Probability & Statistics - 2
Assignment - 2

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

b) Gradient of regression line : $\hat{\beta} = \cancel{1.72}\ 1.724197531$

Q.2 (i) $\hat{\beta} = \dfrac{S_{xy}}{S_{xx}}$ & $\hat{\alpha}_i = \bar{y} - \hat{\beta}\bar{x}$

So, $S_{xx} = \Sigma x_i^2 - \dfrac{(\Sigma x_i)^2}{n}$

$= 12666.58 - \dfrac{(385.2)^2}{12}$

$= 301.66$

$S_{yy} = \Sigma y_i^2 - \dfrac{(\Sigma y_i)^2}{n}$

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

$= 6409.7125$

$S_{xy} = \Sigma x_i y_i - \dfrac{\Sigma x_i \cdot \Sigma y_i}{n}$

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

$= 875.16$

$\hat{\alpha}_1 = 3.748181$ & $\hat{\beta} = 2.901146$

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

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

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

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

$= 387.0744$

$\therefore se(\hat{\beta}) = 1.132$

So, $P\left(-2.228 < \dfrac{2.901 - \beta}{1.132} < 2.228\right) = 0.95$

$= P(-228(1.132) - 2.901 < -\beta < 2.228(1.132) - 2.901) = 0.95$

$\therefore$ 95% CI for $\beta = \{2.901 - 2.228(1.132), 2.901 + 2.228(1.132)\}$
$= \{0.378904, 5.423096$

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

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

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

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

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

Page No.
Date

Q.3 (i) Comments on the plot -

- The centre of the distribution ~~is~~ differs for all the 4 cities. Thus, there is a prime face for suggesting that the underlying means are different.

- The difference between the means time taken to commute to office in peak hours & non- peak hours ~~the~~ are in the order city A to city D.

- The variation in the data for city C is largest compared to city D appears to be highest. However with only 7 observation for each city, we can't be sure that there is a real underlying difference in variance

(ii) Assumptions underlying analysis of variance -
- the population must be normal.
- the population have a common variance
- the observation are independent

(iii) ~~H~~ H_0 : the mean of difference is same for each city
H_1 : the mean of difference is not same.

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

$$= 1116.11$$

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

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

ANOVA table

Sources of variation	df	Sum of squares	mean sum of squares
Regression	3	516.11	172.04
Residual	24	600	25
Total	27	1116.11	

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

Test statistic = 6.88

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

Hence, we have sufficient evidence to reject H_0 at LS of 5%
Thus, there are underlying difference between the cities.

(iv) Analysis of mean difference

$\because \bar{y}_1 = 9.14$

$\bar{y}_2 = 6.29$

$\bar{y}_3 = 1.14$

$\bar{y}_4 = -1.81$

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

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

The least significance difference between any of means is

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

$= 5.52$

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

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 the 3 differences are less than 5.52, we will use these pairs to show that there is no significant difference.

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

Examining to ~~is the~~ see if the first 2 groups can be combine,

$\bar{y}_1 - \bar{y}_2 = 8$

There is significant difference between mean 12.3. So, we can't combine the first two groups.

Examining, to see if last 2 groups can be combined,

$\bar{y}_3 - \bar{y}_4 = 8.15$

~~Hen~~

Hence, we can't combine the last 2 groups

$\therefore$ the diagram remains the same as before.

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

$$Y_{ij} = \mu + t_i + e_{ij} \; ; \; i = 1, 2, \ldots, K$$
$$j = 1, 2, \ldots, m_i$$

where, K = no. of treatment

m_i = no. of responses from i^{th} treatment

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

t_i = i^{th} treatment

μ = over mean response

e_{ji} = error term

Assumption → e_{ij} are i.i.d ; $N(0, \sigma^2)$

b) H_0: mean claim amounts of five companies are equal

H_1: mean claim amounts of five companies are not equal.

