Assignment II

1

a) $\hat{y}_i = \hat{\alpha}_i + \hat{\beta} x_i$

b) Gradient of regression line =)

$$\hat{\beta} = 1.7242$$

2.

i) $\hat{\beta} = \frac{S_{xy}}{S_{xx}}$ & $\hat{\alpha}_0 = \bar{y} - \hat{\beta}\bar{x}$

$$S_{xx} = \Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n} = 12666.5 - \frac{(385.2)^2}{12} = 301.66$$

$$S_{yy} = \Sigma y_i^2 - \frac{(\Sigma y_i)^2}{n} = 119026.9 - \frac{(1162.5)^2}{12} = 6409.1725$$

$$S_{xy} = \Sigma x_i y_i - \frac{(\Sigma x_i)(\Sigma y_i)}{n} = 38191.41 - \frac{(385.2)(1162.5)}{12}$$

$$= 875.16$$

$\hat{\alpha}_i = 3.748181$ $\hat{\beta} = 2.901146$

$\therefore \hat{y}_i = \hat{\alpha}_i\ 3.748181 + 2.901146 x_i$

ii) $PQ = \frac{\hat{\beta} - \beta}{Se(\hat{\beta})} \sim t_{n-2}$

$Se(\hat{\beta}) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$ =) $\hat{\sigma}^2 = \frac{1}{10}\left[6404.7125 - \frac{(878.16)^2}{301.66}\right]$

$$= 387.0744$$

$Se(\hat{\beta}) = 1.132$.

$$\Rightarrow P\left[-2.228 < \frac{2.901 - \beta}{1.132} < 2.228\right] = 0.95$$

$$\Rightarrow P\left[-2.228(1.132) < 2.901 - \beta < 2.228(1.132)\right] = 0.95$$

$$\Rightarrow P\left[2.901 - 2.228(1.132) < \beta < 2.901 + 2.228(1.132)\right] = 0.95$$

95% of CI $\Rightarrow$

$(0.378904, 5.423096)$.

iii) $PQ = \dfrac{\hat{y_i} - y_i}{Se(y_i)} \sim t_{n-2}$

$$\Rightarrow \hat{y_i} = 3.748181 + 2.90146(40)$$
$$= 119.79$$

$$\Rightarrow Se(y_i) = \sqrt{\left[\frac{1}{12} + \frac{(7.9)^2}{301.66}\right] 387.07} = 10.5898$$

$$\Rightarrow P\left[-2.228 < \frac{119.79 - y_i}{10.5988} < 2.228\right] = 0.95$$

$\therefore$ CI of $y_i \Rightarrow$ $(96.17, 143.40)$.

Q.3

i)

→ • The center of the distribution differ from all the 4 cities. Therefore the underlying means are different.

• The difference between the mean time taken to commute to office in peak hours & non peak hours are in order from A to D.

•

ii)

→ Assumptions underlying analysis of variance:

i) it must be a normal distribution.
ii) it should have a common variance.
iii) observation is independent.

iii)

→ H_0 : mean of difference is same for each city
H_1 : mean is different.

$\therefore SS_{TOT} = S_{YY} = \Sigma y_i^2 - \frac{(\Sigma y_i)^2}{n}$

$= 1495 - \frac{103^2}{28}$

$= 1116.11$

$\therefore SS_{REG} = \frac{1}{7}\left[64^2 + 44^2 + 8^2 + (-13)^2\right] - \frac{103^2}{28}$

$= 516.11$

$\therefore SS_{RES} = SS_{TOT} - SS_{REG} = 600$

iv) Analysis of mean difference.

