Prob And Stats Assignment 2 :

1. a) Let the regression line of y on x be :

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

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

$\Sigma xy = 182560$

$\bar{x} = \frac{\Sigma x_i}{n} = 850$

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

$\Sigma x^2 = 92500$

$$\hat{\beta} = \frac{182560 - (850 \times 173.7)}{92500 - (10 \times 85^2)}$$

$\boxed{\hat{\beta} = 1.72421}$

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

$\boxed{\hat{\alpha} = 27.143}$

$\therefore \hat{y} = 27.143 + 1.72421x$

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

2. (i) Least sq. fit regression is given by:

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

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

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

$$\boxed{\hat{\beta} = 2.9011}$$

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

$$= 96.875 - (2.9011 \times 32.1)$$

$$\boxed{\hat{\alpha} = 3.7497}$$

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

(ii) Test statistic $\dfrac{\hat{\beta} - \beta}{\sqrt{\hat{\sigma}^2}} \sim t_{n-2}$

$$\hat{\sigma}^2 = \frac{S_{yy} - \frac{S_{xy}^2}{S_{xx}}}{n-2}$$

$$\hat{\sigma}^2 = \frac{6409.7125 - \frac{(875.16)^2}{301.66}}{10}$$

$$\hat{\sigma}^2 = 387.0745$$

$$95\%\ CI = \hat{\beta} \pm t_{0.025,10}\sqrt{\frac{387.0745}{301.66}}$$

$$CI \Rightarrow (0.3773, 5.4249)$$

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

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

95% CI for $y_0$ = $(\hat{\alpha} + \hat{\beta} x_0) \pm t_{0.025,10} \sqrt{se(y_0)}$

$se(y_0) = \sqrt{387.0745\left(\frac{1}{12} + \frac{(40 - 32.1)^2}{301.66}\right)}$

$se(y_0) = 10.5989$

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

80ln. = $(96.179, 143.408)$

3.

(i) The plot suggests that:

→ The difference in cumiliation in peak curve and non peak curve is maximum in city A and least in city D.

→ The variation in city C is lesser than city D which has maximum variable.

(ii) Assumptions to carry out ANOVA:

→ The data should confer 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 for each city.

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

$$= 1495 - \left[28 \times \left(\frac{103^2}{28^2}\right)\right]$$

$$SS_{Tot} = 1116.111$$

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

$$SS_{Reg} = 516.11$$

$$SS_{Res} = 1116.11 - 516.11$$

$$SS_{Res} = 600$$

ANOVA TABLE

| Source of variation | DOf | Sum of sq | M |
|---|---|---|---|
| Regression | 3 | 516.11 | 172 |
| Residual | 24 | 600 | 25 |
| Total | 27 | 1116.11 | 19 |

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

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

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

$F_{cal} > F$

4. Mathematical model of one-way ANOVA is given by:

a) $y_{ij} = \mu + T_i + \varepsilon_{ij}$, where

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

Assumptions:
→ Errors are assumed to be normally and independently distributed with the mean 0 and constant variance $\sigma^2$.

→ $\mu \rightarrow$ fixed parameter

b) $H_0$: mean claims are equal
$H_1$: mean claims are not equal.

$m_1 = 7$ $m_2 = 8$ $m_3 = 6$ $m_4 = 7$
$m_5 = 5$ $\therefore m_{Tot} = 33$

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

$$\sum_{i=1}^{5} \sum_{j=1}^{m_j} y_{ij}^2 = 27184 + 29430 + 2123 + 84669 + 3091\ldots$$
$$= 174316$$

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

$$\boxed{SS_{Tot} = 69930.06}$$

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

$$= \underline{22325.69}$$

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

ANOVA Table :

| Sources of Variation | Dof | Sum of Sq | MSS |
|---|---|---|---|
| Reg | 4 | 22325.69 | 5581.4225 |
| Res | 28 | 47604.37 | 1700.156 |
| Total | 32 | 69930.06 | |

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

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

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

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

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

$H_1$ : There is association / interdependence

| paper \ 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 |

$\varepsilon_i$

| | 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 | 7.5694 | 5.82278 |

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

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

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

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

→ Salaries & no. of papers are clearly associated

5.

