

PBS-II Assignment -2

Q. 17 a) $\hat{y}_i = \hat{\alpha}_i + \hat{\beta}x_i$

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

Q. 27 i) $\hat{\beta} = \frac{S_{xy}}{S_{xx}}$ and $\hat{\alpha}_i = \bar{y} - \hat{\beta}\bar{x}$

$$S_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n} = 126658 - \frac{(385.2)^2}{12} = 301.66$$

$$S_{yy} = \sum y_i^2 - \frac{(\sum y_i)^2}{n} = 119026.9 - \frac{(1162.5)^2}{12} = 6409.7125$$

$$S_{xy} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} = 38191.41 - \frac{(385.2)(1162.5)}{12} = 875.16$$

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

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

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

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

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

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

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

$$\{0.378904, 5.423096\} = 95\% \text{ CI for } \beta$$

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

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

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

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

$$\Rightarrow 95\% \text{ CI for } \mu_i = \{96.12, 143.40\}$$

Q.37 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

$$SS_{TOT} = SS_Y = \sum y_i^2 - \frac{(\sum y_i)^2}{n} = 1495 - \frac{103^2}{28} = 1716.11$$

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

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

ANOVA table

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

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

$$\text{Test Statistic} = 6.88$$

$$F_{3,24,51} = 3.009$$

iv) Analysis of mean difference

$$\text{Since } \bar{y}_1 = 9.14$$

$$\bar{y}_2 = 5.29$$

$$\bar{y}_3 = 1.14$$

$$\bar{y}_4 = 1.86$$

$$ii) \hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{661.2}{54.9} = 12.04371$$

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

$$iii) SS_{TOT} = SS_y = 8923.6$$

$$SS_{REG} = \frac{S_{xy}^2}{S_{xx}} = \frac{661.2^2}{54.9} = 7963.30492$$

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

$$iv) R^2 = \frac{S_{xy}^2}{S_{xx} S_{yy}} = 89.24\%$$

Comment

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

$$Z_{\text{cal}} = 1.6448$$

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

Hence, enough evidence to reject H_0 .

$$\Rightarrow p < 0$$

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

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

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

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

Hence model explain only 47% of variability

$$\text{vi) } X_0 = 1$$

$$\begin{aligned} \text{Expected Change} &= \hat{\beta} X_0 \\ &= -3.9467 \times 1 \\ &= -3.9467 \end{aligned}$$

10

Source
Variation

Regression
Residual
Total

8.57 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 won't hold.

$$\text{ii) Mean at Rate for } \sqrt{n} = \frac{\sqrt{29} + \sqrt{13} + \sqrt{21}}{3} = 4.52443$$

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

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.

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

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

qual
qual

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

$$\text{Test Statistic} = \frac{MSS_{\text{REG}}}{MSS_{\text{RES}}} = 3.287$$

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

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

$$TS = \sum_{E_i} \frac{(O_i - E_i)^2}{E_i} = 49.91449.9144$$

$$\text{degree of freedom} = 2 \times 3 = 6$$

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

We have sufficient evidence to reject H_0

	01-01	01-02	01-03	01-04	01-05
JP	2	2	05	211	2-8
SN	2	P	05	F	2-4
OH	51	21	8	2	P-4
321	05	05	30	FB	

	01-01	01-02	01-03	01-04	01-05
P	830-11	0811-11	220-85	811-15	8-0
	811-2	1111-1	121-51	121-21	2-4
	552-2	1112-1	121-51	0811-11	P-4

iv) b) For ANOVA

$$H_0 = \bar{Y}_i = 0$$

$$H_1 = \bar{Y}_i \neq 0$$

ln (fit)

Rate	Mean	Variance	Y_i
1	2.992	0.163	80.382
2	4.827	0.121	209.678
3	5.716	0.186	294.063
4	6.575	0.148	389.063
	20.11	0.618	973.373

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

$$SS_{\text{Rate}} = 21153$$

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

$$= 2 \times \text{Variance} \times 2 \times \text{Variance}$$

$$= 2 \times 0.618$$

$$= 1.236$$

$$T.S = \frac{0.45451 - 0}{\sqrt{\frac{22.20136}{4789 \cdot 4221}}} \sim t_{39}$$

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

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

H_0 reject at 5% L.S

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

Assumptions $\rightarrow$

- ① Popⁿ is normal
- ② Pop has a common variance
- ③ The observations are independent

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

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

The least significance difference between means is

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

$$= 5.52$$

8.44a) The mathematical model of one way ANOVA is given by.

