ASSIGNMENT II

Date
Page

Q no 1) Modern email servers and anti-spam filter attempt to identify spam email and direct them to a junk folder. There are various ways to detect spam and research still continues. In this regard, an information security officer tries to confirm that the chance for an email to be spam depends on whether it contains image or not. The following data were collected on n = 1000 random email messages:

spam status	Image Containing status		
	With image	No image	Total
With spam	160	240	400
No spam	140	460	600
Total	300	700	1000

Assess whether being spam and containing images are independent factors at 1% level of significance.

→ Soln, (Using Chi-Square Test for independence)

- Problems to test:

  H_0: Being spam and containing images are independent factor.

  H_1: Being spam and containing images are dependent factors.

- Expected Frequency:

  ~~Spam~~ formula: E_f = (row total * Column total) / Total emails

status	observed frequency (OF)	Expected Frequency
Spam with image	160	400 x 300/1000 = 120
Spam without image	240	400 x 700/1000 = 280
No spam with image	140	600 x 300/1000 = 180
No Spam without image	460	700 x 600/1000 = 420

Date
Page

- Calculating Chi-square (X^2):

$$\text{Chi-Square } (X^2) = \sum \left[\frac{(OF - EF)^2}{EF} \right]$$

$$X^2 = (160-120)^2/120 + (240-280)^2/280 + (140-180)^2/180 + (460-420)^2/420$$

$$= 13.33 + 5.7142 + 8.88889 + 3.8095$$

$$= 31.746$$

- Degree of Freedom (df):

$df = (\text{no of rows} - 1) * (\text{no of columns} - 1)$

$df = (2-1) * (2-1) = 1$

- P-value:

$\alpha = 0.01$ (1%)

Here $\boxed{x^2_{\alpha} < x^2_{calc}}$ in the degree of freedom 1

ie, $6.635 < 31.746$

- Decision:

The calculated chi-square is greater than the table value for $df = 1$ and $\alpha = 0.01$

hence, we reject H_0 at 1% level of significance level.

- Conclusion:

Hence the presence of image is not independent of whether an email is spam. There is a strong relationship between the two factors.

Date
Page

no 2) The following data related to the number of cohildren classified according to the type of feed and the nature of teeth

	Nature of teeth.	
Type of feed	Normal	Defective
Brest	18	12
Bottle	2	13

Do the information provide sufficent evidence to conclude that type of feeding and nature of teeth are dependent? Use chi square test at 5% level of significance.

→ Soln,

Using Chi-square Test for independence:

- Problem to test:

H_0: There is no association

H_1: There is association.

- Datas

Type of feed	Normal Teeth	Defective Teeth	Total
Breast	18	12	30
Bottle	2	13	15
Total	20	25	45

Level of significance : $\alpha = 0.05$ (5%)

- Expected Frequency (EF):

Formula: EF = (row total * Column total) / Total

Date
Page

Type	observed Frequency(OF)	Expected Frequency (EF)
Brest, Normal	18	(20×30)/45 = 13.33
Breast, Defective	12	(30×25)/45 = 16.67
Bottle, Normal	2	(15×20)/45 = 6.67
Bottle, Defected	13	(15×25)/45 = 8.33

- Calculating chi-square (χ^2):

$$\chi^2 = \sum \left[\frac{(OF-EF)^2}{EF} \right]$$

$$= \frac{(18-13.33)^2}{13.33} + \frac{(12-16.67)^2}{16.67} + \frac{(2-6.67)^2}{6.67} + \frac{(13-8.33)^2}{8.33}$$

$$= 1.636 + 1.30827 + 3.2697 + 2.618115$$

$$= 8.832085$$

- Degree of freedom (df)

df = (no of rows - 1) * (no of column - 1)

= 2-1 x 2-1 = 1

- P-value:

$\alpha = 0.05 (5\%)$

Here

$\boxed{\chi^2_\alpha < \chi^2_{calc}}$ in the (df = 1)

ie $3.841 < 8.832$

- Decision:

Since χ^2 (calculated) is greater than the table value, We reject Ho at 5% level of significance.

Date
Page

Conclusion:

Hence there is an association between type of feeding and the nature of teeth.

Qno3) Social media users use a variety of devices to access social networking, mobile phones are increasingly popular. However, is there a difference in the various age groups in the proportion of social media users who use their mobile phone to access social ~~media~~ networking? A study showed the following results for the different age groups.

	Age		
Use mobile phones to access social networking	18-34	35-64	65+
yes	60	37	14
No	40	63	86

At the 0.05 level of significance, is there evidence of a different age group with respect to use of mobile phone for accessing social networking?

→ Soln,

Using Chi-square Test.

- Problem to Test

H_1: Evidence of different age group

H_0: No enough Evidence of different age group.

Date
Page

- Datas:

Age group	yes	No	Total
18-34	60	40	100
35-64	37	63	100
65+	14	86	100
Total	111	189	300

Level of significance $\alpha = 0.05$

- Expected frequency (EF):

Formula: EF = (Total row * Column Total) / all total

Age	Observed frequency (OF)	Expected frequency (EF)
18-34(yes)	60	(100x111) /300 = 37
18-34(No)	40	(100x189)/300 = 63
35-64(yes)	37	(100x111)/300 = 37
35-64(No)	63	(100x189)/300 = 63
65+(yes)	14	(100 x 111)/300 = 37
65+(No)	86	(100x189) / 300 = 63

- Chi-square (X^2):

$$X^2 = \sum \left[\frac{(OF - EF)^2}{EF} \right]$$

$= 14.29729 + 8.396825 + 0 + 0 + 14.29729 + 8.39682$

$= 45.3882304$

Date
Page

- Degree of free-dom (Df)

$df = (row-1)*(Column-1)$
$= (3-1) \times (2-1)$
$= 2$

- P-value:

$\alpha = 0.05$

Here

$\boxed{\text{If } X^2_{\alpha} < X^2_{cal}}$ in $df = 2$

ie, $5.991 < 45.3882$

- Decision:

Since X^2 is greater than the table value at α 0.05 level significance, ~~Accept H_1~~. Reject H_0.

- Conclusion:

Hence, there is sufficient evidence to suggest that there is a different age group in the proportion of social media users who use their mobile phone to access social networking at 0.05 level of significance.

Date
Page

Qno4) A random sample of 200 married men, all retired, were classified according to education and number of children.

Education	No. of children			
	0-1	2-3	Over 3	Total
Elementary	14	37	32	83
Secondary	19	42	17	78
College	12	17	10	39
Total	45	96	59	200

Test the hypothesis, at the 1% level of significance, that the number of children is independent of the level of education attained by the father.

