Assignment

1) a) $S_{xx} = \Sigma x_1^2 - (\Sigma x_1)^2 / n = 92500 - \frac{850^2}{10} = 20250$

$S_{xy} = \Sigma xy - \Sigma x \Sigma y / n = 182560 - \frac{1737}{10} = 34916$

$y = \hat{a} + \hat{b} x$

$\hat{a} = \left(\frac{1737}{10}\right) - 172\left(\frac{850}{10}\right) = 27.14$

$\hat{b} = \frac{S_{xy}}{S_{xx}} = 1.72$

$= \quad y = 27.144 + 1.72x$

b) gradient represents the amount of hours per rupee.

2) 1) $\Sigma x = 3852$, $\Sigma y = 1162.5$, $\Sigma x^2 = 12666.58$
$\Sigma y^2 = 119026.9$, $\Sigma xy = 38191.41$, $n = 12$

$S_{xx} = \Sigma x^2 - n\bar{x}^2 = 12666.58 - 12\left(\frac{385.2}{12}\right)^2$
$= 301.66$

$S_{yy} = \Sigma y^2 - n\bar{y}^2 = 119026.9 - 12\left(\frac{1162.5}{12}\right)^2$
$= 6409.7$

$S_{xy} = \sum xy - n\bar{x}\bar{y} = 38191.41 - 12\left(\frac{385.2}{12}\right)\left(\frac{1162.5}{12}\right)$

$= 878.16$

$\alpha = \bar{y} - \hat{\beta}\bar{x} = \left(\frac{1162.5}{12}\right) - 2.90\left(\frac{385.2}{12}\right)$

$= 3.78$

$y = 3.78 + 2.90x$

ii) $\beta \pm t_{n-2, \alpha/2} \cdot \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$

$= 2.90 \pm 2.228 (1.13) = (0.38, 5.42)$

iii) $\hat{Y}_x \pm t_{n-2, \alpha/2} \sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)}$

$x_0 = 40$

$\hat{Y}_{x_0} = 3.48 + 2.90(40) = 119.78$

$= 119.78 \pm 2.228 \cdot \sqrt{387.07\left(\frac{1}{12} + \frac{(40-32.1)^2}{301.66}\right)}$

$= 119.78 \pm 2.228 \times 10.5989$

$= (96.17, 143.39)$

3) i) The centres of distributions differ from all cities

The difference b/w the mean time taken to commute to office in peak hours and non-peak hours are in order from city A to D.

ii)
- The Observations are independent
- The populations must be normal
- common variance

iii) $SS_T = 1495 - \frac{(103)^2}{28} = 1116.11$

$SS_B = \frac{1}{7}(64^2 + 44^2 + 8^2 + (-13)^2) - \frac{103^2}{28}$

$= 516.11$

Source	df	SS	MS	F
Treatments	3	516.11	172.04	6.88
Residual	24	600	25	
Total	27	1116.11		

Under H_0, $F = \frac{172.04}{25} = 6.88$

$\therefore$ We have sufficient evidence to reject H_0 at 5% LOS.

5) i) Assumptions :

- Population must be normal
- Population must have common variance
- Observations are independent

ii) For the transformation $x \to \sqrt{x}$, the value of sample mean for state 1 will be :

$$\frac{1}{3}\left(\sqrt{29} + \sqrt{13} + \sqrt{21}\right) = 4.52$$

For transformation $x \to \log_e x$, the value of sample variance for state 2 will be :

$$\frac{1}{2}\left[(\log_e 180 - 4.827)^2 + (\log_e 90 - 4.827)^2 + (\log_e 120 - 4.827)^2\right]$$

$= 0.1213$

iii) The scientist was correct in asserting that the $\log_e$ transformation should be done before.