$m_1 = 7$ $\quad$ $m_4 = 7$

$m_2 = 8$ $\quad$ $m_5 = 5$

$m_3 = 6$ $\quad$ $m = 33$

$$\sum_{i=1}^{5} \sum_{j=1}^{m_i} y_{ij} = 1856 \quad \& \quad \sum_{i=1}^{5} \sum_{j=1}^{m_i} y_{ij}^2 = 174316$$

$$CF = \frac{\left[\sum_{i=1}^{5} \sum_{j=1}^{m} y_{ij} \right]^2}{n} = 104385.94$$

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

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

$$= 22325.69$$

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

Sources of variation	~~d.f~~ d.f	Sum of squares	mean sum of sq.
Regression	4	22326	5581
Residual	28	47604	1700
Total	32	69930	

Test statistic $= \dfrac{MSS_{REG}}{MSS_{RES}}$

$= 3.283$

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

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	27.41772	23.08861	14.43038	11.06329
4-6	15.15189	12.75949	7.97468	6.11392
7-9	14.43038	12.15188	7.59494	5.82278

No clubbing for E_i's is required as all E_i's are greater than 5.

Contingency Table,

H_0: Sales are independent

H_1: Sales are dependent.

$$T.S = \frac{\Sigma (O_i - E_i)^2}{E_i}$$

$$= 49.91146$$

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

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

We have sufficient evidence to reject H_0

Q.5 (i) Assumptions underlying ANOVA.

- Populations are normal
- Populations have common variance
- observation are independent

The sample variance are very ~~difficult from each~~ different from each other.

Hence, the common variance assumption won't hold true.

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

$$= 4.52443$$

Variance at rate 2 for $\log_e(x)$ $= \frac{\Sigma (x_i - \bar{x})^2}{n-1}$

$$= 0.12123$$

(iii) The test was correct asserting that the loge 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) $Y_{ij} = \mu + t_i + e_{ij}$; $i = 1, 2, 3, 4$

$j = 1, 2, 3$

- Y_{ij} is the loge transformed value of the j^{th} observation of the no. of germination per square foot observed when the i^{th} rate was applied.
- μ is the overall popⁿ mean.
- t_i is the deviation of the i^{th} rate mean such that $\sum t_i = 0$
- e_{ij} are the independent error terms → i.i.d $\sim N(0, \sigma^2)$

b) For ANOVA

$H_0: t_i = 0$ vs $H_1: t_i \neq 0$

	ln (x)		
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, $Y_i^2 = \left(\sum_{j=1}^{3} Y_{ij}\right)^2 = (3 \times \text{mean}_i)^2$

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

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

$$= 2 \times 0.618 = 1.236$$

Source of variation	df	SS	MSS
Rate	3	21.153	7.051
Residuals	8	1.236	0.155
	11	22.389	

$$TS = \frac{MSS_{RATE}}{MSS_{RES}} = 45.638$$

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

Given, the observed F statistic value is much larger than this, we can state that the P-value for this test is almost 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 $\Sigma X = 39$; $\Sigma X^2 = 207$

$\Sigma Y = 562$; $\Sigma Y^2 = 40508$

$\Sigma XY = 2853$; $n = 10$

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

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

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

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

$$= 0.94466$$

(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) $SS_{TOT} = \cancel{SS_{xy}}\ S_{yy} = 8923.6$

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

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

(iv) $R^2 = \frac{S^2_{xy}}{S_{xx} \cdot S_{yy}} = \cancel{88}\ 89.24\%$

Comment →

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

Q.7 (i) $\Sigma X = 174.13$, $\Sigma X^2 = 7545.9$

$\Sigma Y = 197.84$, $\Sigma Y^2 = 4747.45$

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

$S_{xx} = 7545.9 - \frac{174.13^2}{11} = 4789.4221$

$S_{yy} = 4747.45 - \frac{197.84^2}{11} = 1189.20767$

$S_{xy} = 2176.84$

$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = 0.45451$ & $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} = 10.78056$