→ Soln;

Using Chi-Square Test

- Problem to test:

H_0: The no of children is independent of level of education attained by the father.

H_1: The no of children is dependent of level of education attained by the father.

- Expected Frequency (Ef)

Formula:

$$Ef = (\text{Row total} \times \text{Column total}) / \text{all total}$$

Date
Page

Education	Observed frequency (OF)	Expected frequency (EF)
Elementary (0-1)	14	(83 x 45) / 200 = 18.675
Elementary (2-3)	37	(83 x 96) / 200 = 39.84
Elementary (3+)	832	(83 x 59) / 200 = 24.485
Secondary (0-1)	19	(78 x 45) / 200 = 17.55
Secondary (2-3)	42	(78 x 96) / 200 = 37.44
Secondary (3+)	17	(78 x 59) / 200 = 27.01
College (0-1)	12	(39 x 45) / 200 = 8.775
College (2-3)	17	(39 x 96) / 200 = 18.72
College (3+)	10	(39 x 59) / 200 = 11.505

- Chi square (X^2):

$$X^2 = \sum \left[\frac{(OF - EF)^2}{EF} \right]$$

$$= 1.170314 + 0.202448 + 2.306523 + 0.1198 + 0.55538 + 3.70974 + 1.185256 + 0.158034 + 0.196873$$

$$= 9.60436$$

- Degree of freedom (df):

$$df = (row - 1) \times (column - 1)$$
$$= 2 \times 2 = 4$$

- P-value

$\alpha = 0.01$ (1%)

Here

$$\boxed{X^2_{\alpha} < |X^2_{cal}|} \quad \text{in } df = 4$$

ie, $13.277 > 9.60436$

Date
Page

- Decision:

Hence, we accept H_0 at 1% level of significance

- Conclusion:

Hence, the no of children is independent of level of education attained by the father.

538

Date
Page

Qno5) A psychologist wishes to verify that a certain drug increases the reaction time to given stimulus. The following reaction times (in tenth of seconds) were recorded before and after injection of the drug for each of four subjects.

Subject'		1	2	3	4
Reactiontime	Before	7	2	12	12
	After	13	3	18	13

~~Soln~~ Test at 5% level of significance to determine whether the drug significantly increases reaction time. Use non parametric test.

→ Soln

performing non-parametric (Wilcoxon Signed Rank Test)

- Problems to test

H_0: Drug has no effect on reaction time.
H_1: The drug increases reaction time. (after reaction time greater)

- Data

Subject	Before RT	After RT	Diffrence (After - before)
1	7	13	6
2	2	3	1
3	12	18	6
4	12	13	1

Significance level $\alpha = 0.05$ (5%)
Sample size(n) = 4

Date
Page

- Ranking :

Subject	Difference	Rank
1	6	4
2	1	1
3	6	4
4	1	1

→ Here we only consider +ve ranks (T+) as only increasing or not is asked.

- Calculating T+ :

$T+ = \text{Rank(Sub1)} + \text{Rank(Sub3)}$

$= 4+4$

$= 8$

- Wilcoxon signed-Rank statistic (W):

$W = T+$ (for increasing)

$= 8$

- Critical value:

$\alpha = 0.05$

for 2 two-tail and 4 sample size, $W\alpha = 0$

- Decision

$W_{0.05} < W_{calc}$

ie. $0 < 8$

hence, we reject H_0 at 0.05 level of significance.

Date
Page

- Conclusion:

Hence, we have stronge evidence to suggest that the drug significantly increases reaction time.

Date
Page

- Conclusion:

Hence, we have stronge evidence to suggest that the drug significantly increases reaction time.

Qno 6) What do you mean by ~~para~~ non-parametric test? write down advantages of non-parametric tests over parametric tests.

→ Non-parametric test, also known as distribution-free tests, are statistical method used in hypothesis testing that do not make assumptions about the frequency distribution of the variables being evaluated.

→ It's advantages are:

a) Can be applied to qualitative (rank, odinal, categorical data) aswell as quantitive data.

b) It is the only possible test for respectivaly small sample.

c) It is simple to understand, quicker and easier to apply

d) It is less time consuming.

e) It needs no assumption about the population from which sample size is selected.

f) It has greater rank of applicability because of milder assumption.

g) It does not require complicated sampling theory.

Date
Page

Q.no 7) Bank of Nepal recorded the sex of first 30 customers who appeared last monday with notation MMFMMF MFFMMMFFMMFFMFFMFFMFMMMFF. At the 0.05 level of significance, test the randomness of this sequence.

→ Soln,

no of M(n_1) = 15

no of F(n_2) = 15

no of runs (r) = 16

Level of significance (α) = 0.05

- Problem to test:

H_0: The sequence is random order

H_1: The sequence isn't in random order.

- Test statistic

→ $r = 16$

$$\mu_1 = \frac{2n_1n_2}{n_1+n_2} = \frac{2\times15\times15}{30} = 15$$

$$\sigma_r^2 = \frac{2n_1n_2(2n_1n_2-n_1-n_2)}{(n_1+n_2)^2(n_1+n_2-1)} = \frac{2\times15\times15(2\times15\times15-15-15)}{(15+15)^2\times(15+15-1)}$$

$$\therefore \sigma_r = \sqrt{4.2413} = 2.6909$$

$$z = \frac{r-\mu_r}{\sigma_r} = \frac{16-15}{2.6909} = 0.3716$$

Date
Page

- Critical value:
  At $\alpha = 0.05$ level of significance critical value is
  $Z_{tabulated} = Z_{\alpha}$

- Level of significance
  At $\alpha = 0.05$, $(\underline{r}, \overline{r}) = (10, 22)$, $r = 16$

- Decision
  Accept H_0 at $\alpha = 0.05$. as $\forall\, r \in (\underline{r}, \overline{r})$ ie $16 \in (10, 22)$

- Conclusion:
  Hence, the sample are in random order.

Date:
Page

Qno 8) Define level of significance. Describe run test with some relevant examples

→ Level of significance, denoted by α (alpha), is a critical value used in the hypothesis testing to determine the cutoff point for rejecting the null hypothesis. It represents the maximum probability of making type I error, which occurs when the null hypothesis is incorrectly rejected when it is actually true.

→ Run test is a ~~para~~ non-parametric test used to determine the randomness of the selected sample.
Run is a set of indentical or related symbols contained between two different symbols or none at all.

Let us consider a random sample of size n is selected.
Let $x_1, x_2, x_3, x_4 ---- x_n$ be samples.

Step 1) Problem to test
H_0 : In random order
H_1 : ~~for~~ Not in random order.

