PROBABILITY & STATISTICS-2

Assignment-2

1.

(a) Let regression line y on x be:

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

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

$\sum xy = 182560$

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

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

$\sum x^2 = 92500$

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

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

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

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

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

2.

(i) Least square fit regression line is given by:

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

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{38191.41 - (12 \times 32.1 \times 96.875)}{12666.58 - (12 \times 32.1^2)}$$

$$= \frac{875.16}{301.66}$$

$$= 2.9011$$

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

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

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

$$\hat{\sigma}^2 = \frac{S_{yy} - \frac{S_{xy}}{S_{xx}}}{n-2} = \frac{6409.7125 - \frac{(875.16)^2}{301.66}}{10}$$
$$= 387.0745$$

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

$$= 2.9011 \pm \left(2.228 \times \sqrt{\frac{387.0745}{301.66}}\right)$$

$= 2.9011 \pm 2.5238$

Ans $= (0.3773, 5.4249)$

iii ~~Test statistic~~:

$\mu_0 \sim N\left[\hat{\alpha} + \hat{\beta}x_0, \sigma^2\left[\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right]\right]$

Test Statistic:

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

95% CI for $\mu_0 = (\hat{\alpha} + \hat{\beta}x_0) \pm t_{0.025,10}\sqrt{(se(\mu_0))^2}$

$se(\mu_0) = \sqrt{387.0745\left(\frac{1}{12} + \frac{(40-32.1)^2}{301.66}\right)}$

$se(\mu_0) = 10.5989$

→ 95% CI $= 3.7497 + (2.9011 \times 40) \pm (2.228 \times 10.5989)$

Ans $= (96.179, 143.408)$

3.

(i) The plot suggests that:

→ The time difference in commutation ~~it~~ in peak hours and non-peak hours is ~~most~~ maximum in City A and least in city D

→ The variation in city C is least as compared to D which has the maximum variance.

(ii) Assumptions to carry out ANOVA:

→ The data should conform to a normal distribution
→ The sample items must be independent of each other
→ The populations should have a common variance.

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

$$SS_{Tot} = \Sigma y^2 - n\bar{y}^2$$
$$= 1495 - \left[28 \times \left(\frac{103^2}{28^2}\right)\right]$$
$$= 1116.11$$

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

$$SS_{Res} = SS_{Tot} - SS_{Reg}$$
$$= 1116.11 - 516.11$$
$$= 600$$

ANOVA TABLE

Source of Variation	Dof	Sum of sq	MSS
Regression	3	516.11	172.04
Residual	24	600	25
Total	27	1116.11	

TS :

$$\frac{MSS_{Reg}}{MSS_{res}} \sim F_{3,24}$$

$$F_{cal} = \frac{MSS_{Reg}}{MSS_{res}} = \frac{172.04}{25} = 6.88$$

$$F_{tab} = F_{3,25} = 3.009$$

Since $F_{cal} > F_{tab}$, we reject H_0 at the 5% level. Thus we can conclude that the mean of difference in lines is not same.

4.

(a) Mathematical model of one-way ANOV is given by:

$y_{ij} = \mu + \tau_i + \varepsilon_{ij}$; where

$y_{ij} \rightarrow$ j^{th} observation on the i-th treatment
$\mu \rightarrow$ common effect for the whole experiment
$\tau_i \rightarrow$ i^{th} treatment effect
$\varepsilon_{ij} \rightarrow$ random error in the j^{th} observation on the i^{th} treatment

Assumptions:
→ Errors are assumed to be normally and independently distributed with mean 0 and constant variance σ^2

→ μ : fixed parameter

(b) H_0: Mean Claims are equal
H_1: Mean Claims are not equal

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

$\Rightarrow m_{tot} = 33$

$$\sum_{i=1}^{5}\sum_{j=1}^{m_i} y_{ij} = \cancel{18}\ 354 + 386 + 87 + 645 + 384 = 1856$$

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

$$SS_{Tot} = 174316 - \left[\cancel{36}\ 33 \times \left(\frac{1856}{33}\right)^2\right]$$

$$= 69930.06$$

$$SS_{reg} = \left[\frac{354^2}{7} + \frac{386^2}{8} + \frac{87^2}{6} + \frac{645^2}{7} + \frac{384^2}{3}\right] - \left[33 \times \left(\frac{1856}{33}\right)^2\right]$$

$$= 22325.69$$

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

Source of Variation	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 ~~at~~ 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: $\sum \frac{(O_i - E_i)^2}{\sum E_i} \sim \chi^2_{(3-1)(2-1)}$

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

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

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

→ Salaries and no. of papers cleared are associated.

~~5.~~
~~(i) Assumption:~~
~~Data are normally distributed with mean 0, and a constant variance σ^2.~~

~~But this is not the case as the~~

5.(i) Assumptions:
Data should be normally distributed
Sample items must be independent
Population should have a common variance.

