PBS-II Assignment -2

Q 1) a) $\hat{y}_i = \hat{\alpha}_i + \hat{\beta} x_i$

1 b) Gradient of regression line : $\hat{\beta} = 1.72497531$

Q 2) i) $\hat{\beta} = \frac{S_{xy}}{S_{xx}}$ and $\hat{\alpha}_i = \bar{y} - \hat{\beta}\bar{x}$

$S_{xx} = \Sigma x_i^2 - \frac{(\Sigma y_i)^2}{n} = 12666.58 - \frac{(385.2)^2}{12} = 301.66$

$S_{yy} = \Sigma y^2 - \frac{(\Sigma x_i)^2}{n} = 119026.9 - \frac{(1162.5)^2}{12} = 6409.7125$

$S_{xy} = \Sigma x_i y_i - \frac{\Sigma(x_i)\Sigma(y_i)}{n} = 35191.41 - \frac{(385.2)(1162.5)}{12}$

$= 875.16$

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

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

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

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

$\hat{\sigma}^2 = \frac{1}{10}\left[6409.7125 - \frac{(875.16)^2}{301.66}\right] = 387.0744$

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

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

$\{0.378904, 5.423096\} = 95\%$ CI for β.

iii) PQ $= \dfrac{\hat{\mu}_i - \mu_i}{Se(\hat{\mu}_i)} \sim t_{m-2}$

$\hat{\mu} = 3.748181 + 2.901146(40)$
$= 119.79$

$Se(\mu_i) = \sqrt{\left[\dfrac{1}{12} + \dfrac{(7.9)^2}{301.66}\right] 387.07} = 10.5988$

$P\left[-2.228 < 119.79 - \mu_i < 2.228\right] = 0.95$

$\Rightarrow 95\%$ CI for $\mu_i = \{96.17, 143.40\}$

Q.34 i) Comments for the plot

- The center of the distribution differ for all the 4 cities. Thus, there is a prime face suggesting that their underlying means are different.

- The difference between the mean time taken to commute to office in peak hours & non peak hours are in the order city A to city B

- ii) Assumptions for the analysis of variances

  - The population must be normal

  - The population have a common variance

  - The observations are unbiased

iii) H_0 = The mean difference is same for each city

H_1 : The mean of difference not must the same

$$SS_{TOT} = S_{yy} = \Sigma y_i^2 - \frac{(\Sigma y^2)}{n} = 1495 - \frac{103^2}{28}$$

$$= 1716.11$$

$$SS_{REG} = \frac{1}{7}\left(64^2 + 44^2 + 8^2 + (-13)^2\right) - \frac{103^2}{28} = 516.11$$

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

ANOVA table.

Source of Variation	df	sum of squares	Mean sum of squares
Regression	3	516.11	172.04
Residual	24	600	25
Total	27	1116.11	

$$F = \frac{MSS_{REG}}{MSS_{RES}} = \frac{172.04}{25} = 6.88$$

Test Statistic = 6.88

$F_{3,24,5\%} = 3.009$

IV) Analysis of mean difference

Since $\bar{y}_1 = 9.14$

$\bar{y}_2 = 5.29$

$\bar{y}_3 = 1.14$

$\bar{y}_4 = 1.86$

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

The least significance difference between means is

$$t_{24, 0.025} \times \hat{\sigma} \sqrt{\frac{1}{7} + \frac{1}{7}} = 2.064 \times \sqrt{25} \times \sqrt{\frac{1}{7} + \frac{1}{7}}$$

$= 5.52$

Q.4a) The mathematical model of one way ANOVA is given by

$$Y_{ij} = \mu + \tau_i + e_{ij} \; ; \; i = 1, 2, 3, \ldots, k \; ; \; j = 1, 2, 3, \ldots, n_i$$

where

k = number of treatments

n_i = number of responses from i^{th} treatment

Y_{ij} = j^{th} response from i^{th} treatment

τ_i = is the i^{th} treatment

μ = mean response

e_{ij} = is the error term

Assumption: e_{ij} are iid ϵ $N(0, \sigma^2)$

Page No
Date

H_0: Mean claim amounts of 5 companies are equal
H_1: Mean claim amounts of 5 companies are unequal

$m_1 = 7$ $\qquad$ $m_4 = 7$
$m_2 = 8$ $\qquad$ $m_5 = 5$
$m_3 = 6$ $\qquad$ $n = 33$

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

$$CF = \left\{ \sum_{i=1} \sum_{j=1} y_{ij} \right\} / n = 104385.94$$

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

$$SS_{REG} = \frac{354^2}{7} + \frac{386^2}{8} + \frac{87^2}{6} + \frac{643^2}{7} + \frac{384^2}{3} - 104385.94$$

$$= 22325.69$$

$$SS_{RES} = SS_{TOT} - SS_{REG} = 47604.37$$

Sources of Variation	df	Sum of Squares	Mean sum of Squares
Regression	4	22326	5581
Residual.	28	47604	1700
Total	32	69930.	