Since, $\bar{y}_{1*} = 9.14$

$\bar{y}_{2*} = 6.29$

$\bar{y}_{3*} = 1.14$

$\bar{y}_{4*} = -1.86$

& $\bar{y}_{1*} > \bar{y}_{2*} > \bar{y}_{3*} > \bar{y}_{4*}$

$$\hat{\sigma}^2 = \frac{SS_{RES}}{n-2} = 25$$

$$\therefore t_{24, 0.025} \times \hat{\sigma} \sqrt{\frac{1}{7} + \frac{1}{7}}$$

$$= 2.064 \times \sqrt{25} \times \sqrt{\frac{2}{7}}$$

$$= 5.52$$

Now we can examine the difference between each of the pairs of mean. If the difference is less than the level of significance, then there is no significant difference betⁿ the means.

$\therefore$ we have,

$\bar{y}_{1*} - \bar{y}_{2*} = 2.85$

$\bar{y}_{2*} - \bar{y}_{3*} = 5.15$

$\bar{y}_{3*} - \bar{y}_{4*} = 3$

all of these pairs have difference less than 5.52

$\therefore \bar{y}_{1*} > \bar{y}_{2*} > \bar{y}_{3*} > \bar{y}_{4*}$

Examining to see if the first 2 groups can be combined;

$\bar{y}_{1*} - \bar{y}_{2*} = 8.$

Hence we can't combine the last 2 groups.
Therefore, the diagram remains as before.

Q.4.

a).

→ Model :

$$Y_{ij} = \mu + \tau_i + e_{ij} \quad , \quad i = 1, 2, \ldots, K \quad j = 1, 2, \ldots, M$$

where,

K → no. of treatments.

M → no. of responses from i^{th} treatment.

Y_{ij} → j^{th} response from i^{th} treatment.

τ_i → i^{th} treatment.

μ → over mean response.

e_{ij} → error term,

b)

→ H_0: Mean of claim amounts is same for all compan

H_1: Mean is not same.

$m_1 = 7$

$m_2 = 8$

$m_3 = 6$

$m_4 = 7$

$m_5 = 5$

$m = 33$.

$$\therefore \sum_{i=1}^{5} \sum_{j=1}^{m} y_{ij} = 1856.$$

$$\sum_{i=1}^{5} \sum_{j=1}^{m} y_{ij}^2 = 174316$$

$$\therefore CF = \frac{\left\{ \sum_{i=1}^{5} \sum_{j=1}^{m} y_{ij} \right\}^2}{M} = 104385.94.$$

$$SS_{TOT} = 174316 - CF$$
$$= 69930.06.$$

$$SS_{REG} = \frac{354^2}{7} + \frac{388^2}{8} + \frac{87^2}{6} + \frac{643^2}{7} + \frac{384^2}{3} - CF$$
$$= 22325.69.$$

$$\therefore SS_{RES} = SS_{TOT} - SS_{REG}$$
$$= 47604.37.$$

Test Statistic = $\frac{SS_{REG}}{SS_{RES}}$ = 3.283.

$\therefore F_{4,28,0.05}$ = 2.714.

$\therefore$ We have sufficient evidence to reject H_0.

c) $\rightarrow$. O_i

	3-5	5-8	8-10	10-12		
0-3	45	20	6	5	=	76
4-6	7	20	9	6	=	42
7-9	5	8	15	12	=	40
	57	48	30	23		158

E_i.

	3-5	5-8	8-10	10-12
0-3	27.41772	23.0845	14.4304	12.01234
4-6	15.15189	12.7588	7.978	4.0069
7-9	14.4312	12.1581	4.2228	8.0781

note: no clubbing for E_i is required as all E_i are greater than 5.

H_0 : Sales are independent
H_1 : sales are dependent.

$$TS = \sum \frac{(O_i - E_i)^2}{E_i}$$
$$= 49.91146.$$

∴ Degrees of freedom = 6

We reject H_0

Q.5

i) Assumptions underlying ANNOVA.

=) population is normal
=) it has common variance
=) observations are independent.

