Assignment - 2

By Tanishq Navin Upreti
Roll number - 84.

DATE
PAGE

1) a) $S_{xy} = \cancel{\frac{\sum x_i y_i - \sum x_i y_i}{n}}$ $\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}$

$= 182560 - \left(\frac{850 \times 1737}{10}\right)$

$= 34,916$

$S_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$

$= 92600 - \frac{850^2}{10}$

$= 20250$

$a = \frac{1737}{10} - \frac{1.72(850)}{10} = 27.14$

$b = 1.72 \; (\because S_{xy} / S_{xx})$

$y = 27.14 + 1.72$

b) Gradient represents the amount of hours per rupee spent.

② i, $\Sigma x = 38.52$ $\quad \Sigma y = 162.5$ $\quad \Sigma xy = 88191.41$

$n = 12$

$\Sigma x^2 = 12666.56$ $\quad \Sigma y^2 = 119026.9$

$$S_{xx} = \Sigma x^2 - n\bar{x}^2$$
$$= 12666.68 - 12\left(\frac{385.2}{12}\right)^2 = 301.66$$

$$S_{yy} = \Sigma y^2 - n\bar{y}^2 = 119026.90 - 12\left(\frac{1162.5}{12}\right)^2 = 6909.71$$

$$S_{xy} = \Sigma xy - n\bar{x}\bar{y} = 38191.41 - 12\left(\frac{385.2}{12}\right)\left(\frac{11625}{12}\right)$$

$$\Rightarrow \hat{\beta} = 2.90$$

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

$$= \frac{1162.5}{12} - 2.90\left(\frac{385.2}{12}\right)$$

$$= 3.78$$

$$y = \hat{\alpha} + \hat{\beta}x = 3.78 + 2.90x$$

DATE
PAGE

ii) 95% CI for $\beta \pm t_{n-2}$ (2.5%) $s.e(\hat{\beta})$

$$s.e.(\hat{\beta}) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

$$\hat{\sigma}_1^2 = \frac{1}{n-2}\left[S_{xy} - \frac{S_{xy}^2}{S_{xx}}\right] = 387.07$$

$$\therefore s.e(\hat{\beta}) = \sqrt{\frac{387.07}{301.66}} = 1.13$$

$\therefore$ 95% CI for $\beta = 2.90 \pm 2.228\ (1.13) = (0.38, 5.72)$

iii) $\hat{y}_{x_0} \pm t_{n-2}(2.5\%) \sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}$

$x_0 = 40$ (Dell share price in US)

$y_{x_0} = 3.78 \pm 2.90(40) = 119.78$

95 C.I. $= 119.78 \pm 2.228 \sqrt{387.07\left(\frac{1}{12} + \frac{(40-32.1)}{301.66}\right)}$

$= 119.78 \pm 2.228 \times 10.5987$

$= (96.17, 143.39)$

DATE
PAGE

8) i) The centres of the distribution differ for all 4 cities.
The difference between the mean time taken to communicate to office in peak hours and non-peak hours are in order city A to city D. (highest to lowest)

ii) * Populations must be normal
* The observations are independent.
* The populations have a common variance.

iii) H_0 = The mean of difference is same for each city

V/s.

H_1 = The mean of difference is not same for all cities.

$$SS_T = 1495 - \frac{103^2}{28}$$

$$= 1116.11$$

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

$$= 516.11$$

DATE
PAGE

ANOVA table

Source	df	SS	MS	f
Treatment	3	516.11	172.04	6.88
Residual	24	600	25	
Total	27	1116.11		

The 5% critical point is 3.009, so we have sufficient evidence to reject H_0 @ 5% level. Therefore it is reasonable to conclude that therefore there are underlying difference between the cities.

(4) The mathematical model that describes the relationship between the response & treatment for the one-way ANOVA is given by

$$y_{ij} = \alpha + \beta_i + e_{ij}$$

where "y_{ij}" represents the j^{th} observation ($j = 1,2,\dots n$) on the i^{th} ~~distribut~~ treatment ($i = 1,2, \dots k$ levels)

So "y_{23}" the third observation using level 2 of the factor. "α" is the common effect for the whole experiment

β_i is the i^{th} treatment effect, & e_{ij} represents the random error present is the j^{th} observation on the i^{th} treatment.

