ROLL NO. - 38

DATE
PAGE

Assignment Unit 3 & 4

1. $\Sigma x = 850$ $\quad \Sigma x^2 = 92{,}500$

$\Sigma y = 1737$ $\quad n = 10$ $\quad \Sigma xy = 182560$

a) Eqn of regression line is of the form:

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

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} \quad \& \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

$$S_{xy} = \Sigma xy - n\bar{x}\bar{y}$$
$$= 182560 - (10)(85)(173.7)$$
$$= 34915$$

$$S_{xx} = \Sigma x^2 - n\bar{x}^2$$
$$= 92500 - 72250$$
$$= 20250$$

$$\hat{\beta} = \frac{34915}{20250} = 1.724197531$$
$$\approx 1.724$$

$$\& \quad \hat{\alpha} = 173.7 - (1.724)(85)$$
$$\hat{\alpha} = 27.1432$$

Eqn of regression line: $\boxed{\hat{y} = 27.14 + (1.725)x}$

(b) Gradient of regression line is $\hat{\beta}$ which is 1.724.

$(X \to \text{DCU} \quad Y \to \text{IBM})$

2. $\Sigma x = 385.2 \quad \Sigma x^2 = 12666.58$

$\Sigma y = 1162.5 \quad \Sigma y^2 = 119026.9$

$\Sigma xy = 38191.41$

(i) $S_{xy} = \Sigma xy - n\bar{x}\bar{y}$

$= 875.16$

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

$= 6409.07125$

$S_{xx} = \Sigma x^2 - n\bar{x}^2$

$= 301.66$

$\hat{\beta} = \dfrac{S_{xy}}{S_{xx}} = \dfrac{875.16}{301.66} = 2.9015$

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

$= 3.74809$

eqn of least squares fit regression line :-

$\hat{y} = 3.75 + 2.90x$

ii) Assumptions :

- explanatory variables (x_i's) are independent of eachother
- X and Y follows a normal distbn.

95% CI interval for β :

$$\hat{\beta} \pm t_{10, 2.5\%} \sqrt{\frac{\sigma^2}{S_{xx}}}$$

$$\hat{\sigma} = \frac{1}{n-2}\left(S_{YY} - \frac{S^2_{XY}}{S_{xx}}\right)$$

$$= \frac{1}{10}\left(6409.7125 - \frac{(875.16)^2}{301.66}\right)$$

$$\hat{\sigma} = 3\cancel{8}.\cancel{7}.07447$$

$\therefore$ 95% CI :

$$\hat{\beta} \pm t_{10, 2.5\%} \frac{\hat{\sigma}}{\sqrt{S_{xx}}}$$

$$= 2.90115 \pm 2.52379$$

$$= (0.37736, 5.425)$$

(iii) Let mean of IBM share price be μ_0

$$\frac{\hat{\mu}_0 - \mu_0}{se(\hat{\mu}_0)} \sim t_{n-2}$$

$$x_0 = 40$$

$$\hat{\mu}_0 = \hat{\alpha} + \hat{\beta} x_0$$

$$\hat{\mu}_0 = 119.79409$$

$$se(\hat{\mu}_0) = \sqrt{\left(\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{S_{xx}}\right) \hat{\sigma}^2}$$

$$= 10.59394$$

$$95\% \text{ CI for } \mu_0 = \hat{\mu}_0 \pm t_{10, 2.5\%} \, se(\hat{\mu}_0)$$

$$= 119.79409 \pm 23.61443$$

$$95\% \text{ CI} : (96.18, 143.41)$$

3. (i) Comments :-

- Means for all 4 cities are different.
- Mean time taken to commute to the office is highest for city A and lowest for city D
-

(ii) Assumptions for ANOVA:

→ Population follows normal distbn

→ Population is homoscedastic in nature.

→ The observations are independent

(iii) H_0 : Mean of difference is same for all cities.

H_1 : mean of difference is not the same.

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

$$= 1116.11$$

$$SS_{REG} = 516.11$$

$$SS_{RES} = SS_{TOT} - SS_{REG} = 600$$

ANOVA table:

	df	Sum of squares	MSS
REG	3	516.11	172.04
RES	24	600	25
TOT	27	1116.11	

Under H_0, $F = \dfrac{MSS_{REG}}{MSS_{RES}} = \dfrac{172.04}{25} = 6.88$

Test statistic $= 6.88$

$F_{3,24,5\%} = 3.009$

We reject H_0.

