 212.10$

$SST = \Sigma(y_i - \bar{y})^2$

$= (85-79)^2 + (95-79)^2 + (70-79)^2 + (75-79)^2 + (70-79)^2$

$= 470$

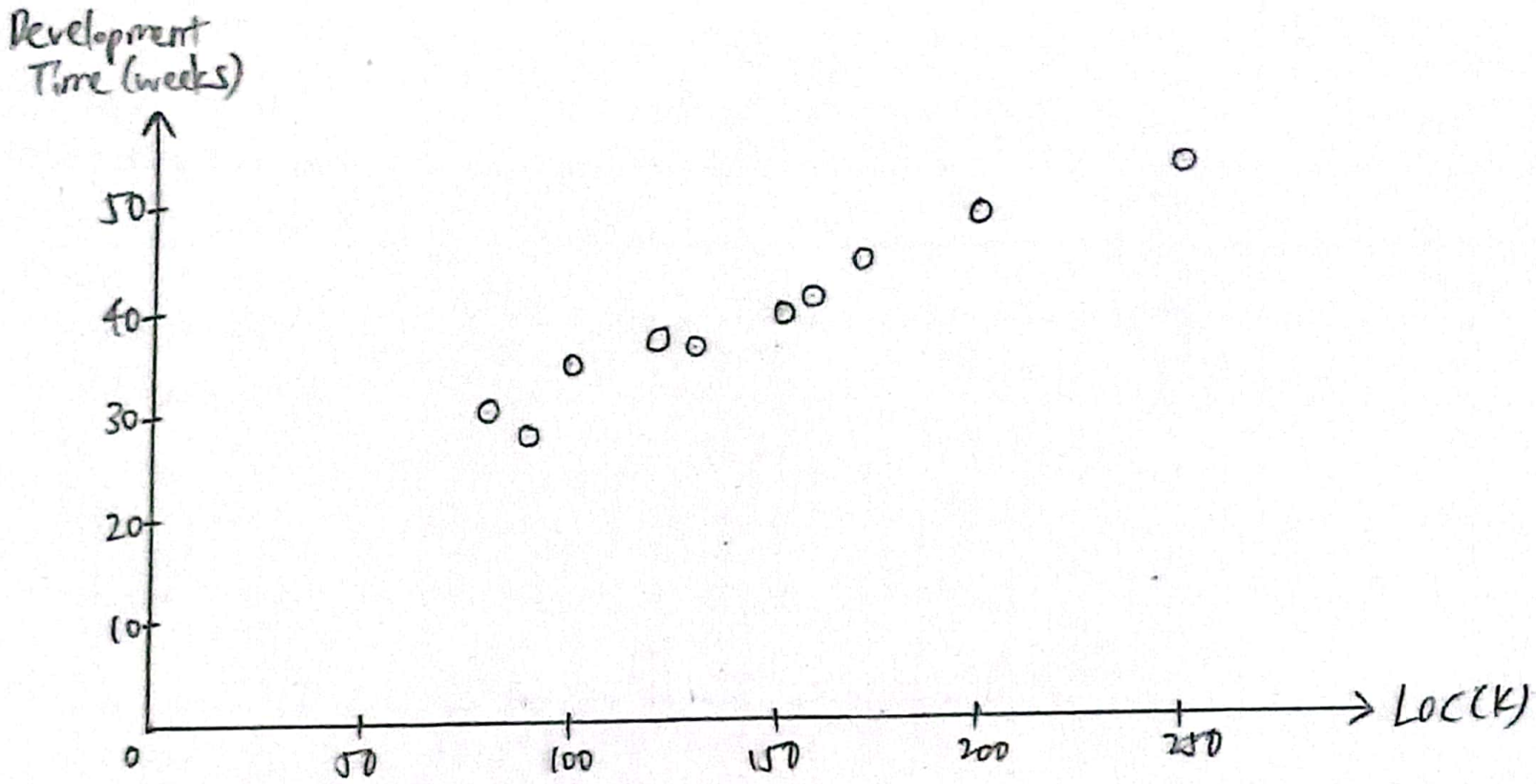
$R^2 = \dfrac{SSR}{SST}$

$= \dfrac{212.10}{470}$

$= 0.45$

g. R^2 value of 0.45 indicates that 45% of variation in statistic grade is explained by math aptitude test scores. It suggest a moderate fit of regression equation to the data.

5a.



Ref: Date:

5b.

LOC (x)	Development Time (y)	xy	x^2	y^2
150	40	6000	22500	1600
100	35	3500	10000	1225
200	50	10000	40000	2500
80	30	2400	6400	900
170	45	7650	28900	2025
120	38	4560	14400	1444
160	42	6720	25600	1764
90	28	2520	8100	784
250	55	13750	62500	3025
130	37	4810	16900	1369
$\Sigma x = 1450$	$\Sigma y = 400$	$\Sigma xy = 61910$	$\Sigma x^2 = 235300$	$\Sigma y^2 = 16036$

$$r = \frac{\Sigma xy - (\Sigma x)(\Sigma y)/n}{\sqrt{[\Sigma x^2 - (\Sigma x)^2/n][\Sigma y^2 - (\Sigma y)^2/n]}}$$

$$= \frac{61910 - (1450)(400)/10}{\sqrt{[235300 - (1450)^2/10][16036 - (400)^2/10]}}$$

$$= 0.98$$

c. $$b_1 = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}$$

$$= \frac{61910 - \frac{(1450)(400)}{10}}{235300 - \frac{(1450)^2}{10}}$$

$$= 0.16$$

$$\bar{y} = \frac{400}{10} = 40$$

$$\bar{x} = \frac{1450}{10} = 145$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

$$= 40 - 0.16(145)$$

$$= 16.8$$

$\therefore \hat{y} = 16.8 + 0.16x$ where 16.8 is the y-intercept and 0.16 is the slope.

d. $R^2 = r^2$

$= 0.98^2$

$= 0.96$

$\therefore R^2$ value of 0.96 indicates that 96% of variation in development time is explained by variation in LOC.

e. $\hat{y} = 16.8 + 0.16(180)$

$= 45.6$ weeks

$\therefore$ The development time for a new project with an estimated 180 K LOC is 45.6 weeks.