PROBABILITY & STATISTICS - 2

ASSIGNMENT - 2

1. (a) Let regression line y on x be

$$\hat{y} = \hat{\alpha} + \hat{\beta} x$$

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{\Sigma xy - n\bar{x}\bar{y}}{\Sigma x^2 - n\bar{x}^2}$$

$\Sigma xy = 182560$

$\bar{x} = \frac{\Sigma x_i}{n} = 85.0$

$\bar{y} = \frac{\Sigma y_i}{n} = \cancel{137}\ 173.7$

$\Sigma x^2 = 92500$

$\Rightarrow \hat{\beta} = \frac{34915}{20250} = 1.72421$

$\therefore \hat{y} = 27.143 + 1.7242x$

(b) Since the slope (β) is positive, y and x have a positive relation.

MAHAGURU Page No.
Date

Q2. (i) $\hat{\beta} = \frac{S_{xy}}{S_{xx}}$ and $\hat{\alpha}_i = \bar{y} + \hat{\beta}\bar{x}_i$

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

$= 301.66$

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

$= 119026.9 - \frac{(1162.5)^2}{12} = 6409.7125$

$S_{xy} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}$

$= 38191.41 - \frac{(385.2)(1162.5)}{12}$

$= 875.16$

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

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

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

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

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

$s(\hat{\beta}) = 1.132$

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

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

95% CI for $\beta = \{2.901 - 2.228(1.32),\ 2.901 + 2.228(1.132)\}$

$= \{0.378904,\ 5.423096\}$

MAHAGURU Page No.
Date

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

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

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

$P\left[-2.228 < \dfrac{119.79 - \mu_i}{10.5988} < 2.228\right] = 0.95$

95% CI for $\mu_i = \{119.79 - 2.228\,(10.5988),\ 119.79 + 2.228\,(10.5)\}$
$= \{96.17, 143.40\}$

Q3. Comments
(i) The plot suggest:
The variation in city C is least as compared to which has the maximum variance

→ The time difference in commutation in peak hours and non-peak hours is maximum in city A and least in city D.

(ii) Assumptions:

→ The population ~~mean~~ has a common variance

→ The observations are independent

→ The population must be normal

(iii) Ho : The mean of difference is same for each city

H_1 : The mean of difference is not the same

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

$$= 1495 - \frac{103^2}{28} = 1116.11$$

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

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

Annova Table

Sr of variation	df	Sum of sq.	mean sum of sq
Regression	3	516.11	172.04
Residual	24	600	25
Total	27	1116.11	197.04

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

Test statistic = 6.88

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

Hence we have sufficient evidence to reject Ho at LS of 5%. Thus, there are underlying difference b/w the cities

MAHAGURU Page No. Date / /

IV Analysis of mean differences

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

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

$y_{3\cdot} = 1.14$

$y_{4\cdot} = -1.86$

$\bar{y}_1 > \bar{y}_2 > \bar{y}_3 > \bar{y}_4$

$\hat{\sigma}^2 = \dfrac{SS_{Res}}{n-2} = 25$

The least significance difference between mean is

$t_{24,\ 0.025} \times \hat{\sigma}^2 \sqrt{\frac{1}{7}+\frac{1}{7}} = 2.064 \times \sqrt{25} \times \sqrt{\frac{1}{7}+\frac{1}{7}} = 5.52$

$\bar{y}_1 - \bar{y}_2 = 2.85$

$\bar{y}_2 - \bar{y}_3 = 5.15$

$y_3 - y_4 = 3$

All are less than 5.52

$\bar{y}_1 > \bar{y}_2 > \bar{y}_3 > \bar{y}_4$

Q4. (a) Mathematical model :

$y_{ij} = \mu + \tau_i + e_{ij}$ where

y_{ij} = j^{th} observation on the ith treatment

μ = common effect for the whole experiment

τ_i = ith treatment effect

ε_{ij} = random error in the j^{th} observation on the i^{th} treatment

Assumptions

→ errors are assumed to be normally & independently distributed with mean 0 & constant variance σ^2