Step 2) Finding median of sample and assign a symbols, number of observations of symbol A denoted by n_1 and number of observations of symbol B denoted by n_2.

Step 3) If,
~~Step~~ For small sample ($n_1, n_2 \leq 20$)
Test statistic
Number of runs (r)

Date
Page

~~Step 1)~~ Level of significance.
$\alpha = 0.05$ is taken unless we are given.

~~Step 2)~~ Critical value:
$\underline{r}$ and $\bar{r}$ are obtained from table according to level of significance, degree of freedom n_1 and n_2, and alternative hypothesis.

Decision
Accept H_0 at α level of significance if
$r \in (\underline{r}, \bar{r})$, reject else.

Step 3) Else,
For Large sample size.
In this case ~~large sa~~ r is aprox normally distributed with mean $\mu_r = \frac{2n_1n_2}{n_1+n_2} + 1$

And variance $\sigma_r^2 = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1+n_2)^2(n_1+n_2-1)}$

Test statistic
$$z = \frac{r - \mu_r}{\sigma_r} \sim N(0,1)$$

Level of significance
$\alpha = 0.05$ is taken unless we are given

Critical value
z is obtained from table for level of significance, and alternative hypo thesis.

Date
Page

Decision.
Reject H_0 if $|z| > z$ tabulated. accept otherwise

Note For α other than 5% even if sample size is small use z test in which $z = \frac{|r - M_r| - 0.5}{\sigma_r}$.

Q.no 9) What do you mean by run? Marks secured by a sample of 15 students in Final exam of statistics II are found to be 27, 34, 48, 21, 7, 56, 44, 32, 25, 42, 33, ~~48~~ 28, 41, 5, 49, Are marks in random order? use 5% level of significance.

→ Run is a set of identical or related symbols contained between two different symbols or none at all.

→ Soln,
$n = 15$
Arranging data in ascenting order
5, 7, 21, 25, 27, 28, 32, 33, 34, 41, 42, 44, 48, 49, 56

$$\text{median} = \left(\frac{n+1}{2}\right)^{th} \text{item}$$

$$= \frac{15+1}{2} = 8^{th} \text{ item}$$

∴ median = 33

Now number greater than 33 be B & less than be A.
ABBAABBAA BBABAB

Date
Page

no of A (n_1) = 7
no of B (n_2) = 8
no of runs (r) = 9

Problem to test
H_0 : The sequence is random.
H_1 : The sequence isn't random.

level of significance.
$\alpha = 0.05$, $(\underline{r}, \overline{r}) = (4, 13)$, $r = 9$

- Decision:
Accept H_0 at $\alpha = 0.05$ as $r \in (\underline{r}, \overline{r})$ ie $9 \in (4, 13)$

- Conclusion:
Hence, the sample are in random order.

Date
Page

Q no 10) What is median test? Following data represents marks secured by students of section A and section B of a college in mid-term exam of statistics II.

Section A	30	27	19	22	28	25	9	13	20
Section B	24	28	16	22	19	29	7	11	

Is there any significant difference in marks of section A and section B? Use median test at 5% level of significance.

→ Median Test is the non-parametric test used to test the significance difference between two independent distribution.

→ Soln

arranging combined data in ascending order

7 9 11 13 16 19 19 20 22 22 24 25 27 28 28 29 30

$n = 12$

median $Md = \left(\frac{n+1}{2}\right)^{th}$ item

$= \frac{18}{2}^{th}$ item

$= 9^{th}$ item

∴ median = 22

Date
Page

no of obs in 2nd sample less or equal to median (a) ≥ 5
1st sample size (n_1) = 9
2nd sample size (n_2) = 8

$$K = \frac{n_1 + n_2}{2} = \frac{9+8}{2} = 8.5 = 8$$

let Md_1 and Md_2 be median of I population and II population respectively.

- Problem to test
  $H_0 : Md_1 = Md_2$
  $H_1 : Md_1 \neq Md_2$ } Two tailed

- Test statistics
  $$P(A=a) = \frac{C(n_1, a)\, C(n_2, K-a)}{C(n_1+n_2, K)} = \frac{C(9,a)\, C(8, K-a)}{C(9+8, K)}, \quad a = 0,1,2,3,4,5,6,7$$

- Critical value:
  $$P = P(A \geq a) = P(A \geq 5)$$
  $$= \sum \frac{C(9,a)\, C(8, 8-a)}{C(9+8, 8)}$$
  $$= \frac{C(9,5)\, C(8,3)}{C(17,8)} + \frac{C(9,6)\, C(8,2)}{C(17,8)} + \frac{C(9,7)\, C(8,1)}{C(17,8)}$$
  $$\approx 0.398849$$

- Decision:
  $2P = 2 \times 0.398849 = 0.7984 > \alpha = 5\% = 0.05$. Accept H_0.

- Conclusion:
  There is no significant difference in marks of section A and section B

Date
Page

Qno 11) Two computer makers, A and B, compete for a certain market. Their users rank the quality of computers on a 4-point scale as "Not satisfied", "satisfied", "Good quality" and "Excellent quality, will recommend to others. The following counts were observed.:

Computer maker	Not satisfied	Satisfied	Good quality	Excellent quality
A	20	40	70	20
B	10	30	40	20

Is there a significant difference in costomer satisfaction of the computers produced by A and by B using Mann-Whitney U test at 5% level of significance.

→ Soln

A	Ranks	B	Ranks
20	3	10	1
40	6.5	30	5
70	8	40	6.5
20	3	20	3
	$R_1 = 20.5$		$R_2 = 15.5$

Sample size of A = 4 (n_1)
Sample size of B = 4 (n_2)
Sum of Ranks of A (R_1) = 20.5
Sum of Ranks of B (R_2) = 15.5

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 = 4 \times 4 + \frac{4(4+1)}{2} - 20.5$$

$$= 5.5$$

$$U_2 = n_1 n_2 - U_1 = 4 \times 4 - 5.5$$

$$= 10.5$$

Date
Page

$U_0 = \min\{U_1, U_2\} = 5.5$

Let Md_1 and Md_2 be median of A and B resp.

- Problem to test:

H_0 : $Md_1 = Md_2$

H_1 : $Md_1 \neq Md_2$

- Test statistics.

$U_0 = 5.5$

- Critical value

$\alpha = 0.05$, $P = 0$ (critical value at (4,4))

- decision.

$P < U_0$ ie $U_{0.05}(4,4) = 0 < U_0 = 5.5$. Accept H_0 at 5% level of significance.

- Conclusion:

There is no significant diffrence in customer satisfactio of computer producted by A & B.