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

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

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

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

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

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

$$= 6.6755 \sim t_9$$

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

$\therefore$ we have sufficient evidence to reject H_0 at 5% LS

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

Assumption →

1) Popⁿ have a common variance

2) Popⁿ follow the normal distribution

3) The observations are independent

(iii) $r = \frac{S_{XY}}{\sqrt{S_{XX} \cdot S_{YY}}}$

$= 0.91213$

(iv) $x_0 = 25$

mean response (μ_0) = P.d $\rightarrow \dfrac{\hat{\mu} - \mu}{S_E(\hat{\mu})} \sim t_{n-2}$

$$S_E(\hat{\mu}) = \sqrt{\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)\hat{\sigma}^2} \qquad [\hat{\mu} = \hat{\alpha} + \hat{\beta} \times 25]$$

$$S_E(\hat{\mu}) = \sqrt{\left(\frac{1}{11} + \frac{(25 - 174.13/11)^2}{4789.4221}\right) \times 22.02136} = 1.55$$

95% CI for the μ_0 = $\hat{\mu} \pm t_{9,\,2.5\%} \times SE(\hat{\mu})$

$= 22.15331 \pm 3.51018$

$= (18.64313,\ 25.66349)$

for individual response,

$\hat{y}_0 = \hat{\alpha} + \hat{\beta} x_0$

$= 22.15331$

$$S.E(\hat{y}_0) = \sqrt{\left[\frac{1}{11} + 1 + \left(\frac{(25 - 174.13/11)^2}{4789.4221}\right)\right] \times 22.02136}$$

$= 4.96079$

$\dfrac{\hat{Y}_0 - Y_0}{S_E(\hat{Y}_0)} \sim t_9$

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 between residuals & the explonatory variable values. Hence, our assumption of normality might be wrong.

∴ The model is not a good ~~fit~~ fit for the data.

Q.8 (i) The main purpose of factor analysis is to reduce many individual items 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 minimising information loss.

While newly identified principle components are choosen to be -
1) Uncorrelated.
2) Linear ~~[illegible]~~ combination of the original variables of the data.
3) which maximise the variance.

(ii)

Principal Component	Diagonal Entries	PCi / sum (PCi over 1 to 5)
PC1	0.456	6.5%
PC2	0.137	19.5%
PC3	0.~~8~~080	11.4%
PC4	0.0165	2.4%
PC5	0.012	1.7%
	0.7015	

(iii) As the 1st 3 P.C explains over 95% of the variance, demonstrating can be reduced to these dataset. The 1st 3 PC can be then used for building further complication on regression modelling purposes.

Page No.
Date

Q9 $\Sigma X = 45$, $\Sigma X^2 = 277.5$
$\Sigma Y = 644$, $\Sigma Y^2 = 43956$
$\Sigma XY = 2602$, $n = 10$

(i) There appears to be a negative correlation between the marks scored & hours spent per day on social media.

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

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

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

$r = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}} = -0.686802$

Hence, we have a weak -ve correlation

(iii) $H_0 : \rho = 0$ vs $H_1 : \rho < 0$

$\tanh^{-1}(r) - \tanh^{-1}(\rho) \sim N\left(0, \frac{1}{n-3}\right)$

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

$= -2.07893$

$Z_{5\%} = 1.6448$

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

Hence we have sufficient evidence to reject H_0

$\therefore \rho < 0$

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

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

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

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

Hence, model explain only 47% of the veriability.

∴ The model is not a good fit for the data.

(vi) $x_0 = 1$

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

$= -3.9467 \times 1$

$= -3.9467$

∴ For every additional 1 hr spent on social media, the expected decrease in marks is 3.9467

Q.10. $\Sigma x = 0.93$, $\Sigma x^2 = 0.2925$

$\Sigma y = 0.23$, $\Sigma y^2 = 0.0189$

$\Sigma xy = 0.0735$, $n = 3$

$S_{xx} = 0.2925 - \frac{0.93^2}{3} = 0.0042$

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

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

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

$$SS_{REG} = \frac{S^2_{XY}}{S_{XX}} = 0.00115$$

$SS_{RES} = 0.00011$

ANOVA table →

Source of variation	df	SS	MSS
Regression	1	0.00115	0.00115
Residual	1	0.00011	0.00011
Total	2	0.00126	

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

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

$$= 10.4545 \sim F_{1,1}$$

$F_{1,1,5\%} = 161.4$;

$F_{cal} < F_{tab}$

∴ we do not have sufficient evidence to reject H_0

∴ Industry has no effect on ~~registration~~ resignation rate.