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

$F_{4,28,0.025} = 2.714$

We have sufficient evidence to reject H_0

c)

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	27.417	23.088	14.430	11.063
4-6	15.151	12.759	7.974	6.113
7-9	14.430	12.151	7.594	5.822

No clubbing require for E_i as all values are greater than 5.

H_0 : Sales are independent
H_1 : Sales are dependent.

$$TS = \frac{\Sigma(O_i - E_i)^2}{E_i} = 49.914\ 49.91146$$

degree of freedom $= 2 \times 3 = 6$

$$\Rightarrow \chi^2_{6,5\%} = 12.59$$

We have sufficient evidence to reject H_0

Q 57 i) Assumptions under ANOVA
- Populations are normal
- Populations have common variance
- Observations are independent

The sample variances are very different from each other. Hence, the variance assumption wont hold.

ii) Mean at Rate for $\sqrt{x}$ $= \frac{\sqrt{29} + \sqrt{13} + \sqrt{21}}{3}$

$= 4.52443$

Variance at Rate-2 $= \frac{\sum(x-\bar{x})^2}{n-1} = 0.12127$

iii) The test was correct asserting that the large transformation must be done before carrying out one way ANOVA as for the transformation ~~it cannot~~ it can be claimed that the assumption of common value variance for the underlying variance proportion holds.

iv) a) $Y_{ij} = \mu + \tau_i + e_{ij}$, $i = 1,2,3,4, \; j = 1,2,3$

Y_{ij} is the $\log_e$ tranformationed value of the j^{th} observation of the no of germnation per square food observed.

iv)b) For ANOVA

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

ln(flu)

Rate	Mean	Variance	Y_i^2
1	2.992	0.163	80.582
2	4.827	0.121	209.678
3	5.716	0.186	294.053
4	6.575	0.148	389.063
	20.11	0.618	973.373

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

$SS_{Rate} = 21153$

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

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

$= 2 \times 0.618$

$= 1.236$

Source of Variation	df	SS	MSS
Rate	3	21.153	7.051
Residuals	8	1.236	0.155
	11	22.387	

$$TS = \frac{MSS_{Rate}}{MSS_{Res}} = 45.638$$

$$F_{3,8,1\%} = 7.951$$

6) $\Sigma X = 39$, $\Sigma X^2 = 207$, $\Sigma Y = 562$, $\Sigma Y^2 = 40508$
$\Sigma XY = 2853$, $n = 10$

i) $S_{XX} = 207 - \frac{39^2}{10} = 54.9$

$$S_{YY} = 40508 - \frac{562^2}{10} = 8923.6$$

$$S_{XY} = 2853 - \frac{562 \times 39}{10} = 661.2$$

$$r = \frac{S_{XY}}{\sqrt{S_{XX} S_{YY}}} = \frac{661.2}{\sqrt{54.9 \times 8923.6}} = 0.94466$$

ii) $\hat{\beta} = \frac{S_{XY}}{S_{XX}} = \frac{661.2}{54.9} = 12.04371$.

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

iii) $SS_{TOT} = SS_Y = 8923.6$.

$SS_{REG} = \frac{S_{XX}^2}{S_{XX}} = \frac{661.2^2}{54.9} = 7963.30492$.

$SS_{RES} = S_{TOT} - SS_{REG} = 8923.6 - 7963.30492$
$= 960.29508$

iv) $R^2 = \frac{S_{XY}^2}{S_{XX} S_{XY}} = 89.24$.

Comment
→ The model is good giving us 89% of variability in the output

7) i) $\Sigma X = 174.13$, $\Sigma Y = 197.84$, $\Sigma X^2 = 7545.9$

$\Sigma Y^2 = 4747.45$, $\Sigma (X-\bar{X})(Y-\bar{X}) = 2176.84$

$n = 11.$

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

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

$S_{XY} = 276.84.$

$\hat{y} = 10.78056 + 0.45451X$

ii) $H_0 : \beta = 0 \qquad H_1 : \beta \neq 0$

$$TS = \frac{\hat{\beta} - \beta}{S_E(\hat{\beta})} \sim t_{n-2}$$

$$\hat{\sigma}^2 = \frac{1}{n-2}\left(S_{YY} - \frac{S_{XY}^2}{S_{XX}}\right) = 22.20136$$

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

$$T.S = \frac{0.45451 - 0}{\sqrt{\frac{22.20136}{4789.4221}}} \sim t_{9}$$

$$= 6.6755 \sim t_{9}$$

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

H_0 reject at 5% LS

$\therefore \beta \neq 0$ hence there is a linear relationship.