At $\alpha = 5\% = 0.05$, critical value is $U_\alpha(n_A, n_B) = U_{0.05}(4,4) = 0$

* Decision

$U_0 = 5.5 > U_{0.05}(4,4) = 0 \Rightarrow$ Accept H_0 at 0.05 level of signific

* Conclusion

There is no significant difference in customer satisfaction of computers produced by A and B.

Q12 A chemist uses 3 catalyst for distilling alcohol and layout were tabulated below

Catalyst	Alcohol (in cc)				
C_1	380	430	410		
C_2	290	350	270	250	270
C_3	400	380	450		

Are there any significant differences between catalyst? Test at 5% level of significance. Use Kruskal Wallis H test.

Soln:

We are given

Catalyst	Alcohol (in cc)				
C_1	380	430	410		
Rank C_1	6.5	10	9		
C_2	290	350	270	250	270
Rank C_2	4	5	2.5	1	2.5
C_3	400	380	450		
Rank C_3	8	6.5	11		

Here,

Sample size of $C_1 = n_1 = 3$

" " $C_2 = n_2 = 5$

" " $C_3 = n_3 = 3$

$n = n_1 + n_2 + n_3 = 3 + 5 + 3 = 11$

And,

t_i :- No. of observations having same rank $[i = 1, 2, 3, \ldots]$

i.e. $t_1 = 2$, $t_2 = 2$ [for 270 and 380].

* Problem to test

H_0: $Md_1 = Md_2 = Md_3$

H_1: At least one Md_i is different $[i = 1, 2, 3]$

Md_1 :- Median of alcohol for C_1 (in cc)

Md_2 :- " " C_2 "

Md_3 :- " " C_3 "

* Test statistic

$$H = \frac{\frac{12}{n(n+1)}\left[\sum_i \frac{R_i^2}{n_i}\right] - 3(n+1)}{1 - \frac{\sum (t_i^3 - t_i)}{n^3 - n}}$$

$$= \frac{\frac{12}{11(11+1)}\left[\frac{25.5^2}{3} + \frac{15^2}{5} + \frac{25.5^2}{3}\right] - 3(11+1)}{1 - \frac{(2^3-2) + (2^3-2)}{11^3 - 11}}$$

$$= 7.5688$$

* Critical value

At $n_1 = 3$, $n_2 = 5$ and $n_3 = 3$ and $H = 7.5688$, critical value is

$P = 0.009$

* Decision

$P = 0.009 < \alpha = 5\% = 0.05$

$\Rightarrow$ Reject H_0 at 5% level of significance

* Conclusion

There is significant difference between catalyst.

Q13 There are 3 brands of computers Dell, Lenovo and HP. The following are lifetime of 15 computers in years

Serial No.	Computer brand	lifetime (in years)
1	Dell	15
2	Lenovo	10
3	HP	9
4	Dell	12
5	Lenovo	6
6	HP	7
7	Dell	4
8	Lenovo	8
9	HP	13
10	Dell	11
11	HP	5
12	Lenovo	7
13	Dell	3
14	HP	5
15	Lenovo	4

Apply appropriate statistical test to identify whether the average life of time (in years) is significantly different across 3 brands of computers at 5% level of significance. You can again tabulate data initially in required format for statistical analysis.

Soln:

We are given

Sample size of Dell brand $(n_1) = 5$

" " Lenovo " $(n_2) = 5$

" " HP " $(n_3) = 5$

& $n = n_1 + n_2 + n_3 = 15$

Computer brands	Lifetime (in yrs)					Rank (R_i)
Dell	15	12	4	11	3	
Rank for Dell (R_1)	15	13	2.5	12	1	43.5
Lenovo	10	6	8	7	4	
Rank for Lenovo (R_2)	11	6	9	7.5	2.5	36
HP	9	7	13	5	5	
Rank for HP (R_3)	10	7.5	14	4.5	4.5	40.5

Here,

$R_1 = 43.5$, $R_2 = 36$, $R_3 = 40.5$

$t_1 = 2\ (4,4)$, $t_2 = 2\ (5,5)$, $t_3 = 2\ (7,7)$

* Problem to test

H_0: $Md_1 = Md_2 = Md_3$

H_1: At least one Md_i is different. $[i = 1,2,3]$

* Test statistic

$$H = \frac{\frac{12}{n(n+1)} \sum_i \frac{R_i^2}{n_i} - 3(n+1)}{1 - \frac{\sum (t_i^3 - t_i)}{n^3 - n}}$$

$$= \frac{\frac{12}{15(15+1)} \left[\frac{R_1^2}{n_1} + \frac{R_2^2}{n_2} + \frac{R_3^2}{n_3}\right] - 3(15+1)}{1 - \frac{(t_1^3 - t_1) + (t_2^3 - t_2) + (t_3^3 - t_3)}{15^3 - 15}}$$

$$= \frac{\frac{12}{15 \times 16} \left[\frac{43.5^2}{5} + \frac{36^2}{5} + \frac{40.5^2}{5}\right] - 3 \times 16}{1 - \frac{(2^3 - 2) + (2^3 - 2) + (2^3 - 2)}{15^3 - 15}}$$

$$= 0.2865$$

* Critical value

At $\alpha = 5\% = 0.05$, critical value is $\chi^2_{\alpha(k-1)} = \chi^2_{0.05(3-1)} = \chi^2_{0.05}$
$= 5.991$

* Decision

$H = 0.2565 < \chi^2_{\alpha(k-1)} = 5.991$

$\Rightarrow$ Accept H_0 at 5% level of significance

* Conclusion

Average lifetime is not significantly different across 3 brands of computers at 5% level of significance.

Q14 Marks secured by 8 students in 3 chapter tests in a subject are as follows:

udent	A	B	C	D	E	F	G	H
. test I	13	11	16	19	6	14	18	5
. test II	14	10	18	11	12	9	18	7
test III	15	19	13	10	11	5	17	4

Is there any significant difference in marks in 3 chapter tests? Use Friedman's two way ANOVA test at 10% level of significance.