iv) ANOVA should be performed on $\log_e x$ data

$$Y_{ij} = \mu + T_i + e_{ij} \quad i = 1, 2, 3, 4 \; ; \; j = 1, 2, 3$$

Rate	Mean	variance	Y^2
1	2.992	0.163	80.582
2	4.827	0.121	209.678
3	5.716	0.186	294.053
4	6.575	0.148	389.060
	20.110	0.618	973.373

$$Y_i^2 = \left(\sum_{j=1}^{3} Y_{ij}\right)^2 = (3 \times mean_i)^2$$

$$SS_{Rate} = \sum_i^4 = \frac{Y_i^2}{3} - \frac{Y^2}{12} = \frac{973.373}{3} - \frac{(3 \times 20.110)^2}{12}$$

$$= 21.153$$

$$SS_{Res} = \sum_i^4 = \left\{ \sum_{j}^{3} - 1 (Y_{ij} - \bar{Y})^2 \right\}$$

$$= \sum_{i=1}^{4} \{2 \times var_i\} = 2 \times 0.618 = 1.236$$

Rate	Observed Mean	95% CI (Transformed Scale) LCI	95% CI (Transformed Scale) UCI	95% CI (Original Scale) LCI	95% CI (Original Scale) UCI
1	2.992	2.469	3.516	11.810	33.68
2	4.827	4.303	5.350	73.954	210.62
3	5.716	5.193	6.239	179.950	512.505
4	6.575	6.052	7.098	424.774	1209.76

Source	d.f	Sum of sq.	Mean sq.	F
Rates	3	21.153	7.051	45.0
Residuals	8	1.236	0.155	
Total	11	22.389		

Critical value = 7.951

$\therefore$ P-value for this test is almost near to zero.

v) 95% CI =

$$\bar{y}_i \pm t_{8;\,0.975} \times \sqrt{\frac{\hat{\sigma}^2}{\hat{r}}} \qquad i = 1, 2, 3, 4$$

Here, $t_{8;\,0.975}$ at 97.5% = 2.306

$\hat{\sigma} = \dfrac{\text{Residual sum of squares}}{8} = 0.155$

6)

i) $S_{xx} = \sum x_i^2 - n\bar{x}^2 = 207 - 10\left(\frac{39}{10}\right)^2 = 54.90$

$S_{xy} = \sum x_i y_i - n\bar{x}\bar{y}_i = 2853 - 10\left(\frac{39}{10}\right)\left(\frac{562}{10}\right) = 661.20$

$S_{yy} = \sum y_i^2 - n\bar{y}^2 = 40508 - 10\left(\frac{562}{10}\right)^2 = 8923.6$

$r = \frac{S_{xy}}{\sqrt{S_{xx}} \cdot \sqrt{S_{yy}}} = \frac{661.20}{\sqrt{54.90}\sqrt{8923.6}} = 0.946$

ii) Coefficients :

$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{661.20}{54.90} = 12.04$

$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = \left(\frac{562}{10}\right) - 12.04\left(\frac{39}{10}\right) = 9.23$

$= 9.23 + 12.04x$

iii) Relation : $SS_{TOT} = SS_{Reg} + SS_{Res}$

$SS_{TOT} = S_{yy} = 8923.60$

$SS_{Res} = S_{yy} - \frac{S_{xy}^2}{S_{xx}} = 8923.60 - \frac{(661.20)^2}{54.90} = 960.30$

$SS_{Reg} = SS_{TOT} - SS_{Res} = 8923.60 - 96030$
$= 7963.30$

iv) Coefficient of determination:

$$R^2 = \frac{S^2xy}{Sxx \cdot Syy} = \frac{SS_{Reg}}{SS_{TOT}} = \frac{7963.30}{8923.60} = 0.8924$$

$$r = \frac{Sxy}{\sqrt{Sxx \cdot Syy}} = \sqrt{R^2} = \sqrt{0.8924} = 0.94472$$

Q7)

i) $Sxx = \Sigma x^2 - (\Sigma x)^2/n = 4789.42$

$Sxy = 2176.84$

$$\hat{\beta} = \frac{Sxy}{Sxx} = \frac{2176.84}{4789.42} = 0.4545$$

$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x} = 17.895 - 0.4545 \times 15.83 = 10.79$

$\hat{y} = 10.79 + 0.4545x$

ii) $S_{xx} = 4789.42$

$S_{yy} = \sum y^2 - \frac{(\sum y)^2}{n} = 1189.21$

$S_{xy} = 2176.84$

$\hat{\sigma}^2 = \frac{1}{n-2}\left(S_{yy} - \frac{S^2_{xy}}{S_{xx}}\right) = \frac{1}{9} \times \left(1189.21 - \frac{2176.84^2}{4789.42}\right)$

