UTM

UNIVERSITI TEKNOLOGI MALAYSIA

FACULTY OF COMPUTING

SEMESTER 2

2023/2024

SECI1143 - PROBABILITY & STATISTICAL DATA ANALYSIS

SECTION 02

ASSIGNMENT 4 - CHAPTER 7

LECTURER: DR. NOORFA HASZLINNA MUSTAFFA

NAME	MATRIC NUMBER
NABIL AFLAH BOO BINTI MOHD YOSUF BOO YONG CHONG	A23CS0252
LUBNA AL HAANI BINTI RADZUAN	A23CS0107
NUR FIRZANA BINTI BADRUS HISHAM	A23CS0156
NAWWARAH AUNI BINTI NAZRUDIN	A23CS0143

Assignment 4

Question 1

a. Two most popular correlation coefficient are
- Pearson's product-moment correlation coefficient
- Spearman's rho rank correlation coefficient

b.

x	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
$\Sigma x = 600$	$\Sigma y = 804$	$\Sigma xy = 46544$	$\Sigma x^2 = 36486$	$\Sigma y^2 = 60104$

$n = 12$

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

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

$$= \frac{46544 - 40200}{\sqrt{(6486)(6236)}}$$

$$= \frac{6344}{6359.7717}$$

$$= 0.998$$

c. There is relatively strong positive linear relationship between the number of items produced and the production cost because the correlation coefficient, r obtained is near to 1 which is 0.998.

NO.

Date

Question 2

a.

x	y	xy	x^2	y^2
0.27	2	0.54	0.0729	4
1.41	3	4.23	1.9881	9
2.19	3	6.57	4.7961	9
2.83	6	16.98	8.0089	36
2.19	4	8.76	4.7961	16
1.81	2	3.62	3.2761	4
0.85	1	0.85	0.7225	1
3.05	5	15.25	9.3025	25
$\Sigma x = 14.6$	$\Sigma y = 26$	$\Sigma xy = 56.8$	$\Sigma x^2 = 32.9632$	$\Sigma y^2 = 104$

$n = 8$

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

$$= \frac{56.8 - [14.6(26)]/8}{\sqrt{[(32.9632) - (14.6)^2/8][104 - (26)^2/8]}}$$

$$= \frac{9.35}{\sqrt{(6.3182)(19.5)}}$$

$$= 0.842$$

b. $H_0 : \rho = 0$

$H_1 : \rho \neq 0$

C.I = 0.95

$\alpha = 1 - 0.95$

$= 0.05$

$d.f = 8 - 2$

$= 6$

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

$$= \frac{0.842}{\sqrt{\frac{1-(0.842)^2}{8-2}}}$$

$$= 3.823$$

$t_{\alpha/2} = t_{0.025, 6}$

$= 2.447$

∴ Since $t > t_{\alpha/2}$ which $3.823 > 2.447$, reject H_0. There is sufficient evidence of a linear relationship between weight of plastic usage and the size of household at the 95 % confidence level.

c. C.I = 0.99

$\alpha = 1 - 0.99$

$= 0.01$

d.f = 6

$H_0 : \rho = 0$

$H_1 : \rho \neq 0$

$t_{\alpha/2} = t_{0.005, 6}$

$= 3.707$

$t > t_{\alpha/2}$

$3.823 > 3.707$

∴ Reject H_0 because there is sufficient evidence to claim the linear relationship between weight of plastic usage and the size of household. Thus, there is no change between the decision in (b) and decision in (c) if the confidence level increased to 99%.

No Date

Question 3

a) Engangement score , sentiment score

b) Number of likes, number of comments

c)

Post ID	Engangement Score	Rank	sentiment score	Rank	d_i	d_i^2
1	85	3	4	3	0	0
2	70	5	3	4.5	0.5	0.25
3	90	1	5	1.5	-0.5	0.25
4	60	6	2	6	0	0
5	88	2	5	1.5	0.5	0.25
6	75	4	3	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. likes and comments

likes (x)	comments (y)	xy	x^2	y^2
150	20	3000	22 500	400
100	10	1000	10 000	100
200	25	5000	40 000	625
80	8	640	6 400	64
170	22	3740	28900	484
120	15	1800	14400	225
$\Sigma x = 820$	$\Sigma y = 100$	$\Sigma xy = 15180$	$\Sigma x^2 = 122200$	$\Sigma y^2 = 1898$