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

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 $\mathcal{N}(0, \sigma^2)$

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

$$\begin{aligned} m_1 &= 7 & m_4 &= 7 \\ m_2 &= 8 & m_5 &= 5 \\ m_3 &= 6 & \bar{m} &= 3.3 \end{aligned}$$

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

$$CF = \left\{ \sum_{i=1}^5 \sum_{j=1}^{m_i} y_{ij} \right\} / n = 104385.94$$

$$\begin{aligned} SS_{TOT} &= 174316 - CF \\ &= 69930.06 \end{aligned}$$

$$\begin{aligned} SS_{REG} &= \frac{354^2}{7} + \frac{386^2}{8} + \frac{87^2}{6} + \frac{643^2}{7} + \frac{384^2}{5} - \frac{104385.94^2}{174316} \\ &= 22325.89 \end{aligned}$$

$$SS_{RES} = SS_{TOT} - SS_{REG} = 47604.17$$

84 i) The main purpose of factor analysis is to reduce many individual 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 _i / Sum (PC _i)
PC1	0.436	6.5%
PC2	0.137	19.5%
PC3	0.080	11.4%
PC4	0.0165	2.4%
PC5	0.012	1.7%
	0.7015	

Page No.
 Date.
 iii) As the first 3 PC explain over 95% of the variance, demonstrating can be reduced to the dataset.

9y $\sum X = 45$, $\sum X^2 = 277.5$, $\sum Y = 644$, $\sum Y^2 = 439.56$, $\sum XY = 2602$, $n = 10$.

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

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

$S_{yy} = \frac{439.56 - \frac{644^2}{10}}{10} = 2482.4$

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

$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = -0.6860002$

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

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

$\rightarrow -2.07893$

104 $\sum X = 0.93$, $\sum X^2 = 0.2926$, $n = 3$, $\sum Y = 0.23$,
 $\sum Y^2 = 0.0189$, $\sum XY = 0.0735$

$$S_{xx} = 0.2926 - \frac{0.93^2}{3} = 0.00126$$

$$S_{yy} = 0.0189 - \frac{0.23^2}{3} = 0.00126$$

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

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

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

$$SS_{RES} = 0.00011$$

Industry has no effect on resignation rate.

ANNOV ANOVA table

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

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

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

$F_{1,1,5\%} = 161.4$ $F_{cal} < F_{tab}$ We do not have sufficient evidence to reject H_0

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

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

$$F_{3,8,11} = 7.951$$

$$\sum X = 39, \sum X^2 = 207, \sum Y = 562, \sum Y^2 = 40608, \sum XY = 2853, n = 10$$

$$i) S_{xx} = \frac{\sum X^2 - \frac{(\sum X)^2}{n}}{n} = 54.9$$

$$S_{yy} = \frac{\sum Y^2 - \frac{(\sum Y)^2}{n}}{n} = 8923.6$$

$$S_{xy} = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{n} = 661.2$$

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{661.2}{\sqrt{54.9 \times 8923.6}} = 0.94466$$

$$\text{7Y i.) } \sum X = 174.13, \sum Y = 197.84, \sum X^2 = 7545.9$$

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

$$n = 11$$

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

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

$$S_{xy} = 2176.84$$

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

$$\text{ii) } H_0: \beta = 0 \quad H_1: \beta \neq 0$$

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

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

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

iv) $X_0 = 25$

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

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

$$= 1.55$$

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

$$= (18.64313, 25.66349)$$

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.