PROB & STATS ASSIGNMENT

Q.1

a) $\Sigma x^2 = 92500$, $\Sigma xy = 182500$

$\bar{x} = 85$, $n = 10$

$\Sigma y^2 = 372717$

$\bar{Y} = 173.7$

$S_{xx} = \Sigma x^2 - n\bar{x}^2 = 20250$

$S_{xy} = \Sigma xy - n\bar{x}\bar{y} = 34915$

$S_{yy} = \Sigma y^2 - n\bar{y}^2 = 71000.1$.

$\hat{\beta} = \dfrac{S_{xy}}{S_{xx}} = \cancel{0.57998}\ 1.724$

$\hat{a} = \bar{Y} - \hat{\beta}\bar{x}$

$= \cancel{124.4}\ 27.14$

a) Regression line :

$\hat{\beta} \times \quad Y = \hat{a} + \hat{\beta}x$

$y = \cancel{124.4} + \cancel{0.57998}x$.

$y = 27.14 + 1.724x$

b)

b) Gradient represents the amt of hours per rupee spent .

Page No
Date

Q.2

$\Rightarrow$ $\Sigma x^2 = 12666.58$

$\Sigma y^2 = 119026.9$

$\bar{x} = 32.1$

$\bar{Y} = 96.875$

$\Sigma xy = 38191.41$

$S_{xx} = \cancel{1222.81\ 38}\ 302.66$

$S_{xy} = 875.16$

$S_{yy} = 6409.7125$

(ii) $\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \cancel{0.2047}\ 2.9011$

$\hat{a} = \bar{Y} - \hat{\beta}\bar{x} = 3.7481$

(i) $y = \cancel{\hat{y}}\ \ Y = 3.7481 + 2.9011x$.

(ii) We know $\frac{\beta - \beta_0}{\sqrt{\hat{\sigma}^2 / S_{xx}}} \sim t_{n-2}$

$\hat{\sigma}^2 = \left(S_{yy} - \frac{S_{xy}^2}{S_{xx}} \right) \times \frac{1}{n-2} = 387.0744$

$\Rightarrow$

$se(\hat{\beta}) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} = 1.132$

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

$P(2.901 - 2.228 \times 1.132 < \beta < 2.901 + 2.201 \times 1.132)$

$\Rightarrow [0.378904 , 5.423096]$

(iii) Mean resp CI :

$$\frac{\hat{\mu} - \mu_0}{se(\hat{\mu})} \sim t_{n-2}$$

$\hat{\mu} = 3.7481 + 2.9011\,(40)$

$= 119.79$

$$se(\hat{\mu}) = \sqrt{\left(\frac{1}{n} + \frac{(x_0 - \bar{x})}{S_{xx}}\right)\hat{\sigma}^2}$$

$= 10.5988$

$$P\left(-2.228 < \frac{119.79 - \mu_0}{10.5988} < 2.228\right] = 0.95$$

95% CI :

$\Rightarrow$ (96.17, 143.40)

3)

(i) Comments :

⇒ The centre of the distribution differ for all the 4 cities. Thus there is a prime face for suggesting that the underlying mean are diff

⇒ The diff btw the mean time taken to commute to office in peak hours & non peak hours are in the order city A to city D

⇒ The variation in the data for city C is largest compared to city D appear to be highest, However with only 7 observations for each city, we cant be sure that there is a real underlying diff in variance

(ii) Assumptions :

⇒ The population must be normal

⇒ The population have common variance

⇒ The observations are independent

(iii) H_0: Mean diff is same for each city

H_1: The mean diff isnt the same.

SS_{TOT} : $1495 - 103^2/28 = 1116.11$

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

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

ANOVA :

Type	D.F	SS	MSS
Regress	3	516.11	172.04
Residual	24	600	25
Total	27	1116.11	

Under $H_0 = \frac{MSS_{REG}}{MSS_{RES}} = \frac{172.04}{25} = 6.88$

Test statistic $= 6.88$

$F_{3,24,5\%} = 3.009$ (right tailed test)

Hence we have sufficient evidence to reject H_0 at LS of 5%.

