PROBABLITY AND STATISTICS ASSIGNMENT

1] a) Let regressions line y on x be.

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

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

$\Sigma xy = 182560$

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

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

$\Sigma x^2 = 92500$

$$\hat{\beta} = \frac{182560 - (10 \times 85 \times 173.7)}{92500 - (10 \times 85^2)}$$
$= 1.7292$

$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$
$= 27.143$

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

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

2].) Least sq. fit regression line is given by

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

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{28141.41 - (12 \times 32.1 \times 96.875)}{12666.58 - (12 \times 32.1^2)}$$

$$= 2.9011$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$
$$= 96.875 - (2.9011 - 32.1)$$
$$= 3.7497$$

$$\hat{y} = 3.7497 + 2.9011x$$

ii) Test Statistic : $\dfrac{\hat{\beta} - \beta}{\sqrt{\hat{\sigma}^2 / S_{xx}}} \sim t_{n-2}$

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

$$= 387.0745$$

95% CI : $\hat{\beta} \pm t_{0.025,10} \sqrt{\hat{\sigma}^2 / S_{xx}}$

$$= 2.9011 \pm \left(2.228 \times \sqrt{\frac{387.0745}{301.60}}\right)$$

$$= (0.3773, 5.4274)$$

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

Test Stadisdic:

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

95% CI for $y_0 = (\hat{\alpha} + \hat{\beta} x_0) \pm t_{2.5\%, 10} \sqrt{(s.e(y_0))^2}$

$$s.e = \sqrt{387.0775\left(\frac{1}{12} + \frac{(70.37.1)^2}{301.60}\right)}$$

$= 10.5989$

95% CI: $3.7797 + (2.9011 \times 48) \pm (2.228 \times 10.5989)$
$= (96.179, 143.408)$

3] i) The plot suggests that:
The time difference in peak hours and non-peak hours is maximum in City A and least in City D

The variation in city P is least as compared to ~~B~~ D which has maxim variance

ii) Assumptions in ANOVA:
- The data follows a Normal distribution
- The sample items must be independent
- The ~~pop~~ population should have a common variance.

iii) H_0: Mean of difference is same for each
H_1: Mean of difference is not same

$SS_{Tot} = \Sigma y^2 - n\bar{y}^2$

$= 1405 - \left(28 \times \left(\frac{103^2}{28^2}\right)\right)$

$= 1116.11$

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

$= 516.11$

$SS_{Res} = SS_{Tot} - SS_{Reg}$

$= 600$

ANOVA TABLE

Source of Variation	d.f	Sum of Sq.	MSS
Regression	3	516.11	172.04
Residual	24	600	25
Total	27	1116.11	

TS : $\frac{MSS_{Reg}}{MSS_{Res}} \sim F_{3,24}$

$F_{cal} = \frac{MSS_{Reg}}{MSS_{Res}}$

$= 6.88$

$F_{tab} : 3.009$

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

4](a) Mathem model of One-way ANOVA is given by:

$y_{ij} : \mu + T_i + \varepsilon_{ij}$

y_{ij} - j^{th} obs. ith ~~experiment~~ treatment
$\mu \rightarrow$ common effect for the whole experiment
$T_i \rightarrow$ i^{th} treatment effect
ε_{ij} - random error in the j^{th} obs.

Assumptions:
- Errors ~~as~~ are normally distributed with mean 0 and constant variance σ^2.
- μ: fixed parameter

b) Ho: Means claims are equal
Hy: Mean claims are not equal

$m_1 = 7$ $m_2 : 8$ $m_3 : 6$ $m_4 : 7$ $m_5 : 5$
$m_{tot} = 33$

$$\sum_{i=1}^{5} \sum_{j=1}^{m_i} y_{ij} : 354 + 386 + 07 + 045 + 3889$$
$$= 1836$$

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

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

$$= 69930.06$$

$$SS_{Reg} = \left[\frac{354^2}{7} + \frac{380^2}{8} + \frac{87^2}{6} + \frac{645^2}{7} + \frac{304^2}{3}\right] - \left[33 \times \left(\frac{1856}{33}\right)^2\right]$$

$$= \cancel{2992.} \rightarrow 22325.69$$

$$SS_{Res} = SS_{Tot} - SS_{Reg}$$
$$= 47604.37$$

Anova Table:

Source of variation	d.f	SS	MSS
Regression	4	22325.69	5581.4225
Residual	28	47604.37	1700.156
Total	32	69930.06	

TS: $\frac{MSS_{Reg}}{MSS_{Res}} \sim F_{4,28}$

$F_{cal} = 3.2829$
$F_{tab} = 2.714$

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

c] H_0: No association between salary and no. of papers cleared
H_1: Interdependence exists

O_i:

	3-5	5-8	8-10	10-12	Tot
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	14.43078	11.06329
4-6	15.15189	12.75949	7.97468	6.11392
7-9	14.43031	12.15189	7.59494	5.82271

$TS = \Sigma \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{(3-1)(4-1)}$

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

$\chi^2_{tab} = 12.59$

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

5] i) Assumptions:
- Data should be normally distributed
- Sample items are independent
- Pop^n should have constant variance

ii) Missing Values are:

$$\text{Mean} = \frac{\sqrt{29} + \sqrt{13} + \sqrt{21}}{3}$$

$$= 4.5244$$

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

$$= 0.1213$$

iii) $\log_e$ transformation is best suited as the common var. assumption holds true in this case.