STANDARD

$$r = \frac{15180 - \left[\frac{(820)(100)}{6}\right]}{\sqrt{\left[122\,200 - \frac{820^2}{6}\right]\left[1898 - \frac{100^2}{6}\right]}}$$

$$= 0.988$$

b) likes and share

likes (x)	share (y)	xy	x^2	y^2
150	30	4500	22 500	900
100	20	2000	10 000	400
200	40	8000	40 000	1600
80	15	1200	6 400	225
170	35	5950	28 900	1225
120	25	3000	14 400	625
$\Sigma x = 820$	$\Sigma y = 165$	$\Sigma xy = 24\,650$	$\Sigma x^2 = 122\,200$	$\Sigma y^2 = 4975$

$$r = \frac{24\,650 - \frac{(820)(165)}{6}}{\sqrt{\left[122\,200 - \frac{820^2}{6}\right]\left[4975 - \frac{165^2}{6}\right]}}$$

$$\doteq 0.997$$

No: Date:

c) Engangement score and sentiment score

Rank ES (x)	Rank SS (y)	xy	x^2	y^2
3	3	9	9	9
5	4.5	22.5	25	20.25
1	1.5	1.5	1	2.25
6	6	36	36	36
2	1.5	3	4	2.25
4	4.5	18	16	20.25
$\Sigma x = 21$	$\Sigma y = 21$	$\Sigma xy = 90$	$\Sigma x^2 = 91$	$\Sigma y^2 = 90$

$$r = \frac{90 - \frac{(21)(21)}{6}}{\sqrt{(91 - \frac{21^2}{6})(90 - \frac{21^2}{6})}}$$

$$= 0.971$$

∴ The strongest positive correlation is between likes and shares which is 0.997 and the strongest negative correlation is between engangement score and sentiment score

DARD

No. Date

e) $H_0 : \rho = 0$

$H_1 : \rho \neq 0$

y	x	xy	x^2	y^2
85	200	17 000	40 000	7225
70	100	7 000	10 000	4900
90	250	22 500	62 500	8100
60	150	9000	22 500	3600
88	220	19 360	48 400	7744
75	180	13 500	32 400	5625
$\Sigma y = 468$	$\Sigma x = 1100$	$\Sigma xy = 88360$	$\Sigma x^2 = 215800$	$\Sigma y^2 = 37194$

$$r = \frac{88360 - \frac{(468)(1100)}{6}}{\sqrt{\left[215800 - \frac{1100^2}{6}\right]\left[37194 - \frac{468^2}{6}\right]}}$$

$$= 0.820$$

test statistic:

$$t = \frac{0.820}{\sqrt{\frac{1 - 0.820^2}{6-2}}}$$

$$= 2.865$$

critical value

$\alpha = 0.05$

$df = 6 - 2 = 4$

$t_{0.025,4} = \pm 2.776$

∴ since $t = 2.865 > t_{0.025,4} = 2.776$, reject H_0. There is sufficient evidence to conclude that there is significant correlation between Engangement score and Post length.

No: ..... Date: .....

QUESTION 4

Last year, five randomly selected students took a math aptitude test before they began their statistic course. The Statistics Department would like to analyze the relationship of the data, make predictions using the regression equation and validate the regression equation. The data of the math aptitude test score and corresponding statistics grade are as shown in Table 4 below:

Score on math aptitude test(x)	Statistic grade(y)
95	85
85	95
80	70
70	75
65	70

a) Calculate Σx, Σy, Σxy and Σx^2.

x	y	xy	x^2
95	85	8075	9025
85	95	8075	7225
80	70	5600	6400
70	75	5250	4900
65	70	4550	4225
$\Sigma x = 395$	$\Sigma y = 395$	$\Sigma xy = 31550$	$\Sigma x^2 = 31775$

$\Sigma x = 395$

$\Sigma y = 395$

$\Sigma xy = 31550$

$\Sigma x^2 = 31775$

b) Calculate value of b_1 and b_0.

$$b_1 = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{\Sigma x^2 - \frac{(\Sigma x)^2}{n}} = \frac{31550 - \frac{(395)(395)}{5}}{31775 - \frac{(395)^2}{5}}$$

$$= \frac{345}{570}$$

$b_1 = 0.61$ #

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

$\bar{y} = 79$

$\bar{x} = 79$

$b_0 = 79 - (0.61)(79)$

$= 30.81$ #

No: ..... Date: .....

c) Based on answers in (a) and (b), what linear regression equation best predicts statistics performance, based on math aptitude scores?

$\hat{y} = b_0 + b_1 x$

$\hat{y} = 30.81 + 0.61x$ #

d) which graph represents the regression equation in (c)?

Scatter plot graph. The regression equation line will be a straight line with the intercept 30.81 and slope 0.61.

e) If a student made a 60 on the aptitude test, what grade would we expect her to make in statistics?

$x = 60$

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

$= 30.81 + 0.61(60)$

$= 67.41$ #

f) Find the value of SSR, SST and R^2. $\bar{y} = 79$

x	y	$\hat{y}$	$(\hat{y} - \bar{y})^2$	$(y_i - \bar{y})^2$
95	85	88.76	95.26	36
85	95	82.66	13.40	256
80	70	79.61	0.37	81
70	75	73.51	30.14	16
65	70	70.46	72.93	81

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

$= 95.26 + 13.40 + 0.37 + 30.14 + 72.93$

$= 212.10$ #

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

$= 36 + 256 + 81 + 16 + 81$

$= 470$ #

$R^2 = \frac{SSR}{SST} = \frac{212.10}{470}$

$R^2 = 0.45$ #

g) Based on answer in (f), how well does the regression equation fit the data?