(iv) Analysis of mean differences

since $\bar{y}_1 = 9.14$
$\bar{y}_2 = 6.29$
$\bar{y}_3 = 1.14$
$\bar{y}_4 = 1.86$

$\bar{y}_1 > \bar{y}_2 > \bar{y}_3 > \bar{y}_4$

$\hat{\sigma}^2 = \frac{SS_{RES}}{n-2} = 25$

Least significant difference between means:

$t_{24, 0.025} \times \sqrt{\frac{1}{7} + \frac{1}{7}} = 2.064 \times \sqrt{25} \times \sqrt{\frac{2}{7}}$

$= 5.52$

Also:

$\bar{y}_1 - \bar{y}_2 = 2.85$
$\bar{y}_2 - \bar{y}_3 = 5.15$
$\bar{y}_3 - \bar{y}_4 = 3$

All the differences are less than 5.52, ∴ they ~~have~~ do not have a significant difference.

ANOVA Table:

	dof	RSS	MSS
REG	3	516.11	172.07
RES	24	600	25
TOT	27	1116.11	

DATE
PAGE

as calculated above

$F = 6.83$

$F_{3,24\ 5\%} = 3.009$

$\therefore$ We reject H_0.

4 (a) The mathematical model of one-way ANOVA:-

$y_i = \mu + T_i + e_{ij}$ $\quad i = 1, 2 \ldots K$
$\quad j = 1, 2, \ldots n_i$

where,
K = no. of treatments
n_i = no. of responses from i^{th} treatment
y_i = j^{th} response from i^{th} treatment
T_i = i^{th} treatment
μ → mean response
e_{ij} → error term

Assumptions : e_{ij} is iid and follows normal : $N(0, \sigma^2)$

(b) H_0 : mean claims amount are equal
H_1 : mean claim amount are not equal

according to the information given in ques :-

$$\sum_{i=1}^{s}\sum_{j=1}^{n_i} y_{ij} = 1856$$

$$\sum_{i=1}^{s}\sum_{j=1}^{n_i} y_{ij}^2 = 174316$$

$n_1 = 7$
$n_2 = 8$
$n_3 = 6$
$n_4 = 7$
$n_5 = 5$
$n = 33$

$$CF = \left(\sum_{i=1}^{s}\sum_{j=1}^{n_i} y_{ij}\right)^2 \Big/ n = \frac{(1856)^2}{33}$$

$$= 104385.9394$$

$$SS_{TOT} = 174316 - CF$$
$$= 69930.0606$$

$$SS_{REG} = \sum_{i=1}^{s}\left(\sum_{j=1}^{n_i} y_{ij}\right)^2 \Big/ n_i - CF$$

$$= 22325.69$$

$$SS_{REG} = SS_{TOT} - SS_{RES}$$

$$= 47604.3706$$

ANOVA table:

	dof	SS	MSS
REG	4	22326	5581.42
RES	28	47604	1700.16
TOT	32	69930	

$$T.S. = \frac{MSS_{REG}}{MSS_{RES}} = \underline{\underline{3.283}}$$

DATE
PAGE

$F_{4,28,0.05} = 2.714$

$\therefore$ We reject H_0.

(c) Chi - square test :-

Observed frequency (O_i)

Papers cleared	Salary per annum 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

Expected frequency (E_i)

Papers cleared	Salary per annum 3 - 5	5 - 8	8 - 10	10 - 12
0 - 3	27.41772	23.08861	14.4303	11.06329
4 - 6	15.15189	12.75949	7.97468	6.11392
7 - 9	14.43038	12.15189	7.59497	5.82278

Hypothesis test:

H_0 : Salaries are independent of no. of papers cleared

H_1 : Salaries are dependent.

DATE
PAGE

$$T.S. = \frac{\Sigma(O_i - E_i)^2}{E_i} \sim \chi^2_{(r-1)(c-1)}$$

$$= 11.275 + 0.413 + \ldots + 6.55324$$

$$T.S = 49.91146$$

degree of freedom = $2 \times 3 = 6$

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

Hence we have sufficient evidence to reject H_0. The salaries are dependent on no. of actuarial papers cleared.