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

$H_0: \tau = 0$
$H_1: \tau \neq 0$

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

$$SS_{Reg} = \sum_{i=1}^{4} y_i^2/3 - y^2/12$$

$$= 21.153$$

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

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

$$= 1.230$$

Source of variation	d.f	SS	MSS
Regression	3	21.153	7.051
Residual	8	1.236	0.155
Total	11	22.316	

TS: $\frac{MSS_{Reg}}{MSS_{Res}} \sim F_{3,8}$

$F_{cal} = 45.638$
$F_{tab} = 4.060$

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

6] i) Correlation coef $= \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}$

$$= \frac{661.2}{\sqrt{54.9 \times 8923.6}}$$

$$= 0.9447$$

ii) Fitted linear reg. eq^n is given by:

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

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{661.9}{54.9} = 12.0437$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$
$$= 9.29951$$

$$\hat{y} = 9.27957 + 12.0437x$$

iii) $SS_{Tot} = S_{yy} = \Sigma y^2 - n\bar{y}^2$
$= 8993.6$

$$SS_{Reg} = \frac{S^2_{xy}}{S_{xy}} = \frac{661.9^2}{54.9}$$
$= 7963.3$

$SS_{Res} = SS_{Tot} - SS_{Reg}$
$= 960.3.$

iv) $R^2 = \frac{SS_{Reg}}{SS_{Tot}} = 0.8924$

7] i) Linear Regression eq:

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

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

$$= 0.4575$$

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

$$= 10.7907$$

$$\hat{y} = 10.7907 + 0.4575x$$

ii) $H_0: \beta = 0$

$H_1: \beta \neq 1$

$$TS: \frac{\hat{\beta} - \beta}{s.e(\hat{\beta})} \sim t_{n-2}$$

$$s.e = \sqrt{\hat{\sigma}^2 / S_{xx}} = \sqrt{\frac{1}{(n-2)}\left(S_{yy} - \frac{S_{xy}^2}{S_{xx}}\right)}$$

$$= 0.0681$$

$$t_{cal} = \frac{0.4575 - 0}{0.0681} = 6.677$$

$$t_{tab} = 3.262$$

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

iii) Corr. coeff $= \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}$

$= 0.9121$

iv) Individual response

$Y_0 \sim N(\hat{\alpha} + \hat{\beta} x_0, \sigma^2 (1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}))$

$TS: \frac{Y_0 - (\hat{\alpha} + \hat{\beta} x_0)}{\hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim t_{n-2}$

$95\%\ CI: Y_0 \pm t_{n-2, 2.5\%} \sqrt{\hat{\sigma}^2 (1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}})}$

$= 22.1532 \pm 2.262 \sqrt{22.2017 (1 + \frac{1}{11} + 0.01780)}$

$= (10.9319, 33.3745)$

Mean Response:

$95\%\ CI: \hat{\alpha} + \hat{\beta} x_0 \pm t_{n-2, 2.5\%} \sqrt{\hat{\sigma}^2 (\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}})}$

$= (18.643, 25.663)$

v) The residual plot appears to follow a pattern hence the model isn't a good fit

8] ii) Variance - Cov matrix of PC1, PC2, PC3, PC4 and PC5 is given by:

$$\begin{bmatrix} 0.458 & 0 & 0 & 0 & 0 \\ 0 & 0.137 & 0 & 0 & 0 \\ 0 & 0 & 0.080 & 0 & 0 \\ 0 & 0 & 0 & 0.0165 & 0 \\ 0 & 0 & 0 & 0 & 0.012 \end{bmatrix}$$

% of total variance explained by each PC:

$PC1 : \frac{0.456}{0.7015} \times 100 = 65\%$

$PC2 : \frac{0.137}{0.7015} \times 100 = 19.53\%$

$PC3 = \frac{0.080}{0.7015} \times 100 = 11.40\%$

$PC4 = \frac{0.0165}{0.7015} \times 100 = 2.35\%$

$PC5 : \frac{0.012}{0.7015} \times 100 : 1.71\%$

iii) we can conclude that

9] i) There is a -ve linear correlation between marks and hrs ~~pe~~ spent / day on social media

ii) $r = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}$

$= \frac{-246}{\sqrt{75 \times 2782.4}}$

$= -0.6860$

iii) $H_0 : \rho = 0$

$H_1 : \rho \neq 0$

Using Fisher's Transformation:

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

$Z_{cal} : \frac{-0.84036 - 0}{\sqrt{1/7}} = -2.2137$

$Z_{tab} = -1.96$

Since $Z_{cal} < Z_{tab}$ we reject H_0

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

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

$= -3.9467$

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

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

$= 82.16$

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

v) ~~Corr~~ $R^2 = \frac{S_{xy}^2}{S_{xx} S_{yy}}$

$= \frac{(-296)^2}{75 \times 2487.4}$

$= 0.4706$

$r^2 = R^2$

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

$= -3.9467$

10] H_0: Industry has no impact on resignation rat
H_1: There is an impact

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

$SS_{reg} = 20[(27-31)^2 + (36-31)^2 + (30-31)^2] = 840$

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

Source of Variation	d.f	SS	MSS
Regression	2	8840	420
Residual	57	3591	63

TS: $\frac{MSS_{Reg}}{MSS_{Res}} \sim F_{2,57}$

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

$F_{tab} = F_{2\,60,\,5\%} = 3.150$

Since $F_{cal} > F_{tab}$, we reject H_0
i.e there is an impact.