45% of the variation in statistics grade is explained by variation in score math aptitude test.

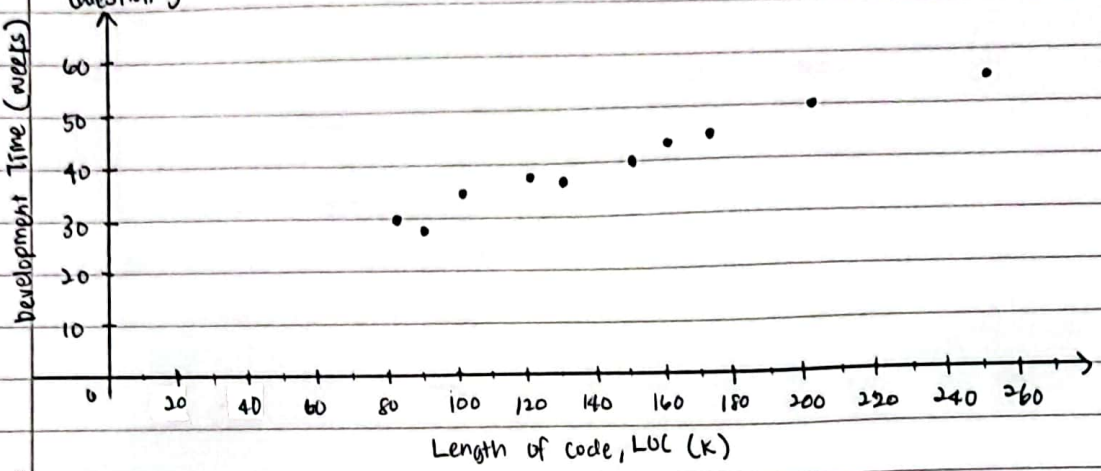
STANDARD

Ref: Date:

PSDA Assignment 4

Question 5

a)



b) Correlation Coefficient:

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

Project ID	LOC (K)	Development Time (weeks)	x^2	y^2	xy
1	150	40	22500	1600	6000
2	100	35	10000	1225	3500
3	200	50	40000	2500	10000
4	80	30	6400	900	2400
5	170	45	28900	2025	7650
6	120	38	14400	1444	4560
7	160	42	25600	1764	6720
8	90	28	8100	784	2520
9	250	55	62500	3025	13750
10	130	37	16900	1369	4810
Total	1450	400	235300	16636	61910

$$r = \frac{61910 - \frac{(1450)(400)}{10}}{\sqrt{\left[(235300) - \frac{(1450)^2}{10}\right]\left[(16636) - \frac{(400)^2}{10}\right]}}$$

$\therefore r = 0.98$, since $r = 0.98$, we ~~only~~ conclude $r = 0.98$ relatively strong positive linear association between LOC and Development Time.

Ref: Date:

c) least square Equation:

$\hat{y}_i = b_0 + b_1 x$, x = LOC, y = development ~~week~~ time

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

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

$$= 0.16$$

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

$\bar{y} = 40$

$\bar{x} = 145$

$b_0 = 40 - (0.16)(145)$

$= 16.8$

∴ the regression equation = $\hat{y} = 16.8 + 0.16x$

Interpreting the coefficient: the slope of the regression line is 0.16

the intercept is 16.8

1. The slope : indicates that the estimated change in the average value is 0.16, thus, $b_1 = 0.16$ tells us that the LOC will change in value of $0.16(1) = 0.16$ of ~~reg~~ length of code on average for each additional week.

2. The intercept = 16.8.

Here, $b_0 = 16.8$ tells us that the average value of development time when the LOC (x) value is zero, is 16.8 weeks (if $x=0$ is in the range of observed value).

~~b~~ Here, no development time has zero LOC, so $b_0 = 16.8$ just indicates that for development ~~we~~ time observed, 16.8 LOC is the portion ~~that~~ not explained by development time.

d) $R^2 = \frac{SSR}{SST}$ $\boxed{0 \leq R^2 \leq 1}$

$= \frac{641.28}{636}$

$= 1.00$

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

$= 636$

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

$= $ ~~1050.88~~ 641.28

∴ since $R^2 = 1.00$, ~~in~~ it is perfect linear relationship between LOC and development time.

PSDA — Ref: Date:

Assignment 4

d) Regression Model :

estimated 180 K LOC

$$\hat{y} = 16.8 + 0.16x$$

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

$$\hat{y} = 45.6$$

∴ The predicted Development Time for a new project with an estimated 180K LOC is 45.6 weeks.