$= 22.20$

$S.e(\hat{\beta}) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} = \sqrt{\frac{22.20}{4789.42}} = 0.0681$

$\frac{\hat{\beta} - 0}{S.e(\hat{\beta})} = \frac{0.4545}{0.0681} = 6.674$

Critical value for t-dist with 9 degrees of freedom $= t_{9, 0.025} = 2.68$

$\therefore$ We have sufficient evidence to reject Ho.

iii) Pearson coefficient : $\frac{S_{xx}}{\sqrt{S_{xx} \times S_{yy}}} =$

$S_{yy} = 1189.21$

$r = \frac{S_{xy}}{\sqrt{S_{xx} \times S_{yy}}} = \frac{2176.84}{\sqrt{4789.42 \times 1189.21}} = 0.91$

iv) Estimated value of y corresponding to $x^3 = 25$ is $10.79 + 0.4545 \times 25 = 22.15$

$\hat{\sigma}^2 = 22.20$

Variance of mean response =

$$\left[\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right]\hat{\sigma}^2 = \left[\frac{1}{11} + \frac{84.089}{4789.42}\right] 22.20$$

$= 2.41$

Var. of individual =

$$\left[1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right]\hat{\sigma}^2 = [1 + 0.1085]\, 22.20 = 24.61$$

95% CI =

Mean = $22.20 \pm 2.262 \cdot \sqrt{2.41} = (18.7, 25.7)$
Ind. = $22.20 \pm 2.262 \cdot \sqrt{24.61} = (11.0, 33.4)$

v) The plot shows a defined pattern. Even the correlation coefficient is high the model does not seem to be appropriate.

Scanned by TapScanner

8) i) Factor analysis seeks to identify key components necessary to model and understand data.

Original variables maybe correlated with each other. While newly identified principal components are chosen to be uncorrelated.

ii)

Principal Component	Diagonal entry	PCi
PC1	0.456	65%
PC2	0.137	19%
PC3	0.08	11.4%
PC4	0.0165	2.4%
PC5	0.12	1.7%
	0.7015	100%

iii) As 1st 3 principal components explain over 95% of total variance, dimensionality can be reduced to 3 for the dataset.

The 1st 3 principal components can then be used for building further classification or regression modelling purpose.

9) i) The scatter plot suggests an inverse relation b/w marks obtained and hours spent on social media per day.

ii) $S_{xx} = 277.5 - \frac{(45)^2}{10} = 75$

$S_{YY} = 43956 - \frac{(664)^2}{10} = 2482.4$

$S_{xY} = 2602 - 664\left(\frac{45}{10}\right) = -296$

$\gamma = \frac{S_{xY}}{\sqrt{S_{xx} \cdot S_{YY}}} = -0.686$

iii) Null hypothesis $H_0 = 0$
$H_1 \; p < 0$

$\gamma = 0.5 \left[\ln \left(\frac{1-0.686}{1.686} \right) \right] = -0.8404$

Fishers $= \frac{(-0.8404-0)}{\sqrt{1/7}} = -2.22$

Pvalue $= P(Z < -2.22) = 0.013$

∴ Pvalue is small so reject H_0 at 95% CI

∴ Negatively correlated

classmate
Date
Page

iv) $\hat{B} = \frac{Sxy}{Sxx} = \frac{-296}{75} = -3.9467$

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

$= \frac{644}{10} + 3.9467\left(\frac{4.5}{10}\right) = 82.16$

$y = 82.16 - 39467x$

v) $R^2 = -0.686^2 = 0.4706$

vi) For every additional hour spent on social media, the total marks reduce by 3.95.

10) H0: There is no difference
H1: There is difference

$SS_R = 19\,((5)^2 + (19)^2 + (8)^2) = 3591$

$Mean = \frac{27+36+30}{28} = 3.1$

$SS_B = 20\,((27-31)^2 + (36-31)^2 + (30-31)^2) = 840$

$F_{2,57} = (840/2) \;/\; (3591/57) = 6.667$

Test statistic is higher than 4.9777, we reject H_0.
∴ The registration rate is different