Thus there are underlying diff in mean of cities

iv) Analysis of Mean differences

$\overline{Y_1} = 9.14$

$\overline{Y_2} = 6.29$

$\overline{Y_3} = 2.14$

$\overline{Y_4} = -7.56$

$\overline{Y_1} > \overline{y_2} > \overline{y_3} > \overline{y_4}$

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

The least significance diff btw any of mean is

$t_{24,0.025} \times \hat{6} \sqrt{\frac{1}{7} + \frac{1}{7}} = 2.064 \times \sqrt{25} \times \sqrt{\frac{2}{7}}$

$= 5.52$

Now, we can examine the diff btw each each of pair of means. If the diff is less than the level of signific. Then there is no significant diff btw the means

We have:

$\overline{Y_1}_* > \overline{Y_2}_* > \bar{y}_{3*} > \bar{y}_{4*}$

Examining to see if the first 2 groups can be combined

$\bar{Y_1} - \bar{Y}_2 = 8$

There is significant diff btw mean 12.3 so we cant combine the first two groups

Examining to see if last 2 groups can be com

$\overline{Y_2}_* - \overline{Y}_{4*} = 8.5$

Hence we cant combine the last 2 groups

Therefore the diagram remains as before.

Page No.
Date

A) The mathematical assumptions model of one way ANOVA is given by :

$y_{ij} = \mu + \tau_i + e_{ij}$, $i = 1, 2, \ldots k$
$j = 1, 2, \ldots m$.

where
k = number of treatments
n_i = no of responses from ith treatm
τ = is the ith treatment
μ = is the over mean response
e_{ij} = is the error term

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

B) H_0 : Mean claims amt of 5 comp are equal.
H_1 : Mean claim amt of 5 comp are diff

$m_1 = 7$
$m_2 = 8$
$m_3 = 6$
$m_4 = 7$
$m_5 = 5$

$$\sum_{i=1}^{5} \sum_{j=1}^{m_i} y_{ij} = 1856$$

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

$$CF = \left\{ \sum_{i=1}^{5} \sum_{j=1}^{m_i} y_{ij} \right\}^2 / m = 104385.94$$

$SS_{TOT} = 174316 - CF$

$= 69930.06$

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

$= 22325.69$

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

Source of Variation	DOF	SS	MSS
Regression	4	22326	5581
Residual	28	47604	1700
Total	32	69930	

$$T.S = \frac{MSS_{REG}}{MSS_{RES}} = 3.283$$

$F_{4,28,0.05} = 2.714$

We have sufficient evidence to reject H_0

c) Observations (Oi)

	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

Expected values: (Ei)

	3-5	5-8	8-10	10-12	Total
0-3	27.42	23.09	14.43	11.06	76
4-6	15.15	12.76	7.94	6.11	42
7-9	14.43	12.15	7.59	5.82	40
Total	57	48	30	23	158

No clubbing for Ei is required as all Ei are greater than 5

contingency table

$$T.S = \frac{\Sigma (Ei - Oi)^2}{Ei} = 49.9114$$

DOF = 6.

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

We have sufficient evidence to reject Ho

Page No.
Date

8.5 Assumptions underlying ANOVA

(i) Populations are normal

⇒ Population have common variance

⇒ observation are independent

⇒

the sample variance are very diff from each other Hence the common variance assumption wont hol.

(ii) Mean Rate 1 $= \frac{\sqrt{29} + \sqrt{13} + \sqrt{21}}{3} = 4.52443$ for $\sqrt{x}$

variance at Rate 2 for $\log_e(x)$

$\Rightarrow \frac{\Sigma(x_i - \bar{x}_i)^2}{m-1} = 0.12127$

(iii) The test was correct asserting that the large transformation must be done before carrying out one-way ANOVA as for the transformation it can be claimed that the assumptions of common variance for the underlying proportion hole.

(iv)

a) $y_{ij} = \mu + \tau_i + e_{ij}$

⇒ y_{ij} is the log transformed value of the jth observation of the no. of germation per square food observed when the i th rate was applied.

