P & S (II)

Date / /
Page

## Assignment - 2

(1) a) let regression line y on x be

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

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

$\Rightarrow \sum xy = 182560$

$\bar{x} = \frac{\sum x_i}{n} = 850$ ; $\bar{y} = \frac{\sum y_i}{n} = 173.7$

$\Rightarrow \sum x^2 = 92500$

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

$\Rightarrow \hat{\beta} = 1.7242$

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

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

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

(2) (i) Here, the least square fit regression line is : $\hat{y} = \hat{\alpha} + \hat{\beta}x$.

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

$$= \frac{875.16}{301.66} = 2.9011$$

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

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

$$= \underline{\underline{3.7497}}$$

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

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

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

On putting values, we get

$$\boxed{\hat{\sigma}^2 = 387.075}$$

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

$$= 2.9011 \pm 2.228 \times \sqrt{\frac{387.075}{301.66}}$$

$$= (0.3773, 5.4249)$$

Date / /
Page

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

Test Statistic → 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 for $\mu_0 = \hat{\alpha} + \hat{\beta} x_0 \pm t_{0.025,10} \sqrt{(se(\mu_0))^2}$

$= se(\mu_0) = \sqrt{387.075 \left\{\frac{1}{12} + \frac{(40 - 32.1)^2}{301.66}\right\}}$

$= \left\{ se(\mu_0) = 10.3989 \right\}$

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

$= (96.178, 143.408)$

(3) (i) The plot suggests that :-
the time differences in commutation in peak hours and non peak hours is minimum in city D and max in city A. whereas the variation for city C is the least as compared to D with max. variance.

(ii) Assumptions for ANOVA :-
→ The data should conform to a Normal distribution
→ The sample items must be independent of each other

Date / /
Page

→ The population should have a common variance

(iii) $H_0$: Mean of difference is same for each city.
$H_1$: Mean is different

= $SS_{TOT} = SS_{yy} = \sum y^2 - n\bar{y}^2$
$= 1495 - \left(28 \times \frac{103^2}{28^2}\right)$
$= \underline{\underline{1116.11}}$ — (1)

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

= ∴ From (1) & (2).

= $SS_{res} = SS_{TOT} - SS_{reg}$
= $SS_{res} = 1116.11 - 516.11$
$\boxed{SS_{res} = 600}$

ANOVA Table

| Source of variation | Dof | Sum of Sq | MSS |
|---|---|---|---|
| Regression | 3 | 516.11 | 172.0 |
| Residual | 24 | 600 | 25 |
| Total | 27 | 1116.11 | |

Date / /
Page

TS :

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

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

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

$= \therefore$ $F_{cal} > F_{tab}$, so we reject $H_0$ at 5% significance level. Thus we can conclude that the mean of difference in hence $\mu$ is not same

④ a) Mathematical Model of one-way ANOVA given by :-

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

$y_{ij}$ = $j^{th}$ observation on the ith treatment

$\mu$ = common effect for the whole experiment

$T_i$ = ith treatment effect

$\varepsilon_{ij}$ = random error in the jth observation on the ith treatment

Assumptions →

1. Errors are assumed to be Normally and independently distribution with mean 0 and constant variance $\sigma^2$
2. $\mu$ is a fixed parameter

(b) $H_0$: Mean claims are equal
$H_1$: Mean claims are unequal

$M_1 = 7$ $\quad M_3 = 6$ $\quad M_5 = 5$

$M_2 = 8$ $\quad M_4 = 7$ $\quad M_{TOT} = 33$

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

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

$$SS_{Tot} = 174316 - \left[\frac{\cancel{33} \times 1856 \times 1856}{33 \times \cancel{33}}\right]$$

$$= 69930.06$$

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

$$= \underline{22325.69}$$

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

$$= 47604.37$$

| Source of variation | DoF | Sum of Sq | MSS |
|---|---|---|---|
| Regression | 4 | 22325.69 | 5581.4227 |
| Residual | 28 | 47604.37 | 1700.156 |
| Total | 32 | 69930.06 | |

Date: / /
Page ____

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 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 | 27.4177 | 23.08861 | 14.43038 | 11.06329 |
| 4-6 | 15.15189 | 12.75949 | 7.97468 | 6.11392 |
| 7-9 | 14.43058 | 12.151~~6~~89 | 7.59494 | 5.82278 |

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

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

$\Rightarrow \chi^2_{tab} = \chi^2_{1,5\%} = 12.59$

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

$\therefore$ Salaries and no. of papers cleaned are associated.