Assumptions →

1. Pop^n is normal
2. Pop^n has a common variance
3. The observations are independent

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

iv) $X_0 = 25$

Mean Response $(\mu_0) = \frac{\hat{\mu} - \mu}{S_e(\hat{\mu})} \sim t_{n-2}$

$SE(\hat{\mu}) = \sqrt{\left(\frac{1}{n} + \frac{(X_0 - \bar{X})}{S_{xx}}\right)\hat{\sigma}^2}$

$= 1.55$

95% CI for $\mu = \hat{\mu} \pm t_{.9,25} \times SE(\hat{\mu})$

$= (\cancel{8}18.64313, 25.66349)$

v) There is clearly some relationship b/w residuals and the explanatory variable values. Hence our assumptions ~~might be~~ of normality might be wrong. The model is not a good fit for the data

8) i) The main purpose of factor analysis is to reduce many individuals items into a fewer number of dimensions.

Factor Analysis can be used to simplify data, such as reducing the number of variables in regression model.

While newly identified principle components are chosen to be →

i) Uncorrelated
ii) Linear combination of original values of data.
iii) Which maximize the variance

ii)

Principal Component	Diagonal Entries	PC_1/Sum (PC
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	

iii) As the first 3 PC explain over 95% of the variance, dimensionstrating can be reduced to the dataset.

9) $\Sigma X = 45$, $\Sigma X^2 = 277.5$, $\Sigma Y = 644$, $\Sigma Y^2 = 43956$, $\Sigma XY = 2602$, $n = 10$.

i) There appears to be negative correlation b/w the marks scored and hours spent per day on social media

ii) $S_{XX} = 277.5 - \frac{45^2}{10} = 75$

$S_{YY} = 43956 - \frac{644^2}{10} = 2482.4$

$S_{XY} = 2602 - \frac{644 \times 45}{10} = -296$

$r = \frac{S_{XY}}{\sqrt{S_{XX} S_{YY}}} = -0.6860062$

iii) $H_0: \rho = 0$; $H_1: \rho < 0$

$$T.S \rightarrow \frac{\tanh^{-1}(r) - \tanh^{-1}(\rho)}{\sqrt{1/n-3}} \sim N(0,1)$$

$\rightarrow -2.07893$

$Z_{5\%} = 1.6448$

$|Z_{cal}| > Z_{tab}$

Hence, enough evidence to reject H_0.

$\Rightarrow \rho < 0$

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

$\hat{\alpha} = \bar{y} - \bar{x}\hat{\beta} = 82.16$

$\hat{y} = 82.16 - 3.9467X$

v) $R^2 = \frac{S_{xy}^2}{S_{xx} S_{yy}} = 47.0598\%$

Hence model explain only 47% of variabi

vi) $X_0 = 1$

Expected Change $= \hat{\beta} x_0$
$= -3.9467 \times 1$
$= -3.9467$

Page No. | Date

104 ii) $\Sigma X = 0.93$, $\Sigma X^2 = 0.2925$, $n = 3$, $\Sigma Y = 0.23$, $\Sigma Y^2 = 0.0159$, $\Sigma XY = 0.0735$

$$S_{xx} = 0.2925 - \frac{0.93^2}{3} = 0.00126$$

$$S_{YY} = 0.0159 - \frac{0.23^2}{3} = 0.00126$$

$$S_{XY} = 0.0735 - \frac{0.93 \times 0.23}{3} = 0.0022$$

$$SS_{TOT} = S_{YY} = 0.00126$$

$$SS_{REG} = S_{XY}^2 / S_{XX} = 0.00115$$

$$SS_{RES} = 0.00011$$

~~ANNOV~~ ANOVA table.

Industry has no effect on resignation rate.

Sources of Variation	DOF	SS	MSS
Regression	1	0.00115	0.00115
Residual	1	0.00011	0.00011
Total	2	0.00126	

$H_0: B = 0$ $\quad H_1: B \neq 0$.

$$T.S = \frac{0.00115}{0.00011} \sim F_{1,1} = 10.4545 \sim F_{1,1}$$

$F_{1,1,0.05} = 161.4$

$F_{cal} < F_{tab}$ We donot have sufficient evidence to reject H_0