Soln:

We are given

Student	A	B	C	D	E	f	G	H	R_i
test I	13	11	16	19	6	14	18	5	
Rank I	1	2	2	3	1	3	2.5	2	16.5
h test II	14	10	18	11	12	9	18	7	
Rank II	2	1	3	2	3	2	2.5	3	18.5
ch test III	15	19	13	10	11	5	17	4	
Rank III	3	3	1	1	2	1	1	1	13

Here,

No. of samples $(k) = 3$

Total no of samples $= 8+8+8 = 24 = n_t$

Sample size of each test $(n) = 8$

Tied cases $= 2$ [18 and 18 in G] $\Rightarrow t_1 = 2$

* Problem to test

H_0: $Md_I = Md_{II} = Md_{III}$

H_1: At least one Md_i is different $[i: I, II, III]$

* Test statistic

$$F_r = \frac{\frac{12}{nk(k+1)}\left[\sum_i R_i^2\right] - 3n(k+1)}{1 - \frac{\sum t_i^3 - t_i}{n(k^3-k)}}$$

i.e

$$Fr = \frac{12}{8 \times 3(3+1)}\left[16.5^2 + 18.5^2 + 13^2\right] - 3 \times 8(3+1)$$

$$1 - \frac{(2^3-2)}{8 \times (3^3-3)}$$

$$= 2$$

* Critical value

From Friedman table, at $k=3$, $n=8$ and $Fr=2$, critical value is $p =$

* Decision

$p =$ $> \alpha = 0.05$. Accept H_0 at 5% level of significance

* Conclusion

There is no significant difference in marks in 3 chapter tests.

Q15 It was reported somewhere that children whenever plays game in computer, they used the computer very roughly which may reduce the lifetime of computer. The random access memory (RAM) of a computer also plays a crucial role on the lifetime of a computer. A researcher wanted to examine how the lifetime of a personal computer which is used by children is affected by the time (in hours) spends by children per day to play games and available random access memory (RAM) measured in mega bytes (MB) of a used computer. Data is provided in the following table:

Lifetime (yrs)	5	1	7	2	3	4	6
Play time (hrs/day)	2	8	1	5	6	3	2
RAM (MB)	8	2	6	3	2	4	7

Identify which one is dependent variable? Solve this problem using multiple linear regression model and provide problem specific interpretation based on the regression model developed.

Soln:

→ Life time (in years) is dependent variable = y (say)
Play time (x_1) [say] and RAM (x_2) [say] are independent variables.

And,

Required linear regression model is

$$y = b_0 + b_1 x_1 + b_2 x_2 \quad \text{— (*)}$$

Here,

y	x_1	x_2	$x_1 y$	$x_2 y$	x_1^2	x_2^2	$x_1 x_2$
5	2	8	10	40	4	64	16
1	8	2	8	2	64	4	16
7	1	6	7	42	1	36	6
2	5	3	10	6	25	9	15
3	6	2	18	6	36	4	12
4	3	4	12	16	9	16	12
6	2	7	12	42	4	49	14

Where,

$\Sigma y = 28$ $\quad \Sigma x_2 y = 154$ $\quad n = 7$

$\Sigma x_1 = 27$ $\quad \Sigma x_1^2 = 143$

$\Sigma x_2 = 32$ $\quad \Sigma x_2^2 = 182$

$\Sigma x_1 y = 77$ $\quad \Sigma x_1 x_2 = 91$

Now,

To fit (*), we solve the following

→ $\Sigma y = n b_0 + b_1 \Sigma x_1 + b_2 \Sigma x_2$

Or, $28 = 7 b_0 + 27 b_1 + 32 b_2 \quad \text{— (i)}$

→ $\Sigma x_1 y = b_0 \Sigma x_1 + b_1 \Sigma x_1^2 + b_2 \Sigma x_1 x_2$

or, $77 = 27 b_0 + 143 b_1 + 91 b_2$ — (II)

→ $\Sigma x_2 y = b_0 \Sigma x_2 + b_1 \Sigma x_1 x_2 + b_2 \Sigma x_2^2$

or, $154 = 32 b_0 + 91 b_1 + 182 b_2$ — (III)

Then,

Using Cramer's rule

coeff. of b_0	coeff. of b_1	coeff. of b_2	constant
7	27	32	28
27	143	91	77
32	91	182	154

i.e.

$$D = \begin{vmatrix} 7 & 27 & 32 \\ 27 & 143 & 91 \\ 32 & 91 & 182 \end{vmatrix}$$

$$D_0 = \begin{vmatrix} 28 & 27 & 32 \\ 77 & 143 & 91 \\ 154 & 91 & 182 \end{vmatrix}$$

$$D_{b1} = \begin{vmatrix} 7 & 28 & 32 \\ 27 & 77 & 91 \\ 32 & 154 & 182 \end{vmatrix}$$

$$D_{b2} = \begin{vmatrix} 7 & 27 & 28 \\ 27 & 143 & 77 \\ 32 & 91 & 154 \end{vmatrix}$$

And,

$b_0 = \frac{D_0}{D} = 6.9613$

$b_1 = \frac{D_1}{D} = -0.7854$

$b_2 = \frac{D_2}{D} = 0.0149$

So,

④ becomes

$y = 6.9613 + (-0.7854) x_1 + 0.0149 x_2$

i.e $y = 6.9613 - 0.7854 x_1 + 0.0149 x_2$

Now,

b_0 : 6.9613 means that the lifetime of computer is 6.9613 years when play time is 0 hours/day and 0 MB RAM is used.

b_1 : - 0.7854 means that lifetime of computer decreases by 0.7854 years when play time increases by 1 hours/day and constant RAM is used.

b_2 = 0.0149 means that the lifetime of computer increases by 0.0149 years when RAM usage increases by 1 MB at constant playtime.

Q16 What are required conditions for error variable in multiple regression analysis? The Internal Revenue Service is trying to estimate the monthly amount of unpaied taxes discovered by its auditing division. The Internal Revenue service estimated this figure on the basis of field auditing labour hours and no. of hours its computers are used. The table given below presents these data for the last 10 months:

Month	Field audit labour hours in 100 (x_1)	Computer hours in 100 (x_2)	Annual unpaid taxes discovered million of dollars (y)
Jan	45	16	29
Feb	42	14	24
Mar	44	15	27
Apr	45	13	25
May	43	13	26

Jun	46	14	28
Jul	44	16	30
Aug	45	16	28
Sep	44	15	28
Oct	43	15	27

Given,

$\Sigma y x_1 = 12005$, $\Sigma y x_2 = 4013$, $\Sigma x_1 x_2 = 6485$,
$\Sigma y^2 = 7428$, $\Sigma x_1^2 = 19461$, $\Sigma x_2^2 = 2173$

i. Develop the estimating equation best describing these data.

ii. Interpret the value of regression coefficients.

iii. Estimate the annual unpaid tax for field audit labour 4200 hours and computer 1600 hours.