⇒ μ is overall population mean

⇒ τ_i is the deviation of the ith rate mean such that $\Sigma \tau_i = 0$

⇒ e_{ij} are independent error terms ($e_{ij} \sim N(0, \sigma^2)$)

For ANOVA

b) $H_0 : \tau_i = 0$

$H_1 : \tau_i \neq 0$

$\ln(x)$

Rate	Mean	Variance	Y_i^2
1	2.992	0.163	80.582
2	4.827	0.14	209.678
3	5.716	0.186	294.053
4	6.577	0.148	389.063
total	20.11	0.618	973.373

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

$$SS\ RATE = \sum_{i=1}^{4} \frac{Y_i^2}{3} - \frac{Y^2}{12} = 21.153$$

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

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

$$= 2 \times 0.618$$

$$= 1.236$$

Source of Variation	DOF	SS	MSS
Rate	3	21.153	7.051
Residuals	8	1.236	0.155
	11	22.389	

$$TS = \frac{MSS\ rate}{MSS\ res} = 45.638$$

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

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 mean are not equal

Page No.
Date

6) $\Sigma xy = 2853$, $n = 10$

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

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

(i) $Sxx = 207 - \frac{39^2}{10} = 54.9$

$Syy = 40508 - \frac{562^2}{10} = 8928.6$

$Sxy = 2853 - 562 \times 39/10 = 661.2$

$r = \frac{Sxy}{\sqrt{Syy \cdot Sxx}}$

(ii) $\hat{\beta} = \frac{Sxy}{Sxx} = \frac{661.2}{54.9} = 12.04371$

$\hat{a} = \overline{Y} - \hat{\beta}\overline{X} = 56.2 - 12.04371 \times 3.9$

$= 9.2295$

(iii) $SS_{TOT} = SSY = 8923.6$

$SSREG = \frac{S^2xy}{Sxx} = \frac{661.2^2}{54.9} = 7963.3049$

$SSRES = SS_{TOT} - SSREG$

$= 960.29508$

(IV) $R^2 = \frac{S^2xy}{Sxx \cdot Syy} = 89.24\%$

comment :

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

Page No.
Date

7) $\Sigma x = 174.13$, $\Sigma x^2 = 7545.9$

(i) $\Sigma y = 197.84$, $\Sigma y^2 = 4747.45$

$\Sigma xy = 2176.84$, $n = 11$

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

$$= 4789.4221$$

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

$$= 1189.20767$$

$$S_{xy} = 2176.84$$

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

$$\hat{a} = \bar{y} - \hat{\beta}\bar{x} = 10.78058$$

$$\hat{y} = 10.78058 + 0.45451x$$

(ii) $H_0 : \beta = 0$

$H_1 : \beta \neq 0$

$$T.S = \frac{\hat{\beta} - \beta_0}{se(\hat{\beta})} \sim t_{n-2}$$

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

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

Page No.
Date

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

$\Rightarrow 6.6755 \sim t_9$

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

$\therefore$ We have sufficient evidence to reject H_0 at 5% LS, $\beta \neq 0$ & hence there is a linear relationship

Assumptions :

$\Rightarrow$ population have a common variance

$\Rightarrow$ population follow the Normal dist

$\Rightarrow$ the observations are independent

(iii) $r = \frac{Sxy}{\sqrt{Sxx \cdot Syy}} = 0.91213$

(iv) $x_0 = 25$

Mean response (μ_0) = $\frac{\hat{\mu} - \mu}{se(\hat{\mu})} \sim t_{n-2}$

$$se(\hat{\mu}) = \sqrt{\left(\frac{1}{n} + \frac{(x_0 - \bar{x})}{Sxx}\right)\hat{\sigma}^2}$$

$$= \sqrt{\frac{1}{11} + \frac{(25 - 174.311) \times 22.02136}{4789.221}} = 1.55$$

$\hat{\mu} = \hat{a} + \hat{\beta} Y$

$= 22.153$

Page No.
Date

95% CI : $\hat{\mu} \pm t_{9\%, 2.5\%} \times se(\hat{\mu})$

$= 22.153 \pm 3.51018$

$= (18.64313, 25.66349)$

For individual response,

$Y_0 = \hat{a} + \hat{\beta} x_0$

$= 22.15331$

$$se(\hat{y}) = \sqrt{\left(\frac{1}{11} + 1 + \frac{(25-174.73)11^2}{4789.422}\right) 22.704}$$