DATE
PAGE

5) i) The following assumptions are required for one way analysis of variance (ANOVA)

- The population must be normal.
- The population have a common variance.
- The observations are independent.

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

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

For the transformation $x \to \log_e x$, the value of sample variance, for rate 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 asserting the log transformation must be done before carrying out one way ANOVA as for this it can be claimed that the assumption of common variance of underlying population holds.

iv) We will perform an ANOVA on $\log_e x$ data.

We would assume the following model.

$$y_{ij} = \mu + \tau_i + e_{ij}$$

$$i = 1, 2, 3, 4$$
$$j = 1, 2, 3.$$

DATE ______
PAGE ______

- μ is overall population mean.

- τ_i is the standard deviation of i^{th} male such that $\Sigma \tau_i = 0$

- e_{ij} are the independent error i

For ANOVA, the null hypothesis being tested here is:

$H_0: \tau = 0$, $i = 1,2,3,4$ against $H_1: \tau \neq 0$ for atleast one i.

Sum of squares tables : (loge(x))

Rate	Mean	Variance	y^2
1	2.992	0.163	80582
2	4.827	0.121	209.676
3	5.716	0.186	294.053
4	6.575	0.148	389.060
	20.010	0.618	973.373.

Here

$$y_i^2 = (\Sigma y^3 \; y_{ij}^2) = (3 \times \text{Mean})^2$$

Now

$$SS(\text{Rate}) = \sum_i^4 = \frac{y_i^2}{3} = \frac{973.373}{3} - \frac{(3 \times 26.110)^2}{12}$$

~~SSE~~ $= 21.153$

$$SS(\text{Residual}) = \Sigma_{i=1}^4 \left\{ \Sigma_j^3 = (y_{ij} - y_i)^2 \right\}$$

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

$$= 2 \times 0.618 = 1.236$$

ANOVA table is as follows.

Source	df	Sum of squares	Mean Squares	F
Rates	3	21.153	7.65	45.638
Residuals	8	1.236	0.155	
Total	11	22.389		

The 1% critical value for $F(3,8)$ distribution is 7.951

Considering the F statistic value is much greater than this, we can observe that p-value for the test is closer to zero. There is an evidence against null hypothesis H_0. Thus it must be concluded that underlying means are not equal.

V) A 95% confidence interval for the i^{th} rate mean is given by

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

Here: $t_{8;\,0.975}$ is the 97.5% critical point of a t-distribution with 8 degrees of freedom & equals 2.306.

The common error variance σ^2 is estimated as:

$$\hat{\sigma}^2 = \frac{\text{Residual sum of squares}}{8} = 0.155$$

The 95% confidence interval ~~are~~ are tabulated in the following table.

Rate	Observed mean	Transformed scale LCI	UCI	95% CI (original scale) LCI	UCI
1	2.992	2.992	2.469	11.810	33.635
2	4.827	4.827	4.303	73.954	210.625
3	5.716	5.716	5.193	179.950	312.505
4	6.575	6.575	6.052	424.770	1204.763

(6) i) $S_{xx} = \Sigma x^2 - n\bar{x}^2 = 207 - 10\left(\frac{39}{10}\right)^2 = 54.9$

$S_{xy} = \Sigma xy - n\bar{x}\bar{y} = 2853 - 10\left(\frac{3.9}{10}\right)\left(\frac{56.2}{10}\right)$

$= 661.20$

$S_{yy} = \Sigma y^2 - n\bar{y}^2 = 40.508 - 10\left(\frac{562}{10}\right)^2$

$= 8923.66$

Correlation coeficient

$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

$$= \frac{661.20}{\sqrt{54.9}\sqrt{89236}}$$

$$= 0.945$$

DATE
PAGE

ii) Fitted linear regression equation
The coefficient of the regression equation are.

$$\hat{\beta} = \frac{Sxy}{Sxx} = \frac{661.20}{54.90} = 12.04$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta} \times X = 9.23$$

$$Y = \hat{\alpha} + \hat{\beta} x$$
$$= 9.23 + 12.04 x$$

iii) Relation = $SS_{tot} = SS_{Reg} + SS_{Res}$

$$SS_{Tot} = SS_{yy} = 8922.6.$$

$$SS_{Res} = SS_{yy} - \frac{Sxy^2}{SS_{xx}} = 8923.6 - \frac{(661.20)^2}{54.9}$$
$$= 960.3.$$

iv) Coefficient of Determination