Soln:

The required linear regression equation is

$$y = b_0 + b_1 x_1 + b_2 x_2 \quad - (*)$$

Where,

y = amount of unpaid taxes discovered (in million $)
x_1 = field audit labour hours (in 100)
x_2 = computer hours (in 100)

Now,

To fit (*) we use the following equations

$$\Sigma y = n b_0 + b_1 \Sigma x_1 + b_2 \Sigma x_2$$

Where,

$\Sigma y = 272$, $\Sigma x_1 = 441$, $\Sigma x_2 = 147$, $n = 10$

i.e.

$$272 = 10 b_0 + 441 b_1 + 147 b_2 \quad - (1)$$

And,

$$\Sigma y x_1 = b_0 \Sigma x_1 + b_1 \Sigma x_1^2 + b_2 \Sigma x_1 x_2$$

where,

$\Sigma y x_1 = 12005$, $\Sigma x_1^2 = 19461$, $\Sigma x_1 x_2 = 6485$

i.e.

$$12005 = 441\, b_0 + 19461\, b_1 + 6485\, b_2 \quad -(II)$$

And,

$$\Sigma y x_2 = b_0 \Sigma x_2 + b_1 \Sigma x_1 x_2 + b_2 \Sigma x_2^2$$

where,

$\Sigma y x_2 = 4013$, $\Sigma x_2 = 147$, $\Sigma x_1 x_2 = 6485$, $\Sigma x_2^2 = 2173$

i.e.

$$4013 = 147\, b_0 + 6485\, b_1 + 2173\, b_2 \quad -(III)$$

Now,

Using Cramers rule to solve (I), (II) and (III)

coeff. of b_0	coeff. of b_1	coeff. of b_2	constant
10	441	147	272
441	19461	6485	12005
147	6485	2173	4013

So,

$$D = \begin{vmatrix} 10 & 441 & 147 \\ 441 & 19461 & 6485 \\ 147 & 6485 & 2173 \end{vmatrix}$$

$$D_0 = \begin{vmatrix} 272 & 441 & 147 \\ 12005 & 19461 & 6485 \\ 4013 & 6485 & 2173 \end{vmatrix}$$

$$D_{b_1} = \begin{vmatrix} 10 & 272 & 147 \\ 441 & 12005 & 6485 \\ 147 & 4013 & 2173 \end{vmatrix}$$

$$D_2 = \begin{vmatrix} 10 & 441 & 272 \\ 441 & 19461 & 12005 \\ 147 & 6485 & 4013 \end{vmatrix}$$

Then,

$$b_0 = \frac{D_0}{D} = -13.8196$$

$$b_1 = \frac{D_1}{D} = 0.5637$$

$$b_2 = \frac{D_1}{D} = 1.0995$$

Thus,

(*) can be written as

(i) $y = -13.8196 + 0.5637 x_1 + 1.0995 x_2$

(ii)

$b_0 = -13.8196$ means annual unpaid taxes discovered is -13.8196 million dollars when field audit labour hour is 0 and computer hours is 0.

$b_1 = 0.5637$ means annual unpaid taxes discovered increase by 0.5637 million dollars when field audit labour hour increa by 1 and computer hour is constant

$b_2 = 1.0995$ means annual unpaid taxes discovered increase by 1.0995 million dollars when computer hour increases by 1 and field audit labour hour is kept constant.

(iii)

$y = ?$ when $x_1 = 42$ (in 100 hrs) and $x_2 = 16$ (in 100 hrs)

i.e.

$$y = -13.8196 + 0.5637 * 42 + 1.0995 * 16$$
$$= 27.4478$$

⇒ Actual unpaid tax for field labour hour is 4200 and computer hour 1600 is 27.4478 million dollars

9.17 A computer manager needs to know how efficiency of her new computer program depends on the size of incoming data and how many tables are used to arrange each data set. Efficiency will be measured on basis of number of processed requests per hour. Applying program to data set of different sizes and number of tables used, she gets the following results:

Processed requests (y)	16	26	17	41	50	55	40
Data size (GB) (x_1)	15	10	10	8	7	7	6
No. of tables (x_2)	1	2	10	10	20	20	4

The regression equation obtained is $y = 52.7 - 2.87 x_1 + 0.95 x_2$

Total sum of square = 1452

Sum of square due to regression = 1143.3

a. Interpret the values of regression coefficients b_1 and b_2.

b. Test significance of regression model at 0.05 level of significance.

c. Is there significant relationship between processed requests and no. of tables at 0.05 level of significance? Given standard error of $b_2 = 0.55$.

d. What percentage of variation of processed requests is explained by data size and number of tables?

e. Compute standard error of estimate.

f. Estimate the number of processed requests if data size is 9 GB and number of tables used are 8.

5.7 A computer manager needs to know how efficiency of her new computer program depends on the size of incoming data and how many tables are used to arrange each data set. Efficiency will be measured on basis of number of processed requests per hour. Applying program to data set of different sizes and number of tables used, she gets the following results:

Processed requests (y)	16	26	17	41	50	55	40
Data size (GB) (x_1)	15	10	10	8	7	7	6
No. of tables (x_2)	1	2	10	10	20	20	4

The regression equation obtained is $y = 52.7 - 2.87 x_1 + 0.85 x_2$

Total sum of square = 1452

Sum of square due to regression = 1143.3

a. Interpret the values of regression coefficients b_1 and b_2.

b. Test significance of regression model at 0.05 level of significance.

c. Is there significant relationship between processed requests and no. of tables at 0.05 level of significance? Given standard error of $b_2 = 0.55$.

d. What percentage of variation of processed requests is explained by data size and number of tables?

e. Compute standard error of estimate.

f. Estimate the number of processed requests if data size is 9 GB and number of tables used are 8.

Soln:

The given regression equation is

$$y = 52.7 - 2.87 x_1 + 0.85 x_2$$

where,

$b_0 = 52.7$, $b_1 = -2.87$, $b_2 = 0.85$

for (a)

$b_1 = -2.87$ means that the no. of processed request decreases by -2.87 ~~GB~~ when data size increases by 1 GB, while number of tables are kept constant

$b_2 = 0.85$ means that the no. of processed request increases by 0.85 when no. of table increases by 1 while ~~number~~ data size is kept constant.

~~For (a)~~

Here,

y	x_1	x_2
16	15	1
26	10	2
17	10	10
41	8	10
50	7	20
55	7	20
40	6	4

Total sum of square (TSS) = 1452

Sum of square due to regression (SSR) = 1143.3

Total no. of observations (n) = 7

No. of independent variables (k) = 2

We know,

$SSR = TSS - SSE$

or $SSE = TSS - SSR$

$\Rightarrow SSE = 308.7$ = Sum of Square due to error

For (b)

* Problem to test

$H_0: \beta_j = 0$ (There is no linear relationship between dependent variable y and independent variables)

H_1: At least one β_j is different from zero $[j = 1, 2]$.