5. (i) Assumptions under ANOVA:
- Population follows normal distbn.
- Common variance - Homoscedastic in nature
- Observations are independent

The sample variance given are different from each other, Hence homoscedastic nature assumption is violated.

(ii) Mean $(\sqrt{x}) \rightarrow \frac{\sqrt{29} + \sqrt{13} + \sqrt{21}}{3}$

$$= 4.52443$$

$$\text{var}\ [\log_e(x)] = \frac{\Sigma(x - \bar{x})^2}{n-1} = 0.12127$$

DATE
PAGE

(iii) (a) $X_{ij} = \mu + T_i + e_{ij}$ $\quad i = 1, 2, 3, 4$
$j = 1, 2, 3$

- X_{ij} is the loge transformed value of the j^{th} observations of the no. of generalw per square foot observed where the i^{th} rate was applied.
- μ is the overall population mean
- T_i is the deviation of the i^{th} mean such that $\Sigma T_i = 0$
- e_{ij} are the independent error terms $e_{ij} \sim N(0, \sigma^2)$

b) for ANOVA
$H_0 : T_i = 0$
$H_1 : T_i \neq 0$

ln (x)

Rate	Mean	Variance	x_i^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
Total	20.11	0.618	973.373

here $y_i^2 = \left(\sum_{j}^{3} x_{ij}\right)^2 = (3 \times \text{Mean})^2$

$$SS_{Rate} = \sum_{i=1}^{4} \frac{x_i^2}{3} - \frac{y^2}{12} = \frac{973.373}{3} - \frac{(3 \times 20.11)^2}{12}$$

$$SS_{Rate} = 21.153$$

DATE
PAGE

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

$$= \sum_{i=1}^{4} [2 \times \text{variance}]$$

$$= 2 \times 0.618$$

$$= \boxed{1.236}$$

ANOVA table:

	Dof	SS	MSS
Rate	3	21.153	7.051
Residual	8	1.236	0.155
Total	11	22.38	

$$T.S. = \frac{MSS_{RES}}{MSS_{RES}} = 45.638$$

$$F_{3,8,1\%} = 7.951$$

$\therefore$ We reject H_0. Means are not equal.

6.

X	5	9	2	3	3	11	6	5	4	
Y	73	120	34	46	35	24	26	93	45	66

$n = 10$

$\Sigma x = 39$ $\quad \Sigma x^2 = 207$

$\Sigma y = 562$ $\quad \Sigma y^2 = 40508$

$\Sigma xy = 2853$

40508 562

DATE
PAGE

(i) Correlation coefficient $(r) = \dfrac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$

$S_{xy} = \Sigma xy - n\bar{x}\bar{y}$
$= 661.2$

$S_{xx} = \Sigma x^2 - n\bar{x}^2$
$= 54.9$

$S_{yy} = \Sigma y^2 - n\bar{y}^2$
$= [redacted]\ 8923.6$

~~$r = \dfrac{661.2}{\sqrt{(54.9)(2994)}}$~~

$r = \dfrac{661.2}{\sqrt{(54.9)(8923.6)}}$

$= 0.9446623904$

(ii) fitted linear regression eqn is of the form:
$\hat{y} = \hat{\alpha} + \hat{\beta} x$

$\hat{\beta} = \dfrac{S_{xy}}{S_{xx}} = \dfrac{661.2}{54.9} = 12.044$

DATE
PAGE

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

$$= 56.2 - (12.044)(3.9)$$

$$= 9.23$$

$$\boxed{\text{eqn.}: \quad \hat{y} = 9.23 + (12.044)x}$$

(iii) $SS_{TOT} = S_{yy} = 8923.6$

$$SS_{REG} = \frac{S_{xy}^2}{S_{xx}} = \frac{(661.2)^2}{54.9}$$

$$= 7963.305$$

$$SS_{RES} = SS_{TOT} - SS_{REG}$$

$$= 960.295082$$

(iv) coefficient of determination $= R^2$

$$= \frac{SS_{REG}}{SS_{TOT}} = 0.107613$$

$$= 10.76\,\%$$

→ model is not a good fit because coefficient of determination is low

DATE
PAGE

7. $\Sigma x = 174.13 \quad \Sigma y = 197.84$

$\Sigma x^2 = 7545.90 \quad \Sigma y^2 = 4747.45$

$\Sigma(x - \bar{x})(y - \bar{y}) = 2176.84$

$\boxed{\text{(i) linear regression eqn: } \hat{y} = \hat{\alpha} + \hat{\beta} x}$