→ μ is fixed parameter

(b) H_0 : mean claims are equal
H_1 : mean claims are not equal
$m_1 = 7$, $m_2 = 8$, $m_3 = 6$ $m_4 = 7$ $m_5 = 5$
$m_{tot} = 33$

$$\sum_{i=1}^{5} \sum_{j=1}^{m_i} y_{ij} = 354 + 386 + 87 + 648 + 384 = 1856$$

$$\sum_{i=1}^{5} \sum_{j=1}^{m_i} y^2_{ij} = 27184 + 29430 + 2123 + 84669 + 30910$$
$$= 174316$$

$$SS_{TOT} = 174316 - \left[33 \times \left(\frac{1856}{33} \right)^2 \right]$$

$$= 69930.06$$

$$SS_{REG} = \left[\frac{354^2}{7} + \frac{386^2}{8} + \frac{87^2}{6} + \frac{645^2}{7} + \frac{384^2}{5} \right]$$
$$- \left[33 \times \left(\frac{1856}{33} \right)^2 \right]$$
$$= 22325.69$$

$$SS_{res} = SS_{TOT} - SS_{reg}$$
$$= 47604.37$$

Src of var.	DOF	Sum of sq	MSS
Regression	4	22325.69	5581.4225
Residual	28	47604.37	1700.156
Total	32	69930.06	

$$TS : \frac{MSS_{REG}}{MSS_{RES}} \sim f_{4,28}$$

$$F_{cal} = \frac{5581.4225}{1700.156} = 3.2829$$

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

Since $F_{cal} > F_{tab}$, we reject H_0

Mean claims are not equal for the 5 companies

(c) H_0 : No association between salaries & no. of papers cleared

H_1 : There is an interdependence between salary and papers cleared

O_i

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

E_i

	3-5	5-8	8-10	10-12
0-3	27.41772	23.08861	14.43038	11.06329
4-6	15.15189	12.75949	7.97468	6.11392
7-9	14.43038	12.15189	7.59494	5.82278

TS : $\sum \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{(3-1)(2-1)}$

$\chi^2_{cal} = 49.91146$

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

Since $\chi^2_{cal} > \chi^2_{tab}$, we reject H_0

Salaries and no. of papers cleared are associated.

Q5. (i) Assumptions

→ Populations are normal
→ Population have common variance
→ Observation are independent

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

(ii) Missing values are:

Mean, → $\frac{\sqrt{29} + \sqrt{13} + \sqrt{21}}{3} = 4.5244$

Var, → $\frac{\Sigma (x_i - \bar{x})^2}{n-1} = 0.1213$

(iii) log transformation is the best suited as the common variance assumption holds true in this case

(iv) (a) $y_{ij} = \mu + \tau_i + \varepsilon_{ij}$; $i = 1, 2, 3, \ldots$
$j = 1, 2, 3, \ldots$

$H_0 : \tau_i = 0$

$H_1 : \tau_i \neq 0$

(b) $y_i^2 = \Sigma y_{ij}^2 = (3 \times \text{mean}_i)^2 = 973.373$

$SS_{reg} = \sum_{i=1}^{4} \frac{y_i^2}{3} - \frac{y^2}{12} = 12.153$

$SS_{reg} = \sum_{i=1}^{4} \left(\sum_{j=1}^{3} (y_{ij} - \bar{y}_i)^2 \right)$

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

$= 2 \times 0.618$

$= 1.236$

Src of vari	DoF	sum of sq	MSS
Regression	3	21.153	7.051
Residual	8	1.236	0.155
Total	11	22.386	

Ts : $\frac{MSS_{reg}}{MSS_{res}} \sim F_{3,8}$

$F_{cal} = 45.638$

$F_{tab} = F_{3,8,5\%} = 4.066$

Since $F_{cal} > F_{tab}$, we reject H_0

6. (i) correlation coeff $(r) = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$

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

$= \frac{661.2}{\sqrt{34.9 \times 8923.6}}$

$= 0.9447$

(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) $SS_{TOT} = SS_{y} = 8923.6$

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

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

(iii) $R^2 = \dfrac{S^2_{XY}}{S_{XX}\cdot S_{YY}} = 89.24\%$

Q7. (i) $\Sigma X = 174.13 \qquad \Sigma Y = 197.84$

$\Sigma X^2 = 7545.9 \qquad \Sigma Y^2 = 4747.45$

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

$n = 1$

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

$= 4789.4221$

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

$= 1189.20767$

$S_{XY} = 2176.84$

$\hat{\beta} = \dfrac{S_{XY}}{S_{XX}} = 0.45451$

$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} = 10.78056$

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

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

$T.S \rightarrow \dfrac{\hat{\beta} - \beta}{se(\hat{\beta})} \sim t_{n-2}$

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

$se(\hat{\beta}) = \sqrt{\dfrac{\hat{\sigma}^2}{S_{XX}}}$

$T.S = \dfrac{0.45451 - 0}{\sqrt{\dfrac{22.20136}{4789.4221}}} \sim t_9$

$6.6755 \sim t_9$

Assumptions

1. Pop^n have a common variance
2. Pop^n follow the normal $dist^n$
3. The obs^n are independent

(iii) $r = \dfrac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = 0.91213$

(iv) $x_0 = 25$

Ind. response

$$y_0 \sim N\left[\hat{\alpha} + \hat{\beta} x_0, \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xy}}\right]\right]$$