But this is not the case as the data items do not have a common variance.

(ii) Missing values are:

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

$\rightarrow Var_2 = \cancel{\sum [\log_e(x)]^2 - n\,Mean^2} \; \dfrac{\sum (x_i - \bar{x})^2}{n-1}$

$= \cancel{70.135 - (3 \times 4.827^2)}$

$= \cancel{\boxed{0.2352}} \; \boxed{0.1213}$

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

(iv)

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

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

(b) $y_i^2 = \sum y_{ij}^2 = (3 \times Mean_i)^2 = 973.373$

$SS_{reg} = \sum_{i=1}^{4} \dfrac{y_i^2}{3} - \dfrac{y^2}{12} = 21.153$

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

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

$= 2 \times 0.618$

$= 1.236$

Source of Variation	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 $(r) = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{\Sigma xy - n\bar{x}\bar{y}}{\sqrt{(\Sigma x^2 - n\bar{x}^2)(\Sigma y^2 - n\bar{y}^2)}}$

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

$= \underline{\underline{0.9447}}$

(ii) Fitted linear regression eqn is given by:

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

$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{661.2}{54.9} = 12.0437$

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

$= 9.22957$

Ans $\hat{y} = 9.22957 + 12.0437x$

(iii) $SS_{tot} = S_{yy} = \sum y^2 - n\bar{y}^2$

$= 40508 - (10 \times 56.2^2)$

$= 8923.6$

$SS_{Reg} = \frac{S^2_{xy}}{S_{xx}} = \frac{661.2^2}{54.9} = 7963.3$

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

(iv) $R^2 = \frac{SS_{Reg}}{SS_{Tot}} = \frac{7963.3}{8923.6} = \underline{0.8924}$

→ Using part (i), we have

$r^2 = R^2$

→ The correlation coeff is the square root of the coeff. of determination

7.

(i) Linear regression eqn:

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

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

$= \frac{2176.84}{7545.9 - (11 \times 250.5889)}$

$= 0.4545$

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

$= 10.7907$

ans $\hat{y} = 10.7907 + 0.4545x$

classmate

Date
Page

ii. $H_0 : B = 0$

$H_1 : B \neq 0$

TS: $\dfrac{\hat{B} - B}{se(\hat{B})} \sim t_{n-2}$

$$se(\hat{B}) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} = \sqrt{\frac{SS_{res}}{(n-2) S_{xx}}} = \sqrt{\frac{S_{yy} - \frac{S^2_{xy}}{S_{xx}}}{(n-2) S_{xx}}}$$

$$= \sqrt{\frac{1189.208 - \frac{2176.84^2}{4789.4221}}{9 \times 4789.4221}}$$

$$= 0.0681$$

$$t_{cal} = \frac{0.4545 - 0}{0.0681} = 6.674$$

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

$\Rightarrow$ Since $t_{cal} > t_{tab}$, we reject H_0. There is a linear relationship between X & Y.

(iii) Corr. coeff $(r) = \dfrac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \dfrac{2176.84}{\sqrt{4789.4221 \times 1189.208}}$

$$= \underline{0.9121}$$

iv Individual response:

~~TS~~ $Y_0 \sim N\left[\hat{\alpha} + \hat{B} X_0, \sigma^2 \left[1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{xx}}\right]\right]$

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

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

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

$$= 22.1532 \pm 2.262\sqrt{22.2014\left(1 + \frac{1}{11} + 0.01756\right)}$$

Ans $$= (10.9319, 33.3745)$$

Mean Response:

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

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

$$= 22.1532 \pm 2.262\sqrt{22.2014\left(\frac{1}{11} + 0.01756\right)}$$

Ans $$= (18.643, 25.663)$$

(v) The residual plot appears to follow a pattern (first increasing and then decreasing), which implies that the model isn't a good fit for the data.

8.

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

→ ~~when~~ the original data is rewritten in terms of new variables that are uncorrelated ; only some are needed to explain most of the variability observed in the data.
These 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 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 the dimentionality could be reduced 3, using the first 3 PC's, as 96% of the variability has been explained by these 3 alone.

4.

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

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

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

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

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

(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 $(R^2) = \frac{S_{xy}^2}{S_{xx}S_{yy}} = \frac{(-296)^2}{75 \times 2482.4}$

$= \underline{0.4706}$

→ $r^2 = R^2$

vi. Expected change in marks $= \beta \Delta x$
$= \underline{-3.9467}$

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

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

$SS_{reg} = 20[(27-31)^2 + (36-31)^2 + (30-31)^2] = 840$
$SS_{res} = 19(5^2 + 10^2 + 8^2) = 3591$

ANOVA TABLE

Source of Variation	Dof	Sum of Sq	MSS
Residual	57	3591	63
Regression	2	840	420

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

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

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

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