$$\hat{\beta} = \frac{S_{xy}}{S_{xy}}$$

$S_{xy} = \Sigma xy - n\bar{x}\,\bar{y} = 2176.7393$

$S_{xx} = \Sigma y^2 - n\bar{x}^2 = 4779.5042$

$\hat{\beta} = 0.4544$

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

$= 10.7896$

$\hat{y}_i = 10.7896 + 0.4544x$

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

$H_0 : \beta = 0$

$H_1 : \beta > 0$

$$TS = \frac{0.4544 - 0}{se(\hat{\beta})}$$

$$se(\hat{\beta}) = \sqrt{var\,\hat{\beta}}$$

$$= \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

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

$$= \frac{1}{9}\left[1189.029 - \frac{(2176.7353)^2}{4789.5042}\right]$$

$$= 22.193561$$

$$se(\hat{\beta}) = 0.06807195$$

$$TS = \frac{0.4544}{0.06807195} \sim t_9$$

$$= 6.67528 \sim t_9$$

$\beta > 0$

$\Rightarrow$ We reject H_0

$\Rightarrow$ there is some positive relationship.

(iii) $$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = 0.91214$$

(iv) $\hat{x}^3 = 25$

mid response

$$\hat{y} = \hat{\alpha} + \hat{\beta} x_0$$
$$\hat{y} = 22.1496$$

$$P = \theta = \frac{\hat{y}_0 - y_0}{sd(\hat{y}_0)} \sim t_{n-2}$$

$$sd(\hat{y}_0) = 4.959912$$

95% CI :

$$P(-2.262 < t_9 < 2.262)$$

$$\left[\hat{y}_0 \pm 2.262\,(sd\,\hat{y}_0)\right]$$

$$= (10.93027, 33.3689)$$

mean response

$$\hat{\mu}_0 = 22.1496$$

$$var(\hat{\mu}_0) = 2.40769$$

$$sd = 1.551505$$

95% CI : $(18.64009, 25.6591)$

DATE
PAGE

(v) Errors are not iid random variables.

ϵ_i does not follow $N(0, \sigma^2)$

8.

9. $S_{xx} = 75$
$S_{yy} = 2482.4$
$S_{xy} = -296$

(i) negative correlation

(ii) $r = \dfrac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = -0.686001$

(iii) $H_0 : \rho = 0$
$H_1 : \rho < 0$

$r = -0.686001$

$TS = \dfrac{\tanh^{-1} r - \tanh^{-1} \rho}{\sqrt{1/n-3}} \sim N(0,1)$

$= -2.22339$

$\Rightarrow$ we reject H_0.

(iv) Response variable = Marks
hours spent = explanatory variables

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

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 82.16$$

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = -3.946666$$

eqn :- $\hat{y}_i = 82.16 - (3.95)x$.

(v) $R^2 = 47.059\%$

→ not a good fit

(vi) $x_0 = 1$

$$\hat{y} = \hat{\alpha} + \hat{\beta} x_i$$
$$= 82.16 - (3.95)(1)$$
$$= 78.21334$$

10. ANOVA Table:

	dof	SS	MSS
REG	1	0.00115238	0.00115232
RES	1	0.00011422	0.00011422
TOT	2	0.0012666	

$SS_{TOT} = S_{yy} = 0.0012666$

$SS_{REG} = \frac{(0.027)^2}{0.0042} = 0.115238$

$SS_{RES} = 0.00011422$

H_0 : no impact
H_1 : impact

$TS = \frac{SSREG}{SSRES/n-2} \sim F_{1,1}$

$10.089126 \sim F_{1,1}$

$\Rightarrow$ we do not reject H_0.