(i) Assumptions:

Data should be normally distributed
sample items will be independent
Populations should be having a common variance.

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

(ii) Missing values are:

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

$var_2 \rightarrow \frac{\sum(x_i - \bar{x})^2}{n-1}$

$= 0.1213$

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

(iv)

a) $y_{ij} = \mu + T_i + \epsilon_{ij}$, $i = 1, 2, 3 \ldots$
$j = 1, 2, 3 \ldots$

$H_0 : T_i = 0$

$H_1 : T_i \neq 0$

6.
(i) Correlation coeff $(r) = \dfrac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \dfrac{\Sigma xy - n\bar{x}\bar{y}}{\sqrt{(\Sigma x^2 - n\bar{x}^2)(\Sigma y^2 - n\bar{y}^2)}}$

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

$$= \underline{0.9447}$$

(ii) fitted linear reg. 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) $SS_{Tot} = S_{yy} = \Sigma y^2 - n\bar{y}^2$

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

$$= \underline{8923.6}$$

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

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

Page No.
Date

(iv) $R^2 = \dfrac{SSReg}{SSTot} = \dfrac{7963.3}{8923.6} = 0.8924$

→ using part (i) we have
$r^2 = R^2$

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

7. (i) Linear Reg. eqn. $\hat{y} = \hat{\alpha} + \hat{\beta}x$

$$\hat{\beta} = \frac{Sxy}{Sxx} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma x^2 - n\bar{x}}$$

$= 0.4545$

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

$= 10.7907$

$\therefore \quad \hat{y} = 10.7907 + 0.4545x$

(ii) $H_0 : \beta = 0$

$H_1 : \beta \neq 0$

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

$$se(\hat{\beta}) = \sqrt{\frac{\hat{\sigma}^2}{Sxx}} = \sqrt{\frac{SSres}{Sxx\,(n-2)}}$$

$$= \sqrt{\frac{S_{yy} - \frac{S^2_{xy}}{S_{xx}}}{(n-2) S_{xx}}} = \sqrt{\frac{1189.208 - \frac{2176.84^2}{4789.4221}}{9 \times 4789.4221}}$$

$$t_{cal} = 0.0681$$

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

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

(iii) correlation coeff. $(r) = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}} = 0.9121$

(iv) Individual response:

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

95% CI

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

$$= (10.9319, 33.3745)$$

Mean Response:

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

$$= (18.643, 25.663)$$

v) first ↑ then ↓ ∴ model is a good fit.

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

var - cov. matrix :

$$\begin{bmatrix} 0.456 & 0 & 0 & 0 & 0 \\ 0 & 0.137 & 0 & 0 & 0 \\ 0 & 0 & 0.08 & 0 & 0 \\ 0 & 0 & 0 & 0.0165 & 0 \\ 0 & 0 & 0 & 0 & 0.012 \end{bmatrix}$$

% of total var. explained by each PC

$PC_1 = \frac{0.465}{0.7015} \times 100 = 65\%$

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

$PC_3 = 11.4\%$

$PC_4 = 2.35\%$

$PC_5 = 1.71\%$

9 (i) Based on the plot, there is a negative linear correlation

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

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

$$= -0.6860$$

(iii) $H_0 : \rho = 0$

$H_1 : \rho < 0$

Using Fisher's Transformation:

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

$Z_{cal} = -2.22348$

$Z_{tab} = -1.98$

$Z_{cal} < Z_{tab}$, we reject $H_0$.

(iv) Let the linear reg. model be:

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

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

$$\hat{y} = 8.216 - 3.9467x$$

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

$= -3.9467$

10. $H_0$ = Inclusive has no impact

$H_1$ = has impact

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

SS reg = 840

SS res = 3591

ANOVA :-

| | Dof | Sum of sq. | MSS |
|---|---|---|---|
| Res. | 57 | 3591 | 63 |
| Reg. | 2 | 840 | 420 |
| Tot. | 59 | | |

$TS = \frac{\cancel{3591}}{\cancel{840}} = \frac{420}{63} = 6.67$

F cal = 6.67

$F_{2,57} \rightarrow$ Ftab = 3.150

Fcal > Ftab, we reject $H_0$.