The sample variance are very different from each other
Hence,
the common variance assumption won't apply

classmate

Date

Page

ii)

$\rightarrow$ Mean at Rate 1 $= \frac{\sqrt{29} + \sqrt{13} + \sqrt{21}}{3}$

$= 4.52441$

Variance of Rate 2 $= \frac{\Sigma (x_i - \bar{x})^2}{n-1}$

$= 0.12127$.

iii)

$\rightarrow$ The test was correct asserting that the large transformation must be done before carrying out one way ANOVA as for the transformation it can be claimed that the assumption of common variance for the underlying proportion holds.

iv)

a)

$\rightarrow Y_{ij} = \mu + T_i + e_{ij}$ $\quad i = 1, 2, 3 \ldots$

$j = 1, 2, 3 \ldots$

- Y_{ij} is the $\log_e$ transformed value of the j^{th} observation of the no. of generation per square food observed when the i^{th} rate was applied

- μ is the overall population mean.

- τ_i is the deviation of the i^{th} rate mean such that $\Sigma \tau_i = 0$.

- e_{ij} are the independent error terms.

b)
→ For ANOVA :

$H_0 : \tau_i = 0$
$H_1 : \tau_i \neq 0$

lm (ne).

Rate	Mean	Variance	Y_i^2
1	2.992	0.163	80.582
2	4.342	0.216	209.678
3	5.687	0.144	294.053
4	6.1124	0.231	389.063
Total	20.11	0.618	973.373

$$\therefore Y_i^2 = \left(\sum_{j=1}^{3} Y_{ij}\right)^2 = (3 \times \text{Mean}_i)^2$$

$$SS_{Rate} = \sum_{i=1}^{n} \frac{Y_i^2}{3} - \frac{Y^2}{12} =$$

$$= 21.153$$

$$SS_{RES} = \sum_{i=1}^{4} \left(\sum_{j=1}^{3} (Y_{ij} - \bar{Y}_{i})^2 \right).$$

$$= \sum_{i=1}^{4} \{2 \times \text{variance}\}$$

$$= 1.236.$$

c)

→ Given the observed F statistic value is much larger than this, we can tell the P-value for this test is almost near to 0 or in other words there is overwhelming evidence against H_0. Thus it can be concluded that the underlying means are not equal.

Q.6.

$\rightarrow$ $\Sigma x = 39$ $\Sigma y = 562$

$\Sigma x^2 = 207$ $\Sigma y^2 = 40508$

$\Sigma xy = 2853$

$n = 10$.

i) $\rightarrow$ $S_{xx} = 207 - \frac{39^2}{10}$

$= 54.9$

$S_{yy} = 40508 - \frac{562^2}{10}$

$= 8923.6$

$S_{xy} = 2853 - \frac{562 \times 39}{10}$

$= 661.2$

$\therefore r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{661.2}{\sqrt{54.9 \times 8923.6}}$

$= 0.94466$

ii) $\rightarrow$ $\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{661.2}{54.9}$

$= 12.04371$

$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 56.2 - 12.04371 \times 3.9$
$= 9.2295$

ii)
→ $R^2 = \frac{S^2_{XY}}{S_{XX} \cdot S_{YY}} = 89.24\%$

Comment :

• Model is a good fit as it explains 89% of the variability in the output.

Q.7.

i)
→ $\Sigma X = 174.13$ $\Sigma Y = 197.84$
$\Sigma X^2 = 7545.9$ $\Sigma Y^2 = 4747.45$

$\Sigma(X-\bar{X})(Y-\bar{Y}) = 2176.84$ $n = 11$

$\therefore S_{XX} = 7545.9 - \frac{174.13^2}{11}$

$= 4789.4221$

$S_{YY} = 4747.45 - \frac{197.84^2}{11}$

$= 1189.207$

$S_{xy} = 2176.84$.

$\therefore \hat{\beta} = \frac{S_{xy}}{S_{xx}} = 0.45451$

$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$
$= 10.7805$

$\therefore \hat{Y} = 10.78056 + 0.45451 \hat{X}$.

ii)

$H_0 : \beta = 0$.
$H_1 : \beta \neq 0$.