(5) i) Assumptions :-

Data should be Normally distributed
sample items must be independent
populations should have common variance.

ii) Missing values are

$$\text{Mean} \rightarrow \frac{\sqrt{29} + \sqrt{13} + \sqrt{24}}{3}$$

$$= 4.5944,$$

$$Var_n = \frac{\Sigma (X_i - \bar{x})^2}{n-1}$$

$$= 0.1213$$

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

(iv) a) $y_{ij} = \mu + T_i + \varepsilon_{ij}$ , $i = j = 1, 2, 3 \cdots$

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

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

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

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

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

$= 2 \times 0.618$

$= 1.236$

| Source of variation | D.o.f | 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 $H_0$

Date / /
Page

(5) i) Correlation coeff (r) = $\frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}$

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

$$= \frac{661.2}{\sqrt{34.9 \times 8923.6}}$$

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

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

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

$$= \underline{\underline{8923.6}}$$

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

Date / /
Page

$SS_{res} = SS_{Tot} - SS_{reg} = 960.3$

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

→ 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 equation :-

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

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{2176.84}{\cancel{7545.9 - (11 \times 250.5887)}}$$

$$= 0.4545$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$
$$= 10.7907$$
$$\hat{y} = 10.7907 + 0.4545x$$

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

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

$$se(\hat{\beta}) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} = \sqrt{\frac{SS_{res}}{(n-2)S_{xx}}} = \sqrt{\frac{1189.208 - \frac{2176.84^2}{4789.422}}{9 \times 4789.4221}}$$

$$= 0.0681$$

$t_{cal} = \frac{0.4545}{0.0681} = 6.674.$

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

Since $t_{cal} > t_{tab}$, we reject $H_0$, There is a linear relationship b/w $x$ & $y$

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

$= \frac{2176.84}{\sqrt{4789.4221 \times 1184.208}}$

$= 0.9121$

(iv) Individual Response :-

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

$\rightarrow$ T.S. $= \frac{y_0 - (\hat{\alpha} + \hat{\beta} x_0)}{\hat{\sigma}\sqrt{\frac{1}{n} + 1 + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim t_{n-2}$

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

$= (10.9319, 33.3745)$

Mean Response $\rightarrow$

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

Date / /
Page

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

$$= (18.643, 25.663)$$

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

③ (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 understa-nd data.

→ 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

$$\begin{bmatrix} 0.436 & 0 & 0 & 0 & 0 \\ 0 & 0.137 & 0 & 0 & 0 \\ 0 & 0 & 0.08 & 0 & 0 \\ 0 & 0 & 0 & 0.0165 & 0 \\ 0 & 0 & 0 & 0 & 0.012 \end{bmatrix}$$

% of total variance explained by each PC

Date / /
Page

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

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

$PC3 = \frac{0.08 \times 100}{0.7015} = 11.40\%$

$PC4 = \frac{0.018 \times 100}{0.7015} = 2.35\%$

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

(iii) We can conclude that the dimensionali could be reduced 3, using the first 3PC as 96% of variability has been explain by these 3 alone

(9) (i) Based on the plot, there is a negative linear correlation b/w marks and hours spent daily on social Media

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

$= \frac{-290}{\sqrt{75 \times 2428.4}}$

$= -0.686$

(iii) $H_0: \rho = 0$

$H_1: \rho < 0$

Using fisher's transformation

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

Date / /
Page

$= Z_{cal} = -0.84036\sqrt{7} = -2.2234$

$= Z_{tab} = -1.98$

→ Since $Z_{cal} < Z_{tab}$; we reject $H_0$. there is negative correlation b/w the variables

(iv) Let the linear regression model:

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

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

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$
$$= 64.4 - (-3.9467)(4.5)$$
$$= 82.16$$

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

(v) Coefficient of determination $(R^2) = \frac{S^2xy}{Sxx \cdot Syy} = \frac{(-296)^2}{75 \times 2482.4}$

$= 0.4706$

→ $r^2 = R^2$

(vi) Expected change in marks $= B\Delta x$
$= -3.9467$

(10) $H_0$: Inclusion has no impact on resignation rate
$H_1$: There is an impact

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

Date / /
Page

$SS_{Reg} = 20\left[(27-31)^2 + (36-31)^2 + (30-31)^2\right] = 840$

$SS_{Res} = 19\,(5^2 + 10^2 + 8^2) = 3591$

ANOVA TABLE

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

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

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

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

→ As $F_{cal} > F_{tab}$ we reject $H_0$, there is an impact of industry on the resignation ratio.