* Test statistic

$$F = \frac{MSR}{MSE}$$

where,

$MSR = SSR/k = 1143.3/2 = 571.65$

$MSE = SSE/(n-k-1) = 308.7/(7-2-1) = 77.175$

So,

$$F = \frac{57.165}{77.175} = 7.4072 = F_{calculated}.$$

* Critical value

At $\alpha = 0.05$, $k = 2$, $n-k-1 = 4$, critical value is $F_{\alpha}(k, n-k-1)$

$= F_{0.05}(2, 4) = 6.944$

* Decision

$F = 7.4072 > F_{0.05}(2,4) = 6.944$

Reject H_0 at 0.05 level of significance

* Conclusion

Regression model is significant at 0.05 level of significance.

For (c)

Check for significant relationship between y and x_2 [$\alpha = 0.05$ and $S_{b_2} = 0.55$ = standard error of b_2].

Here,

* Problem to test

$H_0: \beta_2 = 0$

$H_1: \beta_2 \neq 0$

* Test statistic

$$t = \frac{b_2}{S_{b_2}} = \frac{0.85}{0.55} = 1.5455$$

* Critical value

At $\alpha = 0.05$, $(n-k-1) = 4$, critical value (2-tail) is $t_{\alpha}(n-k-1)$

$= t_{0.05}(4) = 2.776$

* Decision

$t = 1.5455 < t_{0.05}(4) = 2.776$

$\Rightarrow$ Accept H_0 at 0.05 level of significance

* Conclusion

There is no significant relationship between processed request and no. of tables at 0.05 level of significance.

For (d)

$$R^2 = \frac{SSR}{TSS} = \frac{1143.3}{1452} = 0.7874$$

$\Rightarrow$ 78.74 %. Variation in processed request is explained by data size and no. of tables.

For (e)

Standard error of estimate is given by

$$Se = \sqrt{\frac{SSE}{n-k-1}} = \sqrt{MSE} = \sqrt{77.175} = 8.7849$$

For (f)

Data size (x_1) = 9 GB

No. of tables used (x_2) = 8

No. of processed request (y) = ?

where,

$$y = 52.7 - 2.87 x_1 + 0.85 x_2$$
$$= 52.7 - 2.87 \times 9 + 0.85 \times 8$$
$$= 33.67$$

i.e.

No of processed request is 33.67.

Q18 A computer manager is keenly interested to know how efficiency of her new computer program depends on size of incoming data

and data structure. Efficiency will be measured by no. of processed requests per hour. Data structure may be measured on how many tables were used to arrange each data set. All the information was put together as follows:

Data size (GB)	6	7	7	8	10	10	15
No. of tables	4	20	20	10	10	2	1
Processed requests	40	55	50	41	17	26	16

Identify which one is dependent variable? Fit appropriate multiple regression model and provide problem specific interpretations of the fitted regression coefficients.

Soln:

→ ~~Processed request~~ ~~Data size~~ (GB) is dependent variable.

Here,

Required linear regression equation is

$$y = b_0 + b_1 x_1 + b_2 x_2 \quad \text{— (*)}$$

where,

y = processed requests

x_1 = Data size (GB)

x_2 = No. of tables processed.

Now,

To fit (*) we require certain values for certain equations which are calculated as:

y	x_1	x_2	$x_1 y$	$x_2 y$	x_1^2	x_2^2	$x_1 x_2$
40	6	4	240	160	36	16	24
55	7	20	385	1100	49	400	140
50	7	20	350	1000	49	400	140
41	8	10	328	410	64	100	80
17	10	10	170	170	100	100	100
26	10	2	260	52	100	4	20
16	15	1	240	16	225	1	15

Here,

$\Sigma y = 245$ $\quad \Sigma x_2 y = 2908$

$\Sigma x_1 = 63$ $\quad \Sigma x_1^2 = 623$ $\quad n = 7$

$\Sigma x_2 = 67$ $\quad \Sigma x_2^2 = 1021$

$\Sigma x_1 y = 1973$ $\quad \Sigma x_1 x_2 = 519$

Then,

$$\Sigma y = n b_0 + b_1 \Sigma x_1 + b_2 \Sigma x_2$$

or, $245 = 7 b_0 + 63 b_1 + 67 b_2$ —(I)

And,

$$\Sigma y x_1 = b_0 \Sigma x_1 + b_1 \Sigma x_1^2 + b_2 \Sigma x_1 x_2$$

or, $1973 = 63 b_0 + 623 b_1 + 519 b_2$ —(II)

And,

$$\Sigma y x_2 = b_0 \Sigma x_2 + b_1 \Sigma x_1 x_2 + b_2 \Sigma x_2^2$$

or, $2908 = 67 b_0 + 519 b_1 + 1021 b_2$ —(III)

Now,

Solving (I) (II) and (III) using Cramer's rule

Coeff. of b_0	Coeff. of b_1	Coeff. of b_2	Constant
7	63	67	245
63	623	519	1973
67	519	1021	2908

$$D = \begin{vmatrix} 7 & 63 & 67 \\ 63 & 623 & 519 \\ 67 & 519 & 1021 \end{vmatrix}$$

$$D_1 = \begin{vmatrix} 245 & 63 & 67 \\ 1973 & 623 & 519 \\ 2908 & 519 & 1021 \end{vmatrix}$$

$$D_1 = \begin{vmatrix} 7 & 245 & 67 \\ 63 & 1973 & 519 \\ 67 & 2908 & 1021 \end{vmatrix}$$

$$D_2 = \begin{vmatrix} 7 & 63 & 245 \\ 63 & 623 & 1973 \\ 67 & 519 & 2908 \end{vmatrix}$$

Then,

$$b_0 = \frac{D_0}{D} = 51.1685$$

$$b_1 = \frac{D_1}{D} = -2.7309$$

$$b_2 = \frac{D_2}{D} = 0.8786$$

i.e.,

(*) can be written as

$$y = 51.1685 - 2.7309 x_1 + 0.8786 x_2$$

Here,

$b_0 = 51.1685$ means processed request is 51.1685 when data size is 0 GB and no. of tables processed is 0.

$b_1 = -2.7309$ means no. of processed request decreases by 2.7309 as data size increases by 1GB keeping no. of tables processed constant.

$b_2 = 0.8786$ means no. of processed request increases by 0.8786 as no. of tables increase by 1 keeping data size consta

Q19 What is Multiple linear Regression (MLR)? From following information of variables x_1, x_2 and y, fit a regression equation of y on x_1 and x_2.

$\Sigma x_1 = 272$, $\Sigma x_2 = 441$, $\Sigma y = 147$, $\Sigma x_1^2 = 7428$
$\Sigma x_2^2 = 19461$, $\Sigma y^2 = 2137$, $\Sigma x_1 y = 4013$, $\Sigma x_1 x_2 = 12005$
$\Sigma x_2 y = 6485$, $n = 10$.

