A50

PROBABILITY & STATISTICS - 2.

ASSIGNMENT - 2.

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

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

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

$$\Sigma xy = 182560.$$

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

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

$$\Sigma x^2 = 92500$$

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

$$= \frac{34915}{20250}$$

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

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 173.7 - (1.72421)(850$$

$$= 27.143$$

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

(b) Since slope is positive, y and x have a positive relationship.

classmate

Date
Page

2. (i) least square fit regression line is given by

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

$$\hat{\beta} = \frac{S_{XY}}{S_{XX}} = \frac{38191.41 - (12)(32.1)(96.875)}{12666.58 - (12)(32.1)^2}$$

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

$$= 2.9011$$

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

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

$$= 3.7497$$

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

(ii) $$\frac{\hat{\beta} - \beta}{\sqrt{\hat{\sigma}^2 / S_{xx}}} \sim t_{n-2}.$$

where $\frac{\hat{\beta} - \beta}{\sqrt{\hat{\sigma}^2 / S_{xx}}}$ is the test statistic.

$$\hat{\sigma}^2 = \frac{S_{YY} - S^2_{XY}/S_{XX}}{n-2}$$

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

$$= 387.0745$$

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

$$= 2.9011 \pm \left[2.228 \times \sqrt{\frac{387.0745^2}{301.66}}\right]$$

$$= (0.3773,\ 5.4249)$$

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

Test statistic:

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

$$95\%\ CI = (\hat{\alpha} + \hat{\beta}x_0) \pm t_{0.025,10} \times \sqrt{[se(y_0)]^2}$$

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

$\therefore\ se(y_0) = 10.5989$

$\therefore\ 95\%\ CI = 37497 + (2.9011 \times 40) \pm (2.228 \times 10.5989)$

$= (96.179,\ 143.408)$

3.(i) The plot suggests that:
- the time difference in cumulatio[n] in peak hours and non-peak hou[rs] is maximum in city A and leas[t] in city D.

- the variance in city C is least ~~as~~ as compared to that of D which has maximum variance.

(ii) Assumptions for ANOVA -
- data should ~~be~~ follow normal distribution
- sample items must be independent of ~~ea~~ each other.
- population should have a common variance i.e. homoscedasticity.

(iii) H0: Mean of difference is same for each city.
H1: Each city has a different mean of difference.

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

$$= 1495 - 28 \times \frac{(103)^2}{28^2}$$

$$= \underline{1116.11}$$

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

$$= 516.11.$$

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

$$= 1116.11 - 516.11$$

$$= 600$$

ANOVA TABLE.

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

Test statistic :

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

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

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

$\because$ $F_{cal} > F_{tab}$, we reject H0 at 5% LOS. Thus we can conclude that the mean of difference is not same for all cities.

4. (a) Mathematical model of one-way ANOVA is given by

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij} \quad , \text{ where}$$

$y_{ij} \Rightarrow j^{th}$ observation on i^{th} treatment
$\mu \Rightarrow$ common effect for whole exp.
$\tau_i \Rightarrow i^{th}$ treatment effect.
$\varepsilon_{ij} \Rightarrow$ random error in the j^{th} obs. for i^{th} treatment.

Assumptions:
- Errors are assumed to be normally and independently distributed with mean 0 and common variance σ^2.
- μ is a fixed parameter.

(b) H_0: Mean claims are equal
H_1: Mean claims are not equal.

$n_1 = 7$, $n_2 = 8$, $n_3 = 6$, $n_4 = 7$, $n_5 = 5$

$n_{tot} = 33$

$$\sum_i^5 \sum_j^{n_i} y_{ij} = 354 + 386 + 87 + 645 + 384 = 1856$$

$$\sum_i^5 \sum_j^{n_i} y^2_{ij} = 27184 + 29430 + 2123 + 84669 + 30910$$
$$= 174316$$

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

$$= 69930.06$$

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

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

$$= 22325.69$$

$$SS_{res} = SS_{tot} - SS_{reg}.$$

$$= 47604.37$$

ANOVA TABLE.

Sources of Variation	Dof	Sum of Sq.	MSS.
Regression	4	22325.69	5581.4225
Residual	28	47604.37	1700.156
Total	32.	69830.06	

$$F\,cal = \frac{5581.4225}{1700.156} = 3.2829$$

$$F\,tab = F_{4,28} = 2.714.$$

$\because$ $F\,cal > F\,tab$

we reject H0 i.e. mean claims are not equal for companies

(c) H0: No association between salaries and no. of papers cleared
H1: There is interdependence between salary and papers cleared.