$\therefore$ Test Statistic :-).

$$\frac{\hat{\beta} - \beta}{SE(\hat{\beta})} \sim t_{n-2}.$$

$$\hat{\sigma}^2 = \frac{1}{n-2}\left(S_{yy} - \frac{S_{xy}^2}{S_{xx}}\right)$$
$$= 22.20136.$$

$$SE(\hat{\beta}) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}.$$

$$\therefore TS = \frac{0.45451 - 0}{\sqrt{\frac{22.2013}{4789.422}}} \sim t_9.$$

$t_9 \sim 6.6755.$

$\therefore t_{9, 2.5\%} = 2.262.$

$\therefore$ Reject H_0

Assumptions:

i) popⁿ has common variance
ii) it follows normal distribution.
iii) obs is independent

iii) $\rightarrow r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = 0.91213.$

v) $\rightarrow$ There is clearly some relationship between residuals & the explanatory variable values. Hence, our assumption of normality might be wrong.

$\therefore$ The model is not a good fit for the data.

Q.8.

i)

→ The main purpose of factor analysis is to reduce any individual item into a fewer no. of dimensions.

Factor Analysis can be used to simplify data such as reducing the no. of variables in regression model.

Principal component analysis is a technique for reducing the dimensionality of such data sets increasing interpretability but at the same time ~~min~~ information loss.

ii)

→ Principal Component	Diagonal entries	PC_i over I_{tos}
PC1	0.456	6.5%
PC2	0.137	19.5%
PC3	0.080	11.4%
PC4	0.0165	2.4%
PC5	0.012	1.7%
	0.7015	

classmate

Date

Page

Q.9

→ $\cdot \Sigma X = 45$

$\Sigma X^2 = 277.5$

$\Sigma Y = 644$

$\Sigma Y^2 = 43956$

$\Sigma XY = 2602$

$n = 10$

i)

→ There appears to be a negative correlation between marks score & hours spent on social media.

ii)

→ $\cdot S_{xx} = 277.5 - \frac{45^2}{10}$

$= 75$

$S_{YY} = 43956 - \frac{644^2}{10}$

$= 2482.4$

$S_{xy} = 2602 - \frac{45 \times 644}{10}$

$= -296$

$\therefore r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = -0.686002$

iv)

→ $\hat{\beta} = \frac{S_{xy}}{S_{xx}} = -3.9467$

$\hat{\alpha} = \bar{y} - \bar{x}\hat{\beta}$

$= 82.16 -$

$\therefore \hat{y} = 82.16 - 3.9467\hat{x}$

v)

→ $R^2 = \frac{S^2_{xy}}{S_{xx} S_{yy}} = 47.0598\%$

∴ model is not a good fit for the data as it covers only 47% of variability.

vi)

→ $x_0 = 1$

Expected Change $= \hat{\beta} x_0$

$= -3.9467 \times 1$

$= -3.9467$

∴ for additional 1 hour spent on social media, the expected decrease in marks is 3.9467.

classmate

Date
Page

Q.10

→ $\Sigma X = 0.93$

$\Sigma X^2 = 0.2925$

$\Sigma Y = 0.23$

$\Sigma Y^2 = 0.0189$

$\Sigma XY = 0.0735$

$n = 3$

$$\therefore S_{xx} = 0.2925 - \frac{0.93^2}{3}$$

$$= 0.0042$$

$$S_{yy} = 0.0189 - \frac{0.23^2}{3}$$

$$= 0.00126$$

$$S_{xy} = 0.0735 - \frac{0.93 \times 0.23}{3}$$

$$= 0.0022$$

$$\therefore SS_{TOT} = S_{yy} = 0.00126$$

$$SS_{REG} = \frac{S_{xy}^2}{S_{xx}} = 0.00115$$

$$SS_{RES} = SS_{TOT} - SS_{REG}$$

$$= 0.00011$$