Soln:

Required linear regression equation is

$$y = b_0 + b_1 x_1 + b_2 x_2 \quad \text{—} (*)$$

Here,

To fit (*), we use the following equations

$$\Sigma y = n b_0 + b_1 \Sigma x_1 + b_2 \Sigma x_2$$

or, $147 = 10 b_0 + 272 b_1 + 441 b_2$ — (I)

And,

$$\Sigma y x_1 = b_0 \Sigma x_1 + b_1 \Sigma x_1^2 + b_2 \Sigma x_1 x_2$$

or, $4013 = 272 b_0 + 7428 b_1 + 12005 b_2$ — (II)

And,

$$\Sigma y x_2 = b_0 \Sigma x_2 + b_1 \Sigma x_1 x_2 + b_2 \Sigma x_2^2$$

or, $6485 = 441 b_0 + 12005 b_1 + 19461 b_2$ — (III)

Then,

solving (I) (II) and (III) using Cramer's rule

Coeff. of b_0	Coeff. of b_1	Coeff. of b_2	Constant
10	272	441	147
272	7428	12005	4013
441	12005	19461	6485

$$D = \begin{vmatrix} 10 & 272 & 441 \\ 272 & 74028\ 19461 & 12005 \\ 441 & 12005 & 19461 \end{vmatrix}$$

$$D_0 = \begin{vmatrix} 147 & 272 & 441 \\ 4073 & 74028\ 19461 & 12005 \\ 6485 & 12005 & 19461 \end{vmatrix}$$

$$D_1 = \begin{vmatrix} 10 & 147 & 441 \\ 272 & 4073 & 12005 \\ 441 & 6485 & 19461 \end{vmatrix}$$

$$D_2 = \begin{vmatrix} 10 & 272 & 147 \\ 272 & 74028\ 19461 & 4073 \\ 441 & 12005 & 6485 \end{vmatrix}$$

Now,

$$b_0 = \frac{D_0}{D} = 10.4934$$

$$b_1 = \frac{D_1}{D} = 0.5801$$

$$b_2 = \frac{D_2}{D} = -0.2624$$

Thus,

(*) can be written as

$$y = 10.4934 + 0.5801 x_1 - 0.2624 x_2$$

Multiple linear regression (MLR)

MLR can be defined as linear functional relationship of a variable with 2 or more variables.

Q20 Suppose we are given the following information with $n = 7$. Multiple regression model is

$$\hat{y} = 8.15 + 0.56 x_1 + 0.54 x_2$$

Here, Total sum of square = 1493

Sum of square due to error = 91

Find

(i) R^2 and interpret it.

(ii) Test overall significance of model.

Soln:

We are given

Total sum of square (TSS) = 1493

Sum of square due to error (SSE) = 91

$n = 7$,

x_1 & x_2 are 2 independent variables $\Rightarrow k = 2$

Then,

Sum of square due to regression is given by

$SSR = TSS - SSE$

$= 1493 - 91$

$\Rightarrow SSR = 1402$

Now,

$$R^2 = \frac{SSR}{TSS} = \frac{1402}{1493} = 0.9390 = 93.9\%.$$

It means 93.9% variation in y is explained by x_1 & x_2

Here,

Mean Square Regression (MSR) = SSR/k

$\Rightarrow MSR = 1402/2 = 701$

Mean Square Error (MSE) = $SSE/n-k-1$

$$\Rightarrow MSE = \frac{91}{7-2-1} = 22.75$$

For (ii)

* Problem to test

H_0: $\beta_j = 0$ [Dependent and independent variables are insignificant]

H_1: At least one β_j is different from zero [$j = 1, 2$] i.e. There is linear relationship between dependent and at least one independent variable ⇒ Dependent and independent variables are significa

* Test statistic

$$F = \frac{MSR}{MSE} = \frac{701}{22.75} = 30.8132$$

* Critical value

Let $\alpha = 5\% = 0.05$. At $\alpha = 0.05$, $k = 2$ and $(n-k-1) = 4$, critical value is $F_\alpha(k, n-k-1) = F_{0.05}(2,4) = 6.944$

* Decision

$F = 30.8132 > F_{0.05}(2,4) = 6.944$

Reject H_0 at 0.05 level of significance

* Conclusion

Regression model is significant.

Q22 The following ANOVA summary table was obtained from a multiple regression model with 2 independent variable

SV	SS	df	MS	F-ratio
Regression	12.62	2	?	?
Error	0.78	12	?	
Total	13.4	14		

i. Determine Mean sum of square due to regression, the mean sum of square due to error and F-value.

ii. Test significance of overall regression model at 5% level of significa

iii. Compute coefficient of determination and interpret its value.

iv. Find standard error of estimate.

Soln:

From ~~the~~ given ANOVA table

Sum of square due to regression (SSR) = 12.62

" " error (SSE) = 0.78.

df (regression) = 2 = k

df (error) = 12 = n-k-1

& n = 15

(i) Then,

Mean Sum of Square due to Regression is given by

$$MSR = \frac{SSR}{k} = \frac{12.62}{2} = 6.31$$

And,

Mean Sum of Square due to error is given by

$$MSE = \frac{SSE}{n-k-1} = \frac{0.78}{12} = 0.065$$

Now,

F-value is given by

$$F = \frac{MSR}{MSE} = \frac{6.31}{0.065} = 97.0769$$

(ii)

* Problem to test

$H_0: \beta_i = 0$

H_1: At least one β_i is different from zero $[i = 1,2]$

* Test statistic

$F = 97.0769$

* Critical value

At α = 5% = 0.05, k = 2 and n-k-1 = 12, critical value is

$F_\alpha (k, n-k-1) = F_{0.05}(2,12) = 3.885$

* Decision

$F = 97.0769 > F_{0.05}(2,12) = 3.885$

Reject H_0 at 5% level of significance.

* Conclusion

The overall regression model is significant at 5% level of significance.

(h)

We know,

$TSS = SSR + SSE = 13.4$

Then,

Coeff. of determination $= R^2 = \frac{SSR}{TSS} = \frac{12.62}{13.4} = 0.9418 = 94.18$

It means 94.18% variation in y is explained by x_1 and x_2.

(i)

Standard error of estimate $(Se) = \sqrt{MSE} = \sqrt{0.065} = 0.2550$

Hence,

(i) MSR = 6.31

MSE = 0.065

F = 97.0769

(ii) The overall regression model is significant at 5% level of signific

(iii) $R^2 = 0.9418 = 94.18\%$

(iv) Standard error of estimate = 0.2550.