O_i

Salary / papers	0 - 3	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

Salary / papers	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

Test statistic:

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

$\chi^2_{cal} = 49.91149.$

$\chi^2_{tab} = \chi^2_{2, 5\%} = 5.9912$

$\because \chi^2_{tab} < \chi^2_{cal}$, we reject our null hypothesis ~~than~~ that there is no association between salary and papers completed.

5. (i) Assumptions.
- Data should be normally distributed.
- Sample items should be independent.
- Populations should have common variance.

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

PTO.

(ii) Missing values are.

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

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

$$= \boxed{0.1213}$$

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

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

$H0: \tau_j = 0$
$H1: \tau_j \neq 0.$

(b) $y_{i.}^2 = \Sigma y_{ij} = (3 \times Mean_i)^2 = 973.373$

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

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

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

$$= 1.236$$

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

$F_{cal} = 45.638$

$F_{tab} = F_{3,8,\,5\%} = 4.066$.

$F_{tab} < F_{cal}$, therefore we reject our null hypothesis.

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

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

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

$$= 0.9447$$

(ii) Fitted regr. line is given by

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

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

$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 9.22957$.

$\hat{y} = 9.22957 + 12.0437\,x$.

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

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

$= 8923.6$

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

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

$= 960.3$

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

- using ~~pa~~ data from part (i) we have, $r^2 = R^2$.
- The correlation coefficient is the square root of the coeff. of determination

7. (i) Linear regression eqn.

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

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

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

$= 0.4545$

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

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

(iii) $H0: \beta = 0$

$H1: \beta \neq 0.$

$$TS = \frac{\hat{\beta} - \beta}{Se(\hat{\beta})} \sim t_{n-2}.$$

$$Se(\hat{\beta}) = \sqrt{\frac{\sigma^2}{Sxx}} = \sqrt{\frac{SS\,Res}{(n-2)\,Sxx}}$$

$$= \sqrt{\frac{Syy - S^2xy/Sxx}{(n-2)\,Sxx}}$$

$$= 0.0681$$

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

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

$\because$ $t_{cal} > t_{tab}$, we reject H0.
i.e. there is a linear relationship between x and y.

classmate

Date
Page

(iv) Individual response:

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

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

$$= (10.9319, 33.3745)$$

(v) The residual plot appears to follow a pattern (first increasing 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 i.e. it seeks to identify the key components necessary to model and understand data.

When original data set is written in terms of new variables that are uncorrelated, only some are needed to explain most of the variability observed in the data.

(ii) variance - covariance 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 variance explaind by each pc :

$PC_1 = \frac{0.456}{0.7015} \times 100 = 65\%$

$PC2 = \frac{0.137}{0.7015} \times 100 = 19.537\%$

$PC3 = \frac{0.08}{0.7015} \times 100 = 11.407\%$

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

$PC5 = \frac{0.012}{0.7015} \times 100 = 1.717\%$

(iii) Dimensionality could be reduced by using the first 3 PCs as 96% of variability has been explained by first 3 PCs alone.

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

(ii) corr. coeff $= \dfrac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}$

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

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

(iii) $H_0: \rho = 0$
$H_1: \rho < 0$
using fisher's transformation

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

$Z_{cal} = -2.2234$

$Z_{tab} = -1.98$

$\because$ $Z_{cal} > Z_{tab}$, we reject H_0.
$\therefore$ There is a negative correlation between the variables.

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

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

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

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

$$= 82.16$$

$$\therefore \hat{y} = 82.16 + (-3.9467)x.$$

(v) coeff of determination $(R^2) = \frac{S^2_{xy}}{S_{xx} \cdot S_{yy}}$

$$= 0.4706$$

$$r^2 = R^2.$$

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

$$= -3.9467$$

PTO

10. H0: Industry has no impact on resignation rate.
H1: There is impact.

Mean = 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	Df	Sum of Sq.	MSS
Residual	57	3591	63
Regression	2	840	420
Total	59	4431	483

TSS : $\frac{MSS\ Reg}{MSS\ Res} \sim F_{2,57}$

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

$F_{tab} = F_{2,60,5\%} = 3.15$

$\because$ $F_{cal} > F_{tab}$, we reject H0.
i.e. there is an impact of industry on resignation rate.