$= 4.96079$

$$\frac{\hat{y} - y_0}{SE(y_0)} \sim t_9$$

95% CI for Y_0 = $\hat{y} \pm t_{9, 2.5\%} \times se(\hat{y}_0)$

$= 22.1531 \pm 11.22131$

$= (10.932, 33.3746)$

v) There is clearly some relationship b/w residuals & the explanatory variable values. Hence our assumption of homodity may be wrong the model is not a good fit for the data

Page No.
Date

9)
(i) The main purpose of factor analysis is to reduce many individual items into a fewer no of dimension. Factor analysis can be used to simplify data such as reducing the no of variables in a regression model

PCA is a technique for reducing the dimensionality of such data sets increasing in temperability but at the same time minimising data loss.

Newly identified principle comp analysis are chosen to be:

- ~~Uncorrelated~~
- linear combination of the original variables of data
- which maximise the variance

(I)

Principle component	Diagnol entries	PC / sum (PC)
PC 1	0.452	6.5 %.
PC 2	0.137	19.5 %.
PC 3	0.080	11.4 %.
PC 4	0.0165	2.4 %.
PC 5	0.012	1.7 %.
	0.7015	

(iii) At the 1st PC explains over 95 %. of the variance, demonstrating can be reduced to the dataset

The 1st 3 PC can be used for building further classifical or regression modelling purposes

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 btw the marks scored & how time spent per day on social media

(ii) $S_{xx} = 277.5 - 10(45/10)^2 = 75$

$S_{yy} = 43956 - 10(64.4)^2 = 2482.4$

$S_{xy} = 2602 - 10(64.4)(4.5) = -296$

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}} = -0.686$$

Hence we have a moderate negative correlation

(iii) $H_0 : \rho = 0$

$H_1 : \cancel{\rho \neq 0}\ \rho < 0$

$\cancel{\tanh^{-1}(r) \sim \tanh^{-1}(\rho)\ N(0,}$

$\tanh^{-1}(r) \sim N(\tanh^{-1}(\rho), 1/n-3)$

$$T.S \Rightarrow \frac{\tanh^{-1}(r) - \tanh^{-1}(\rho)}{\sqrt{1/n-3}}$$

$= -2.07893$

$z_{5\%} = 1.6448$

$|z_{cal}| > z_{tab}$.

Hence we have sufficient evidence to Reject H_0.

$\Rightarrow \rho < 0$

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

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

$= 82.16$

$Y_0 = 82.16 - 3.9467 x_0$

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

Hence model explain only 47% of the variability. The model is not a good fit for the data

(vi) $x_0 = 1$

Expected change : $\hat{\beta} x_0$

$= -3.9467 \times 1$

$= -3.9467$

$\therefore$ For every additional 1 hour spent on social media, the expected decrease in marks is -3.9467.

Page No.
Date

10) $\Sigma x = 0.93$

$\Sigma x^2 = 0.2925$

$\Sigma y = 0.23$

$\Sigma y^2 = 0.0189$

$\Sigma xy = 0.0735$

$n = 3$

$S_{xx} = \Sigma x^2 - n\bar{x}^2 = 0.0042$

$S_{yy} = \Sigma y^2 - n\bar{y}^2 = 0.00126$

$S_{xy} = \Sigma xy - n\bar{x}\bar{y} = 0.0022$

$SS_{TOT} = S_{yy}$

$= 0.00126$

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

$SS_{RES} = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$

$= 0.00011$

ANOVA TABLE :

source of variation	DOF	SS	MSS
Regression	1	0.00115	0.00115
Residual	1	0.00011	0.00011

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

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

$F_{1,1} = 161.4$; $F_{tab} > F_{cal}$

$\therefore$ we don't have any evidence to reject H_0

Industry has no effect on resignation rate.

Scanned with Oken Scanner