$$TS : \frac{y_0 - (\hat{\alpha} + \hat{\beta} x_0)}{\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim t_{n-2}$$

$$95\% \; CI = y_0 \pm t_{n-2, 0.025} \sqrt{\hat{\sigma}^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}$$

$$= (\alpha + \beta x_0) \pm t_{9, 0.025} \sqrt{\hat{\sigma}^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}$$

$$= (10.9319, 33.3745)$$

Mean response

$$TS : \frac{\mu_0 - (\hat{\alpha} - \hat{\beta} x_0)}{\hat{\sigma}^2 \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xy}}}} \sim t_{n-2}$$

$$95\% \; CI : \alpha + \hat{\beta} x_0 \pm t_{9, 0.025} \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xy}}\right)}$$

$$= (18.643, 25.663)$$

(v) The residual plot appears to follow a pattern (first increasing and then decreasing), which implies that the model is a good fit for the data.

8. (i) Factor analysis provides a method for reducing the dimensionality of the data set i.e it seeks to identify the key components necessary to model and understand data.

→ the original data is rewritten in terms of few variables that are unrelated ; only some are needed to explain most of the variability observed in the data.
The new variables are called principle components.

(ii) variance - cov matrix of PC1, PC2, PC3, PC4, PC5 is given by :

$$\begin{bmatrix} 0.456 & 0 & 0 & 0 & 0 \\ 0 & 0.137 & 0 & 0 & 0 \\ 0 & 0 & 0.080 & 0 & 0 \\ 0 & 0 & 0 & 0.0165 & 0 \\ 0 & 0 & 0 & 0 & 0.012 \end{bmatrix}$$

% of total vaiance explained by each PC

$$PC_1 = \frac{0.456}{0.7015} \times 100 = 65\%$$

$$PC_2 = \cancel{0.0}\ \frac{0.137}{0.7015} \times 100 = 19.53\%$$

$$PC_3 = \frac{0.080}{0.7015} \times 100 = 11.40\%$$

$$PC_4 = \frac{0.0165 \times 100}{0.7015} = 2.35\%$$

$$PC_5 = \frac{0.012 \times 100}{0.7015} = 1.71\%$$

iii) We can conclude that the dimensionality could be reduced 3. using the first 3 PC's, as 96% of the variability has been explained by these 3 alone.

9. (i) Based on the plot, there is a negative linear correlation between marks and hours spent /day on social media.

(ii) Corr Coeff $(r) = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$

$$= \frac{\sum xy - n\bar{x}\bar{y}}{\sqrt{(\sum x^2 - n\bar{x}^2)(\sum y^2 - n\bar{y}^2)}}$$

$$= \frac{-296}{\sqrt{75 \times 2482.4}}$$

$$= \underline{\underline{-0.6860}}$$

iii) $H_0 = \rho = 0$

$H_1 : \rho < 0$

Using Fisher's Transformation,

$$w = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \sim N\left(\frac{1}{2}\ln\frac{1+\rho}{1-\rho}, \frac{1}{n-3}\right)$$

$$Z_{cal} = \frac{-0.84036 - 0}{\sqrt{\frac{1}{7}}} = -2.2234$$

$Z_{tab} = -1.98$

Since $Z_{cal} < Z_{tab}$; we reject H_0. Thus there is negative correlation between the variables.

(iv) Let the linear regression model :

$\hat{y} = \hat{\alpha} + \hat{\beta} x$

$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{-296}{75} = -3.9467$

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

$= 64.4 - (-3.9467)(4.5)$

$= 82.16$

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

(v) Coeff of determination $(R^2) = \frac{S^2_{xy}}{S_{xx} S_{yy}} = \frac{(-246)^2}{75 \times 2\cancel{4}824}$

$= 0.4706$

$r^2 = R^2$

vi) Expected change in marks $= \beta \Delta x$

$= -3.9467$

8

Page No.
Date

10. H_0 : Industry has no impact on resignation rate
H_1 : There is an impact

$$\text{Mean} = \frac{27+36+30}{3} = 31$$

$$SS_{Reg} = 20\left[(27-31)^2 + (36-31)^2 + (30-31)^2\right] = 840$$

$$SS_{Reg} = 19(5^2 + 10^2 + 8^2) = 3591$$

ANOVA TABLE

Source of variation	DOF	Sum of Sq	MSS
Residual	57	3591	63
Regression	2	840	420

$$TS : \frac{MSS_{Reg}}{MSS_{res}} \text{ in } F_{2,57}$$

$$F_{cal} = \frac{420}{63} = 6.67$$

$$F_{tas} \approx F_{2,60,.5}] = 3.150$$

→ Since $F_{cal} > F_{tabs}$, we reject H_0. There is an impact of industry on the resignation rate.